HAL
open science

Evaluation des risques de crise, appliquée à la détection des conflits armés intra-étatiques

Thomas Delavallade

To cite this version:

Thomas Delavallade. Evaluation des risques de crise, appliquée à la détection des conflits armés intra-étatiques. Autre [cs.OH]. Université Pierre et Marie Curie - Paris VI, 2007. Français. NNT : . tel-00230663

HAL Id: tel-00230663

https://theses.hal.science/tel-00230663v1

Submitted on 31 Jan 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITE PIERRE & MARIE CURIE
LA SCIENCE A PARIS

Évaluation des risques de crise, appliquée à la détection des conflits armés intra-étatiques

Thèse de doctorat de l'Université de Paris 6

présentée pour obtenir le grade de

Docteur de l'Université Paris 6
(spécialité informatique)

par

Thomas Delavallade

soutenue le 06 décembre, devant le jury composé de

Mme Bernadette Bouchon-Meunier (Directeur de recherche, CNRS)	Directrice de thèse
M. Philippe Capet (Thales)	Co-encadrant de thèse
M. Christophe Marsala (Maître de conférences, Université Paris 6)	Co-encadrant de thèse
M. Georges Hébrail (Professeur, ENST Paris)	Rapporteur
M. Louis Wehenkel (Professeur, Université de Liège)	Rapporteur
M. Bruno Crémilleux (Professeur, Université de Caen Basse-Normandie)	Examinateur
M. Jean-François Marcotorchino (Professeur, Université Paris 6)	Examinateur

Table des matières

Résumé **vii**

Introduction **1**

I Évaluation des risques à moyen terme **8**

1 État de l'art **10**
1.1 Approches qualitatives 10
1.2 Approches quantitatives 14
1.3 Complémentarité des deux approches 28

2 Un premier modèle d'évaluation des risques **30**
2.1 *Salammbô* : construction d'arbres de décision flous 30
2.2 Description des données 36
2.3 Premières expérimentations 38

3 Améliorations du modèle **42**
3.1 Un algorithme génétique pour la sélection d'attributs 42
3.2 Résultats expérimentaux 46

4 Discussion **49**

II Étude de la chaîne d'apprentissage dans son ensemble **52**

5 Comparaison de classifieurs **54**
5.1 Évaluation d'un classifieur 55
5.2 Deux classifieurs évalués sur une seule base de données 55
5.3 Deux classifieurs évalués sur n bases de données 59
5.4 k classifieurs évalués sur une seule base de données 61
5.5 k classifieurs évalués sur n bases de données 65
5.6 Conclusion 67

6 Traitement des données manquantes **69**
6.1 Position du problème 69
6.2 Mécanismes de génération des données manquantes 70
6.3 Importance de la répartition des données manquantes 73
6.4 État de l'art sur le traitement des données manquantes 74
6.5 Technique de substitution basée sur l'entropie 88
6.6 Analyse comparative empirique 92

6.7 Conclusion 110

7 Sélection d'attributs **113**
7.1 Position du problème 113
7.2 Définitions du problème 116
7.3 État de l'art sur les techniques de sélection d'attributs 123
7.4 Filtrage basé sur le test de Kolmogorov-Smirnov 157
7.5 Substitution et filtrage 164
7.6 Analyse comparative empirique 168
7.7 Conclusion 179

8 Discussion **182**

III Un nouveau modèle d'évaluation des risques **184**

9 Système global d'évaluation des risques **186**
9.1 Apprentissage du modèle 186
9.2 Utilisation du modèle 188

10 Application aux conflits armés intra-étatiques **192**
10.1 Théories sur l'émergence des conflits 193
10.2 Base de données sur les conflits armés intra-étatiques 205
10.3 Résultats expérimentaux 217
10.4 Analyse et interprétation des résultats 234

11 Discussion **250**

Conclusion **252**
11.1 Conclusion 252
11.2 Originalité de nos travaux 253
11.3 Perspectives 257

Bibliographie **260**

A Notations **275**

B Bases de données UCI **279**

C Caractéristiques générales des bases de données étudiées **281**

D Liste des pays étudiés **285**

E Liste des variables utilisées **307**

F Liste des sources utilisées **319**

G Résultats de la sélection de modèles **321**

Index des sigles et noms de méthodes **323**

Remerciements[1]

Écrire ces dernières lignes après plus de trois ans passées en compagnie de mon sujet de thèse est loin d'être aussi simple que je ne l'escomptais. Bien que que peu enclin au sentimentalisme, je ne saurais mettre un terme à cette thèse sans exprimer ma profonde reconnaissance à toutes celles et ceux qui ont contribué d'une manière ou d'une autre à sa réalisation. Je tâcherai donc d'être sobre dans la forme, emphase et autres hyperboles ne pouvant de toute façon suffire à transcrire fidèlement ma gratitude envers celles et ceux qui m'ont accompagné durant ces années.

Je remercie en premier lieu ma directrice de thèse, Bernadette Bouchon-Meunier, Christophe Marsala et Philippe Capet, mes encadrants aujourd'hui amis, qui m'ont fait confiance dès mon DEA. C'est essentiellement à eux que je dois la rigueur et l'honnêteté scientifique dont je crois avoir su faire preuve durant ma thèse. Bernadette, toujours disponible malgré un emploi du temps surchargé, par ses conseils mais aussi par l'environnement de travail qu'elle a su créer dans son équipe, fut une directrice que nombre de camarades thésards dans d'autres labos m'enviaient. Christophe a mis à ma disposition non seulement son logiciel *Salammbô* qui est au cœur de mes travaux, mais également son temps et sa précieuse expérience dans le domaine de l'apprentissage automatique, sans laquelle je n'aurais su éviter de nombreux cul-de-sacs dans lesquels je m'aventurais parfois allégrement. Les fréquentes discussions que j'ai pu avoir avec Philippe ont contribué de manière décisive à guider mes choix méthodologiques, à définir l'orientation générale de cette thèse. Je lui dois en particulier ma découverte de l'épistémologie qui eut une influence déterminante sur mon travail. Ses talents d'alchimistes pour extraire des idées prometteuses d'expérimentations anodines me furent également d'un grand secours durant les périodes de stagnation qui ont émaillé l'avancement de mes recherches.

Je tiens également à remercier les autres membres de mon jury, Georges Hébrail, Louis Wehenkel, Bruno Crémilleux et Jean-François Marcotorchino. Leurs nombreuses remarques et questions ont été pour moi l'occasion, la dernière peut-être, d'argumenter en toute franchise les choix techniques et méthodologiques que j'ai pris durant ces trois dernières années.

L'orientation géopolitique de mes travaux depuis mon DEA n'a été possible que grâce à Claude Michel. C'est lui qui a initié et développé depuis plus de 10 ans, au sein de THALES, les recherches sur la détection des crises. Avec Philippe il a constamment suivi mes travaux au sein de THALES.

Laure Mouillet a réalisé le premier démonstrateur de détection des crises qui a servi de base à mes premières implémentations. C'est également elle qui a guidé mes premiers pas dans le domaine de la prévision des conflits.

Mes séjours au sein de l'équipe LOFTI de Bernadette, chaque jeudi, m'ont énormément apporté tant sur le plan scientifique que sur le plan moral. Outre Bernadette et Christophe,

[1] Cette thèse a été cofinancée par THALES et le CNRS.

je tiens à remercier les permanents Maria, Herman, Nicolas et Marcyn, sans oublier l'irremplaçable Marie-Jeanne. Je suis également redevable à tous les thésards de l'équipe et en particulier à Marco qui a eu la lourde tâche de me supporter trois ans dans son bureau, Jason grâce à qui les cours de Maple finissaient toujours sur une bonne note, Vincent grâce à qui j'ai laissé une cheville à Orléans, Thanh Ha avec qui j'ai eu le plaisir d'écrire deux articles et Adrien avec qui j'ai croisé le fer sur les terrains de rugby. Je n'oublie pas non plus Javier qui me supporte depuis le DEA, Olivier, Thomas, Romain, Tri Duc et Lionel. Guillaume, Jean-François et Nicolas de l'équipe CONNEX de Patrick Gallinari m'ont également beaucoup soutenu, ainsi que Cédric à THALES.

C'est à Clara et Tom que je dois la relecture attentive de la dernière partie de cette thèse. Leurs conseils avisés et leurs explications économétriques m'ont permis de mieux comprendre la littérature relative aux causes des guerres civiles. Clara m'a également permis de rencontrer le professeur Collier dont les travaux ont joué un rôle très important dans la rédaction de la dernière partie de cette thèse.

Les discussions que j'ai pu avoir sur mon sujet avec Christian Mullon furent plus qu'enrichissantes. Je regrette seulement de n'avoir pas su trouver le temps d'exploiter toutes les pistes de recherche qu'il m'a suggérées. Je remercie également Jacek Biesiada d'avoir pris le temps de m'éclairer sur l'utilisation du test de Kolmogorov-Smirnov pour la sélection d'attributs.

Nico ami de toujours et grand amateur de la langue française a eu le courage de relire mon premier chapitre et c'est à Ren que je dois l'idée de mon étude de cas sur le Rwanda en fin de thèse. Enfin Ann-Cécile a contribué à la relecture de la quasi-totalité du manuscrit et à la réalisation d'une grande partie des graphes et tableaux de cette thèse.

L'aide de mes amis et cousins fut également inestimable. Ils m'ont contraint à expliciter, clairifier mes idées sur la détection des crises, et ils ont su me divertir lorsque cela devenait nécessaire, de Paris à La Réole en passant par Bordeaux, Toulouse ou Antibes, au comptoir des bistrots ou sur les prés de l'ovalie.

Mais c'est sans aucun doute à mes parents, ma soeur Quitterie et mon frère Martin, ainsi qu'à Ann-Cécile, que je suis le plus redevable. Cette thèse n'aurait jamais pu être achevée sans leur soutien. Ann-Cécile a en outre eu le mérite de supporter au quotidien mes humeurs parfois exécrables pendant ces trois années.

Je tiens en dernier lieu à remercier mes grand-mères et avoir une pensée pour mes grand-pères. C'est évidemment en grande partie à eux quatre que je dois l'affection de tout le reste de la famille qui si fut importante durant ces années. Je voudrais évoquer plus particulièrement celui que nous appellions bon papa. Sa soif de connaissance, son besoin constant de comprendre rationnellement le monde ont, je crois, grandement influencé ma personnalité et mon orientation vers la recherche. Je le tenais régulièrement au courant de mes travaux. Malheureusement il n'aura pas pu en connaître l'aboutissement.

Résumé

Dans de nombreux domaines, l'analyse rationnelle des risques fait partie intégrante du processus de décision. Il s'agit d'un outil méthodologique essentiel pour les acteurs politiques et économiques qui leur permet d'anticiper le déclenchement de crises potentielles. Dans certains secteurs d'activité les conséquences de telles crises sont parfois telles que le recours à l'analyse de risque correspond à une contrainte réglementaire imposée par le législateur. L'objectif d'une telle analyse est de parvenir à identifier les situations à risque ainsi que les principaux facteurs de risque de manière à pouvoir mettre en place les politiques de prévention adéquates.

Si de nombreuses cellules de veille ont été mises en place, tant au niveau de l'entreprise, qu'au niveau des institutions nationales et internationales, la quantité d'information potentiellement pertinente pour un sujet donné est parfois telle que la mise à disposition d'outils automatisant tout ou partie du traitement de cette information répond à un besoin réel, sinon à une nécessité.

Dans cette optique, dans cette thèse, nous proposons un système générique d'aide à l'anticipation de crises. Notre objectif est de fournir une synthèse d'une situation donnée, d'un point de vue structurel et non événementiel, via l'identification des crises potentielles ainsi que des principaux facteurs de risque associés. Le système que nous proposons repose sur l'apprentissage supervisé de règles de décision floues.

La qualité des données d'apprentissage étant problématique dans de nombreuses applications, nous proposons, dans nos travaux, une étude approfondie sur la chaîne de prétraitement, et en particulier sur le traitement des valeurs manquantes et sur la sélection d'attributs. Nous avons également mis l'accent sur l'évaluation et la sélection de modèles afin de pouvoir adapter les modèles de détection au problème à traiter, ainsi qu'aux besoins de l'utilisateur final.

La synthèse des résultats fournis par notre système étant destinée à des utilisateurs en charge de la veille stratégique, des outils d'aide au raisonnement et à la compréhension de cette synthèse sont également proposés.

Pour juger de l'intérêt de notre méthodologie nous détaillons son application à un problème concret : la détection des conflits armés intra-étatiques.

Introduction

Le calcul des probabilités, appliqué à la vie des nations, aux cas de guerre et de révolution, est le fondement de toute haute politique. Gouverner, c'est prévoir.

Émile de Girardin –

Cette citation d'Émile de Girardin résume parfaitement les motivations de cette thèse. Le processus décisionnel qui est à l'œuvre en politique ne saurait se passer d'outils d'aide à l'anticipation des crises telles que les guerres ou les révolutions. La référence au calcul des probabilités n'est pas anodine. Elle met en avant l'importance de la mise en place d'une méthodologie rationnelle et systématique d'évaluation des risques. C'est elle, et non l'intuition, qui est le « fondement de toute haute politique ». Les enjeux de la politique sont tels qu'il ne saurait être question de laisser la seule subjectivité des décideurs guider la conduite de l'État. D'une part, il est impossible d'appréhender intuitivement la complexité des phénomènes en jeu lors de la prise de décision. D'autre part, la position du décideur, en cas d'échec, est indéfendable si ses choix ne sont pas justifiables, l'intuition ne pouvant pas servir de justification acceptable.

Le recours à une méthodologie rationnelle d'évaluation des risques de crise n'est pas l'apanage des décideurs politiques. De nombreux autres domaines sont également concernés. Émile de Girardin l'avait bien compris comme en témoigne le début de sa citation que nous avons volontairement omis de présenter : « Le calcul des probabilités, appliqué à la mortalité humaine, a donné naissance à une science nouvelle : celle des assurances ». Il est étonnant de constater que ces propos datent du milieu du XIXe siècle. En effet, ce n'est que dans la première moitié du XXe siècle, dans les sciences économiques, que les travaux de Knight, et surtout ceux de Von Neumann et Morgenstern, ouvrirent la voie de la formalisation de l'évaluation des risques, en posant les bases de la théorie des jeux dans les années 40. À partir des années 60-70 l'industrie initia le développement d'outils méthodologiques afin de limiter les coûts liés aux défaillances techniques et fonctionnelles.

Les recherches sur le risque, qui ont connu leurs premiers développements dans ces deux domaines, se poursuivent désormais dans de multiples branches. En effet, au gré de crises majeures, le besoin s'est fait sentir de disposer d'outils techniques ou méthodologiques pour appréhender les risques. Ce fut le cas pour les risques technologiques avec les accidents de Three Miles Island et Tchernobyl. Citons également les risques de catastrophes naturelles qui préoccupent fortement les pouvoirs publics, et ce de manière accrue depuis le tsunami qui frappa l'Asie du Sud-Est en 2004. Les applications potentielles de l'analyse de risque sont donc nombreuses.

Si elle a été développée en premier lieu pour répondre à une demande spécifique, ce sont à présent souvent des exigences réglementaires qui imposent le recours à de telles méthodes. Cette évolution est flagrante dans le secteur de la santé ou celui de la banque. En ce qui concerne ce dernier, les institutions bancaires doivent respecter de plus en plus d'impératifs

en matière de gestion des risques, aussi bien sur le plan méthodologique à l'instar du monde de la santé, que sur le plan technique dans la formalisation de l'évaluation des risques. Les accords de Bâle puis de Bâle II entre les banques centrales des pays du G10 ont spécifié des standards de bonne pratique, afin de systématiser et rationaliser l'approche du risque. Outre les risques de marché et de crédit, spécifiques à ce secteur, les risques opérationnels[2] doivent désormais eux aussi faire l'objet d'études approfondies. Par contagion, les établissements bancaires concernés accentuent la pression sur le reste des entreprises, qui pour emprunter se voient contraintes à leur tour de respecter des normes de qualité plus exigeantes. Au vu de ces divers exemples il apparaît important, voire indispensable de créer des supports méthodologiques ou technologiques permettant de mieux gérer les risques.

Avant de poursuivre il est essentiel de s'attarder sur la terminologie et d'éclaircir cette notion de gestion des risques que nous venons de mentionner. Elle renvoie au processus de mise en balance des différentes politiques possibles visant à faire face aux risques identifiés. On distingue généralement quatre types de politiques de gestion des risques :

- évitement : ne pas se lancer dans une activité présentant un risque trop important
- réduction : prévention, renforcement du dispositif d'alerte, élaboration de stratégies de maîtrise des risques...
- acceptation : le risque ne peut plus être réduit, mais les profits espérés sont supérieurs aux pertes potentielles
- transfert : sous-traitance, assurance...

On parle de mise en balance car il s'agit de choisir, en termes de coûts, la politique adaptée à un ensemble de risques. On cherche la politique qui minimise la différence entre le coût de sa mise en place et l'espérance de la réduction de coût qu'elle permet. Ces coûts sont à prendre au sens large et sont parfois difficilement quantifiables. Ils peuvent intégrer des dimensions culturelles, sociales ou encore émotionnelles. Pour qu'un tel processus de gestion des risques puisse être mis en œuvre, il apparaît donc fondamental de procéder en amont à l'identification des risques, à leur évaluation et à leur hiérarchisation, ce qui constitue la phase d'analyse des risques. Durant cette phase il est essentiel de parvenir à une quantification ou qualification des risques la plus fine possible.

Cette thèse, réalisée sous l'égide d'une collaboration entre THALES Systèmes Terre et Interarmées et le LIP6, laboratoire d'informatique de l'Université Pierre et Marie Curie, Paris 6, s'inscrit pleinement dans ce cadre, son objectif principal étant réalisation d'un outil d'aide à l'anticipation des conflits géopolitiques intra-étatiques. Dans notre cas, les risques étudiés sont d'un type assez particulier. En effet, les événements à risque susceptibles de se produire correspondent à des crises, c'est-à-dire « des manifestations brusques et intenses, de durée limitée pouvant entraîner des conséquences néfastes »[3]. Les adjectifs « brusques » et « intenses » mettent l'accent sur leurs deux caractéristiques principales, à savoir leur imprévisibilité apparente (qui est en partie due au fait que leur probabilité d'occurrence est très faible) et l'ampleur des dommages qu'elles peuvent engendrer.

Ce second point permet de comprendre pourquoi il est crucial d'appréhender ce genre de phénomène de façon systématique et de disposer d'estimations aussi précises que possible des risques encourus. Si l'imprévisibilité des crises peut sembler, de prime abord, décourager toute velléité d'anticipation, les enjeux sont tels qu'il n'en est rien. Bien au contraire, celle-ci pousse plus que jamais les décideurs à investir dans la compréhension et la surveillance

[2] Ensemble des risques non liés aux fluctuations du marché ni au non-remboursement d'un emprunt de la part d'un client : dysfonctionnements techniques, problèmes dans la relation client, etc.

[3] Trésor de la Langue Française Informatisée : http://atilf.atilf.fr/tlf.htm

de ces crises, l'accent étant mis sur la détection la plus précoce possible du moindre élément crisogène. Du fait de la soudaineté du déclenchement des crises, les décideurs sont souvent pris de court. Aussi toute aide permettant de gagner du temps pour mieux préparer la phase de gestion de crise est-elle d'un apport inestimable.

D'une part, une meilleure compréhension de ces phénomènes peut permettre de mettre en exergue les fragilités du système étudié, suffisamment longtemps à l'avance pour pouvoir instaurer des politiques de prévention, et opérer ainsi une réduction aussi drastique que possible du risque. D'autre part, une surveillance accrue et continue des événements susceptibles de jouer un rôle dans la détérioration de la situation doit permettre d'identifier dès que possible les premiers symptômes de la crise. Le laps de temps ainsi gagné est certes insuffisant pour que des stratégies préventives puissent être effectives et efficaces, mais il permet d'envisager assez tôt les différentes options de maîtrise de la crise. On diminue alors l'effet de surprise, et l'improvisation n'est plus la seule réaction à disposition en période de crise.

Nous avons ici mis en avant l'importance de la détection des crises. L'analyse de risque ne se résume évidemment pas à cela. Si l'on s'en réfère à la norme (ISO/CEI 73), un risque est la « combinaison de la probabilité d'un événement et de ses conséquences ». Évaluer un risque nécessite donc de prendre en compte deux dimensions : l'incertitude quant à l'occurrence d'un phénomène néfaste d'une part, et la magnitude dudit phénomène d'autre part. La tâche de détection des crises, qui est le cœur de cette thèse, ne s'intéresse qu'à la première des deux dimensions et néglige la seconde. Estimer l'incertitude sous-jacente de l'occurrence d'une crise constitue cependant un passage obligé, capital comme nous venons de le voir, qui peut fort bien être découplé de l'étude des conséquences de la crise. Aussi avons-nous décidé, dans notre thèse, de nous focaliser sur cette étape.

Les risques politiques sont au cœur des préoccupations des institutions nationales et internationales, qui ont pour la plupart mis sur pied des cellules de veille stratégique, dont l'une des principales missions est de surveiller l'évolution de la situation dans un certain nombre de pays du monde. Il s'agit de repérer au plus tôt ceux qui risquent de poser problème, soit parce que des conflits potentiels avec d'autres pays se profilent, soit parce que de graves difficultés intérieures pourraient dégénérer en guerre civile, catastrophe humanitaire ou entraîner l'effondrement de toutes les structures étatiques... Depuis la fin de la Guerre froide, les conflits inter-étatiques sont moins nombreux et surtout la crainte d'une régionalisation des conflits est nettement moins forte. L'attention s'est donc portée vers les conflits intra-étatiques. Ils ne sont certes pas plus fréquents[4], mais plus médiatisés qu'auparavant.

Du fait de la mondialisation de l'accès à l'information et de sa diffusion la couverture des événements est désormais mondiale et instantanée. Chaque guerre civile est immédiatement suivie, et la passivité des différents pays et institutions internationales est vite pointée du doigt. Ainsi l'Organisation des Nations Unies (ONU) est parfois critiquée pour son incapacité à endiguer la violence. L'inaction des pays occidentaux dans de nombreux conflits a également été dénoncée. Ce fut le cas par exemple à propos de la guerre civile népalaise (1996-2006) ou encore à propos de la crise du Darfour qui fut déclenchée en 2003. C'est la prévention de ces conflits qui est instamment demandée, ce qui passe obligatoirement par un exigeant travail sur la détection et la compréhension de ces conflits.

Notre thèse s'attaque à ce problème spécifique, mais les retombées des travaux dans ce domaine sont nombreuses. Dans le domaine de l'énergie par exemple, l'instabilité politique

[4]On lit souvent le contraire, mais l'étude approfondie de ces conflits par Lacina et Gleditsch (2005) montre qu'il n'en est rien.

est hautement surveillée car elle peut engendrer d'importantes baisses d'activité, voire la fermeture d'installations. Mais ce ne sont pas là des difficultés propres au monde de l'énergie. Toute société cherchant à s'internationaliser y est confrontée, ce qui est désormais monnaie courante dans le cadre de la libéralisation des échanges qui gagne peu à peu l'ensemble de la planète. Ceci explique pourquoi les agences de notation comme Moody's, Standard & Poor's, évaluant le risque pays, c'est-à-dire le risque pour une entreprise de s'installer dans un pays donné, connaissent un tel succès. Or le risque politique fait partie intégrante de l'analyse du risque pays. Le monde de l'assurance est également concerné. Certains organismes comme la COFACE proposent d'assurer contre ce type de risque les entreprises qui cherchent à s'internationaliser.

Ces risques politiques sont traditionnellement l'affaire des seuls spécialistes, qui en tant qu'experts essaient d'évaluer aussi précisément que possible la situation. Ils se basent sur leur connaissance du terrain, sur leur culture géopolitique, ainsi que sur les données qu'ils ont à leur disposition. Depuis quelques années, avec le développement des réseaux de télécommunication et l'accroissement des capacités de stockage des données, on assiste à une explosion de la quantité d'information disponible. Ceci est vrai également pour les données sur lesquelles travaillent les experts en sciences politiques. Dans le même temps les recherches en fouille de données ont fait d'importants progrès, s'appuyant sur la statistique classique, ainsi que sur la théorie de l'apprentissage automatique. Aussi souhaitons-nous dans cette thèse mettre au profit de la veille stratégique un outil automatisant l'analyse systématique des données.

La problématique de la détection de crises, quoique spécifique, est prégnante dans nombre de domaines comme la finance ou la santé qui ont tous deux connus des crises de grande ampleur à la fin du vingitème siècle. Une vague de crises monétaires et financières a ainsi touché, dans les années 90, non seulement l'Asie du sud-Est mais également la Russie, le Mexique, le Brésil ou encore l'Argentine. Dans le domaine de la sécurité alimentaire, les crises de la vache folle, de la fièvre aphteuse, et l'épizootie de grippe aviaire ont eu d'importantes répercussions économiques et politiques. Du fait de cette multiplicité des applications potentielles, nous sommes fixé d'inscrire notre approche dans une méthodologie d'analyse des risques de crise aussi générique que possible, et non exclusivement centrée sur la question de l'instabilité politique. Celle-ci ne doit servir que de fil conducteur et d'illustration applicative.

La détection des conflits armés intra-étatiques, dans le cadre d'une collaboration entre THALES et le LIP6, a déjà fait l'objet d'une thèse de doctorat. Durant sa thèse, Mouillet (2005) s'est intéressée plus spécifiquement à la détection automatisée de signaux faibles, annonciateurs de crise, au sein d'un flux de données événementielles structurées[5]. Comme nous l'avons évoqué précédemment en parlant des cellules de veille, ce genre d'approche est essentiel dans un système de détection de crise, du fait de l'imprévisibilité du phénomène. Cependant l'analyse peut être fort utilement complétée, en identifiant plus en amont les mécanismes profonds influant sur le déclenchement de la crise. Cette phase de compréhension du contexte dans lequel une crise est susceptible de se déclarer est, répétons-le, fondamentale dans une optique de prévention. On distingue donc deux grandes catégories de méthodes de détection de crise. Dans les deux cas il s'agit de prédire l'occurrence d'une crise, mais elles diffèrent par la précision et l'horizon de leurs prédictions.

- L'**alerte rapide** est une approche événementielle qui consiste à suivre en temps réel un flux d'événements et à repérer dans cette masse ceux qui sont révélateurs d'un basculement de la situation. Le nombre d'événements à prendre en compte pouvant

[5] Ces données peuvent par exemple être issues du filtrage d'un fil d'agence de presse.

être très important, de nombreuses recherches sur l'automatisation de cette tâche ont été menées. Grâce à un suivi quotidien de l'évolution de la situation, des prédictions sur le court-terme quant à l'avènenement de crises potentielles sont effectuées par comparaison avec le déroulement des crises passées. Les méthodes se différencient alors essentiellement par la façon dont les crises du passé sont modélisées. Schrodt (2000) propose par exemple d'apprendre automatiquement des modèles de Markov cachés, tandis que Mouillet (2005) choisit de s'appuyer sur l'expertise pour établir des scénarios typiques de déclenchement de crise.

– L'**évaluation des risques** est une approche structurelle. Le but est de parvenir à une bonne compréhension de la situation vis-à-vis des crises potentielles. Cela se fait par l'analyse du contexte, en cherchant à déterminer les caractéristiques d'un contexte propice au déclenchement de ces crises. Le contexte est ici synthétisé par un ensemble d'indicateurs, comme par exemple le Produit Intérieur Brut (PIB) ou le niveau des réserves en hydrocarbures pour des crises politiques ou économiques. Le nombre d'indicateurs pouvant avoir une influence sur la crise étudiée peut s'avérer énorme. Ici aussi l'emploi de techniques de traitement automatisé peut donc être utilement mis à profit. Comme pour l'*alerte rapide*, on s'attachera à prédire l'émergence de situations critiques, mais cette fois sur le long terme, les données à partir desquelles seront réalisées ces prédictions évoluant beaucoup plus lentement.

Ces deux familles de méthodes ne doivent pas être vues comme concurrentes. Elles sont complémentaires, et devraient être incluses toutes deux dans une procédure globale d'analyse de risque de crise. En effet, lorsque l'on s'intéresse aux crises, un suivi continu de l'actualité est indipensable pour pouvoir alerter les décideurs des moindres signes de détérioration de la situation, ce que seule l'*alerte rapide* est capable d'offrir. En matière de conflits armés, de nombreux organismes ont compris l'importance de cette tâche et ont constitué un réseau d'experts couvrant la planète, chargés de repérer ces signes et de faire remonter les éventuelles alertes. Un des plus avancés dans ce domaine est la Fondation suisse pour la paix avec son programme FAST International[6]. Parmi ceux dont l'analyse repose sur un réseau d'experts, ils sont les premiers, à notre connaissance, à avoir commencé à intégrer les approches automatiques (Hämmerli *et al.*, 2006).

La prévention est également fondamentale. Or elle ne peut passer que par une évaluation fine des conditions pouvant favoriser l'apparition des crises, ce que l'on se propose de faire durant l'*évaluation des risques*. Pour ce qui est des conflits armés, des institutions internationales comme l'ONU ou la Banque mondiale s'intéressent vivement à cette problématique. Cette dernière menant d'ailleurs des études poussées pour essayer de déterminer empiriquement quels sont les facteurs prépondérants expliquant l'émergence de tels conflits (Collier et Hoeffler, 1998).

Ces deux méthodes poursuivent des buts distincts, mais complémentaires, tous deux essentiels à la mise en place d'un outil de détection des crises. De même que le sens d'un terme ne se révèle qu'au sein d'une phrase et de son contexte d'élocution, la portée d'un événement ne peut se comprendre véritablement que si le contexte dans lequel il se produit est lui-même bien compris. Ainsi une manifestation importante dans un pays donné est en soi un signe à prendre en compte pour la détection de crises politiques, humanitaires ou économiques, mais ce signe doit être interprété différemment suivant que le pays est en récession ou non, que la liberté d'expression est respectée ou non dans ce pays, etc. L'évaluation des risques, constitue en soi une étape importante de l'analyse de risque. Si par ailleurs on l'utilise pour contextualiser l'analyse événementielle conduite lors de l'alerte

[6] http://www.swisspeace.org/fast/

rapide, alors le couplage des deux approches prend tout son sens. C'est dans cette optique de complémentarité que nous avons décidé de construire un outil de veille à la suite de celui qui a été développé par Laure Mouillet pour l'*alerte rapide* (Delavallade *et al.*, 2007).

Avant de rentrer dans le vif du sujet et de présenter plus en détail notre vision de l'évaluation des risques, quelques précisions méthodologiques s'imposent. L'analyse de risque, en fonction des disciplines dans lesquelles elle fut étudiée, a pris de multiples formes. Nous y reviendrons en détail au cours de la partie 1 dans laquelle nous dresserons un panorama des différentes techniques existantes. Derrière ces diverses approches, se cachent des conceptions différentes de la notion même de risque. Thompson et Dean (1996), à la suite de Schrader-Frechette (1991), ont mis en évidence le fait que ces multiples conceptions s'organisent le long d'un continuum dont les positions extrêmes correspondent respectivement à une vision positiviste et relativiste du risque[7]. Les tenants du positivisme font de la quantification de l'incertitude l'élément central de l'analyse de risque, tandis que pour les relativistes le risque est multidimensionnel, la prépondérance de telle ou telle dimension dépendant du contexte dans lequel on se place. Pour les plus extrêmes, un risque n'est qu'une construction sociale, certains allant jusqu'à considérer qu'il n'existe pas de risque à proprement parler, mais seulement des perceptions de risque propres à chaque individu. Ces positions ne constituent évidemment que des bornes et sont rarement adoptées dans la pratique. Ce sont cependant des points de repère auxquels se réfèrent souvent les chercheurs de ce domaine, de manière plus ou moins explicite.

En ce qui nous concerne, nous adopterons un point de vue plutôt positiviste, puisque nous nous focalisons sur la détection de crise et donc sur la prédiction de l'occurrence d'une crise. Précisons cependant que notre propos n'est pas non plus de nier le caractère multidimensionnel du risque. La réversibilité d'un risque (le fait qu'il soit permanent ou non), le fait qu'il soit pris volontairement ou non, le caractère intentionnel de la menace, le contrôle que l'on peut ou non exercer sur ce risque, sont tous des facteurs importants dont va dépendre la perception que l'on a du risque comme le souligne Slovic (1987), ce qui influera sur l'acceptabilité du risque. Mais il s'agit là, selon nous, de questions propres à la gestion des risques et à la sélection de la politique adéquate pour faire face aux risques. À la différence de Slovic et de Zimmerman et Bier (2002) nous pensons que gestion et analyse de risque, du moins gestion et évaluation des risques, peuvent être découplées.

De plus, en matière de crise la position relativiste extrême est tout simplement intenable. Avancer que le déclenchement d'une guerre civile, ou que l'effondrement d'un système monétaire, n'est un risque pour un décideur particulier que s'il le perçoit comme tel est certes possible. Il n'en reste pas moins vrai que la crise politique ou financière en question ainsi que ses conséquences sont bien réelles. Aussi la problématique de la détection de crise en tant que telle, dissociée de l'évaluation de ses impacts psychologiques et médiatiques, nous semble-t-elle fondamentale. Comme dernière justification, arguons que la tâche est en soi déjà suffisamment vaste et complexe.

Nous écartons donc de notre champ d'investigation ces questions sociales, psychologiques ou encore anthropologiques, mais nous le faisons consciemment, estimant que ce sont là des points à traiter dans une étape ultérieure de gestion des risques.

Pour synthétiser l'ensemble de ces remarques liminaires, nous nous focaliserons sur la détection des crises. Nous proposerons une démarche aussi générique que possible, que

[7] Nous reprenons ici la terminologie de Scharder-Frechette, Thompson et Dean parlant plus volontiers de vision probabiliste et contextualiste. À l'instar de Schrader-Frechette (1997), la distinction nous semble superfétatoire. Aussi préférons-nous revenir à la terminologie originale, qui nous semble sur le plan philosophique plus porteuse de sens.

nous appliquerons dans le cadre spécifique de la prévision de conflits armés intra-étatiques, l'objectif final étant la réalisation d'un outil d'aide à la détection automatisée de ces conflits. Nous parlons d'aide à la détection car en matière de crises, il est évidemment hors de question de laisser l'homme hors de la boucle. Ce sont les décideurs qui prennent les choix en dernier recours. Étant donné le caractère hautement stratégique de ces décisions, celles-ci doivent pouvoir être prises sur la base d'analyses expertes. L'outil doit faciliter le travail des cellules de veille et n'a aucunement la prétention de se substituer à elles. Il n'a pas vocation à réaliser l'ensemble de la tâche d'analyse de risque, mais à en être un élément central entre les mains d'experts. Nous reportons sur les experts toute la phase de synthèse de l'analyse et de prise en considération des autres dimensions du risque en vue de la préparation de l'étape de gestion qui est, elle, de la responsabilité du décideur.

Pour ne pas biaiser l'interprétation des résultats et ne pas interférer avec la subjectivité de l'analyste, subjectivité qui sera mise à contribution pour tenir compte des paramètres psycho-sociologiques, il nous semble important de veiller à ce que le modèle de détection soit le plus objectif possible.

L'outil étant destiné à interagir avec un utilisateur humain, cela impose certaines contraintes méthodologiques. En effet il conviendra de choisir une technique de détection qui permette d'obtenir des résultats clairs, facilement interprétables par l'utilisateur, pour que celui-ci puisse les remettre en question et aussi se les approprier, sans quoi il n'aura aucune confiance dans l'outil et ne l'utilisera jamais. D'autre part, rappelons qu'un des objectifs, et non des moindres, de l'évaluation des risques, est de mettre en avant les faiblesses structurelles d'un système afin de pouvoir discerner des leviers d'actions préventives. Si l'on veut aider l'expert à identifier de tels leviers, il nous faut donc ne pas simplement lui donner une probabilité (ou autre mesure d'incertitude) de crise, mais lui expliquer comment cette probabilité a été calculée et quels facteurs justifient le niveau de risque ainsi établi. Cela requiert également que les sorties de notre outil soient aussi compréhensibles que possible.

Ces deux contraintes fixées, objectivité dans la modélisation et clarté dans les résultats présentés, nous pouvons maintenant passer à la partie I dans laquelle seront présentés les atouts et faiblesses de différentes méthodes d'évaluation des risques. Cela nous permettra de justifier le choix technique que nous avons arrêté, à savoir l'apprentissage d'arbres de décision flous, avant de voir comment nous l'avons mis œuvre en construisant un premier modèle, que nous avons ensuite affiné pour combler les lacunes qu'ont fait ressortir les premières expérimentations. Dans la partie II nous aborderons les questions liées au prétraitement des données qui jouent un rôle fondamental dans l'élaboration de tout modèle d'apprentissage. Nous nous concentrerons plus particulièrement sur le problème des données manquantes, ainsi que sur celui de la sélection d'attributs, ces deux points étant cruciaux dans notre contexte. La formalisation complète de notre méthodologie, ainsi que le modèle final et son analyse empirique seront alors détaillés dans la partie III. Nous aurons terminé la description du modèle de prévision sur lequel s'appuie notre outil. Enfin nous conclurons par une synthèse de la méthodologie retenue et de ses apports dans le domaine de la détection de crise. Nous discuterons alors des perspectives ouvertes par l'introduction de cette nouvelle méthodologie.

Première partie

Évaluation des risques à moyen terme

Analyser les risques, ou plus précisément les évaluer, est une problématique qui a été abordée dans de nombreuses disciplines, de cultures théoriques et pratiques variées. Aussi n'y a-t-il pas une méthodologie et une technique d'évaluation, mais une multitude, chacune étant plus ou moins adaptée à des besoins particuliers, aucune ne faisant l'unanimité, et ce y compris au sein d'une même discipline. Ce dernier point s'explique par la diversité des manières dont les modélisateurs appréhendent la notion de risque (voir le paragraphe sur les différentes conceptions du risque dans l'introduction).

Dans cette partie nous proposons de dresser un état de l'art des différentes méthodes d'évaluation des risques. Nous ne prétendons pas couvrir l'ensemble des techniques existantes, cependant nous tâcherons de rendre compte de la plupart des grandes familles de méthodes. Si nous avons choisi pour application les conflits intra-étatiques, rappelons tout de même que nous souhaitons développer un outil sur la base d'une méthodologie aussi générique que possible. Pour cette raison, les techniques qui seront présentées dans le chapitre 1, ne seront pas exclusivement tirées des sciences politiques mais plutôt des divers domaines dans lesquels l'évaluation des risques est pratiquée. À travers cet état de l'art nous mettrons en avant les caractéristiques des différentes approches. L'analyse de ces caractéristiques sous le prisme de notre problème spécifique (la détection de crises) en tenant compte des contraintes que nous nous sommes imposées (objectivité et clarté), nous conduira alors à introduire notre modèle de détection.

Chapitre 1

État de l'art

Afin de ne pas procéder à une simple énumération, plutôt rébarbative, des différentes méthodes d'évaluation des risques, nous avons décidé de ne présenter que les principales familles de méthodes. Nous les avons regroupées dans deux catégories : les méthodes qualitatives et quantitatives, suivant que l'on cherche à décrire à l'aide de variables linguistiques le niveau de risque (qualitatif) ou que l'on essaie d'obtenir une évaluation chiffrée de l'incertitude (quantitatif). Ce critère de discrimination est assez naturel. D'une part, les références explicites à l'une ou l'autre de ces deux approches sont monnaie courante dans la littérature. Notons à titre d'exemple que Cullen et Small (2004), dans un article tiré d'un ouvrage collectif destiné à donner une vue d'ensemble de la problématique du risque, adoptent explicitement cette même catégorisation pour caractériser les différentes méthodes d'évaluation des risques. D'autre part, ce critère permet de distinguer les différentes conceptions du risque qui prévalent dans le domaine de l'évaluation. Ces deux approches se focalisent sur la seule notion d'incertitude et correspondent donc plutôt à des visions positivistes, selon la terminologie de Schrader-Frechette. Cependant les tenants du qualitatif, par l'intégration de l'expertise humaine, ont une position moins extrême le long du continuum positivisme-relativisme que ceux qui privilégient le quantitatif.

1.1 Approches qualitatives

Estimer les risques de manière qualitative consiste selon Cullen et Small (2004) à envisager les problèmes qui risquent d'affecter le système étudié, au moyen de la constitution de scénarios. Il faut ensuite, pour tenir compte des deux dimensions principales du risque, estimer dans quelle mesure ils sont susceptibles de se réaliser et évaluer leur gravité. Ce processus est le fruit du travail d'experts, qui, compte tenu de leur expérience, de leur intuition et de leur connaissance du domaine, identifient les scénarios possibles et les évaluent. La façon dont les jugements des experts sont recueillis influe sur la qualité des estimations produites (Dufour *et al.*, 2002), mais nous n'aborderons pas ce point ici, l'objectif de cet état de l'art n'étant pas de détailler chacune des méthodes présentées, mais de les décrire succinctement et de souligner leurs caratéristiques.

1.1.1 Analyse vulnérabilités/menaces

Dans une optique de prévention des risques, il importe de pouvoir mettre en avant, durant la phase d'identification et d'évaluation des risques, les faiblesses structurelles du système qui peuvent le mettre en péril, afin de pouvoir concentrer les efforts en matière de réduction de risque sur ces faiblesses. C'est dans cette perspective, pour rapprocher analyse et gestion de risque, que l'analyse vulnérabilités/menaces a été développée. Comme

le suggère la terminologie employée, c'est essentiellement dans les domaines afférant à la sécurité qu'elle est utilisée (Bass et Robichaux, 2001; Baybutt, 2002).

Elle comporte deux volets :

- **Étude des vulnérabilités** dans laquelle les failles potentielles du système susceptibles d'être exploitées doivent être repérées. Ce sont tous les points faibles qui exposent le système à des risques. Des experts évaluent alors le niveau d'exposition. L'échelle de notation comprendra plus ou moins de niveaux (« exposition faible » , « exposition modérée » , etc.) selon la précision des estimations souhaitées. Bass et Robichaux par exemple proposent d'en utiliser quatre. Il est cependant souhaitable de ne pas avoir recours à des échelles trop fines, mal adaptées à l'imprécision du jugement humain, et que les experts auraient donc du mal à utiliser. Ceci constitue une première étape, donnant une première mesure approximative de l'incertitude. En effet plus l'exposition sera forte, plus il sera vraisemblable qu'un problème surgisse. L'impact d'un tel problème doit ensuite être lui aussi estimé, ce qui conduit les experts à attribuer un niveau de sévérité à chacune des vunérabilités. Cette fois c'est la seconde dimension du risque, relative à la magnitude des dommages, qui est estimée. Enfin pour chacune des vulnérabilités, les deux indices (exposition et sévérité) sont combinés pour former un indice global et synthétique de vulnérabilité. L'agrégation des mesures qualitatives se fait généralement au moyen de la définition, souvent *ad hoc*, d'une matrice croisant les deux indices et dont l'élément (i, j) contient le résultat de la combinaison du niveau i de sévérité et du niveau j d'exposition. En épidémiologie par exemple, Zepeda Sein (1998) ont défini la matrice décrite dans le tableau 1.1, pour pouvoir combiner les mesures qualitatives de deux paramètres en un seul indice de risque[1] :

Tab. 1.1 – Construction d'un indice de risque à partir l'estimation qualitative de deux paramètres

	Évaluation du paramètre 1			
Évaluation du paramètre 2	Négligeable	Faible	Moyen	Élevé
Négligeable	Négligeable	Faible	Faible	Moyen
Faible	Faible	Faible	Moyen	Moyen
Moyen	Faible	Moyen	Moyen	Élevé
Élevé	Moyen	Moyen	Élevé	Élevé

Il pourrait être intéressant de formaliser cette phase et de recourir pour cela aux techniques éprouvées de l'agrégation multicritère (Grabisch et Perny, 1999).

- **Étude des menaces** : elle se focalise sur les éléments, humains ou non (virus informatiques par exemple), susceptibles d'exploiter les vulnérabilités. Ces éléments, une fois identifiés, sont jugés selon un processus identique à celui qui est mis en place dans l'étude des vulnérabilités. Des experts attribuent deux notes à chacune des menaces, appréciant d'une part leurs motivations et d'autre part leurs capacités. Ces deux notes sont ensuite combinées en un indice global de menace. Cette étape permet de raffiner l'approximation de la mesure de l'incertitude. Plus une entité, considérée comme une menace, sera encline à attaquer le système, et plus elle en aura les moyens,

[1] Cette matrice a été initialement construite, dans le cadre de l'évaluation du risque zoosanitaire, pour agréger le niveau d'exposition à une maladie avec la probabilité d'apparition de cette maladie. Voir par exemple (Moutou *et al.*, 2001) pour une application dans le cas de la fièvre aphteuse.

plus il sera alors problable que le système soit mis en danger. Notons que lorsque les menaces ne correspondent pas à des entités humaines, la notion de motivation perd quelque peu de son sens.

L'intérêt de cette démarche réside dans la recherche explicite des points faibles d'un système. L'analyse des menaces complète utilement la démarche et assure une meilleure compréhension de la situation. C'est là un point de départ important pour la définition de politiques préventives. L'analyse des menaces, en tenant compte de l'intentionnalité d'adversaires potentiels, est assez spécifique aux questions de sécurité, mais dans ce domaine c'est justement l'intégration de cette dimension intentionnelle qui fait la force de cette technique.

1.1.2 Analyse des modes de défaillances, de leurs effets et de leur criticité (AMDEC)

Afin de rationaliser la production et de garantir la qualité des produits, le secteur industriel a mis en place des méthodologies d'analyse systématique du risque. Issue de travaux dans l'aéronautique, l'analyse des modes de défaillances, de leurs effets et de leur criticité (AMDEC) a été développée pour répondre à ces besoins. Son usage s'est ensuite assez vite répandu au reste du monde industriel et gagne de l'influence dans d'autres secteurs, en particulier celui de la santé.

Comme le soulignent Kmenta *et al.* (1999) l'AMDEC a pour objectif de répondre aux trois questions suivantes :

- **Quels problèmes le système étudié peut-il rencontrer ?**
  Il convient donc en premier lieu d'identifier les défaillances potentielles dont peut être victime le système. Dans le monde industriel, une telle analyse est menée à chaque lancement de nouveau produit. Le produit en lui-même ainsi que son processus de fabrication constituent le système à étudier.
- **À quel point est-il vraisemblable que ces problèmes se produisent et quelles seraient alors leurs conséquences ?**
  On retrouve les deux composantes du risque à évaluer : incertitude et sévérité.
- **Que peut-on faire pour éviter ces problèmes, ou au moins pour en limiter les conséquences ?**
  Cette partie de l'analyse correspond plutôt à une phase préparatoire de gestion de risque, qu'il s'agit d'entamer le plus tôt possible.

Les estimations de la probabilité d'occurrence d'une défaillance et des dommages qu'elle est susceptible d'engendrer si elle survient, sont classiquement l'affaire d'experts. Cependant cette pratique évolue. Des chercheurs s'intéressent à l'automatisation de cette procédure d'estimation (Papadopoulos *et al.*, 2004; Rhee et Ishii, 2003). Leur objectif est d'une part de tirer profit des immenses bases de données historiques que nombre d'entreprises constituent pour recenser les incidents survenus. La masse des données peut être telle que les limites cognitives d'un individu, serait-ce un expert, l'empêchent de l'appréhender complètement. D'autre part, Rhee et Ishii (2003) l'expliquent clairement : il s'agit d'éviter le biais de la subjectivité inhérent à l'évaluation qualitative par expertise. Leurs motivations sont donc très proches des nôtres.

L'AMDEC est finalement synthétisée dans un tableau dont chaque ligne correspond à la description d'une défaillance potentielle et regroupe des éléments d'estimation et de gestion de risques. Ce tableau assure la mise en parallèle de l'espérance du coût associé

à une défaillance et du coût des solutions envisagées, ce qui permet une hiérarchisation des solutions et constitue l'attrait principal de cette méthode. Voici quelques exemples des champs que l'on peut trouver dans un tableau de synthèse d'AMDEC :

- Modes de défaillance
- Causes
- Effets
- Degré d'incertitude quant à l'occurrence d'un événement non souhaité (Occ) : par exemple un entier compris entre 1 et 10
- Capacité de détection (Det) : par exemple un entier compris entre 1 et 10
- Gravité (G) : par exemple un entier compris entre 1 et 10
- Niveau de risque : agrégation de Occ, Det et G. Le produit est fréquemment utilisé comme opérateur d'agrégation : $R = Occ \times Det \times G$
- Actions envisagées

1.1.3 Systèmes à base de connaissances

Les deux approches précédentes ne sont pas à proprement parler des techniques d'évaluation des risques, mais plutôt des méthodologies générales d'analyse des risques. Elles insistent surtout sur la démarche à suivre pour mener à bien une telle analyse et faire en sorte que celle-ci intègre le maximum d'éléments permettant de faciliter la gestion de risque. L'estimation en elle-même est le fruit de jugements d'experts et ne constitue qu'un des points de cette méthodologie. Peu d'indications sont données pour savoir comment les experts attribuent leurs notes, sur quels critères ils se basent. Pour pouvoir les aider, il serait intéressant de comprendre les mécanismes cognitifs à l'œuvre, afin d'en suivre les principes ou au moins de ne pas être en totale contradiction avec ceux-ci.

Issus des recherches en intelligence artificielle des années 70, les systèmes à base de connaissances sont des systèmes d'inférence qui ont été créés en s'inspirant justement du mode de raisonnement expert. À partir d'une base de faits et d'une base de règles « **Si** ... **alors** ... », qui synthétise la connaissance du domaine, des inférences peuvent être réalisées (voir figure 1.1). Pris comme outils d'aide à la décision, ces systèmes présentent l'avantage d'être transparents pour l'utilisateur, c'est-à-dire que celui-ci comprend à chaque instant ce qui est fait. Toute conclusion à laquelle parvient le moteur d'inférences est en effet accompagnée de la séquence de règles, aisément compréhensibles, qui ont été utilisées pour y aboutir. Le diagnostic médical, qui peut être vu comme un problème d'évaluation des risques, est l'une des grandes applications de ces systèmes à base de connaissances. En analyse de risque, dans divers domaines comme par exemple l'environnement (Potter *et al.*, 2000), des chercheurs les ont ensuite également employés.

Lorsque les faits sur lesquels repose l'inférence correspondent à des données quantitatives, comme c'est le cas en économie ou en finance, il peut être souhaitable de disposer d'une certaine souplesse au niveau des règles, afin de se rapprocher du raisonnement de l'expert. Comme tout individu, il manipule plus facilement des termes vagues, qualitatifs, que des données précises et chiffrées. Ainsi il serait préférable de pouvoir prendre en compte une règle de la forme « **Si** le taux de change baisse significativement **alors** le risque est élevé » plutôt que « **Si** le taux de change baisse de plus de 12,43% **alors** le risque est élevé ». Disposer d'une telle souplesse facilite par ailleurs le recueil de l'expertise, qui se fait souvent par entretiens. Pour faire ce pont entre symbolique et numérique, la logique floue, par la formalisation de l'utilisation de variables linguistiques est parfaitement adaptée. Aussi des systèmes experts flous ont-ils vu le jour et sont actuellement mis en place pour estimer les risques, en économie et en finance par exemple (Dahal *et al.*, 2005).

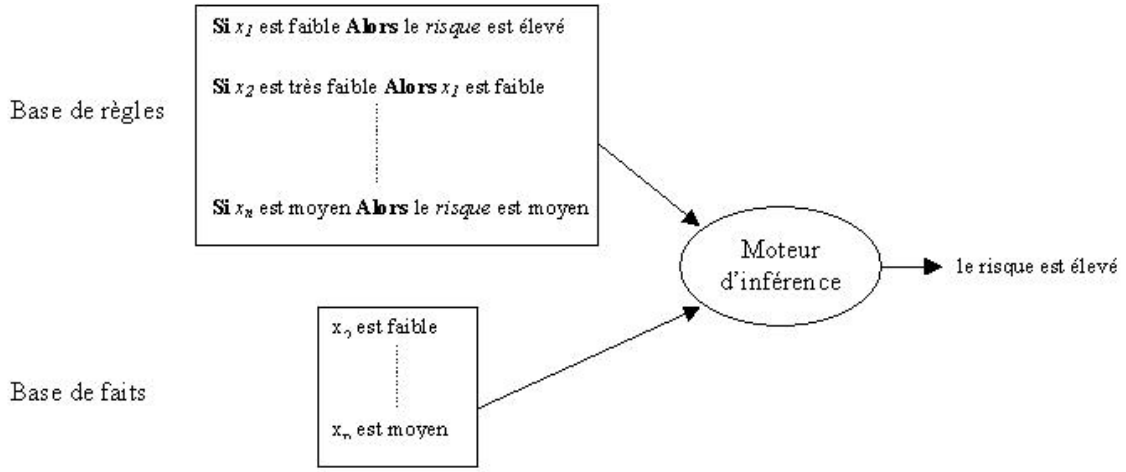

FIG. 1.1 – Architecture simplifiée d'un système expert

L'intégration explicite des connaissances d'experts dans un moteur de raisonnement automatisé est le grand atout des systèmes à base de connaissances. Ils combinent automatisation du processus d'évaluation des risques et clarté de l'analyse produite. Ils restent ainsi avant tout au service des experts. Il faut tout de même tempérer ces remarques, car tout repose sur un recueil de connaissances long et délicat, l'expert n'étant pas forcément conscient de toutes les règles qu'il manipule lorsqu'il a à évaluer un risque. D'autre part, le biais de subjectivité que nous essayons d'éviter est transmis au moteur d'inférence lors de la création de la base de règles. Le problème est plus grave encore lorsque l'utilisateur du système expert est celui qui a rentré les règles dans le système, car il est vraisemblable qu'il se sente conforté dans son intuition par le système, qui ne fait que reproduire son propre biais subjectif.

Pour pallier ces difficultés, il est possible de poursuivre l'effort d'automatisation et d'apprendre automatiquement les règles à inclure dans la base. L'extraction de règles d'associations peut être utilisée à cet effet. C'est ce que font Spanos *et al.* (1999) pour construire un système expert flou d'aide à l'évaluation des risques financiers. Ils notent cependant que le nombre de règles générées est très important, ce qui nuit d'une part à l'efficacité de l'outil, et d'autre part à la lisibilité des résultats.

1.2 Approches quantitatives

L'évaluation quantitative des risques a pour objectif principal le calcul d'un indice de risque et non plus simplement d'un niveau de risque comme c'était le cas pour les méthodes qualitatives. Le calcul de cet indice se fait à partir d'un certain nombre de variables d'entrée, qu'il faut combiner entre elles par la création d'un modèle de risque. Plus formellement, il faut trouver une fonction f telle que $R = f(v_1, v_2, ..., v_p)$, où R est l'indice de risque, v_i est la i-ième variable d'entrée et p est le nombre de ces variables. Les différentes techniques que nous allons présenter dans la suite de cette section diffèrent essentiellement par le choix de modélisation qui est fait, c'est-à-dire par la façon dont la fonction f est construite.

On distingue deux grandes approches selon que la fonction est établie en partant d'une théorie explicative du domaine ou de données empiriques, auquel cas on parlera d'apprentissage automatique. Il s'avère que de nombreuses techniques d'estimation ont recours à ces deux approches, la théorie permettant de spécifier un modèle générique dont seuls les paramètres sont ensuite appris automatiquement.

1.2.1 Analyse multicritère

Face à un problème complexe, une démarche analytique classique consiste à le décomposer en sous-problèmes plus simples à traiter, et à réitérer ce processus sur chacun des sous-problèmes, jusqu'à ce que l'on parvienne à un ensemble de problèmes élémentaires que l'on saura tous résoudre[2]. L'AMDEC (voir section 1.1.2) et l'analyse vulnérabilités/menaces (voir section 1.1.1) peuvent être considérées comme des instanciations particulières de cette démarche. En effet, toutes deux s'efforcent de mettre en évidence ce que l'on pourrait appeler des facteurs élémentaires de risque (défaillances pour l'une et vulnérabilités pour l'autre). Ceux-ci sont suffisamment simples pour être appréhendés directement. L'évaluation finale du risque est alors la combinaison des évaluations élémentaires.

L'analyse multicritère repose sur une méthodologie semblable. À la différence des techniques qualitatives, l'évaluation des risques élémentaires est quantitative et la combinaison de ces estimations s'appuie sur un cadre formel, celui de l'agrégation multicritère. Les travaux de Butler et Fischbeck (2002) dans le domaine de la sécurité illustrent parfaitement ces remarques. Ils procèdent tout d'abord à une analyse vulnérabilités/menaces, et se focalisent ensuite sur l'évaluation du risque lié à chacune des menaces identifiées. Mais contrairement à ce qui est fait par les techniques qualitatives, ils adoptent l'approche fréquentiste, pour quantifier directement incertitude et ampleur des dommages, à partir de données historiques. De plus, ils ont explicitement recours au formalisme de l'agrégation multicritère pour combiner leurs évaluations élémentaires.

Suivant le niveau de finesse du modèle choisi, deux types d'analyse multicritère émergent des différentes recherches qui ont été menées à ce sujet. Les deux partent d'une décomposition du risque global en facteurs de risques élémentaires.

- La première, plus simple, se contente de cette décomposition et évalue directement chacun de ces facteurs avant d'agréger les résultats pour obtenir l'indice de risque global R.

  $$R = Agg\left(r_1, r_2, ..., r_p, a_1, a_2, ..., a_q\right)$$

  Les r_i correspondent aux facteurs de risque, p étant leur nombre. Agg est l'opérateur d'agrégation choisi et les a_i sont les paramètres de cet opérateur, q étant leur nombre. Nous noterons Agg_q un opérateur ayant q paramètres. On a ainsi

  $$R = Agg_q\left(r_1, r_2, ..., r_p\right)$$

  Remarquons que sous cette forme nous retrouvons exactement la formalisation de la tâche à remplir par les techniques quantitatives. Les r_i correspondent aux *variables d'entrée*, les v_i, et Agg_q est la fonction f que l'on cherche. Ici il ne s'agit pas d'apprendre cette fonction mais de choisir celle qui convient le mieux en fonction du domaine et des contraintes que l'on se fixe. Pour plus de détails sur cette question du choix de l'opérateur, on pourra se reporter à la thèse de Detyniecki (2000).

- La seconde méthode n'évalue pas directement les facteurs de risque. Pour chacun d'eux, probabilité d'occurrence P_i et magnitude des conséquences C_i de chacun des facteurs r_i, doivent être estimées et ensuite combinées. Cela revient à développer l'équation précédente, en remplaçant les r_i par cette combinaison.

  $$R = Agg_q\left(Agg^1_{q_1}\left(P_1, C_1\right), ..., Agg^1_{q_1}\left(P_p, C_p\right)\right)$$

[2] Le développement logiciel dans le secteur de l'informatique par exemple, met constamment en pratique cette démarche.

$Agg^1_{q_1}$ est l'opérateur permettant d'agréger probabilité et conséquence d'un événement. Traditionnellement on prend le produit. Ensuite, les différentes conséquences possibles de l'événement étudié doivent être identifiées. L'ampleur globale des conséquences est alors calculée en agrégeant les dommages élémentaires. De manière plus formelle, en notant $Agg^2_{q_2}$ l'opérateur réalisant cette agrégation et c^i_j la j-ième conséquence parmi les p_i possibles, associée au facteur de risque r_i, on peut écrire :

$$R = Agg_q \left(Agg^1_{q_1} \left(P_1, Agg^2_{q_2} \left(c^1_1, c^1_2, ..., c^1_{p_1} \right) \right), ..., Agg^2_{q_2} \left(P_p, Agg^2_{q_2} \left(c^p_1, c^p_2, ..., c^p_{p_n} \right) \right) \right)$$

La technique utilisée par Butler et Fischbeck (2002), brièvement décrite au début de cette section, est très proche de cette seconde méthode. La première, plus simple, est fréquemment utilisée, en particulier pour évaluer les risques d'instabilité politique. Les experts du CIFP (Ampleford *et al.*, 2001) (Country Indicators for Foreign Policy) par exemple commencent par déterminer des grands domaines d'inquiétude, tels que la démographie ou l'économie. Pour chacun de ces domaines, un certain nombre d'indicateurs sont sélectionnés et évalués pour tous les pays étudiés. Ensuite un score par domaine d'inquiétude est obtenu en prenant la moyenne, calculée sur l'ensemble des indicateurs du domaine en question. Enfin ces scores sont eux-mêmes agrégés par moyenne pondérée pour construire l'indice de risque global.

Cette démarche de décomposition du risque nécessite en pratique le recours à l'expertise. Le choix des facteurs élémentaires de risque, leur évaluation et leur agrégation, sont autant d'étapes durant lesquelles un biais important peut être introduit. L'opérateur d'agrégation choisi, sans justification explicite la plupart du temps, est quasiment toujours la moyenne pondérée[3]. Nous l'avons vu pour le CIFP, mais c'est également le cas avec nombre de méthodologies d'analyses risque pays, comme par exemple l'International Country Risk Guide (ICRG) du groupe Political Risk Services (PRS) (Linder et Santiso, 2002). Or, les chercheurs en agrégation multicritère insistent sur la nécessité d'expliciter les propriétés que l'on attend de l'opérateur d'agrégation, sous peine de faire de mauvais choix (Grabisch et Perny, 1999; Marichal, 2000). Il est dommage de disposer d'un cadre formel d'analyse mûr, et de ne pas en tirer pleinement parti. Aussi serait-il bon que les méthodologies d'évaluation des risques basées sur l'analyse multicritère intègrent une phase de sélection de l'opérateur d'agrégation, plutôt que de partir du principe que cet opérateur sera forcément la moyenne pondérée.

1.2.2 Approches graphiques

Les méthodes graphiques d'évaluation des risques prennent en compte les relations entre facteurs de risque et les utilisent pour construire un graphe. Celui-ci modélise l'ensemble des interactions qui permettent de décrire le phénomène étudié. Ces interactions sont symbolisées par les arcs du graphe, tandis que les nœuds représentent les concepts, variables à partir desquelles l'analyse peut être menée.

Un formalisme mathématique est ensuite utilisé pour faire des inférences quant à l'occurrence d'un des éléments du graphe à partir d'informations sur les autres nœuds. La théorie probabiliste, bayésienne, est bien adaptée pour propager les incertitudes. Aussi est-elle fréquemment intégrée aux approches graphiques. Nous en verrons des exemples dans les deux sections suivantes.

[3] Les poids doivent alors être définis par les experts.

1.2.2.1 Arbres d'événements et de défauts

Les arbres d'événements et de défauts, techniques graphiques assez proches, ont principalement été utilisés dans l'industrie pour estimer la probabilité qu'un produit connaisse une défaillance (Zimmerman et Bier, 2002). Le graphe sur lequel repose l'analyse est un arbre, comme on peut le voir sur la figure 1.2. Il correspond à un scénario décrivant la façon dont un produit peut tomber en panne. Nous employons des termes spécifiques aux problématiques de l'industrie, car ce sont elles qui ont motivé les développements de ces techniques graphiques. Cependant, on peut tout à fait imaginer des applications en détection des crises. Dans ce cas les scénarios doivent décrire la façon dont les crises se déclenchent.

La racine de ces arbres correspond à la défaillance du produit, tandis que les feuilles représentent des défaillances au niveau des composants élémentaires du produit. Les autres nœuds symbolisent des défaillances de composants intermédiaires. On retrouve donc la même idée de décomposition d'un macro-phénomène en éléments plus simples à appréhender. Les arcs sont orientés[4] (des feuilles vers la racine) et leur présence signifie que la défaillance d'un composant peut entraîner celle d'un autre. À chacun de ces arcs est associée une probabilité conditionnelle qui évalue à quel point l'influence du nœud source sur le nœud cible est grande.

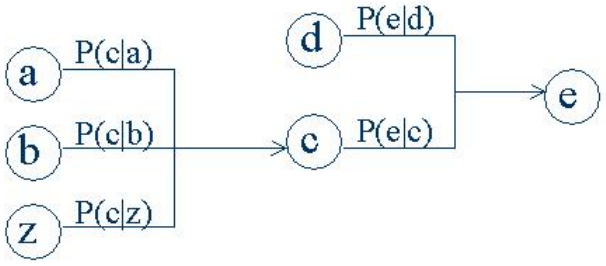

FIG. 1.2 – Exemple d'arbres de défauts ou d'événements

On se sert alors de l'inférence bayésienne pour déterminer la probabilité jointe du scénario, à partir des probabilités marginales et conditionnelles. Dans l'exemple de la figure 1.2, appelons S le scénario correspondant à cet arbre. On a

$$P(S) = P(a,b,c,d,e,z) = P(a)\,P(b)\,P(z)\,P(d)\,P(c|a)\,P(c|b)\,P(c|z)\,P(e|c)\,P(e|d)$$

Les arbres de défauts et d'événements diffèrent par leur mode de construction. Pour les arbres de défauts, on part de la racine. Des experts cherchent ensuite quels sont les composants qui peuvent connaître des défaillances, puis les sous-composants qui peuvent être à l'origine de défauts dans les composants identifiés précédemment et ainsi de suite, jusqu'à parvenir à des composants élémentaires.

Contrairement au processus abductif à l'œuvre dans la construction des arbres de défauts, les arbres d'événements sont établis de manière déductive. Les experts commencent par identifier l'ensemble des composants élémentaires susceptibles de connaître des problèmes. Puis ils analysent les conséquences de telles pannes sur des composants intermédiaires et ainsi de suite jusqu'à ce qu'ils repèrent les composants qui peuvent provoquer la défaillance du produit dans son ensemble.

Les deux méthodes sont assez lourdes à mettre en œuvre et demandent un gros travail d'expertise. Mais une fois construits, les arbres sont facilement utilisables. Ils présentent en

[4] Il ne s'agit donc pas à proprement parler d'arbres mais plutôt de graphes connexes orientés et acycliques, à racine unique.

outre l'avantage, du fait de leur représentation visuelle, d'assurer une bonne compréhension de l'ensemble de la situation. Mais cet avantage n'est plus aussi flagrant lorsque le problème est complexe et que les arbres deviennent trop grands, car ils perdent alors de leur lisibilité. Notons à ce sujet que les arbres d'événements, du fait de leur mode de construction déductif sont généralement beaucoup plus larges que les arbres de défauts (Zimmerman et Bier, 2002).

1.2.2.2 Réseaux bayésiens

Les arbres de défauts ou d'événements sont bien adaptés à l'évaluation du risque de pannes, problème dans lequel les interactions entre les différentes variables d'intérêt sont assez simples. En revanche, du fait de leur mode de construction et de leur structure arborescente avec une racine unique, ils ne permettent pas de modéliser correctement des phénomènes plus complexes, dans lesquels les relations entre variables doivent être décrites de manière assez fine.

Les réseaux bayésiens constituent une autre technique graphique d'aide à la décision. Ils reposent également sur la construction et l'utilisation d'un graphe orienté acyclique qui, pour un risque donné, doit décrire l'ensemble des relations entre les variables susceptibles de jouer un rôle dans la réalisation de ce risque. Mais contrairement aux techniques présentées dans la section précédente, ces graphes peuvent avoir plusieurs racines. De plus aucune contrainte n'est imposée quant à la façon de les contruire. Des experts identifient l'ensemble des facteurs de risque qui seront les nœuds du graphe, ainsi que les liens causaux qui existent entre eux, qui seront les arcs. Comme précédemment une probabilité conditionnelle est associée à chaque arc. Le graphe final ainsi obtenu représente explicitement les dépendances et indépendances conditionnelles[5]. Dans le cadre de l'inférence bayésienne, qui est celle que l'on utilise avec ces réseaux, ce point est essentiel car il permet de factoriser la probabilité jointe des variables du graphe, grâce à la propriété de Markov, ce qui autorise une réduction substantielle du nombre de paramètres du modèle. Si l'on appelle $v_i,\ i = 1..p$ ces variables, on peut écrire :

$$P(v_1, v_2, ..., v_p) = \prod_{i=1}^{p} P\left(v_i | parents\left(v_i\right)\right)$$

où $parents\left(v_i\right)$ correspond à l'ensemble des nœuds parents de v_i. Pour réaliser des inférences, il faut pouvoir estimer les probabilités conditionnelles $P\left(v_i | parents\left(v_i\right)\right)$. Selon le cadre d'interprétation des probabilités dans lequel se place le modélisateur, plusieurs approches sont envisageables.

- interprétation **subjectiviste** : la probabilité est une mesure de la croyance d'acteurs en une proposition donnée. Dans notre cas, cela signifie que ce seront des experts qui fixeront subjectivement les valeurs des différentes probabilités qui apparaissent dans le modèle.
- interprétation **fréquentiste** : la probabilité correspond à la fréquence d'observation d'un événement. Cela implique qu'elle peut être estimée à partir de données empiriques. Lorsque celles-ci sont disponibles cette approche est souvent adoptée. Lorsque les données sont complètes, l'estimateur du maximum de vraisemblance ou du maximum *a posteriori* peuvent être utilisés (Leray et François, 2004). Mais dans la pratique, ceci est rarement le cas. On se tourne alors vers l'algorihtme itératif

[5] Ceci était également vrai pour les arbres de défauts ou d'événements qui peuvent être considérés comme des réseaux bayésiens particuliers.

Expectation-Maximisation (EM) introduit par Dempster *et al.* (1977). Pour plus de détails sur ces différentes méthodes le lecteur pourra se reporter à (Jordan, 1998).

Une autre interprétation de la probabilité, que nous qualifierons de **logiciste**, est possible. Elle la considère comme une preuve inférentielle. C'est-à-dire que la probabilité d'un événement, d'une proposition permet d'en inférer la valeur de vérité. Cette interprétation n'a pas d'implications sur la façon dont sont estimées les probabilités dans le modèle. En revanche elle permet de justifier que les réseaux bayésiens soient également appelés systèmes experts probabilistes. Un arc entre deux nœuds A et B correspond à une règle du type : « **Si** A **alors** B avec telle probabilité ».

Il est donc possible de tracer les résultats obtenus par application des réseaux bayésiens *via* l'ensemble des règles utilisées. De plus, comme toute méthode graphique, la visualisation est un atout non négligeable pour comprendre la situation. Notons également que les recherches sur les réseaux bayésiens sont très actives. Des méthodes robustes sont disponibles pour réaliser automatiquement l'apprentissage des probabilités ainsi que l'inférence elle-même. Ceci explique qu'ils constituent une technique intéressante d'évaluation des risques, répandue dans de nombreux domaines comme par exemple le diagnostic médical (Leray et François, 2004) ou encore la détection des catastrophes naturelles (Straub, 2005).

Ils sont bien adaptés lorsque le domaine de connaissance est bien délimité et maîtrisé, comme cela est le cas dans les deux applications que nous venons de citer. Lorsque tel n'est pas le cas et que la complexité du graphe devient trop importante, l'atout visuel disparaît. D'autre part, lorsque les relations de cause à effet sont loin d'être facilement identifiables, le choix de la structure du graphe devient critique. Les experts doivent en effet s'appuyer, de manière explicite ou non, sur une théorie particulière pour construire le graphe. Or dans les cas complexes, aucune théorie ne fait l'unanimité. Il arrive bien souvent que diverses théories soient parfaitement contradictoires. Il est difficile de rejeter totalement l'une ou l'autre de ces théories. Faire le choix d'un modèle est alors problématique et revient à prendre parti. On retombe sur le problème de la subjectivité du modélisateur, que nous voulons limiter au maximum. Dans cette optique de nombreuses recherches commencent à se développer pour apprendre automatiquement, à partir de données historiques, la structure des réseaux bayésiens. Les méthodes développées sont prometteuses, mais encore très coûteuses en temps de calcul et pas suffisamment performantes (François et Leray, 2004).

1.2.3 Apprentissage automatique

Une manière de s'affranchir du choix d'une théorie est de recourir à des techniques d'apprentissage automatique. Cette solution, que nous avons déjà évoquée à la section 1.1.3 à propos du choix des règles d'un système expert, peut également s'appliquer aux systèmes experts probabilistes. Cela nous a amené à parler de construction automatique de graphes à la section précédente. Abordons maintenant plus spécifiquement ce problème.

Rappelons-le, l'objectif des techniques de quantification de l'incertitude liée à un risque est de parvenir à modéliser ce risque. C'est-à-dire qu'elles cherchent à déterminer une fonction f de l'ensemble des facteurs influant sur l'occurrence de celui-ci et à valeurs[6] dans $[0; 1]$. Par l'analyse multicritère (voir section 1.2.1), nous avons vu comment des experts pouvaient définir entièrement une telle fonction. Les arbres de défauts et réseaux bayésiens de la section 1.2.2 sont quelque peu différents sur le plan méthodologique. La modélisation du risque repose sur la construction d'un graphe et l'estimation de probabilités. Si la première étape est souvent manuelle, les recherches sur l'automatisation de la détermination

[6] Si l'on construit une fonction f à valeurs dans un sous-ensemble borné de $\mathbb{R}$, on peut toujours se ramener à l'intervalle $[0; 1]$.

de la structure du graphe n'étant pas encore suffisamment au point, la seconde est en revanche réalisée de manière automatique. Cela revient à attribuer une forme générique à la fonction f et à apprendre ensuite les paramètres de cette fonction sur des données historiques.

Accepter que des experts fixent manuellement f revient à considérer que la théorie explicative du risque à laquelle ils se réfèrent, explicitement ou non, est potentiellement la bonne. En revanche apprendre automatiquement cette fonction revient à supposer qu'il existe, dans le temps, des régularités dans la façon dont un risque se produit. L'apprentissage consiste alors à repérer ces régularités et à les généraliser pour ne pas créer un modèle qui soit trop dépendant des données sur lesquelles il a été appris. Ceci correspond à une phase d'**induction**.

Le processus d'estimation quantitative du risque, que nous illustrons figure 1.3, est le même quelle que soit la façon dont f est choisie, que ce soit par induction ou par l'application d'une théorie : f est utilisée sur des données actuelles pour calculer l'indice de risque (phase de **déduction**)[7]. Il est également possible de ne pas construire explicitement la fonction f et de calculer cet indice directement à partir des données, ce que Vapnik (1995) appelle **transduction**.

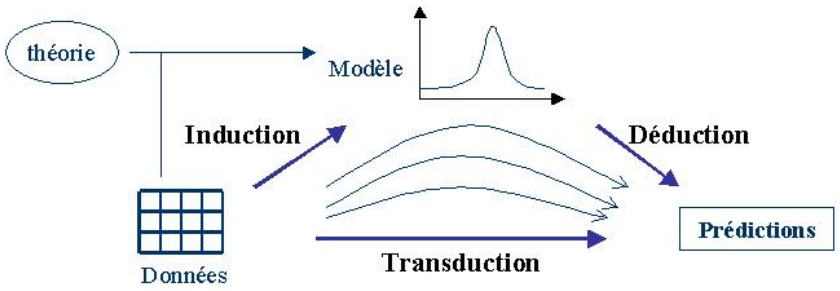

FIG. 1.3 – Schématisation du processus d'estimation quantitative des risques

1.2.3.1 Notations

Avant de présenter les deux grandes techniques d'apprentissage supervisé utilisées en évaluation des risques, introduisons les notations dont nous ferons usage tout au long de ce manuscrit. Par la suite nous ne préciserons que les notations nouvellement introduites, mais il sera possible de se reporter à tout moment à l'annexe A qui rassemble l'ensemble des notations.

Les données à partir desquelles l'apprentissage est effectué forment la **base d'exemples** $\mathcal{E} = \{e_i\}_{i=1..n}$, où e_i est le i-ième exemple de la base, qui en comporte n.

Chacun de ces exemples, ou instances, est décrit par un ensemble $\mathcal{V} = \{v_i\}_{i=1..p}$ de p **attributs** que nous avions précédemment nommés *variables d'entrée*.

En apprentissage supervisé, il est une variable qui joue un rôle particulier. Nous l'appellerons **classe** et nous la noterons y. On présuppose l'existence d'une relation g qui relie les variables **descriptives** v_i à la variable **cible** : $y = g(v_1, v_2, ..., v_p)$. L'objectif est de trouver une relation f qui s'en approche au maximum.

y et les v_i sont considérées comme des variables aléatoires et les valeurs prises par celles-ci pour chacun des exemples correspondent à autant de réalisations. Ainsi on dira que y_j est la j-ième réalisation de y, c'est-à-dire la valeur prise par y pour e_j. De même la

[7] Cette figure est fortement inspirée de celle de Galindo et Tamayo (2000) synthétisant le processus d'apprentissage.

valeur prise par v_i pour ce même exemple e_j sera appelée j-ième réalisation de v_i et sera notée v_{ji}.

Dans le domaine de la statistique, auquel nous ferons parfois appel, la variable *classe* y est désignée par les termes de **variable dépendante** ou **variable à expliquer** tandis que les termes de **variables indépendantes** ou **variables explicatives** sont employés pour parler des attributs v_i. Enfin, les exemples e_i sont appelés observations.

Lorsque y est une variable continue l'apprentissage est un problème de **régression**. On parle de **classification supervisée** ou de **catégorisation** lorsque y est discrète. Dans ce cas on note $\mathcal{C} = \{c_i\}_{i=1..K}$ l'ensemble des classes, c'est-à-dire les valeurs ou modalités que peut prendre y. Nous désignerons par $Classe$ la fonction qui associe un exemple à sa classe. Le modèle appris par un algorithme de classification sera appelé **classifieur**. Par abus de langage nous utiliserons parfois ce même terme pour désigner la chaîne d'apprentissage ayant permis de construire ce modèle.

L'ensemble des données peut également s'écrire sous forme matricielle. On notera V la matrice correspondant à la base d'exemples, de dimension $(n \times p)$ dont les lignes sont les exemples et les colonnes sont les variables.

$$V = \begin{bmatrix} v_{11} & \dots & v_{1p} \\ \vdots & & \vdots \\ v_{n1} & \dots & v_{np} \end{bmatrix}$$

1.2.3.2 Régression

Issues des travaux en statistique, les techniques de régression s'attachent à expliquer le comportement d'une variable dépendante continue y à l'aide des variables explicatives v_i. Formellement elles cherchent à construire une fonction f qui permette de prédire le mieux possible y à partir des v_i. Un des critères retenus pour évaluer la qualité des prédictions est celui des moindres carrés. On cherche f telle que $\sum_{i=1}^{n} \left[f\left(v_{i1}, v_{i2}...v_{ip}\right)) - y_i\right]^2$ soit minimale.

Appliquées à l'évaluation des risques, de telles techniques sont très utiles pour quantifier en termes probabilistes l'incertitude liée à l'occurrence d'un événement particulier. Dans ce contexte la variable dépendante y est binaire et prend la valeur 1 si l'événement en question se produit et 0 sinon. En supposant que les y_i sont des réalisations indépendantes, nous pouvons considérer que y suit une loi de Bernoulli, dont le seul paramètre est $\pi = P\left(y = 1 | v_1, v_2...v_p\right)$.

Les modèles de régression consistent à estimer ce paramètre π qui est une variable continue, à valeurs dans $[0; 1]$, en trouvant la fonction f qui s'en rapproche le plus, par exemple au sens des moindres carrés. Le modélisateur, en posant certaines hypothèses sur la distribution des données, spécifie la forme générale de f et ses paramètres sont ensuite appris automatiquement à partir de la base d'exemples $\mathcal{E}$.

Parmi les différentes méthodes de régression, nous avons choisi d'en présenter deux. La première est la plus simple qui soit et permet de comprendre comment les autres sont formées, tandis que la seconde est l'une des plus usitées en matière d'évaluation des risques ou détection de crises (O'Brien, 2001) :

- **régression linéaire** : on suppose que la fonction f est linéaire, c'est-à-dire que la variable dépendante peut être estimée par une combinaison linéaire des variables

explicatives. Dans notre cas, avec y binaire, on a :

$$\pi_i = P\left(y_i = 1 | v_{i1}, v_{i1} ... v_{ip}\right) = \sum_{j=1}^{p} b_j v_{ij} + \epsilon_i \quad \forall i = 1..n$$

où ϵ_i est le résidu[8] et les b_j sont les paramètres du modèle qui pondèrent l'influence des variables v_j sur la variable π.
On peut également l'écrire sous forme matricielle :

$$\Pi = VB + E$$

où E est le vecteur colonne des n résidus, Π est le vecteur colonne contenant les valeurs de π_i pour les n exemples de la base et B est le vecteur colonne des p paramètres.

– **régression logistique** : la contrainte de linéarité est très forte et rarement réaliste. L'autre problème de la régression linéaire tient au fait que les valeurs prédites par le modèle ne sont pas forcément entre 0 et 1. Or ce sont des probabilités que l'on veut estimer. La régression logistique n'a pas ces défauts. Elle peut prendre en compte des non-linéarités. Elle suppose l'existence d'une variable latente continue y^*, qu'il est possible d'estimer et qui est liée à y par les règles de décision suivantes (Greene, 2003) :

$$\begin{array}{lcl} y^* \geq 0 & \Rightarrow & y = 1 \\ y^* \leq 0 & \Rightarrow & y = 0 \end{array}$$

On suppose l'existence d'une relation linéaire entre y^* et les v_i : $Y^* = VB + E$. Le modèle logistique consiste alors à considérer que la fonction de répartition des résidus F_ϵ est une fonction logistique. On a, sous forme matricielle :

$$\begin{array}{rcl} \Pi & = & P\left(Y = 1 | V\right) = 1 - P\left(Y^* \leq 0\right) \\ \Pi & = & 1 - P\left(E \leq -VB\right) = 1 - F_\epsilon\left(-VB\right) \\ \Pi & = & 1 - \dfrac{\exp\left(-VB\right)}{1 + \exp\left(-VB\right)} = \dfrac{1}{1 + \exp\left(-VB\right)} \end{array}$$

La relation entre π et les variables explicatives est donc log-linéaire et plus précisément c'est $\log\left(\frac{\pi}{1-\pi}\right)$ qui peut être exprimé comme une combinaison linéaire des v_i. Notons également qu'avec cette équation, les estimations qui sont faites sont cette fois dans l'intervalle $[0; 1]$.

Ce ne sont là que deux exemples parmi tous les types de régression possibles, mais quel que soit le type choisi, le processus de modélisation est le même : un expert fixe la forme générale du modèle (linéaire, logistique, Poisson...) et sélectionne les variables explicatives qui doivent entrer dans ce modèle. Elles correspondent à ce que nous avons appelé les facteurs élémentaires de risque. Ensuite des données historiques sont utilisées pour apprendre les paramètres du modèle (le vecteur B).

La régression est un outil puissant. Des modèles sophistiqués, fruits de dizaines d'années de recherche dans le domaine, sont disponibles en fonction des spécificités du problème. Vis-à-vis de notre problème, ces modèles ont également l'avantage d'exhiber directement les probabilités d'occurrence d'une crise. Cependant la spécification du modèle est loin

[8] Un résidu correspond à l'erreur de prédiction commise. Il peut être interprété comme une perturbation sur les données, qui fait que le modèle de prédiction n'est pas totalement correct.

d'être aisée. Cela suppose de faire un certain nombre d'hypothèses sur les données, qui sont difficilement vérifiables, voire non vérifiées : les résidus suivent une loi normale, la relation entre les variables est linéaire, celles-ci sont indépendantes...

Le choix des variables à prendre en compte est lui aussi délicat, car cela se fait sous le prisme d'une certaine théorie. Comme nous l'avons évoqué précédemment, considérer une théorie plutôt qu'une autre revient à faire un choix hautement subjectif qui peut être facilement sujet à controverse. De telles controverses sont fréquentes, en sciences politiques par exemple.

À propos de l'origine des guerres civiles par exemple, certains, à la suite des travaux de Moore et Gurr (1998); Gurr et Harff (1998) mettent en avant l'importance des velléités de rébellion, motivées par les discriminations subies par un groupe minoritaire. D'autres, au contraire, considèrent que ce n'est pas un facteur essentiel, et que ce sont surtout les occasions rendant possible la rébellion qui priment, comme par exemple l'existence de richesses naturelles pouvant être détournées afin de financer la rébellion (Collier et Hoeffler, 2004; Fearon, 2005).

Pour chacune de ces théories, un modèle spécifique de régression doit être construit et validé expérimentalement. Dans un outil d'aide à la détection de guerres civiles, opter pour l'un de ces modèles nous fait courir le risque d'avoir choisi un cadre théorique en contradiction avec celui de l'analyste, ce qui peut ôter toute crédibilité à notre outil. On peut défendre qu'il serait salutaire d'ouvrir l'expert à d'autres façons de penser, mais c'est alors le choix d'un cadre théorique en accord avec le sien qui pose problème, car cela le conforterait peut-être à tort dans son analyse. Les chercheurs de ce domaine ne sont pas véritablement confrontés à ce problème, car ils utilisent majoritairement la régression comme outil de validation d'hypothèses théoriques. Ils cherchent avant tout à mettre en évidence des facteurs permettant d'expliquer un phénomène comme par exemple l'émergence de la violence intra-étatique. Ils prétendent rarement que le modèle auquel ils sont parvenus est un modèle qui, tel quel, pourra détecter le phénomène en question[9].

Pour pallier le problème de la spécification du modèle, il est possible de sélectionner automatiquement les variables explicatives et donc de se passer du recours à une théorie particulière. Dans le domaine des conflits armés intra-étatiques cette méthode a été mise en place par l'une des plus importantes équipes travaillant sur le sujet, à la demande du gouvernement américain, la *State Failure Task Force* (Goldstone *et al.*, 2000). Aujourd'hui cette équipe se nomme *Political Instability Task Force*. L'ancienne désignation étant encore très répandue dans la littérature, c'est elle que nous utiliserons par la suite. Le projet du même nom a pour but de parvenir à un modèle de prédiction fiable de l'occurrence de crises intra-étatiques. En ne retenant que cinq variables (ouverture du marché, mortalité infantile, population, conflits dans les états voisins et niveau de démocratie), le modèle appris arrive à prédire correctement 72% des crises de leur base de test (135 cas), ce qui est assez prometteur.

Au regard de notre projet visant à aider un expert en veille stratégique, l'inconvénient majeur de cette approche, outre la question des hypothèses sur lesquelles repose le modèle, reste cependant que les résultats obtenus sont difficilement compréhensibles par un non-initié des techniques de régression. Il est vrai que l'étude des paramètres b_i permet de comprendre le rôle de chaque variable explicative v_i vis-à-vis de la variable dépendante y. Mais cette étude peut difficilement être faite par un non-spécialiste. Il faut en effet savoir

[9] La simulation multi-agents est également utilisée dans cette optique de validation d'hypothèses théoriques, en particulier en sciences politiques (Epstein, 2002; Situngkir, 2004; Caselli et Coleman, 2006). Ne constituant pas à proprement parler une méthode d'évaluation des risques nous n'avons pas jugé utile de lui consacrer un paragraphe.

à quoi correspondent exactement ces coefficients et donc connaître le type de régression utilisé, ainsi que les hypothèses qui ont été faites. Aussi est-ce le modélisateur lui-même qui interprète les résultats. Le découplage que nous cherchons entre modélisation et analyse des résultats est donc peu évident avec de telles techniques.

Si tous les exemples que nous avons pris jusqu'à présent sont tirés du domaine des sciences politiques, cette technique n'en reste pas moins générique et applicable dans bien d'autres domaines. En économie par exemple, Galindo et Tamayo (2000) comparent différentes méthodes d'évaluation du risque de crédit, dont le *Probit*, un modèle de régression non linéaire, qu'ils considèrent comme une technique statistique classique.

1.2.3.3 Classification supervisée

Plutôt que de s'intéresser directement à la probabilité d'occurrence $\pi = P(y = 1)$ d'un événement y, comme le font les techniques de régression, on peut imaginer que l'on s'occupe en premier lieu de savoir si cet événement se produira ou non. Nous mettons ainsi l'accent sur la prévision de y plutôt que sur celle de π, et donc sur la détection de crises plutôt que sur l'estimation chiffrée du risque. Cela est tout à fait compatible avec notre objectif centré sur l'aide à l'anticipation des crises.

Nous avons deux classes, 1 si l'événement y a bien lieu et 0 sinon. Le changement de point de vue que nous venons d'introduire implique donc le passage de la régression à la classification supervisée. Ceci peut se faire simplement en prenant un seuil au-delà duquel on estime que π est suffisamment élevée pour pouvoir considérer que l'événement se produira. Dans les problèmes bi-classes, toutes les techniques visant à estimer la probabilité conditionnelle de y sachant la valeur des variables descriptives v_i, peuvent être vues comme des classifieurs (le classifieur bayésien naïf par exemple). Leur principe est le suivant :

- Estimer $\pi = P(y = 1|v_1, v_2, ..., v_n)$
- Règle de décision :

$$y = \begin{cases} 1 & \text{si } \pi > seuil, \\ 0 & \text{sinon.} \end{cases}$$

De façon plus générale, lorsque le problème est multiclasse, les techniques probabilistes de classification reposent sur la démarche suivante :

- Estimer $\pi^k = P(y = k|v_1, v_2, ..., v_n) \, \forall k = 1..K$
- Règle de décision (maximum *a posteriori*) : $y = \arg\max_{k=1..K} \left(\pi^k\right)$

Cette règle du maximum *a posteriori* est équivalente à celle que nous avions énoncée dans le cas bi-classe lorsque le seuil choisi est 0.5.

Deviennent également candidates pour détecter un événement à risque toutes les techniques de discrimination et de catégorisation autres que les méthodes purement probabilistes, qui ont été développées en analyse et fouille de données. Ces techniques commencent à être utilisées dans divers types d'applications. Les travaux de Galindo et Tamayo (2000) sur le risque de crédit, que nous avons mentionnés précédemment, comparent différentes méthodes de classification[10] développées au sein de la communauté d'intelligence artificielle, insistant sur l'intérêt qu'il peut y avoir à recourir à de telles techniques.

En sciences politiques, Beck *et al.* (2000) soulignent également ce point. Conscients du problème que pose la spécification d'un modèle de régression, technique la plus cou-

[10] Arbres de décision, réseaux de neurones, k plus proches voisins et Probit, cette dernière étant une technique de régression appliquée en classification.

rante dans leur domaine, ils proposent d'utiliser des réseaux de neurones particuliers : les perceptrons multi-couches (PMC), qu'ils présentent comme des extensions des modèles de régression. Nous avons vu que les modèles logistiques étaient des modèles log-linéaires :

$$P(Y = 1|V) = \frac{1}{1 + \exp(-VB)}$$

ce que l'on peut réécrire

$$Y = Logistique\,(Lineaire\,(V))$$

Il s'agit donc d'une extension du modèle linéaire classique.

La puissance d'approximation des PMC autorise une modélisation bien plus fine des phénomènes non linéaires que la régression logistique. La figure 1.4 illustre l'architecture d'un tel réseau de neurones. Sans rentrer dans les détails, signalons qu'à chaque couche, l'état d'un nœud est calculé en fonction de l'état des nœuds pères sur la couche précédente. La fonction de transition correspond à l'application d'une fonction seuil sur la combinaison linéaire de l'état des nœuds précédents, pondérée par les poids de chaque arc. Une fonction seuil couramment employée est la sigmoïde, qui est une fonction logistique. Entre deux couches successives on applique donc une régression logistique pour évaluer l'état de chaque nœud. Cette opération est ensuite répétée autant de fois qu'il y a de couches dans le réseau. Ainsi, pour un réseau comportant les v_i sur la couche d'entrée, Y comme cellule de sortie et une couche cachée, on peut écrire

$$Y = Logistique\,(Lineaire\,(Logistique\,(Lineaire\,(V))))$$

C'est pour cela que l'on peut dire que les perceptrons multi-couches étendent le principe de la régression logistique. La critique selon laquelle la régression, même logistique, impose une structure peu réaliste à la fonction f recherchée, n'est donc plus valable avec ce nouveau modèle par réseau de neurones. En revanche, l'interprétation des résultats qui était déjà peu évidente pour des non-spécialistes avec les techniques de régression, est encore plus difficile. Conscients de l'importance capitale de l'interprétation et de la faiblesse à cet égard des réseaux de neurones, Beck *et al.* (2000) ont développé des outils graphiques pour faciliter la compréhension du modèle produit. Cela reste cependant une étape lourde et délicate, surtout si celui qui interprète n'est pas le modélisateur.

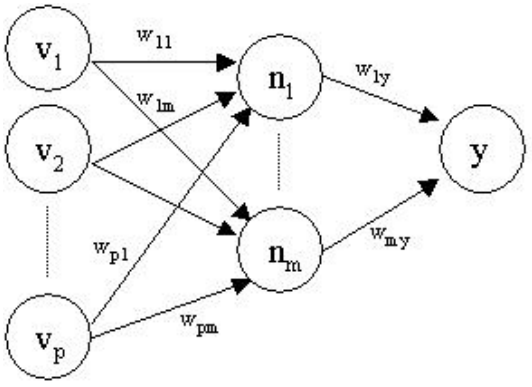

FIG. 1.4 – Exemple de perceptron multi-couches pour l'apprentissage de la dépendance fonctionnelle entre les v_i et la classe y

Une méthode fréquemment employée par les experts humains chargés d'évaluer une situation donnée consiste à la comparer à des situations de référence du passé. Si les contextes

sont suffisamment proches, l'analyste va en déduire qu'il est fort vraisemblable que l'évolution de la situation actuelle sera similaire à celle des situations de référence voisines. Les techniques de raisonnement à partir de cas, comme les k plus proches voisins, s'appuient sur des considérations analogues. Il faut disposer d'une base de cas et rechercher, pour tout nouvel exemple $e \notin \mathcal{E}$, le ou les k exemples $e' \in \mathcal{V}_k(e)$ qui s'en rapprochent le plus parmi ceux de $\mathcal{E}$. $\mathcal{V}_k(e)$ désigne l'ensemble des k plus proches voisins de e, dont la classe ($Classe(e')$) sert à déterminer la classe de e. On prend généralement la classe majoritaire, ce qui s'écrit :

$$Classe(e) = \arg\max_{i=1..K} \left(\left| \left\{ e'/e' \in \mathcal{V}_k(e), Classe(e') = c_i \right\} \right| \right)$$

où $|S|$ est le cardinal d'un ensemble S.

Par exemple Petrark *et al.* (1994) emploient ce type de méthodes pour prédire l'occurrence d'un conflit entre États. Notons que cette technique ne passe pas par l'apprentissage d'un modèle explicite. Ce sont les données du passé qui sont directement utilisées pour anticiper la classe d'un exemple actuel. Ce processus correspond à ce que nous avons appelé **transduction**. La difficulté de cette méthode réside dans le choix d'une distance ou d'une mesure de similarité qui soit efficace et pertinente au regard de la tâche à effectuer. Dans une optique d'aide à la décision, remarquons que les prédictions réalisées sont accompagnées de la liste des cas les plus proches. Ceci est une première étape guidant l'analyste dans la compréhension de la situation, grâce à son expérience. Cette étape est fort utile, mais les facteurs de risque prépondérants et leurs relations ne sont pas mis en évidence.

Les mêmes auteurs, toujours dans le domaine des crises inter-étatiques, ont utilisé des arbres de décision pour effectuer leur classification (Trappl *et al.*, 1996). La figure 1.5 donne un exemple fictif permettant de comprendre la structure de ces arbres, qui sont une représentation graphique du processus de classification. Chaque nœud de l'arbre représente un test sur la valeur d'un attribut, par exemple : « Si $v_i < \alpha$ ». Les arcs sont orientés et étiquetés avec l'un des résultats possibles du test correspondant au nœud père. Enfin les feuilles contiennent l'ensemble des exemples ayant passé les différents tests depuis la racine. Elles sont étiquetées par la classe majoritaire parmi les différentes classes des exemples qui leur sont rattachés.

Un chemin dans l'arbre, de la racine à une feuille, peut donc être interprété comme la conjonction d'un certain nombre de tests (autant qu'il y a de nœuds dans le chemin, sans compter la feuille) aboutissant à une classe donnée (celle qui étiquette la feuille considérée). Chaque chemin est donc une règle et l'ensemble des chemins qui constituent l'arbre forme une base de règles. Ceci constitue un atout non négligeable dans une perspective d'aide à l'analyse. D'une part, les performances des arbres de décision sont en général tout à fait satisfaisantes. D'autre part, la correspondance entre arbres de décision et bases de règles de classification rend les résultats facilement interprétables. À chaque décision du système est associée un chemin dans l'arbre et donc une règle. Cela permet d'identifier immédiatement les facteurs élémentaires de risque. Cette phase est essentielle pour guider les décideurs dans la marche à suivre pour prévenir les conflits potentiels.

Cette présentation des classifieurs utilisés en évaluation des risques ne se veut pas exhaustive. D'autres techniques telles que l'analyse discriminante (Guler *et al.*, 2001) ou encore les machines à vecteurs supports (SVM) (Sepulveda-Sanchis *et al.*, 2002) sont également utilisées. Elles peuvent s'avérer très performantes, comme les SVM par exemple, mais souffrent à l'instar des réseaux de neurones d'un manque de transparence. Les modèles induits ne sont pas aisément analysables par des personnes autres que les modélisateurs.

Cette carence se retrouve dans le classifieur FASE (*Fuzzy Analysis of Statistical Evidence*) développé par Chen (2000) et appliqué par O'Brien (2001) pour prévoir les conflits

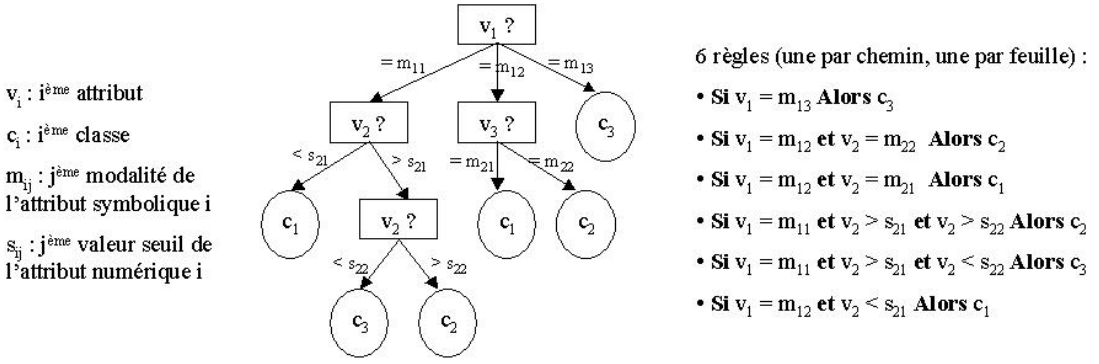

FIG. 1.5 – Un exemple d'arbre de décision

intra-étatiques. Le principe de FASE reprend les idées de l'inférence bayésienne, à ceci près que les incertitudes ne sont pas modélisées dans le cadre probabiliste classique, mais dans celui de la logique possibiliste, mêlant logique floue et théorie des possibilités. Ce changement de paradigme permet de gagner en robustesse, de ne pas être trop sensible à la présence de données erronées ou manquantes et de se passer des hypothèses souvent trop restrictives des modèles probabilistes.

La classification automatique supervisée est bien adaptée pour extraire des modèles visant à anticiper l'occurrence d'un phénomène. Ces modèles sont appris sur les cas du passé et sont ensuite appliqués sur les cas du présent pour savoir si l'événement qui nous intéresse va se produire ou non. Cette phase de détection est essentielle dans le cadre de l'analyse et de la gestion du risque. Cependant pour rester dans le cadre précis de l'évaluation des risques, rappelons que nous souhaitons quantifier l'incertitude liée à cette occurrence.

Les techniques de régression ou les classifieurs bayésiens le font intrinsèquement avant de procéder à la classification. Cela n'est pas le cas de toutes les autres méthodes. Il faut alors dans ces cas-là rajouter un post-traitement pour quantifier l'incertitude liée à la décision de classification et ainsi construire un indice de risque. Pour résumer, l'idée est de se concentrer sur la classification pour en déduire un indice de risque, alors que les techniques comme la régression construisent cet indice avant d'en déduire la classe à affecter. La figure 1.6 synthétise ces deux mécanismes.

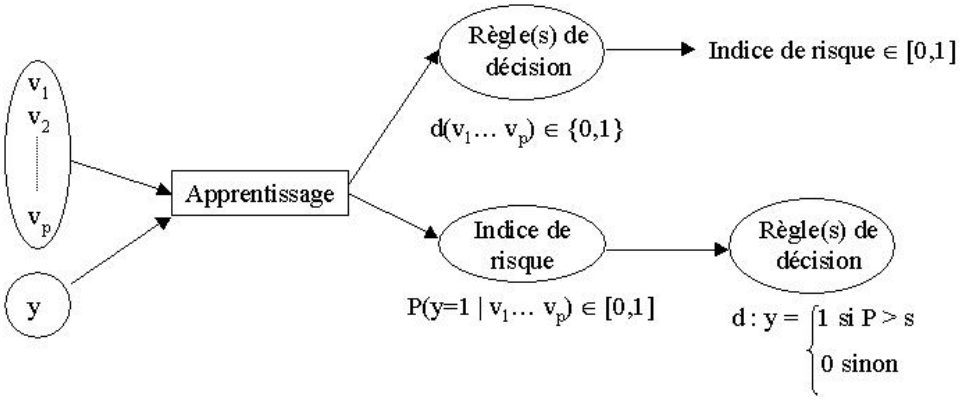

FIG. 1.6 – Apprentissage de classifieurs : deux façons de parvenir à un modèle d'évaluation des risques

1.3 Complémentarité des deux approches

De l'état de l'art que nous venons de dresser il est possible de faire ressortir les traits communs partagés par les diverses techniques introduites et ainsi de définir une méthodologie générique d'évaluation des risques. Celle-ci peut être décomposée en cinq phases :

1. Identification des facteurs élémentaires de risque (manuelle ou automatique)
2. Évaluation de ces facteurs (données chiffrées disponibles ou jugements d'experts)
3. Construction d'un modèle reliant les différents facteurs au risque global (agrégation ou apprentissage automatique)
4. Application du modèle ⇒ évaluation du risque
5. Synthèse claire des résultats pour préparer la gestion de risque

Sans chercher à tendre vers l'exhaustivité, nous avons essayé, au travers de la présentation de quelques-unes des principales méthodes d'évaluation des risques, d'en dégager les caractéristiques essentielles, du moins au regard de notre intérêt pour l'aide à la décision. Le tableau 1.2 récapitule l'ensemble de ces caractéristiques pour les deux types d'approches que nous avons distingués.

- Les **approches qualitatives** visent à attribuer une appréciation qualitative aux deux dimensions principales du risque : incertitude et sévérité. Elles présentent l'intérêt d'intégrer l'expertise humaine dans le processus d'estimation, bénéficiant ainsi de l'expérience de spécialistes. De ce fait elles peuvent être appliquées quelle que soit la quantité de données dont on dispose pour effectuer l'analyse de risque. Elles produisent également des résultats clairs qui permettent une bonne compréhension de la situation, qui va de paire avec l'identification de leviers d'actions dans une optique de prévention.
  En revanche, le recours à l'expertise présente des contre-parties et non des moindres : la subjectivité de l'analyse et le nombre limité d'informations qui peuvent être prises en compte du fait de la saturation des capacités cognitives des experts. De plus de nombreux spécialistes doivent travailler ensemble, ce qui est long, coûteux et exige la mise en place d'un protocole strict de recueil de jugements et d'agrégation de ces jugements.
- Les **approches quantitatives** se focalisent sur la modélisation quantitative du risque, cherchant à déterminer la relation fonctionnelle qui le lie à un certain nombre de facteurs élémentaires de risque. Elles nécessitent de disposer de bases de données importantes, ce qui peut poser problème. Du fait de l'automatisation de tout ou partie du processus d'estimation, cette faiblesse devient cependant un avantage lorsque de telles bases sont accessibles. Elles peuvent tenir compte de toute l'information disponible et fournir ainsi des estimations de risque robustes et « objectives ». Cependant nous avons vu que l'automatisation, si elle permet le traitement d'un plus grand volume de données, s'accompagne d'une perte de lisibilité des modèles construits.

Au vu du tableau récapitulatif 1.2, il apparaît que les deux méthodes sont plus ou moins adaptées en fonction de la quantité de données disponibles et de la part que l'on souhaite accorder à la subjectivité. Sur la figure 1.7, qui précise le positionnement des deux approches selon ces deux axes, on peut observer qu'il existe un certain nombre d'applications pour lesquelles les deux approches sont envisageables.

TAB. 1.2 – Principales Caractéristiques des techniques d'évaluation des risques

	Atouts	**Faiblesses**
Qualitatif	Intégration de l'expertise humaine	Subjectivité de l'analyse
	Clarté des résultats	Coût de mise en œuvre
	Peu de données nécessaires	Impossibilité de traiter beaucoup de données
Quantitatif	Possibilité de traiter beaucoup de données	Beaucoup de données nécessaires
	Objectivité de l'analyse	Opacité des modèles

Plutôt que de les mettre en concurrence, afin de déterminer laquelle des deux est la plus appropriée, il nous semble plus intéressant de mettre l'accent sur leur complémentarité en essayant de voir ce qu'elles peuvent respectivement apporter à l'analyse de risque. Pour l'aide à la détection de crises il apparaît en effet important de pouvoir dans un premier temps traiter de grands volumes de données, le plus *objectivement* possible. Mais face à la complexité des phénomènes étudiés, il serait bon, dans un second temps, de pouvoir intégrer les jugements qualitatifs d'experts du domaine.

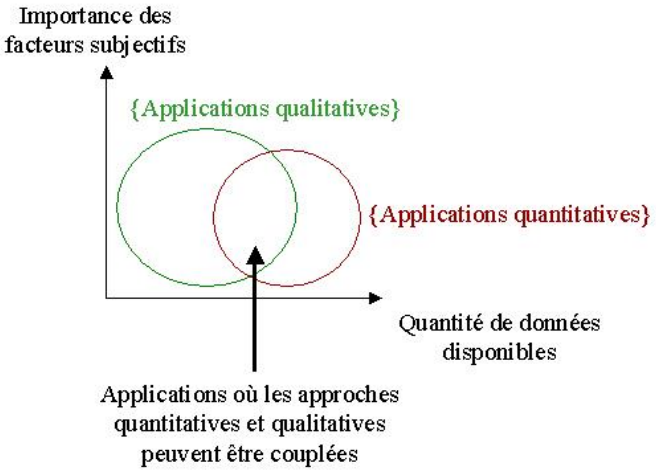

FIG. 1.7 – Positionnement des techniques quantitatives et qualitatives d'évaluation des risques selon la quantité de données disponibles et l'importance des facteurs subjectifs

Notons que nombre des techniques que nous avons abordées précédemment peuvent être considérées comme hybrides. Des méthodes comme l'AMDEC, pourtant essentiellement qualitatives, sont actuellement associées à des modules d'estimation numérique automatiques. À l'inverse, des méthodes, en apparence purement quantitatives comme la régression ou les réseaux bayésiens, s'appuient sur une interprétation qualitative de la réalité pour construire leurs modèles (structure des réseaux bayésiens, choix du mode de régression et des variables explicatives).

Le type de couplage que nous souhaitons réaliser est cependant quelque peu différent. En effet, nous ne voulons pas utiliser l'expertise, subjectivement biaisée, pour guider la construction automatique de notre modèle de détection. Nous souhaitons que cette phase d'automatisation reste la plus neutre possible. En revanche, nous espérons que l'intégration de l'expérience et de l'intuition de spécialistes permettra d'affiner le modèle appris automatiquement.

Chapitre 2

Un premier modèle d'évaluation des risques

Dans le chapitre 1, au fil de la revue critique des différentes méthodes d'évaluation des risques, nous avons identifié leurs caractéristiques ainsi que celles que notre futur outil de détection des crises se doit d'avoir. Au regard de cette analyse, nous allons voir dans ce chapitre pourquoi nous avons choisi de modéliser les risques de crise au moyen d'arbres de décision flous. Nous décrirons ensuite les premières expérimentations que nous avons menées, ce qui permettra de justifier empiriquement l'intérêt de l'approche que nous proposons, et de mettre en évidence les faiblesses de ce premier modèle. Nous présenterons alors les améliorations que nous lui avons apportées ainsi que les axes de recherche que nous avons identifiés.

2.1 *Salammbô* : construction d'arbres de décision flous

2.1.1 Pourquoi *Salammbô* ?

Afin de pouvoir sélectionner une technique d'évaluation des risques, parmi toutes celles qu'on peut envisager, il nous faut repérer ce qui fait leur force et leur faiblesse et voir si cela est compatible avec nos objectifs et nos contraintes. Rappelons que nous nous plaçons dans le cadre de l'aide à la détection de crises à moyen terme. La méthode proposée doit pouvoir traiter de grandes masses de données numériques et symboliques afin de produire un indice de risque global. La plus-value apportée par notre système se situe précisément dans cette capacité à tenir compte d'un grand nombre de facteurs. Ceci nous amène à privilégier les techniques numériques et plus particulièrement les techniques automatiques.

Nous chercherons à apprendre automatiquement un modèle de prévision des crises le plus performant possible, c'est-à-dire qui approxime au mieux la relation fonctionnelle sous-jacente, supposée relier les facteurs de risque élémentaires au risque global. Cette relation étant inconnue, il nous faudra préciser clairement la façon dont nous évaluerons notre modèle. Nous reviendrons une première fois sur ce point à la section 2.3, avant de le détailler à la section 10.3.1. Précisons cependant dès maintenant que cela se fera de manière empirique, en appliquant notre modèle sur des données réelles. Ne pouvant disposer de mesures fiables du risque pour des exemples réels, nous contrôlerons la qualité de notre modèle sur une tâche de prédiction des crises. L'occurrence d'une crise est en effet une information factuelle qui est, elle, accessible. Ainsi nous nous plaçerons dans le cadre de la classification supervisée. Mais cela ne disqualifie pas pour autant les techniques de régression qui comme nous l'avons indiqué à la section 1.2.3.3 peuvent facilement être considérées comme

des classifieurs, par l'introduction d'une règle de décision simple consistant à seuiller la probabilité de crise estimée.

Chercher le modèle le plus performant, au sens d'un critère que nous préciserons plus loin, revient à considérer que nous sommes face à un problème d'optimisation. Étant donné que nous souhaitons construire un modèle, aussi objectif et transparent que possible, nous pouvons affiner notre caractérisation et dire qu'il s'agit d'un problème d'optimisation sous contraintes. Ce sont ces contraintes qui vont orienter notre recherche de la technique la plus appropriée.

- **Objectivité** : afin de ne pas entrer en conflit avec la subjectivité de l'analyste, nous refusons de partir d'une théorie particulière et partisane d'explication du phénomène à modéliser. Ainsi, sont exclues les méthodes qui reposent sur ce principe, soit au niveau de la sélection des paramètres à prendre en compte, soit au niveau de la forme même du modèle. La régression sans sélection automatique des variables, les systèmes experts sans apprentissage des règles d'inférence ou encore les approches graphiques sans détermination automatique de la structure du graphe sont donc inadéquates.
- **Transparence** : afin d'être accepté par l'utilisateur, l'outil que nous proposons doit produire des résultats qui puissent être remis en cause. De plus, dans une optique préventive, il faut pouvoir faire ressortir clairement les faiblesses du système étudié. Le modèle construit doit donc être aussi intelligible que possible et interprétable facilement par un analyste autre que le modélisateur. Ce dernier point nous conduit à rejeter les techniques de régression y compris celles qui procèdent en amont à une sélection automatique des variables explicatives. Elles sont envisageables uniquement si l'utilisateur est également le modélisateur. Les classifications par réseaux de neurones, SVM ou analyse discriminante sont, elles, difficilement interprétables, même par ceux qui les mettent en place et sont donc également rejetées.

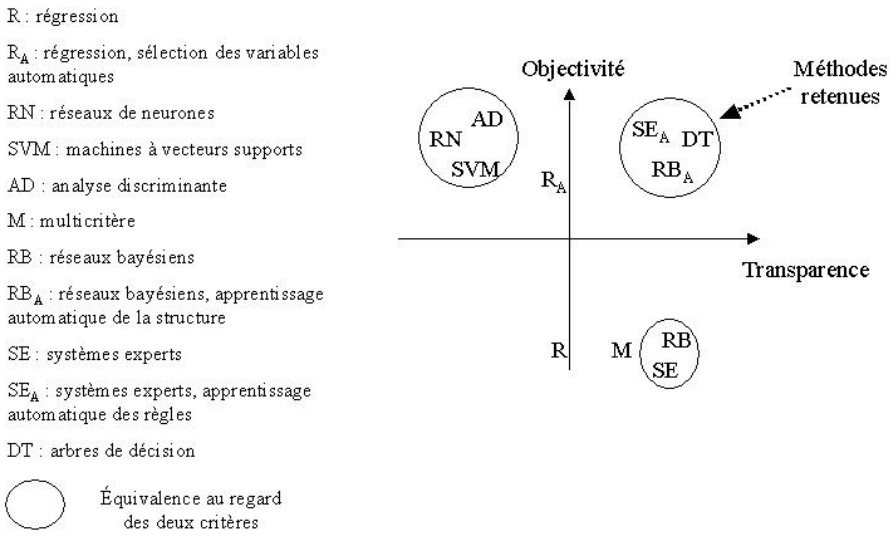

FIG. 2.1 – Répartition des différentes techniques d'évaluation des risques, selon l'objectivité et la transparence des modèles construits

Comme l'illustre la figure 2.1, les réseaux bayésiens dont la structure est apprise automatiquement ainsi que les systèmes à base de connaissances dont la base de règles est construite automatiquement ou encore les arbres de décision respectent bien les contraintes

que nous nous sommes fixées et semblent donc correspondre à notre besoin.

Les systèmes experts ont été rangés dans les approches qualitatives, et peuvent donc paraître mal adaptés pour traiter de grandes bases de données. Lorsque les règles ne sont pas recueillies auprès d'experts mais apprises automatiquement, intégrer de grandes masses d'informations dans le modèle n'est plus un problème. L'apprentissage de la structure des graphes causaux étant encore loin d'être satisfaisant, nous avons décidé de nous orienter vers les systèmes à base de règles.

Afin de réaliser l'apprentissage de ces règles, nous avons finalement retenu les arbres de décision. Certes la structure arborescente limite le type des règles que l'on peut trouver, mais il s'agit d'une technique efficace qui évite l'explosion du nombre de règles, phénomène que l'on observe avec d'autres méthodes comme l'apprentissage de règles d'associations ou encore celle qui est basée sur les algorithmes génétiques (Spanos *et al.*, 1999). De plus la base de règles peut se visualiser directement sous la forme d'un arbre, ce qui rend l'ensemble du modèle plus facilement compréhensible. Ajoutons enfin que la construction d'arbres de décision se base sur une recherche des variables les plus discriminantes, ce qui leur permet de sélectionner, sans *a priori*, uniquement à partir des données, les paramètres à inclure dans le modèle.

Nous avons opté pour *Salammbô*, logiciel de construction d'arbres de décision flous développé par Marsala (1998). Comme l'a montré Marsala, dans le cas flou, de petites fluctuations dans les données d'entrée ne provoquent pas de changement brutal de classe (continuité de la décision). Les arbres flous sont donc plus robustes que les arbres non flous, c'est-à-dire moins sensibles au bruit.

Olaru et Wehenkel (2004) ont par ailleurs montré que l'introduction du flou dans la construction des arbres de décision permettait de réduire la variance des modèles, ce qui traduit également une plus grande robustesse. Outre cette plus grande souplesse d'analyse, les arbres flous, par l'introduction de variables linguistiques, produisent des règles plus facilement manipulables par les individus, ce qui est un atout important pour un système d'aide à la décision.

Avant de décrire ce logiciel autour duquel nous avons bâti notre outil d'aide à la détection des crises, synthétisons notre approche. Pour cela, reprenons la méthodologie générique d'évaluation des risques introduite à la section 1.3 et voyons comment notre modèle l'instancie.

1. Identification des facteurs de risque élémentaires : *sélection automatique des variables discriminantes*
2. Évaluation de ces facteurs : *bases de données disponibles*
3. Construction d'un modèle reliant les différents facteurs au risque global : *apprentissage automatique d'arbres de décision flous, qui peuvent être vus comme des bases de règles*
4. *Inférence déductive à partir des règles apprises* $\Rightarrow$ évaluation du risque
5. Synthèse claire des résultats pour préparer la gestion de risque : *à chaque décision est associé un ensemble de chemins dans l'arbre, c'est-à-dire, un ensemble de règles floues, facilement interprétables*

2.1.2 Caractéristiques de *Salammbô*

Salammbô est un outil de construction de classifieurs par induction d'arbres de décision flous. L'algorithme d'induction utilisé est descendant et fait partie de la classe des algorithmes TDIDT (Top Down Induction of Decision Trees). Il peut être vu comme une version

floue de l'algorithme C4.5 de Quinlan (1986). Aussi allons-nous rappeler brièvement le principe de la construction descendante d'arbres de décision, ce qui nous permettra d'aborder ensuite les spécificités de *Salammbô* au regard de cette méthode générique.

Pour des raisons de cohérence, nous préférons reprendre les notations de la section 1.2.3.1, plutôt que les notations classiquement adoptées pour présenter les arbres de décision. Ainsi nous notons $\mathcal{E}$ la base d'exemples d'apprentissage et $\mathcal{V}$ l'ensemble des attributs qui permettent de décrire les exemples de $\mathcal{E}$.

2.1.2.1 Principales caractéristiques des algorithmes TDIDT

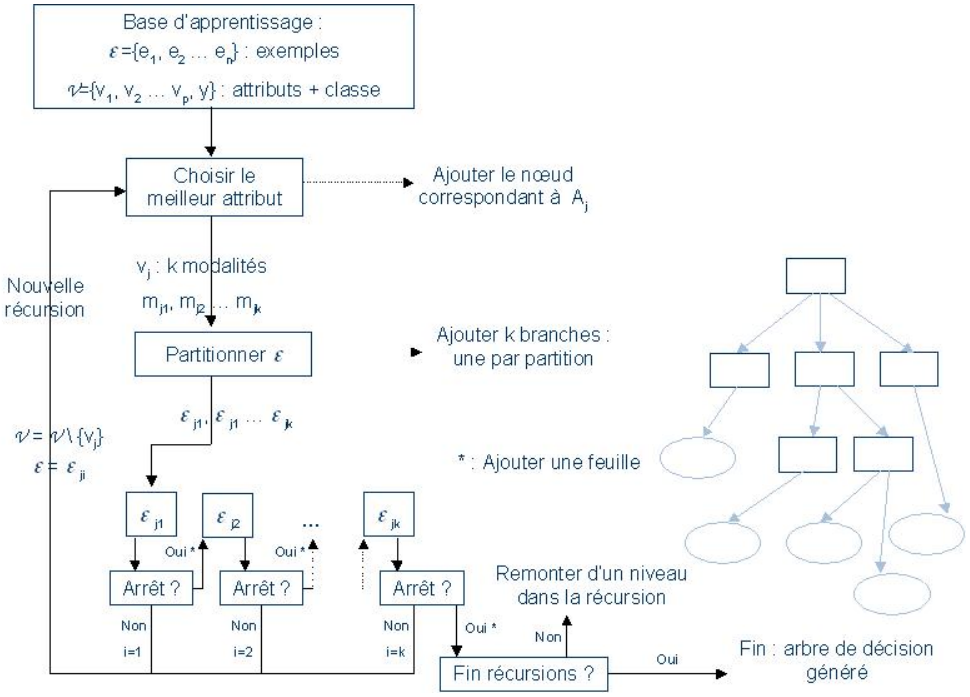

FIG. 2.2 – Processus de construction d'arbres de décision

Les algorithmes TDIDT sont qualifiés d'approches descendantes, car l'arbre induit est construit en commençant par la racine et en « descendant » jusqu'aux feuilles. La figure 2.2 en décrit le processus pour des attributs symboliques. La construction se fait de manière récursive. On commence par choisir le meilleur attribut, au sens d'un certain critère à préciser. Il sera à la racine de notre arbre. Ensuite il faut partitionner la base d'exemples en fonction des modalités de cet attribut. Pour chacune des classes (sous-bases d'exemples) de la partition on recommence la procédure en ne tenant plus compte de l'attribut que l'on vient de sélectionner, à condition qu'un certain critère d'arrêt ne soit pas vérifié.

L'algorithme s'arrête une fois que toutes les récursions sont terminées. Notons cependant que ce processus doit être modifié dans le cas d'attributs numériques pour lesquels il faut rajouter une étape de discrétisation. Un attribut numérique pourra en revanche être considéré à nouveau dans les récursions suivantes afin de pouvoir affiner sa discrétisation.

Sur la figure 2.2, trois étapes essentielles de la construction d'arbres de décision se distinguent. Ce sont autant de degrés de liberté sur lesquels on peut jouer pour définir un nouvel algorithme TDIDT : sélection d'un attribut, partitionnement, arrêt. Dans le cas des attributs numériques, la phase de discrétisation correspond à un quatrième degré de liberté.

Choix d'un attribut Qu'entend-on par « meilleur » attribut ?

On se base sur une mesure de *discrimination* pour trier les attributs. Elle permet d'évaluer la capacité d'un attribut à discriminer les différentes classes. Le « meilleur » attribut sera donc celui qui y parvient le mieux.

De nombreuses mesures ont été élaborées, chacune correspondant à une conception particulière de la notion de discrimination. Les plus répandues reposent sur une quantification de l'impureté d'une base d'exemples, au regard de la variable *classe* et au sens de la théorie de l'information. L'impureté d'une base correspond au degré de mélange entre les différentes classes. Elle est d'autant plus importante que les exemples sont équidistribués dans les différentes classes et d'autant plus faible qu'une même classe regroupe tous les exemples. L'index de Gini et l'entropie de Shannon sont à la base des algorithmes de construction d'arbres de décision les plus connus et les plus utilisés, respectivement CART et ID3 puis son extension C4.5.

$$\begin{aligned} Gini\left(\mathcal{E}\right) &= 1-\sum_{i=1}^{K} P\left(y=c_i\right)^2 \\ Entropie\left(\mathcal{E}\right) : I\left(\mathcal{E}\right) &= -\sum_{i=1}^{K} P\left(y=c_i\right) log_2\left(P\left(y=c_i\right)\right) \end{aligned}$$

Les probabilités de classe $P\left(y=c_i\right)$ sont estimées par maximum de vraisemblance. Il s'agit de la fréquence du nombre d'exemples de $\mathcal{E}$ qui appartiennent à cette classe c_i.

Le pouvoir discriminant d'un attribut est alors calculé comme la différence entre l'impureté de la base d'exemples courante et l'impureté moyenne de l'ensemble des bases d'exemples résultant de la partition induite par l'attribut. Autrement dit, on regarde quel serait le gain d'information si l'on décidait de partitionner la base courante à partir de cet attribut. On obtient ainsi le gain d'information, mesure utilisée par Quinlan et introduite par Picard (1972). Sans rien présumer de la façon dont est construite cette partition, le gain G pour une base d'exemples $\mathcal{E}$ et un attribut v_i s'écrit :

$$G\left(\mathcal{E}, v_i\right) = I\left(\mathcal{E}\right) - E\left[I\left(\mathcal{E}|v_i, S\right)\right]$$

où E est l'espérance mathématique, lorsque l'on considère la variable aléatoire v_i. $(\mathcal{E}|v_i, S)$ désigne un ensemble d'exemples (sous-ensemble de $\mathcal{E}$), déterminé par la valeur de v_i en fonction d'une certaine stratégie de partitionnement S.

Stratégie de partitionnement Quelle partition de la base d'exemples créer, une fois qu'un attribut a été sélectionné ?

Une telle stratégie n'est appliquée qu'une fois un attribut retenu, mais elle est également utilisée de manière prospective pour évaluer le pouvoir discriminant des différents attributs. Pour les attributs symboliques, la stratégie consistant à créer une sous-base d'exemples par modalité est la seule qui soit appliquée. Le traitement des attributs numériques ne diffère que par l'ajout d'une phase de discrétisation. Une fois que celle-ci est accomplie l'attribut est considéré comme symbolique.

Maintenant que le paramètre S de notre équation précédente est connu, nous pouvons réécrire l'expression du gain, ou plutôt de sa version normalisée : le *gain ratio* GR, qui tient compte du nombre de modalités de l'attribut considéré et qui est la mesure de discrimination de l'algorithme C4.5 alors que le gain n'est utilisé que dans ID3.

Critère d'arrêt Quand décide-t-on d'arrêter une récursion ?

Une première approche naïve consiste à mettre un terme au partitionnement récursif d'une base d'exemples lorsque tous les éléments de la base appartiennent à une même classe. Celle-ci sera alors utilisée pour étiqueter la feuille correspondante. Cette approche présente cependant un inconvénient majeur. Des feuilles ne contenant que très peu d'éléments auront en effet tendance à être créées.

De plus, les chemins construits risquent de ne pas correspondre à l'apprentissage d'une régularité observée que l'on généraliserait, mais plutôt à la copie des particularités de la base d'apprentissage, ce qui s'accompagne de mauvaises performances sur un échantillon de test indépendant. On parle alors de *sur-apprentissage* ou *over-fitting*. Pour éviter ce phénomène, on peut procéder à un élagage des arbres appris en supprimant les branches trop spécifiques de la base d'apprentissage. On peut également mettre en place des critères d'arrêt qui limitent la multiplication des ramifications. Nous en verrons deux exemples lorsque nous préciserons les caractéristiques de *Salammbô*.

Discrétisation Comment passer du numérique au symbolique ?

Les variables continues ne sont pas traitables telles quelles par un algorithme classique de construction d'arbres de décision. Il est difficilement imaginable de mettre en place des tests d'égalité sur un domaine continu. En revanche, des tests d'inégalités, ou d'appartenance à des intervalles, sont tout à fait envisageables.

Pour cela il convient d'établir une partition de l'ensemble continu sur lequel est défini l'attribut. Chaque valeur de l'attribut peut alors être remplacée par l'intervalle de la partition dans laquelle elle se situe. Ainsi l'attribut initialement continu peut être considéré comme discret, les intervalles de la partition constituant ses modalités.

Sans rentrer dans les détails des diverses méthodes, signalons simplement que l'on distingue généralement les techniques dites supervisées qui utilisent l'information de classe pour construire la partition, tandis que les techniques non supervisées n'ont recours qu'à la distribution de l'attribut à discrétiser. Pour les arbres de décision, étant donné que l'on souhaite obtenir une partition qui soit la plus en phase possible avec notre variable *classe*, ce seront plutôt les techniques supervisées qui seront mises en œuvre.

Notons enfin que la discrétisation, dans le cas des arbres de décision, peut tirer profit du caractère récursif de la construction. En effet il est possible de ne pas créer une partition par attribut, une fois pour toutes, en amont de la construction de l'arbre. Au contraire, on peut imaginer que l'on procède à une première discrétisation grossière, que l'on raffinera éventuellement plus tard. Cet aspect incrémental de la discrétisation est intéressant, car il est toujours très difficile de faire une discrétisation définitive *a priori*, puisqu'on ne connaît pas le nombre d'intervalles à créer.

2.1.2.2 Spécificités de *Salammbô*

Dans le logiciel *Salammbô*, le flou est intégré à trois niveaux de l'algorithme, dans trois étapes différentes. C'est sur ces points précis que *Salammbô* diffère des algorithmes classiques d'induction d'arbres de décision :

- discrétisation des variables continues (induction)
- mesure du pouvoir discriminant des attributs (induction)
- utilisation floue de l'arbre induit pour classer de nouveaux exemples (déduction)[1].

[1] *Salammbô* est un outil générique qui peut également être utilisé classiquement, sans tenir compte du flou introduit dans la phase de discrétisation.

Dans *Salammbô* la discrétisation est faite à la volée, à chaque fois que l'on a besoin d'évaluer une variable. Un algorithme de filtrage inspiré des techniques de la morphologie mathématique est utilisé pour créer une partition floue de l'espace d'entrée en deux intervalles. L'intérêt d'introduire du flou à ce niveau est d'assouplir les tests que l'on crée sur un attribut continu. On pourra se reporter à (Marsala et Bouchon-Meunier, 1996) pour plus de précisions sur cette méthode.

Le second point concerne la mesure de discrimination. Du fait de la discrétisation utilisée, cette mesure doit pouvoir prendre en compte des attributs dont les modalités sont floues. Le *gain ratio* est défini à partir de l'entropie d'événements flous, qui généralise l'entropie classique de Shannon. Les deux entropies ne diffèrent que dans la définition de la probabilité utilisée. L'entropie classique repose sur une estimation fréquentiste de la probabilité d'un événement, tandis que dans le cas flou nous avons $P^*(\Gamma) = \sum_{i=1}^{n} \mu_\Gamma(e_i) P(e_i)$ où μ_Γ désigne la fonction d'appartenance associée à l'ensemble d'événements $\Gamma = \{e_i\}_{i=1..n}$. Ces événements correspondent aux exemples de la base d'apprentissage. On les suppose généralement équidistribués et donc on a $P(e_i) = \frac{1}{n}$

Enfin, le formalisme de la logique floue est mis en œuvre pour effectuer les déductions permettant de classer de nouveaux exemples. Chaque test de l'arbre fait référence à une partition floue. Aussi, pour chaque entrée, tous les chemins de l'arbre sont-ils activés, contrairement à ce qui est fait dans les cas non flous. Toutes les classes sont donc partiellement reconnues avec un certain degré. La règle de décision consiste simplement à affecter à l'exemple la classe maximisant ce degré, mais d'autres types d'agrégation sont envisageables. Dans notre contexte applicatif, nous avons certes besoin de pouvoir classer des exemples, mais nous souhaitons également leur attribuer un indice de risque reflétant l'incertitude liée à l'occurrence d'une crise. Ceci peut être fait en agrégeant les degrés de reconnaissance de chacune des classes.

Dans le cas classique, un seul chemin est actif à la fois. De plus les tests non flous introduisent des ruptures qui font que de faibles variations sur les paramètres d'entrée peuvent entraîner l'activation d'un chemin différent et donc une conclusion totalement différente. Dans le cas flou, tous les chemins sont actifs. Les faibles variations en entrée modifient les degrés de reconnaissance de chacune des classes et ce de manière continue. C'est pour cela que nous qualifions de robuste l'algorithme d'induction de *Salammbô*.

Nous avons vu pourquoi nous avions opté pour *Salammbô*, d'un point de vue théorique. Voyons maintenant au travers des résultats expérimentaux si ce choix se justifie d'un point de vue pratique, si les performances de *Salammbô* dans un contexte d'évaluation des risques sont suffisamment satisfaisantes.

2.2 Description des données

La prévision des conflits armés intra-étatiques constitue l'application sur laquelle nous nous sommes focalisé pour valider notre approche méthodologique. Sachant que nous cherchons à prévoir une tendance, à estimer un indice de risque à moyen terme par pays et non à déclencher des alertes en fonction de l'évolution quotidienne de la situation d'un ensemble donné de pays, il faut que nous disposions en entrée de notre système, pour chaque pays, d'indicateurs structurels reflétant la situation de ce pays. Ces indicateurs doivent évidemment être estimés avant l'année à partir de laquelle nous essayons de prédire l'occurrence des conflits, sans quoi il ne s'agit plus d'un problème de prévision.

Le choix et le recueil des indicateurs est une phase essentielle de la construction de notre modèle. Nous avons tout d'abord identifié un certain nombre de domaines d'inquiétude, en

nous inspirant de la démarche du CIFP (Ampleford *et al.*, 2001), avant de chercher pour chacun d'eux des descripteurs élémentaires.

Nous ne voulons recourir, autant que faire se peut, qu'à des indicateurs objectifs et non pré-agrégés. En effet, utiliser des indicateurs tels que le niveau de corruption ou de démocratie, qui résultent d'un processus d'agrégation dont les modalités ne sont pas clairement précisées, serait en contradiction avec notre souci d'objectivité. Le processus d'agrégation intègre le biais des experts, dont l'impact est impossible à quantifier. Cela impliquerait de plus que l'alimentation de notre système soit dépendante du travail des experts ayant réalisé l'agrégation, ce qui nuirait à son automatisation. Enfin, il est difficile de justifier l'utilisation de tels indicateurs dans des sytèmes de prévision basées sur l'extraction de régularités dans les données du passé. Comme le notent Linder et Santiso (2002), leur mode d'agrégation varie d'année en année.

Nous avons essayé de faire en sorte que les indicateurs retenus couvrent au mieux les domaines suivants :

- démographie et migrations
- écologie et agriculture
- économie et finance
- infrastructures et moyens de communication
- ressources naturelles et énergie
- développement humain, santé et éducation
- puissance militaire
- régime politique
- hétérogénéité de la population (diversité religieuse, ethnique, linguistique)
- histoire et géographie du pays
- contexte international

Nous nous sommes contenté des ressources librement disponibles sur Internet, en nous focalisant sur la période 1980-1998. Aussi tous ces domaines ne sont-ils pas également bien couverts. La plupart des indicateurs sont issus des ressources de la Banque mondiale[2], du World Factbook de la CIA [3], du Haut Commissariat aux Réfugiés de l'ONU : UNHCR [4]. Notre approche se plaçant dans le cadre de la classification automatique supervisée, nous avons besoin de connaître la classe de chaque pays. Au vu des informations disponibles en sources ouvertes, nous avons choisi de définir la variable binaire classe de la façon suivante :

$$classe\,(pays) = \begin{cases} crise & \text{si } pays \text{ a connu des affrontements armés,} \\ & \text{mettant aux prises des troupes gouvernementales} \\ & \text{et des groupes rebelles en 1999 ou 2000} \\ & \text{faisant au moins 1000 morts sur l'une de ces deux années,} \\ non-crise & \text{sinon.} \end{cases}$$

Pour renseigner cette variable, nous nous sommes appuyé sur des bases de données relatives aux conflits armés : Correlates of War (COW) [5] et *State Failure Task Force* [6].

Ce recueil de données nous a permis de disposer d'une description statique du contexte structurel de chaque pays sur diverses années. Pour rajouter une dimension dynamique à

[2] http://www.worldbank.org/data/
[3] http://www.cia.gov/cia/publications/factbook/
[4] http://www.unhcr.org/
[5] http://www.correlatesofwar.org/
[6] http://www.cidcm.umd.edu/inscr/stfail/sfdata.htm

cette description, nous avons considéré la variation annuelle moyenne de chaque indicateur sur les années disponibles. Nous sommes parvenu à rassembler des informations sur 144 pays, chacun étant décrit par 284 indicateurs plus la classe. Sur ces 144 pays, 106 appartiennent à la classe *non-crise* et 38 à la classe *crise*. En moyenne par pays, les valeurs de près de 65 indicateurs sur 284 ne sont pas renseignées.

Ces premiers éléments d'analyse descriptive des données mettent en relief trois de leurs caractéristiques qui constituent autant de problèmes auxquels il faudra s'attaquer pour réussir à bâtir un modèle de prévision fiable et performant.

- **Données déséquilibrées** : la classe *crise* qui nous intéresse est bien moins représentée dans le corpus. Il faudra en tenir compte dans l'évaluation du classifieur.
- **Grande dimension** : le nombre de variables explicatives est grand. La tâche est d'autant plus délicate que ce nombre est nettement supérieur au nombre d'exemples d'apprentissage.
- **Données manquantes** : plus du quart des données ne sont pas renseignées. Or *Salammbô* ne peut pas apprendre à partir de bases de données incomplètes. Si l'on veut l'utiliser, il faut donc adopter une méthode de substitution de ces données qui soit la moins pénalisante possible pour le classifieur.

2.3 Premières expérimentations

Avant de voir comment se comporte notre outil empiriquement, il est important de décrire le mode opératoire que nous avons suivi pour l'ensemble de nos tests.

Afin de permettre une visualisation synthétique de nos résultats, nous nous sommes appuyé sur les matrices de confusion car elles sont facilement interprétables. Elles permettent de juger de la qualité de la reconnaissance de chacune des deux classes, que l'on quantifie par les taux de rappel. Les taux de précision estiment quant à eux la qualité des prédictions effectuées (mesure de la confiance que l'on peut accorder à ces prédictions). Par la suite, nous donnerons les résultats de chacun de nos tests sous la forme d'une matrice de confusion telle que celle du tableau 2.1.

TAB. 2.1 – Matrice de confusion

classe prédite / classe réelle	*non-crise*	*crise*
non-crise	A	B
crise	C	D

Dans ce tableau, les marginales en lignes correspondent aux nombres d'exemples de la base de test $n_0 = A + B$ et $n_1 = C + D$ des deux classes *non-crise* et *crise*. Le taux de bonnes classifications ainsi que les taux de rappel et précision de chacune des deux classes

se déduisent facilement d'une telle matrice :

$$\begin{aligned}
Reco &= \frac{A+D}{A+B+C+D} \\
Rappel\,(non\text{-}crise) &= \frac{A}{A+B} \\
Précision\,(non\text{-}crise) &= \frac{A}{A+C} \\
Rappel\,(crise) &= \frac{D}{C+D} \\
Précision\,(crise) &= \frac{D}{B+D}
\end{aligned}$$

Savoir comment évaluer un classifieur est un problème en soi et nous y reviendrons plus en détail à la section 10.3.1. Signalons simplement, à ce stade de l'exposé, que le seul taux de reconnaissance est souvent insuffisant pour juger de la qualité d'un classifieur. Cela est particulièrement vrai dans les tâches d'apprentissage pour lesquelles les classes sont déséquilibrées. En effet, dans ces cas-là, la règle de décision simpliste consistant à prédire systématiquement la classe majoritaire s'avère très performante, ce qui n'est pas vraiment satisfaisant. Prenons notre problème : la classe *non-crise* est largement majoritaire (73% des exemples). Un classifieur qui prédirait toujours cette classe aurait donc des performances acceptables...

N'oublions pas le contexte dans lequel se place ce travail : il faut arriver à identifier les crises potentielles pour permettre la mise en place d'une politique de prévention. Il est donc très important d'omettre un minimum de crises. Nous sommes dans une situation où les erreurs sur chacune des deux classes n'ont pas le même poids, le coût des faux négatifs étant plus important que le coût des faux positifs. Outre de bonnes performances globales, nous attendons donc de notre classifieur qu'il se trompe le moins possible sur les exemples de la classe *crise*.

Ainsi, lorsque nous comparerons différents classifieurs, nous les considérerons dans un premier temps comme des solutions d'un problème multicritère dans lequel il convient de maximiser simultanément le taux de bonnes classifications et le rappel de la classe *crise*. Nous privilégierons les solutions non dominées. Pour chaque expérience nous avons construit une matrice de confusion par validation croisée avec 10 sous-ensembles.

Afin de justifier empiriquement notre choix de modèle, à savoir que *Salammbô*, outre sa capacité à produire des résultats clairs, est suffisamment performant pour constituer la base de notre système d'évaluation des risques, nous l'avons comparé à d'autres algorithmes standards de classification supervisée. Le tableau 2.2 synthétise les résultats de cette comparaison. Les algorithmes que nous avons testés sont les suivants :

- ***kppv*** (k plus proches voisins) : pour chaque nouvel exemple à classer, on cherche les k plus proches pays de la base d'apprentissage et on lui attribue la classe majoritaire parmi celles des k pays trouvés. La notion de *plus proche* nécessite la définition d'une distance ou d'une similarité. Seule la distance euclidienne et la similarité basée sur le cosinus de l'angle formé par deux vecteurs (corrélation) ont été envisagées. L'objectif n'était pas d'approfondir cette méthode, mais d'obtenir des résultats de référence pour la comparaison avec *Salammbô*. Des expériences préliminaires, non reportées ici, nous ont conduit à prendre $k = 2$.

- ***Rocchio*** : à partir de la base d'apprentissage, deux prototypes sont construits, un par classe. En pratique il s'agit des barycentres de chacune des deux classes. On affecte à chaque pays de la base de test la classe correspondant au prototype le plus proche. Là aussi nous avons utilisé la distance euclidienne et le cosinus pour définir la notion de proximité.
- ***naïve Bayes*** : algorithme de classification probabiliste, basé sur le principe de l'inférence bayésienne. Les probabilités conditionnelles *a posteriori* de chaque classe sont estimées en faisant l'hypothèse, délibérément simpliste, que les variables d'entrée sont indépendantes, afin de limiter le nombre de paramètres à estimer. Nous avons utilisé la version de cet algorithme implémentée dans Weka 3.4.7 (Witten et Frank, 2005).
- ***C4.5*** : algorithme de construction d'arbres de décision (non flous) développé par Quinlan. Nous avons testé la version de cet algorithme implémentée dans Weka 3.4.7.
- ***Salammbô***

Certains de ces algorithmes étant particulièrement sensibles aux questions d'échelle (plage de valeurs prises par une variable donnée), surtout ceux qui reposent sur une mesure de distance ou de similarité, nous avons eu recours à la normalisation min-max. Ceci nous a permis de ramener l'ensemble des variables dans l'intervalle $[0;1]$. À chaque valeur v_{ij} prise par une variable v_j, on applique la transformation suivante :

$$v_{ij} = \frac{v_{ij} - \min_{k=1..n}(v_{kj})}{\max_{k=1..n}(v_{kj}) - \min_{k=1..n}(v_{kj})}$$

Afin d'homogénéiser la procédure de comparaison entre les divers algorithmes, nous avons appliqué ce pré-traitement à chacun d'eux.

TAB. 2.2 – Performances de cinq classifieurs sur la base de données pays, estimées par validation croisée à 10 sous-ensembles. P : Précision, R : rappel, c : classe *crise*, *nc* : classe *non-crise*. Le paramètre L concerne les arbres de décision. Il correspond au nombre minimum d'exemples que doit contenir un nœud pour pouvoir être partitionné.

Classifieur	kppv	Rocchio	naïve Bayes	C4.5	Salammbô
Matrice de confusion	88 | 18 25 | 13	81 | 25 18 | 20	60 | 46 13 | 25	91 | 15 28 | 10	89 | 17 21 | 17
Paramètres	Distance euclidienne, k=2	Distance euclidienne	Défaut	L=12	Opérateurs de Zadeh, L=15
Reco	70.1%	70.1%	59.1%	70.1%	73.6%
$P(nc)$	77.9%	81.8%	82.2%	76.5%	80.9%
$R(nc)$	83%	76.4%	56.6%	85.8%	84%
$P(c)$	41.9%	44.4%	35.2%	40%	50%
$R(c)$	34.2%	52.6%	65.8%	26.3%	44.7%

Au vu des résultats du tableau 2.2, on constate que *Salammbô* a le meilleur taux de bonnes classifications, mais rappelons que nous sommes également intéressé par une bonne reconnaissance de la classe *crise*. De ce point de vue *naïve Bayes* est le plus performant (meilleur taux de rappel), mais son taux de bonnes classifications est bien trop faible. Le comportement de *Salammbô* sur cette classe est tout de même assez satisfaisant : il la reconnaît bien mieux que *C4.5* ou *2ppv*. Globalement il offre le meilleur compromis entre une bonne reconnaissance de la classe *non-crise* et de la classe *crise*. *2ppv* et *C4.5* ont une matrice de confusion beaucoup trop asymétrique pour être vraiment intéressants.

L'apport du flou semble bien réel. La classification réalisée par *Salammbô*, qui peut être considéré comme une version floue de *C4.5*, est nettement plus proche de nos attentes que celle qui est effectuée sans flou (*C4.5*). Pour l'instant, à l'instar de ce que fait Weka pour les données numériques, nous avons opéré nous-même la substitution des valeurs manquantes d'une variable donnée par sa moyenne, pour *kppv* et *Rocchio*, *Salammbô*. Nous n'avons pas justifié ce choix de méthode de substitution. Aussi aborderons-nous un peu plus en détail ce sujet qui est loin d'être anodin, au chapitre 6. Rappelons en effet que près d'un quart des données ne sont pas renseignées.

En l'état les performances sont loin d'être acceptables. Le taux de bonnes classifications n'est que de 73.6%. En prédisant la classe *non-crise* majoritaire, nous obtenons exactement le même score. De plus, nous ne reconnaissons même pas une crise sur deux. Dans le chapitre suivant nous verrons comment la sélection d'attributs peut améliorer les performances de *Salammbô*.

Chapitre 3

Améliorations du modèle

Bien qu'opérant intrinsèquement une sélection des attributs pertinents, les arbres de décision voient leurs performances se dégrader en présence d'un grand nombre d'attributs non pertinents (Blum et Langley, 1997). Les algorithmes de construction d'arbres de décision se basent sur la recherche des facteurs pertinents, mais cette recherche est ralentie et sous-optimale en présence d'un grand nombre d'indicateurs. Les attributs à chaque nœud sont choisis suivant un critère entropique. Mais lorsque plusieurs attributs obtiennent des valeurs identiques pour la mesure de discrimination, ce qui arrive fréquemment lorsque le nombre d'attributs est grand, l'un d'eux est choisi arbitrairement. D'une part, ceci n'est guère satisfaisant. D'autre part, ce choix peut s'avérer sous-optimal. Il est en effet vraisemblable que le choix d'un autre attribut eût conduit à des ramifications ultérieures bien plus discriminantes. Pour cette raison il est apparu judicieux de procéder, en amont de l'apprentissage, à une réduction de la dimensionnalité du problème *via* une sélection explicite d'attributs. Cette étape est essentielle. Elle permet d'améliorer la qualité du processus de généralisation, d'accélérer l'algorithme d'induction, et également de simplifier les arbres générés, ce qui rend les résultats plus simples à analyser.

3.1 Un algorithme génétique pour la sélection d'attributs

Nous avons développé une première méthode basée sur un algorithme génétique, que nous allons détailler dans cette section. Au travers d'une série d'expériences, nous mettrons en évidence les progrès accomplis, ainsi que les lacunes qui restent à combler. Nous présenterons les grandes familles d'algorithme de sélection, ainsi que les nouvelles techniques que nous envisageons de mettre en place, dans le chapitre 7, entièrement dévolu au problème de la sélection d'attributs.

Notre objectif est de trouver le sous-ensemble d'attributs qui optimise les performances de notre classifieur. Formulée ainsi, il apparaît clairement que la tâche de sélection d'attributs à laquelle nous nous attaquons est un problème d'optimisation combinatoire. Les algorithmes génétiques sont bien adaptés à ce genre de problèmes et leur application dans des tâches de sélection d'attributs a été beaucoup étudiée (Raymer *et al.*, 2000; Morita *et al.*, 2003; Yang et Hononvar, 1998). Nous ne rappellerons pas les fondements de ces algorithmes, que le lecteur pourra retrouver par exemple dans (Man *et al.*, 1999; Michalewicz, 1996).

Pour préciser un peu le vocabulaire, disons simplement qu'un algorithme génétique consiste à faire évoluer de génération en génération une population d'individus, chaque individu correspondant à une solution du problème à traiter (phénotype), selon un processus darwinien de sélection naturelle. Chaque individu est représenté par son génotype

sur lequel sont appliqués les opérateurs génétiques : la mutation et le croisement génétique (*crossover*). Ces opérateurs ainsi que le processus de sélection naturelle, qui est guidé par une fonction mesurant la qualité d'une solution (fonction d'évaluation ou *fitness*), définissent la façon dont la population évolue, c'est-à-dire la façon dont de nouvelles solutions sont construites et évaluées.

Pour définir un algorithme génétique, il suffit d'en préciser les caractéristiques principales.

- **Taille de la population** : nombre de solutions qui sont testées à chaque génération. Plus ce nombre est important, plus la complexité algorithmique est grande, mais la couverture de l'espace des solutions sera également plus grande.
- **Codage du génome** : le génome est représenté par un vecteur de bits. Le codage est ce qui permet de faire le lien entre génotype et phénotype. Il décrit la façon dont une solution du problème sera représentée sous la forme d'un vecteur.
- **Opérateur de mutation** : modification d'un élément du génome, selon un processus stochastique. La mutation est ce qui permet d'introduire de la nouveauté dans la population et donc d'assurer sa diversité, ce qui est essentiel pour une bonne exploration de l'espace de recherche.
- **Opérateur de croisement** : il assure la recomposition de nouvelles solutions à partir des solutions existantes. Par analogie avec les processus d'évolution naturelle, les deux solutions de la génération actuelle qui seront utilisées pour en créer deux nouvelles sont appelées les parents, tandis que les deux nouvelles sont dénommées les enfants.
- **Fonction d'évaluation d'un individu (*fitness*)** : elle évalue la qualité d'une solution ; c'est cette fonction qui est optimisée par l'algorithme.
- **Processus de sélection des individus** : il s'appuie sur la fonction d'évaluation pour décider, de manière stochastique, quelles solutions conserver au sein de la nouvelle génération. Pour assurer la convergence de l'algorithme et par analogie avec le principe darwinien de la sélection naturelle, la phase de sélection doit favoriser les individus les plus performants (exploitation des bonnes solutions) sans pour autant ne retenir qu'eux sous peine d'appauvrir la diversité de la population et gêner l'exploration.
- **Critère d'arrêt** : il s'agit de la règle spécifiant le moment où l'on peut considérer que l'algorithme est terminé.

3.1.1 Codage

Dans notre contexte, le phénotype correspond à un sous-ensemble d'indicateurs. Le codage du génome, vecteur de bits, doit permettre de retrouver le sous-ensemble en question. Une manière simple de procéder consiste à considérer que chaque bit du génome dénote la présence (1) ou l'absence (0) d'un attribut. La taille du génome est alors égale au nombre total d'attributs présents dans la base de données. C'est ce codage que nous avons adopté.

Cherchant à améliorer les performances de *Salammbô*, nous l'avons intégré dans l'algorithme de telle sorte qu'un modèle soit construit pour chaque individu, à partir du sous-ensemble d'indicateurs correspondant à cet individu. La fonction d'évaluation est alors une mesure des performances de ce modèle.

Lors de la construction d'un arbre de décision il est possible de préciser le nombre

minimum L d'exemples que doit contenir un nœud de l'arbre pour pouvoir être partitionné[1]. Suivant les valeurs de L les résultats peuvent être très différents. Aussi avons-nous décidé d'inclure ce paramètre dans le génome afin qu'il prenne part à l'optimisation. Pour cela nous avons rajouté 4 bits en tête du vecteur pour coder des valeurs de L comprises entre 1 et 16. Le tableau 3.1 donne un exemple de génome.

TAB. 3.1 – Exemple de génome à 5 attributs. $L = 7$. Seuls les 1er et 4e attributs sont sélectionnés.

0111	1	0	0	1	0

3.1.2 Croisement génétique

Nous avons utilisé le croisement à un point. Le principe est le suivant : on choisit un point de coupure dans le génome, l'un des deux enfants hérite de la partie du génome du premier parent située avant ce point de coupure et de la partie du génome du second parent située après. Le génome du second enfant se construit de manière symétrique. Ne souhaitant pas découper arbitrairement la partie correspondant à L, nous avons imposé que ce point de coupure se situe après le 4e bit.

3.1.3 Mutation

Afin de mieux parcourir l'espace de recherche, sans s'arrêter au premier optimum local trouvé, on introduit de l'aléa *via* l'opérateur de mutation. Une probabilité P_m de mutation étant fixée, pour chaque bit d'un génome on tire aléatoirement un nombre entre 0 et 1. S'il est inférieur à P_m on inverse la valeur du bit considéré. Fixer définitivement et dès le début une valeur pour P_m n'est pas aisé et pas forcément souhaitable. En effet introduire trop d'aléatoire ralentit fortement la convergence de l'algorithme, mais il en faut tout de même suffisamment pour pouvoir sortir des optima locaux. Il faut trouver un compromis entre les deux. On retrouve le traditionnel conflit en optimisation combinatoire entre exploration (fortes valeurs de P_m) et exploitation (faibles valeurs de Pm).

Pour surmonter ce dilemme, nous avons fait en sorte que P_m ne soit pas constant mais évolutif. L'initialisation de la population étant aléatoire, il n'est pas besoin d'avoir une valeur de P_m forte pour explorer l'espace de recherche. Ce n'est que lorsque la population s'homogénéise, à proximité d'un maximum local par exemple, qu'il faut réintroduire de l'aléatoire. Une fois la population redevenue hétérogène et donc apte à explorer efficacement, P_m peut redevenir faible. Nous nous sommes basé sur l'écart-type de la fonction d'évaluation pour juger de l'homogénéité de la population, en considérant qu'au départ elle était hétérogène. Lorsque cet écart-type se réduit trop, de 20% d'une génération sur l'autre ou de 50% par rapport à la valeur initiale[2], nous augmentons la valeur de P_m et nous la diminuons lorsque cet écart-type augmente à nouveau.

3.1.4 Fonction d'évaluation

La *fitness* étant la fonction qui sera optimisée par l'algorithme génétique, nous avons naturellement choisi une fonction mesurant pour chaque individu les performances de la

[1] C'est là un des critères d'arrêt de *Salammbô*.
[2] Diverses expérimentations non reportées ici nous ont permis de choisir ces valeurs.

classification réalisée par *Salammbô* à partir du sous-ensemble d'attributs correspondant à cet individu.

À la section 2.3 nous avons déjà évoqué la question de l'évaluation des performances d'un classifieur. Nous avons précisé que nous nous placions dans un problème d'agrégation multicritère, dans lequel nous voulions maximiser simultanément le taux de bonnes classifications, ainsi que le rappel de la classe *crise*. La fonction d'agrégation doit donc être croissante selon ces deux paramètres et nous souhaitons privilégier les solutions *non dominées*.

Supposons que nous ayons q critères à maximiser simultanément. Une solution $x = (x_1, x_2, ..., x_q)$ sera dite dominée par une autre solution $z = (z_1, z_2, ..., z_q)$ si on a :

$$\begin{cases} \forall i \in \{1..q\} & x_i \leq z_i \\ \exists j \in \{1..q\} & x_j < z_j \end{cases}$$

Afin de favoriser les solutions non dominées, en nous inspirant de ce qui a été proposé par Man *et al.* (1999) pour l'optimisation multicritère, et après expérimentations, nous avons défini la fonction d'évaluation suivante

$$Fit_1(i) = \frac{Rappel_i(crise)}{\sqrt{1+N}}$$

où $Rappel_i(crise)$ est le taux de rappel de la classe *crise* pour le sous-ensemble d'attributs associé à l'individu i et N et le nombre de solutions qui dominent celle qui correspond à i. Parmi les solutions les moins dominées, nous favorisons ainsi celles qui ont un bon taux de rappel de la classe *crise*. Nous avons également effectué des tests avec Fit_2 qui correspond à une variante de la mesure Fit_1 dans laquelle nous avons remplacé $Rappel_i(crise)$ par $Reco_i$. Ceci permet de mettre davantage l'accent sur le taux de bonnes classifications. Nous avons également envisagé la somme pondérée, pour différents poids w_1 et w_2 : $Fit_3(i) = w_1 Rappel_i(crise) + w_2 Reco_i$.

3.1.5 Sélection des individus

Nous avons employé la sélection par roulette biaisée (Man *et al.*, 1999). Soit Fit la somme de toutes les évaluations des différents individus et n_{pop} la taille de la population. On a : $Fit = \sum_{i=1}^{n_{pop}} Fit(i)$, où $Fit(i)$ désigne la *fitness* de l'individu i. Si on désigne par T_l la population à la génération l, l'algorithme de sélection des individus est le suivant :

1. Initialisation : $i = 0$, $T_{l+1} = \emptyset$
2. Tirer un nombre r aléatoirement entre 0 et Fit
3. $T_{l+1} = T_{l+1} \cup \{k\}$ Sélectionner l'individu k tel que

$$\sum_{i=1}^{k-1} Fit(\sigma(i)) < r \leq \sum_{i=1}^{k} Fit(\sigma(i))$$

   où σ est la permutation sur les individus telle que

$$Fit(\sigma(1)) \geq Fit(\sigma(2)), ..., \geq Fit(\sigma(n_{pop}))$$

4. $i = i + 1$
5. Arrêt ? : si $i < n_{pop}$ revenir en 2, sinon FIN

Nous avons également eu recours à la sélection par tournoi, à deux joueurs. Deux individus sont tirés au hasard dans la population. Celui dont *fitness* est la plus élevée est sélectionné. Cette approche permet d'avoir moins de pression sélective qu'avec la roulette biaisée et de mieux préserver la diversité génétique. En effet, lorsqu'un individu est bien meilleur que tous les autres il sera très souvent sélectionné par roulette biaisée, ce qui aura tendance à trop pénaliser l'exploration au profit de l'exploitation.

Cependant la sélection par tournoi ne prend pas du tout en compte l'écart relatif entre les performances des individus, ce qui peut être gênant. C'est pourquoi nous nous sommes laissé la possibilité de tester les deux techniques. Des tests préliminaires, que nous ne présenterons pas ici, nous ont permis de choisir la roulette biaisée pour laquelle les résultats sont plus intéressants : matrice de confusion mieux équilibrée, moins d'attributs sélectionnés.

Avec ces deux méthodes, certains individus vont être sélectionnés plusieurs fois tandis que d'autres ne le seront jamais. Il est donc possible de perdre le meilleur individu même s'il a la plus grande probabilité de survie, et ce d'autant plus qu'il peut subir une ou des mutations qui peuvent dégrader ses performances. Une telle perte est dommageable dans la mesure où elle ralentit la convergence de l'algorithme. Pour éviter ce problème nous avons adopté une stratégie de reproduction élitiste. Le meilleur individu est recopié tel quel dans la nouvelle population, sans qu'aucun des opérateurs génétiques ne lui soit appliqué. La régénération du reste de la population se fait de la façon suivante : sélection, croisement et mutation.

3.1.6 Critère d'arrêt

Notre critère d'arrêt est double. D'une part, nous considérons que la convergence est atteinte lorsque le meilleur individu reste le même durant un certain nombre de générations. D'autre part, pour limiter les temps de calcul lorsque la convergence est difficile à obtenir, nous avons fixé une borne maximale sur le nombre de générations possibles.

3.2 Résultats expérimentaux

Les différents paramètres de notre méthode ont été choisis de manière empirique. Pour chacun de nos tests, nous avons fait évoluer une population de 100 individus, sur un nombre maximal de 50 générations. Par ailleurs, la convergence est supposée atteinte lorsque le meilleur individu ne change pas durant 10 générations. Suite à des expériences préliminaires nous avons décidé de fixer la probabilité de croisement P_c à 0.6 et la probabilité de mutation P_m[3] à 0.05. Les autres paramètres de l'algorithme varient d'une version à l'autre et seront précisés chaque fois que cela sera utile. Le tableau 3.2 présente les résultats des tests réalisés sur l'algorithme génétique pour différentes fonctions d'évaluation.

Au premier abord on remarque que la moyenne pondérée (Fit_3) se révèle moins intéressante que les deux autres mesures que nous avons mises en place. Fit_1, que nous appellerons *Pareto Rappel* pour insister sur l'importance du taux de rappel de la classe *crise*, est au vu de ces expériences la mieux adaptée à notre problème[4]. La matrice de confusion associée est en effet moins asymétrique que les autres et surtout elle permet de retenir moins d'attributs. C'est là un point important, car plus ce nombre est faible, plus notre système sera rapide et plus les arbres de décision seront faciles à interpréter.

[3] Il s'agit de la valeur initiale. Elle évolue ensuite suivant l'hétérogénéité de la population.
[4] Par analogie, nous appellerons Fit_2 *Pareto Reco*.

TAB. 3.2 – Performances de l'algorithme génétique selon la fonction d'évaluation choisie (sélection par roulette biaisée)

Fitness	Fit_1	Fit_2	Fit_3
Matrice de confusion	97 9 9 29	97 9 12 26	94 12 12 26
Paramètres de *Salammbô*	L=2	L=15	L=14
Nombre d'attributs sélectionnés	120	160	135
Reco	87.5%	85.4%	91.5%
$P(nc)$	91.5%	89%	88.7%
$R(nc)$	91.5%	91.5%	88.7%
$P(c)$	76.3%	74.3%	68.4%
$R(c)$	76.3%	68.4%	68.4%

Plutôt que de mettre en concurrence nos fonctions d'évaluation, nous avons par la suite cherché à tirer profit de leurs avantages respectifs, à savoir que *Pareto Reco* assure une meilleure modélisation de la classe *non-crise* alors que *Pareto Rappel* permet de mieux reconnaître la classe *crise*. Pour ce faire nous avons lancé un premier processus d'optimisation avec *Pareto Rappel* comme fonction d'évaluation. Nous avons ainsi retrouvé l'optimum local du tableau 3.2. Ensuite nous avons relancé une seconde phase de sélection d'attributs en prenant *Pareto Reco* comme *fitness*. Mais au lieu de prendre une initialisation aléatoire pour cette seconde étape, nous nous sommes placé au voisinage du premier optimum local trouvé.

Une fois qu'un bon taux de rappel a été obtenu pour la classe crise, on s'attache à maximiser le taux de bonnes classifications, ce qui conduit aux résultats du tableau 3.3, légèrement meilleurs que ceux que l'on obtient sans coupler les fonctions d'évaluation.

Nous avons également appliqué, en amont de notre phase d'optimisation, un filtrage des attributs trop mal renseignés pour être considérés comme pertinents. Nous avons ainsi décidé de supprimer tous les attributs contenant plus de 50% de valeurs manquantes. 43 sur les 283 indicateurs ont ainsi été retirés avant de lancer l'algorithme génétique. Cela a permis de réduire la dimension de l'espace de recherche et donc d'accélérer l'optimisation, tout en réduisant le nombre de maxima locaux susceptibles de perturber l'algorithme.

Comme on peut le voir dans le tableau 3.3, ce pré-filtrage s'est révélé bénéfique. Pour un même taux de bonnes classifications, le taux de rappel de la classe crise est meilleur que précédemment.

Notre algorithme opère une bonne réduction du nombre d'attributs : on passe de 283 à 123. Cependant il est vraisemblable que nous en ayons encore trop. Il suffit de regarder ce qui a été fait par la *State Failure Task Force* (Goldstone *et al.*, 2000) pour s'en convaincre : 73% de bonnes classifications avec seulement 5 indicateurs. Pour cette raison, nous avons fait en sorte que l'algorithme génétique ne prenne en compte qu'un nombre restreint d'indicateurs. Cela a demandé quelques modifications. Après application de chaque opérateur, nous supprimons aléatoirement des attributs si besoin est, pour n'en avoir que le nombre voulu. Cette suppression s'opère également au niveau de l'initialisation.

Nous avons réalisé plusieurs expériences en faisant varier ce nombre limite et les résultats les plus satisfaisants sont donnés au tableau 3.4. Ce tableau met en évidence une amélioration des performances au niveau de la reconnaissance de la classe crise, même si le taux de bonnes classifications n'a pas augmenté. Cette dernière solution correspond à

TAB. 3.3 – Performances de l'algorithme génétique en combinant Fit_1 et Fit_2, avec et sans pré-filtrage des attributs mal renseignés

Pré-filtrage	Non		Oui	
Matrice de confusion	98	8	97	9
	9	29	8	30
Paramètres de *Salammbô*	L=13		L=13	
Nombre d'attributs sélectionnés	123		123	
Reco	88.2%		88.2%	
$P(nc)$	91.6%		92.4%	
$R(nc)$	92.5%		91.5%	
$P(c)$	78.4%		76.9%	
$R(c)$	76.3%		78.9%	

la version stabilisée de notre outil. À titre de comparaison, nous présentons également les résultats obtenus avec la méthode séquentielle de sélection par *beam search* avant. Cette méthode sera décrite dans le chapitre 7, dédié aux méthodes de sélection d'attributs (voir algorithme 4). Les expériences menées par Aha et Bankert (1996) suggèrent qu'elle est très efficace lorsque la base de données d'apprentissage comporte peu d'exemples et beaucoup d'attributs, ce qui est notre cas (144 pays à classer, chacun étant décrit par 284 indicateurs).

TAB. 3.4 – Performances comparées de la sélection par *beam search* avant (Fit_1 étant la fonction d'évaluation utilisée) avec notre algorithme génétique, combinant Fit_1 et Fit_2, avec pré-filtrage des attributs mal renseignés et limitation du nombre d'attributs sélectionnés à 40

Sélection d'attributs	beam search		algorithme génétique	
Matrice de confusion	96	10	96	10
	8	30	7	31
Paramètres de *Salammbô*	L=2		L=15	
Nombre d'attributs sélectionnés	6		40	
Reco	87.5%		88.2%	
$P(nc)$	92.3%		93.2%	
$R(nc)$	90.6%		90.6%	
$P(c)$	75%		75.6%	
$R(c)$	78.9%		81.6%	

Cette phase de sélection s'avère donc très efficace puisqu'elle nous a permis d'améliorer de façon drastique la qualité de la reconnaissance, et ce sur les deux classes en même temps, le taux de bonnes classifications ayant crû de près de 15%. Notons que le *beam search* avant fournit des résultats très proches de ceux de notre algorithme génétique. Au niveau des performances du classifieur, notre gain est minime et ne concerne que la reconnaissance de la classe crise. Ainsi, notre étude empirique ne réfute pas les observations de Aha et Bankert (1996), à savoir que la technique *beam search* est particulièrement bien adaptée aux problèmes comportant peu d'instances et beaucoup de variables.

Chapitre 4

Discussion

La méthodologie que nous proposons pour contruire un outil d'aide à l'anticipation des crises est, au regard de nos premières expérimentations, tout à fait satisfaisante du point de vue des performances. Elle s'appuie sur une phase amont de sélection d'attributs. Une fois que les attributs les moins pertinents ont été supprimés, le moteur d'induction *Salammbô* est utilisé pour apprendre une base de règles *via* la construction d'un arbre de décision flou. Ces règles peuvent ensuite être appliquées dans un raisonnement déductif pour inférer la classe de nouveaux exemples.

Lorsque la classe réelle de ces *nouveaux pays* est connue (entre un et deux ans plus tard selon le modèle actuel), la base d'apprentissage peut être étendue en intégrant ces nouveaux exemples. Un nouvel apprentissage sur cette base étendue permettra alors la construction d'un nouveau modèle plus fin. L'ensemble de ce processus est synthétisé sur la figure 4.1.

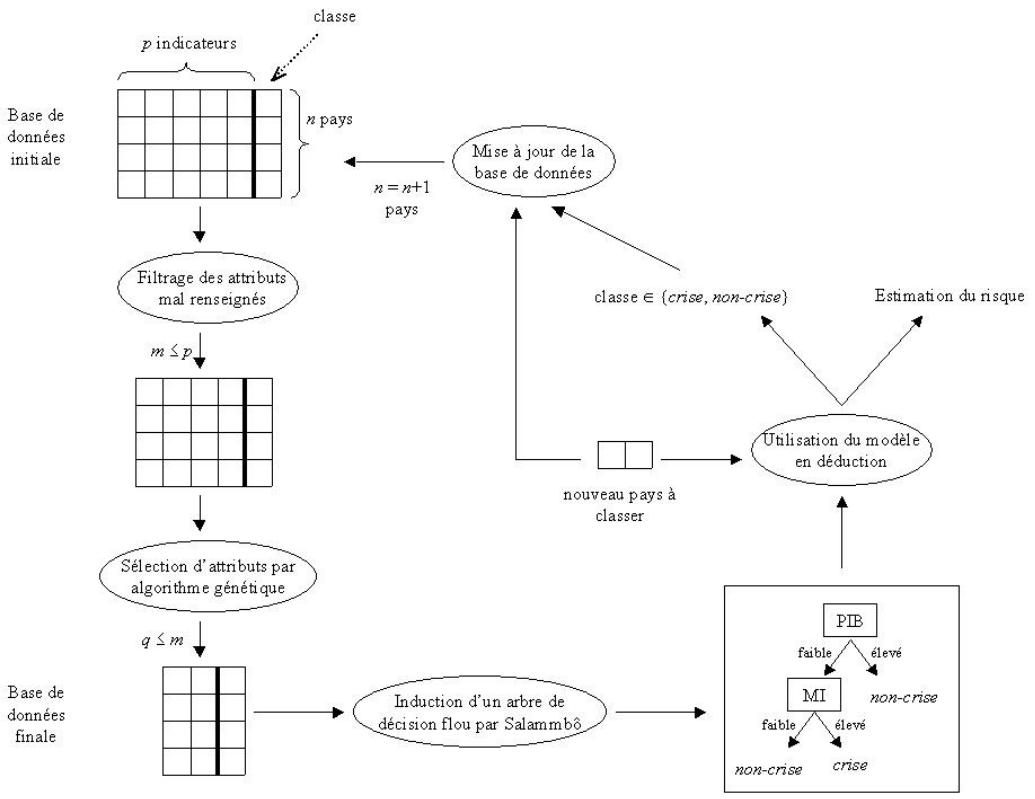

FIG. 4.1 – Fonctionnement général de l'outil de détection à moyen terme des crises intra-étatiques

À ce stade de notre exposé, si les premières expériences nous ont permis de montrer la force de la méthodologie que nous avons mise en place, une analyse critique s'impose, afin de dégager les faiblesses et lacunes de notre approche. Nous pourrons ainsi identifier les points clés sur lesquels nous pouvons agir pour pouvoir pallier ces faiblesses. L'objectif est double :

- améliorer les performances du modèle
- améliorer la compréhension que nous avons de ce modèle de façon à pouvoir mieux justifier ses performances

La principale critique que l'on peut faire au modèle qui a été présenté dans cette partie concerne le traitement des données manquantes. Plus d'un quart des valeurs sont manquantes et pour chacun des pays et quasiment tous les indicateurs nous avons au moins une valeur manquante. Or *Salammbô* a besoin de disposer d'une matrice complète pour pouvoir construire un arbre de décision. La façon dont nous traitons ce problème est donc crucial et influe sur les performances de notre outil. Jusqu'à présent nous avons opté pour une substitution de ces valeurs par la moyenne de l'attribut correspondant.

Ce choix, s'il est simple à mettre œuvre, n'en est pas moins fortement critiquable. Le principal inconvénient réside dans le biais qui est introduit dans l'estimation de la variance. Elle est en effet artificiellement réduite. Un tel choix a besoin d'être justifié. Envisager d'autres méthodes pourrait s'avérer fructueux sur le plan des performances, et serait nettement plus satisfaisant sur le plan méthodologique. C'est ce que nous tâcherons de faire au chapitre 6.

Cette même remarque peut également s'appliquer, dans une moindre mesure, à la phase de sélection d'attributs. Seules deux techniques ont été envisagées. Elles sont certes réputées dans la littérature comme particulièrement efficaces, mais ce n'est qu'au travers d'une étude plus poussée du domaine que nous serons en mesure de justifier nos choix. Ce point sera abordé au chapitre 7.

Pour chacun des points que nous venons de mentionner, nous procéderons comme dans cette première partie. Nous effectuerons ce va-et-vient entre théorie et pratique typique de l'épistémologie popérienne. L'analyse théorique du problème doit permettre de développer un certain nombre d'hypothèses qui seront ensuite soumises à un certain nombre de tests expérimentaux à partir desquels il sera possible de juger la capacité de ces hypothèses à expliquer convenablement certains faits. Les forces et faiblesses des hypothèses théoriques pourront être mises en avant, suscitant ainsi la reformulation, l'adaptation des hypothèses initiales qui seront à nouveau confrontées aux données expérimentales et ainsi de suite.

Un aspect essentiel de nos expérimentations réside dans la comparaison de divers modèles. Pour pouvoir tirer quelque conclusion substantielle des tests empiriques, les questions de l'évaluation d'un classifieur et de la comparaison de deux classifieurs méritent assurément d'être posées. Nous avons déjà entrevu leur importance lors de la construction de la *fitness* de l'algorithme génétique à la section 3.1.4, mais nous y reviendrons à la section 10.3.1.

Améliorer les performances de notre classifieur est une chose, mais il est tout aussi important d'évaluer la fiabilité de ses performances. Si l'on veut pouvoir vanter les mérites de notre approche il faut d'une part, pouvoir justifier nos choix techniques. C'est ce que nous venons de mettre en avant. D'autre part, il faut pouvoir garantir que les résultats obtenus ne sont pas trop spécifiques à la base de données utilisée. À cet effet nous avons construit de nouvelles bases de données contenant bien plus d'indicateurs, estimés sur des

périodes beaucoup plus longues. Nous les décrirons au chapitre 10.2. Nous préciserons plus particulièrement les quelques indicateurs qui sont présents dans la majorité des études économétriques sur les guerres civiles.

Du fait de l'accroissement du nombre d'indicateurs, il sera bon de se pencher également plus attentivement sur la question de la complexité des algorithmes de sélection d'attributs, ce que nous ferons au cours du chapitre 7

Ce chapitre sera également l'occasion de recentrer notre travail sur la partie applicative. Nous nous intéresserons au protocole expérimental qui doit être respecté, non seulement pour que les résultats, ou plutôt leur interprétation, soient valides, mais aussi pour qu'ils soient plus pertinents vis-à-vis de notre contexte applicatif. Cette formalisation de la procédure expérimentale que nous appelons protocole est fondamentale et nous y reviendrons souvent dans le reste de cet exposé. Par exemple l'amélioration remarquable des performances de notre modèle grâce à la sélection d'attributs n'est-il pas dû au fait que ledit modèle est spécifique à notre base de données ? Auquel cas rien ne garantit qu'en généralisation il soit aussi performant. Éclaircir ce point et ainsi améliorer la qualité de notre argumentation pour soutenir notre outil nécessite ce travail de formalisation de la procédure de test.

Nous n'avons pour l'instant évoqué que des critères quantitatifs pour valider notre outil. Rappelons tout de même qu'un des éléments clés nous ayant permis de justifier notre préférence pour les arbres de décision flous concerne l'interprétabilité des résultats. Aussi consacrerons-nous une partie de la section 10.3 à l'analyse qualitative des résultats expérimentaux.

Deuxième partie

Étude de la chaîne d'apprentissage dans son ensemble

Dans la partie précédente nous avons proposé un premier modèle d'évaluation des risques, basé sur l'induction d'arbres de décisions flous. En le mettant à l'épreuve de données réelles, deux sous-problèmes majeurs ont pu être identifiés. Ils concernent la qualité des données à partir desquelles notre modèle peut être appris.

D'une part, ces données sont incomplètes. De nombreuses valeurs sont manquantes, ce qui nécessite la mise en place d'une stratégie spécifique de traitement de ces valeurs. Notre algorithme d'apprentissage nécessite en effet de disposer d'une base d'apprentissage complète pour pouvoir être utilisé. D'autre part, les données sont décrites par un grand nombre de variables qui ne sont pas toutes pertinentes pour la tâche de classification. Il s'est avéré important de procéder à une sélection de ces variables afin d'améliorer les performances de notre modèle.

Dans cette partie nous souhaitons aborder ces deux problèmes, plus finement que ce qui a été fait jusqu'à présent. Étant donné que nous avons l'ambition de construire une méthodologie d'évaluation des risques aussi générique que possible, nous ne voulons pas nous restreindre à trouver les solutions les mieux adaptées à notre contexte applicatif de détection des guerres civiles. Aussi étudierons-nous les différentes approches permettant de résoudre ces deux problèmes, de façon la plus générique possible. Afin de rester cohérent avec notre cadre méthodologique, nous restreindrons tout de même notre analyse au contexte de la classification supervisée.

Dans ce contexte, le traitement des données manquantes et la sélection d'attributs constituent deux maillons de la chaîne d'apprentissage. Nous étudierons chacun de ces maillons indépendamment, afin d'identifier pour chacun les principales techniques existantes ainsi que leurs forces et faiblesses au regard d'une tâche de classification supervisée. Cependant, il nous semble essentiel de ne pas occulter le fait que notre objectif est la construction d'un modèle global d'évaluation des risques et non l'optimisation locale de telle ou telle partie du modèle. Notre approche se doit d'être holistique. En conséquence nous nous intéresserons également aux interactions entre traitement des données manquantes, sélection d'attributs et apprentissage, afin d'analyser la chaîne de traitement dans son ensemble.

Les méthodes que nous présenterons permettent de nettoyer[1] les données, de les préparer en vue de la génération d'un modèle de classification le plus performant possible. Aussi ramènerons-nous l'évaluation de leurs qualités respectives à la comparaison des classifieurs qu'elles permettent d'induire.

L'étude comparative de différentes techniques est un point crucial dans la construction d'un modèle global. Parce qu'elle permet de justifier empiriquement certains choix théoriques, nous y aurons fréquemment recours dans cette partie. Nous commencerons donc par la présentation, dans le chapitre 5, des différents tests statistiques utilisés dans la comparaison de classifieurs, et que nous mettrons en œuvre dans le reste de cette partie. Nous détaillerons ensuite, au chapitre 6, les différentes techniques de traitement des données manquantes. Enfin au chapitre 7, nous aborderons les questions relatives à la sélection d'attributs.

[1] Les anglo-saxons parlent de *data cleansing*.

Chapitre 5

Comparaison de classifieurs

La méthode scientifique d'analyse et de résolution des problèmes peut se décomposer en deux grandes familles.

- Formaliser le problème de façon à se ramener à un cadre d'analyse existant. Le travail consiste alors essentiellement à choisir la technique la mieux adaptée aux spécificités du problème ou à améliorer une technique existante pour mieux tenir compte des exigences particulières du problème.
- Développer un nouveau cadre d'analyse, ce qui conduit généralement à l'établissement de nouvelles normes et de techniques radicalement différentes de celles qui existent. On retrouve la notion de changement de paradigme qui occupe une place centrale dans l'histoire et la sociologie des sciences (Kuhn, 1970).

En formalisant la détection des crises comme un problème de classification supervisée, nous avons implicitement opté pour la première approche. Le principal travail de recherche consiste alors à adapter, à améliorer les méthodes existantes, pour respecter les exigences propres au problème traité. Parmi l'ensemble des méthodes disponibles, il faut ensuite être capable d'identifier la meilleure, selon un critère à définir. Il faut donc être en mesure de comparer objectivement différentes méthodes.

Dans cette partie II, dévolue à l'étude de la partie amont de la chaîne globale d'apprentissage, nous serons amené à comparer diverses techniques de substitution des valeurs manquantes et de sélection d'attributs. Ayant pris le parti de nous focaliser sur la chaîne globale d'apprentissage, ces techniques seront évaluées indirectement et comparées par le biais de la performance des classifieurs construits à partir des données dont elles auront assuré le prétraitement.

Théoriquement, si nous disposons d'une base de données couvrant intégralement un domaine particulier, la comparaison de classifieurs relativement au problème correspondant est immédiate. Celui qui sera choisi est simplement celui qui est associé aux meilleures performances. Les statisticiens nomment généralement *population* cette base de données globale. Mais ce cas de figure est bien évidemment utopique. En pratique la taille de la base de données disponible est limitée. Elle ne représente qu'un échantillon de la population, que l'on espère aussi représentatif que possible. Aussi n'est-il pas possible de conclure directement quant à la supériorité de l'un des classifieurs.

Il est essentiel de s'assurer que les différences observées entre les classifieurs ne sont pas des artefacts liés au processus d'échantillonage. Pour cela nous proposons de nous appuyer sur une méthodologie stricte et bien formalisée, permettant de juger du caractère significatif ou non des différences observées. Les tests statistiques sont parfaitement adaptés

à ce problème et permettent de décider, une fois fixée une probabilité maximale d'erreur, si les différences observées sont suffisamment importantes pour ne pas être attribuables à la variabilité introduite par l'échantillonnage.

Selon le nombre de classifieurs à comparer (2 ou $k > 2$) et le nombre de bases de données utilisées pour l'évaluation (1 ou $n > 1$), nous pouvons distinguer quatre cas d'application dans lesquels les besoins relatifs à la comparaison de classifieurs imposent l'utilisation de tests spécifiques. La suite de cette partie est consacrée à la description de ces tests. L'ensemble des notations qui seront introduites dans cette section sont regroupées à l'annexe A.

5.1 Évaluation d'un classifieur

Afin de rester le plus générique possible, nous ne ferons pas d'hypothèses particulières quant aux mesures employées. L'objectif de cette partie est en effet de proposer des méthodologies adaptées, indépendamment du choix de cette mesure. Nous supposerons cependant qu'un minimum de précautions ont été prises pour évaluer sainement ces mesures.

- Ce sont les performances en généralisation qui nous intéressent. Cela suppose que chaque classifieur est testé sur une sous-base indépendante de la sous-base sur laquelle il a été appris. Les deux sous-bases doivent donc être disjointes.
- Ne considérer qu'une seule paire de bases disjointes (apprentissage,test) ne permet pas de tenir compte de la variabilité liée à l'échantillonnage. Aussi est-il préférable de multiplier les échantillons d'apprentissage et de test. Nous supposerons donc que pour chaque base de données considérée, les performances de chaque classifieur seront le résultat de l'agrégation[1] des performances de ce classifieur estimées sur un ensemble de m paires de bases (apprentissage,test). Les mesures obtenues seront ainsi plus robustes. En revanche nous ne faisons pas d'hypothèse quant à la façon dont sont générées ces m sous-bases ; les techniques les plus fréquentes étant le *bootstrap* et la validation croisée.
- Pour que la variabilité de l'échantillonnage ne perturbe pas l'analyse comparative, il est important que chaque classifieur soit évalué à partir des mêmes échantillons. Pour cette raison, nous supposerons que ce sont les mêmes m paires de bases (apprentissage,test) qui servent à l'estimation des performances de tous les classifieurs.

5.2 Deux classifieurs évalués sur une seule base de données

Pour des applications spécifiques, une seule base de données du domaine étudié est disponible. Bien qu'il soit généralement plus fréquent d'avoir à comparer divers algorithmes sur cette base, nous commencerons par le cas plus simple où seuls deux classifieurs sont évalués.

5.2.1 Test de Student

Pour comparer deux classifieurs, la technique la plus fréquemment employée repose sur le test de Student avec échantillons appariés (Mitchell, 1997). Le test de Student est classiquement utilisé pour vérifier si les moyennes de deux distributions sont identiques

[1] L'agrégation consiste bien souvent en une simple moyenne arithmétique.

ou non. Dans notre contexte, nous pouvons considérer que les performances des deux classifieurs étudiés C_1 et C_2, sont des variables aléatoires, X_1 et X_2, pour lesquelles nous disposons de m observations : une par paire de bases (apprentissage,test). Le test de Student s'applique alors pour comparer les performances moyennes $\overline{X_1}$ et $\overline{X_2}$. Les m bases de données d'apprentissage et de test étant identiques pour C_1 et C_2, cela implique que les observations de X_1 et X_2 sont dépendantes deux à deux. C'est pour cette raison qu'il faut utiliser la version du test de Student qui s'applique aux échantillons appariés.

Si l'on note d la variable aléatoire correspondant à la différence entre les performances des deux classifieurs, on a $d = X_1 - X_2$. Nous disposons pour d de m observations $d_1, ..., d_m$. Notre objectif est de savoir si la différence moyenne $\overline{d} = \overline{X_1} - \overline{X_2}$, observée au cours d'expérimentations sur un nombre limité d'échantillons, est révélatrice d'une différence entre les performances moyennes μ_1 et μ_2 de C_1 et C_2 sur l'ensemble de la population. Pour y parvenir, nous pouvons procéder à la manière des démonstrations par l'absurde.

Nous faisons l'hypothèse qu'il n'y a pas de différence entre les deux classifieurs. Sous cette hypothèse, la distribution théorique de $\overline{d}$ est connue. On peut alors calculer la probabilité qu'un tirage aléatoire selon cette distribution donne une valeur au moins aussi éloignée de 0 que ne l'est le $\overline{d}$ que nous avons observé. Si celle-ci est suffisamment faible, alors nous pouvons remettre en cause l'hypothèse initiale.

Plus formellement, l'hypothèse initiale, notée $H0$ est dite hypothèse nulle. Nous posons $H0 : \mu_1 = \mu_2$ ce qui peut encore s'écrire $H0 : \mu_d = 0$. L'hypothèse concurrente se note $H1 : \mu_1 \neq \mu_2 : \mu_d \neq 0$. Si l'on admet $H0$ et sous certaines conditions que nous préciserons ultérieurement, la variable $T = \frac{\overline{d}}{s_{\overline{d}}}$ suit une loi de Student à $m-1$ degrés de liberté, où $s_{\overline{d}}$ désigne l'écart-type empirique de $\overline{d}$.

$$\begin{cases} \overline{d} & = & \frac{1}{m}\sum_{i=1}^{m} d_i \\ s_{\overline{d}}^2 & = & \frac{1}{m(m-1)}\sum_{i=1}^{m}\left(d_i - \overline{d}\right)^2 \end{cases}$$

La statistique de Student t correspond alors à la valeur de T que les expériences permettent d'observer. Soit p la probabilité que T ne prenne pas une valeur dans l'intervalle $[-t, t]$.

$$p = P\left(T \notin [-t, t]\right) = 1 - \int_{-t}^{t} p_t(x)dx$$

où p_t désigne la densité de probabilité d'une variable suivant la loi de Student. p est également appelée *p-valeur*.

Soit α la probabilité maximale de rejeter à tort $H0$. Selon nos exigences, α sera plus ou moins proche de 0. Les valeurs les plus couramment employées dans la littérature sont 0.1, 0.05 et 0.01. α est généralement appelé le risque de première espèce ou probabilité de l'erreur de type I. On note β la probabilité de l'erreur de type II ou risque de deuxième espèce. Il s'agit de l'erreur qui est faite lorsque $H0$ est acceptée à tort. La *puissance* d'un test est égale à $1 - \beta$. Idéalement on souhaite que α et β soient aussi proches de 0 que possible, mais il convient en pratique de trouver un compromis entre les deux. Il est en effet possible de montrer que la réduction du risque de première espèce s'accompagne d'une baisse de la puissance et réciproquement (Saporta, 2006).

Pour savoir quelle hypothèse accepter, on applique la règle de décision suivante :

$$\begin{array}{lcl} p \leq \alpha & \Rightarrow & \text{Rejeter H0} \\ p > \alpha & \Rightarrow & \text{Accepter H0} \end{array}$$

Pour que ce test soit valide d'un point de vue théorique, les $\overline{d}_i$ doivent suivre une loi normale et être indépendants.

Lorsque le critère de performance est le taux de bonnes classifications, nous avons $\forall j = 1, 2\ X_{ji} = \sum_{l=1}^{k_i} x_{jl}$, où x_{jl} est une variable aléatoire valant 1 si C_j classe correctement l'exemple l de la i-ième base de test et 0 sinon. k_i est le nombre d'exemples de cette base. Chaque X_{ji} correspond à la performance du classifieur C_j sur la i-ième base de test et est une somme de k_i variables de Bernoulli. Elle suit donc une loi binomiale si l'on suppose que ces variables sont indépendantes. D'après le théorème de la limite centrale, lorsque k_i tend vers l'infini, la loi normale constitue une bonne approximation de cette loi binomiale.

En pratique on considère que ce théorème s'applique dès que $k_i \geq 30$. Si chacune des m bases de test contient plus de 30 exemples à classer, on peut considérer que tous les X_{ji} suivent une loi normale. Les d_i, qui sont des différences entre deux variables suivant une loi normale, suivent alors également une loi normale. Le premier critère de validité du test de Student est donc vérifié. Pour que le second le soit également il faut encore que les d_i soient indépendants entre eux. Ce point est beaucoup plus délicat. Il impose en effet que les m bases de test, ainsi que les m bases d'apprentissage, soient indépendantes entre elles. Cela est envisageable si l'on dispose de suffisamment de données. Mais cela est rarement le cas en pratique.

Se pose alors la question du choix de la technique d'échantillonnage utilisée pour créer les m paires de bases. Les deux principales techniques sont la validation croisée et le rééchantillonnage aléatoire avec remise (bootstrap). L'intérêt de la validation croisée est qu'elle assure l'indépendance des bases de test. En revanche les bases d'apprentissage se recouvrent partiellement et ne sont donc pas indépendantes.

5.2.2 Test de McNemar

Si le test de Student est largement utilisé dans la littérature (Demsar, 2006), il est également fort critiqué du fait des conditions requises pour sa validité. De nombreux autres tests ont été développés pour comparer deux taux de bonnes classifications, mais nous ne détaillerons que celui qui est mis en avant par Salzberg (1997) dans sa critique des pratiques expérimentales usuelles dans le domaine de l'apprentissage supervisé.

Le test de McNemar se construit à partir de la matrice suivante, qui permet de décrire les performances de deux classifieurs C_1 et C_2 sur une même base de données.

n_{00}	n_{01}
n_{10}	n_{11}

n_{00} est le nombre d'exemples mal classés par C_1 et C_2, n_{01} le nombre d'exemples mal classés par C_1 mais bien classés par C_2, n_{10} le nombre d'exemples bien classés par C_1 mais pas par C_2 et n_{11} le nombre d'exemples correctement classés par les deux.

Sous $H0$, C_1 et C_2 ont les mêmes performances. On devrait donc avoir $n_{01} = n_{10}$. Le test de McNemar revient alors à effectuer un test du χ^2 pour comparer la distribution observée des n_{ij} avec celle que l'on observerait sous l'hypothèse nulle. Cette dernière peut se représenter par la matrice suivante :

n_{00}	$(n_{01} + n_{10})/2$
$(n_{01} + n_{10})/2$	n_{11}

Sous $H0$, la statistique $M = \frac{(|n_{01}-n_{10}|-1)^2}{n_{01}+n_{10}}$ suit une loi du χ^2 à 1 degré de liberté. Une fois un niveau de confiance $(1-\alpha)$ fixé, la règle de décision du test de McNemar est alors :

$$M > \chi^2_{1,\alpha} \quad \Rightarrow \quad \text{Rejeter H0}$$
$$M \leq \chi^2_{1,\alpha} \quad \Rightarrow \quad \text{Accepter H0}$$

où $\chi^2_{1,\alpha}$ est la valeur telle que $P\left(x \sim \chi^2_1 \notin \left[-\chi^2_{1,\alpha}, \chi^2_{1,\alpha}\right]\right) = \alpha$, $x \sim \chi^2_1$ indiquant que x suit une loi du χ^2 à 1 degré de liberté. Les tables donnant les valeurs critiques $\chi^2_{1,\alpha}$ pour la distribution du χ^2 à 1 degré de liberté sont disponibles dans la plupart des ouvrages de statistique.

Dietterich (1998), dans un article précurseur sur l'évaluation de classifieurs, compare les performances de différents tests statistiques vis-à-vis du risque de première espèce. Les performances de deux classifieurs, théoriquement identiques, sont simulées et chaque test est appliqué avec un niveau de confiance de 0.95 pour savoir si ce test considère qu'il y a une différence significative entre les deux ou non. Cette expérience est répétée 1000 fois, le risque de première espèce de chaque test est alors estimé par la fréquence relative du nombre d'erreurs qu'il a commises. Le test ayant le plus faible taux d'erreurs est évidemment celui qui sera préféré. Ajoutons qu'un test dont le taux d'erreurs est supérieur au risque de première espèce ($\alpha = 0.05$), ne sera pas jugé fiable.

De ces expérimentations il ressort que le taux d'erreurs de type I du test de Student combiné au rééchantillonnage aléatoire est bien trop élevé. La validation croisée à 10 ensembles, si elle donne de meilleures performances, ne peut pas non plus être considérée comme fiable. Pour diminuer le biais lié à la dépendance entre les bases d'apprentissage qui est vraisemblablement la cause de la faiblesse de la validation croisée à 10 sous-ensembles, Dieterrich propose de procéder à 5 validations croisées à 2 sous-ensembles. Il s'avère que cette procédure, associée au test de Student, donne des résultats fiables, comparables à ceux du test McNemar.

5.2.3 Adaptation des tests à une mesure de performance quelconque

Nous avons vu jusqu'ici comment appliquer ces deux tests pour comparer deux taux de bonnes classifications. Ce critère est cependant très controversé. En effet, il ne permet pas de tenir compte du fait que les erreurs faites sur deux classes distinctes n'ont pas forcément les mêmes coûts. Dans une application de diagnostic médical par exemple, lorsqu'il s'agit de déterminer si tel patient est porteur de telle maladie, il est essentiel de ne pas se tromper sur les patients malades. Le coût d'erreur de la classe des malades est nettement supérieur à celui de la classe des patients sains. De même, dans le domaine de la détection des crises, il est essentiel de ne pas passer à côté de crises potentielles.

Pour prendre en compte cette asymétrie, de nombreuses mesures de performance ont été élaborées, comme par exemple l'aire sous la courbe ROC (AUC) ou les F-mesures. En section 2.3, nous avons également proposé notre propre mesure combinant le taux de bonnes classifications et le rappel de la classe *crise*, celle dont le coût d'erreur est le plus élevé. Il est donc important d'envisager la comparaison de classifieurs de manière générique, indépendamment du critère qui aura été retenu pour mesurer les performances d'un classifieur.

Pour le test de Student, il n'est plus possible de considérer que les X_{ji} suivent une loi binomiale que l'on peut approcher par une loi normale. Les d_i suivent *a priori* une loi quelconque. Mais la statistique qui nous intéresse $\overline{d} = \frac{1}{m}\sum_{i=1}^{m} d_i$ est la moyenne de m variables indépendantes. Ici on suppose que la seconde condition de validité du test est respectée. D'après le théorème de la limite centrale, lorsque m tend vers l'infini, nous pouvons conclure que $\overline{d}$ suit une loi normale. La première condition de validité du test est donc bien respectée si le nombre m de paires de bases issues du rééchantillonnage est suffisamment grand. En pratique il faut $m \geq 30$. Notons que nous n'avons plus besoin de la condition portant sur le nombre minimum d'exemples que doit contenir chacune des m bases de test : $k_i \geq 30$.

En ce qui concerne la seconde condition, relative à l'indépendance des d_i, les difficultés sont les mêmes que précédemment. La répétition de validations croisées à 2 ensembles étant la technique d'échantillonnage la mieux à même de garantir la fiabilité du test de Student, il convient donc, lorsque le critère de performance est quelconque, d'itérer 15 fois une telle validation croisée, afin de respecter la condition $m \geq 30$.

Le test de McNemar n'est pas applicable lorsqu'un critère autre que le taux de bonnes classifications est utilisé. En revanche, le test binomial (Salzberg, 1997), proche du test de McNemar, peut être étendu à n'importe quelle mesure de performance. Soit z une variable qui vaut 1 si C_1 a de meilleures performances que C_2 et 0 sinon. z est une variable de Bernoulli.

Notons s le nombre de fois, parmi les m expériences, où C_1 est meilleur que C_2 (sous l'hypothèse non restrictive que C_1 obtient plus de meilleurs résultats que C_2) : $s = \sum_1^m z_i$. Si les observations z_i de z sont indépendantes (ceci correspond à la seconde condition de validité du test de Student), alors s suit une loi binomiale de paramètres m et $p = P(z_i = 1)$. Sous $H0$, les deux classifieurs ayant les mêmes performances, on doit avoir $p = 0.5$ et par conséquent $E[s] = 0.5 \times m$, où E désigne l'espérance mathématique. Si l'on note s_{obs} la valeur observée de s, on obtient

$$q = P(s \geq s_{obs} | p = 0.5) = \sum_{s=s_{obs}}^{m} \frac{m!}{s!\,(m-s)!} (0.5)^m$$

Une fois fixé le niveau de confiance $(1 - \alpha)$, on applique la règle de décision suivante :

$q \leq \alpha \quad \Rightarrow \quad$ Rejeter H0
$q > \alpha \quad \Rightarrow \quad$ Accepter H0

Ce test permet de se passer de la condition de normalité du test de Student, mais ne prend pas en compte la magnitude des différences observées entre C_1 et C_2. Pour $m > 30$, d'après le théorème de la limite centrale, on peut considérer que s suit une loi normale et appliquer le test des signes (*sign test*) (Dietterich, 1998). Dietterich, suite à ses expérimentations, déconseille cependant l'utilisation de ce test.

Jusqu'à présent nous avons considéré que nous ne disposions que d'une seule base de données pour comparer C_1 et C_2. Les tests que nous avons présentés nous permettent de savoir si l'un des deux est significativement meilleur que l'autre. Mais ces conclusions ne s'appliquent qu'au problème particulier correspondant à la base de données utilisée. En apprentissage automatique, il est fréquent de chercher à savoir si tel classifieur obtient de meilleures performances que tel autre, non pas sur un problème particulier, mais sur un ensemble de problèmes. Les deux classifieurs sont alors testés non plus sur une seule mais sur n bases de données. Se pose alors la question de savoir si les méthodes de comparaison envisagées jusqu'ici sont toujours valides.

5.3 Deux classifieurs évalués sur n bases de données

5.3.1 Limites des tests de McNemar et Student

Appliquer les tests précédents lorsque nous disposons des performances de deux classifieurs sur n problèmes de classification ne pose pas de problème à première vue. La seule différence avec ce qui précède concerne les observations de X_j dont nous disposons. Nous ne considérons plus m observations correspondant aux performances sur chacune des m sous-bases issues du rééchantillonnage de la base de données d'origine. Dans cette section, les X_{ji} désignent les performances de C_j sur chacun des n problèmes de classification. Il faut

donc remplacer m par n dans toutes les formules précédentes pour pouvoir les appliquer dans ce contexte.

L'impact de cette modification est cependant loin d'être aussi négligeable qu'il y paraît. La seconde condition de validité liée à l'indépendance des X_{ji} n'est plus problématique puisque les classifieurs sont évalués sur des bases de données indépendantes. Rappelons cependant que pour satisfaire la condition de normalité, il nous faut $n \geq 30$ mesures de performance. Or il est beaucoup plus facile de générer par rééchantillonnage 30 sous-bases de données que d'en disposer de 30 distinctes.

Mais quand bien même serions-nous capable de réunir ces 30 bases, nous sommes confronté au problème d'incommensurabilité, souligné par Demsar (2006). Le nombre de classes, leur distribution, les coûts associées aux différentes erreurs sont propres à chaque problème. Il est donc quelque peu illusoire de chercher à comparer ou additionner les performances d'un même classifieur obtenus sur des bases de données correspondant à des problèmes distincts. Or le test de Student repose sur le calcul des performances moyennes et met donc au même niveau des scores qui ne sont pourtant pas comparables. Ceci implique qu'il est peu vraisemblable, pour un même critère, que les performances estimées sur les différentes bases de données soient distribuées selon la même loi. Le théorème de la limite centrale qui permet de justifier l'hypothèse de normalité ne peut donc plus s'appliquer.

Demsar met en avant une dernière faiblesse du test de Student : sa sensibilité aux valeurs anormales, elle aussi liée à l'agrégation par la moyenne des performances.

5.3.2 Test de Wilcoxon

Pour pallier ces faiblesses il est possible de se tourner vers la version non paramétrique du test de Student avec échantillons appariés : le test de Wilcoxon. Conservant les notations précédentes, d_i désigne la différence entre les performances de C_1 et C_2 sur la i-ième base de données : $d_i = X_{1i} - X_{2i}$. Contrairement au test de Student, le test de Wilcoxon n'utilise pas directement les d_i, mais plutôt leur rang. La procédure à réaliser pour mettre en place ce test est la suivante.

- Trier les n d_i par ordre croissant de leur valeur absolue. Le rang de d_i est notée r_i. Lorsque plusieurs d_i ont même valeur absolue, on leur affecte à tous le même rang : la moyenne des rangs qui leur auraient été affectés s'il n'y avait pas eu égalité.
- Calculer R^+ et R^-, les sommes des rangs des bases de données sur lesquelles C_1 est meilleur que C_2 ($d_i > 0$) et respectivement sur lesquelles C_2 est meilleur que C_1 ($d_i < 0$). Les rangs des bases de données sur lesquelles les deux classifieurs ont des performances identiques, sont répartis entre R^+ et R^-.

$$\begin{aligned} R^+ &= \sum_{i/d_i>0} r_i + \frac{1}{2} \sum_{i/d_i=0} r_i \\ R^- &= \sum_{i/d_i<0} r_i + \frac{1}{2} \sum_{i/d_i=0} r_i \end{aligned}$$

- Soit $T = \min(R^+, R^-)$. On peut appliquer la règle de décision suivante :

  $T \geq W_{n,\alpha} \quad \Rightarrow \quad$ Rejeter H0
  $T < W_{n,\alpha} \quad \Rightarrow \quad$ Accepter H0

  où $W_{n,\alpha}$ est la valeur telle que $P(x \sim W_n \notin [-\infty, W_{n,\alpha}]) = \alpha$, $x \sim W_n$ indiquant que x est la statistique de Wilcoxon pour n échantillons. Les tables donnant les

valeurs critiques $W_{n,\alpha}$ lorsque $n < 30$ sont disponibles dans la plupart des ouvrages de statistique. Lorsque $n \geq 30$, on calcule

$$z = T - \frac{\frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

Sous $H0$, z suit une loi normale pour laquelle les valeurs critiques sont toutes connues.

Ce test, tout comme celui de Student, suppose que les performances sont commensurables, mais uniquement d'un point de vue qualitatif. Seules des comparaisons entre les performances sont effectuées. Celles-ci ne sont jamais agrégées comme dans le test de Student. De plus, il n'est aucunement fait l'hypothèse que les d_i sont distribuées selon une loi gaussienne, hypothèse qui, on l'a vu, est hautement contestable. Enfin, seuls les rangs étant utilisés, ce test est plus robuste que celui de Student vis-à-vis des valeurs anormales.

La magnitude des différences n'est en revanche pas prise en compte directement comme c'est le cas pour le test de Student. Mais, contrairement au test binomial, elle l'est tout de même indirectement, par l'intermédiaire des rangs. Pour ces raisons, le test de Wilcoxon apparaît comme l'une des meilleures solutions pour comparer les performances de deux classifieurs évalués sur $n > 1$ bases de données.

Précisons que pour que ce test soit valide théoriquement, deux conditions doivent être remplies.

- Les d_i doivent être des observations d'une variable continue.
- Les d_i doivent être indépendantes.

Rappelons que la deuxième condition vaut également pour le test de Student. La première peut paraître plus problématique puisque de nombreux critères de performance sont des mesures discrètes, comme par exemple le taux de bonnes classifications. Cependant, même pour ces critères, lorsque le nombre d'exemples à classer est suffisamment grand, ils peuvent être considérés comme des variables continues.

Ce test peut très bien s'appliquer lorsque les classifieurs ne sont évalués que sur une seule base de données. Nous n'avons cependant pas jugé utile de l'inclure dans la section précédente, car l'intérêt du test de Wilcoxon, vis-à-vis de celui de Student, se justifie surtout lorsque nous testons deux classifieurs sur des problèmes indépendants. Lorsqu'une seule base de données est utilisée, l'hypothèse de normalité faite par le test de Student peut facilement se justifier. Or l'avantage du test de Wilcoxon sur celui de Student n'est effectif que lorsque cette hypothèse n'est pas vérifiée, ce qui est le cas dans cette section. C'est l'hypothèse d'indépendance qui était problématique dans la section précédente. Or le test de Wilcoxon repose également sur cette même hypothèse.

Toutes les techniques que nous avons abordées dans ces deux premières sections ne permettent de comparer que deux classifieurs. Or il est souvent utile de pouvoir mettre en balance $k > 2$ classifieurs. C'est d'ailleurs ce que nous ferons constamment par la suite.

5.4 k classifieurs évalués sur une seule base de données

La plupart des études expérimentales en apprentissage automatique cherchent à comparer une nouvelle technique avec celles de la littérature pour évaluer l'intérêt relatif de cette technique. Lorsque des tests statistiques sont utilisés pour justifier les conclusions des auteurs, ce qui n'est pas toujours le cas, les auteurs ont majoritairement recours au test

de Student avec échantillons appariés, tel que nous l'avons décrit à la section 5.2 (Demsar, 2006; Salzberg, 1997).

Ce test, comme tous ceux que nous avons vus jusqu'ici, ne peut pas s'appliquer directement pour comparer plus de deux méthodes à la fois (Hull, 1993). La solution qui est alors retenue consiste à procéder aux comparaisons entre les $\frac{k(k-1)}{2}$ paires de méthodes, où k est le nombre de méthodes étudiées. Pour chacune de ces comparaisons, seules deux méthodes entrent dans la comparaison. Les tests précédents peuvent donc s'appliquer sans problème.

Cette solution naïve est cependant fortement décriée (Demsar, 2006; Salzberg, 1997; Hull, 1993). Elle conduit en effet à des interprétations fallacieuses. Supposons que nous ayons n paires de classifieurs à comparer. Soit α^* le taux d'erreurs que l'on s'autorise pour chacun des tests. α^* correspond à la probabilité de se tromper sur un test lorsque l'on rejette l'hypothèse nulle. On nommera cette probabilité le taux d'erreurs de comparaison. Soit α la probabilité de faire au moins une erreur sur les n tests. On parlera de taux d'erreurs global[2]. Si les n comparaisons à effectuer sont indépendantes, nous avons $\alpha = 1-(1-\alpha^*)^n$. Salzberg (1997) note que pour $n = 154$, fixer α^* à 5%, qui est une pratique courante, conduit à $\alpha = 99.66\%$. En procédant ainsi, nous sommes donc quasiment sûr que l'un des 154 tests donnera une réponse fausse, ce qui n'est évidemment pas souhaitable.

Il convient de commencer par poser correctement le problème. Nous cherchons à savoir si les performances de k classifieurs sont toutes identiques ou non. Si tel est le cas il faut ensuite identifier quels classifieurs diffèrent effectivement. Il y a donc deux étapes distinctes, que les statisticiens ont depuis longtemps étudiées. Dans cette section nous considérons le cas où seule une base de données est à disposition.

5.4.1 Analyse de la variance

Nous avons vu à la section 5.3 que dans ces conditions l'hypothèse de normalité faite par le test de Student est tenable. On peut donc légitimement envisager l'extension de ce test. Il s'agit de l'analyse de la variance, dite ANOVA. Elle permet de comparer non plus deux mais k moyennes. L'hypothèse nulle $H0$ s'écrit alors $\mu_1 = \mu_2 = ... = \mu_k$, où μ_i désigne la moyenne des performances du classifieur C_i. Chaque échantillon utilisé pour évaluer un classifieur est également utilisé pour tous les autres classifieurs[3]. Ceci impose d'utiliser la version de l'ANOVA dite *ANOVA à mesures répétées* (Zar, 1999).

Cette technique repose sur la décomposition de la variance globale qui est une mesure de la variabilité observée entre les performances moyennes des k classifieurs. Plusieurs phénomènes sont sources de variabilité : l'échantillonnage qui introduit de la variabilité entre les m paires de base, les différences intrinsèques entre les classifieurs, l'interaction entre l'échantillonnage et la classification que l'on appelle généralement variabilité résiduelle. La variabilité globale mesurée par la somme des erreurs quadratiques peut donc se décomposer en une somme de trois mesures de variabilité.

$$SS_{tot} = SS_{ech} + SS_{cl} + SS_{res}$$

où SS_{tot} désigne la somme totale des erreurs quadratiques. SS_{ech} correspond à la somme des erreurs quadratiques liée à l'échantillonnage. Elle est classiquement appelée variabilité inter-sujets[4]. SS_{cl} correspond à la variabilité inter-classifieurs, classiquement appelée inter-groupes, tandis que SS_{res} désigne la variabilité résiduelle. Notons A un terme correctif qui

[2] Dans la terminologie anglo-saxonne, α^* correspond au *familywise error rate* (FWER).
[3] Dans le cas de deux classifieurs, nous parlions d'échantillons appariés.
[4] Les sujets correspondent ici aux m bases de test.

TAB. 5.1 – Décomposition de la variance effectuée par l'ANOVA

SS	ddl	V
$SS_{tot} = \sum_{i=1}^{m} \sum_{j=1}^{k} X_{ji}^2 - A$	$ddl_{tot} = k \times m - 1$	$V_{tot} = \frac{SS_{tot}}{ddl_{tot}}$
$SS_{ech} = \frac{\sum_{i=1}^{m} \left(\sum_{j=1}^{k} X_{ji}\right)^2}{k} - A$	$ddl_{ech} = m - 1$	$V_{tot} = \frac{SS_{ech}}{ddl_{ech}}$
$SS_{cl} = \frac{\sum_{j=1}^{k} \left(\sum_{i=1}^{m} X_{ji}\right)^2}{m} - A$	$ddl_{cl} = k - 1$	$V_{tot} = \frac{SS_{ech}}{ddl_{ech}}$
$SS_{res} = SS_{tot} - SS_{ech} - SS_{cl}$	$ddl_{res} = (k-1)(m-1)$	$V_{res} = \frac{SS_{res}}{ddl_{res}}$

interviendra dans le calcul des variances V_{tot}, V_{ech} et V_{cl}. On a :

$$A = \left(\sum_{i=1}^{m} \sum_{j=1}^{k} X_{ji} \right)^2$$

Le tableau suivant récapitule les formules donnant les sommes des erreurs quadratiques (SS), les degrés de liberté (ddl) et la variance (V) correspondant à chaque source de variabilité possible.

Si les classifieurs ont des performances identiques, la variance inter-classifieurs devrait être du même ordre que la variance résiduelle. Elle devrait être plus élevée dans le cas contraire. Une fois ces calculs réalisés, l'ANOVA à mesures répétées se ramène à un test de Fisher pour savoir si les variances V_{cl} et V_{res} sont identiques ou non. Pour cela il suffit de calculer $F = \frac{V_{cl}}{V_{res}}$ et de le comparer à la valeur critique associée au test de Fisher, que l'on peut elle aussi trouver dans les tables usuelles de statistique. Pour un niveau de confiance $(1 - \alpha)$ donné, la règle de décision qui s'applique peut s'écrire sous la forme suivante :

$$\begin{array}{lcl} F \geq F_{k-1,(k-1)(m-1),\alpha} & \Rightarrow & \text{Rejeter H0} \\ F < F_{k-1,(k-1)(m-1),\alpha} & \Rightarrow & \text{Accepter H0} \end{array}$$

où $F_{k-1,(k-1)(m-1),\alpha}$ vérifie $P\left(x \sim F_{k-1,(k-1)(m-1)} \notin \left[-\infty, F_{k-1,(k-1)(m-1),\alpha}\right]\right) = \alpha$, $x \sim F_{k-1,(k-1)(m-1)}$ indiquant que x est la statistique de Fisher avec $k-1$ degrés de liberté pour le numérateur et $(k-1)(m-1)$ pour le dénominateur.

Pour garantir la validité théorique de l'ANOVA, plusieurs conditions sont requises.

- Pour chaque classifieur C_j, les X_{ji} doivent suivre une distribution normale et être indépendants.
- La variance de cette distribution normale doit être la même pour tous les classifieurs.

Nous avons déjà évoqué la première condition qui est la même que celle du test de Student. Nous avons vu que l'hypothèse de normalité était tenable lorsque les tests sont faits sur m sous-bases provenant d'une même base d'origine. La seconde, dite hypothèse d'homoscédasticité, est en revanche beaucoup plus difficile à justifier, mais l'ANOVA est tout de même suffisamment robuste pour pouvoir être utilisée sainement, même lorsque l'homoscédasticité n'est pas vérifiée (Zar, 1999).

5.4.2 Tests *post-hoc* associés à l'ANOVA

Lorsque l'hypothèse nulle est rejetée, nous pouvons conclure que les k classifieurs n'ont pas les mêmes performances. Il convient alors de procéder à une série de tests *post-hoc* pour déterminer quels classifieurs diffèrent.

5.4.2.1 Test de Tukey

Si l'on souhaite comparer les $\frac{k(k-1)}{2}$ paires de classifieurs, nous avons vu que l'approche naïve consistant à utiliser le test de Student, ou tout autre dont l'objectif est de comparer deux moyennes, est mal adapté car elle ne tient pas compte du fait que l'on est face à un problème de comparaisons multiples. Cette approche n'assure le contrôle que du taux d'erreurs de comparaison α^*. Conçu pour être un post-traitement de l'ANOVA, le test de Tukey est l'une des méthodes les plus fréquemment employées pour contrôler le taux d'erreurs global α. Pour chaque paire de classifieurs à comparer C_i et C_j, on calcule la statistique $q = \frac{\overline{X_i}-\overline{X_j}}{\sqrt{\frac{V_{ech}}{m}}}$. V_{ech} est la variance résiduelle calculée au cours de l'ANOVA et m est le nombre de bases de test issues du rééchantillonnage. Sous l'hypothèse nulle $H0 : \mu_i = \mu_j$, q est distribuée selon la loi des écarts studentisés, ce qui nous permet d'adopter la règle de décision suivante :

$$\begin{array}{lcl} |q| \geq q_{k,ddl_{res},\alpha} & \Rightarrow & \text{Rejeter H0} \\ |q| < q_{k,ddl_{res},\alpha} & \Rightarrow & \text{Accepter H0} \end{array}$$

où $q_{k,ddl_{res},\alpha}$ vérifie $P\left(x \sim q_{k,ddl_{res}} \notin [-q_{k,ddl_{res},\alpha}, q_{k,ddl_{res},\alpha}]\right) = \alpha$, $x \sim q_{k,ddl_{res}}$ indiquant que x suit la loi des écarts studentisés dont les paramètres sont le nombre k de classifieurs comparés ainsi que le nombre de degrés de liberté associé à l'erreur quadratique résiduelle calculée au cours de l'ANOVA : $ddl_{res} = (k-1)(m-1)$.

5.4.2.2 Test de Dunnet

En apprentissage automatique, l'expérimentation est souvent utilisée pour évaluer si une nouvelle méthode est meilleure que celles qui existent, ce qui permet de justifier le développement de cette nouvelle approche. Dans ces conditions, toutes les paires de classifieurs n'ont pas à être comparées, mais seulement les $k-1$ où intervient le classifieur correspondant à la nouvelle technique. On l'appellera classifieur de référence. Le test de Tukey contrôle le taux d'erreurs global, en supposant que $\frac{k(k-1)}{2}$ comparaisons seront effectuées. Avec seulement $k-1$ comparaisons, le risque que l'une d'entre elles au moins soit jugée à tort significative est bien moins grand. Le test de Tukey est donc mal adapté. Le taux d'erreurs de type I sera certes bien garanti, mais la puissance du test en sera affectée. De nombreuses différences qui sont effectivement significatives ne seront pas jugées comme telles. Les statisticiens disent que, dans ces conditions, le test est trop conservatif.

Le test de Dunnet a été spécifiquement conçu pour traiter ce genre de problèmes. Son principe est voisin de celui de Tukey. La statistique q à évaluer est légèrement différente : $q = \frac{\overline{X_{ref}}-\overline{X_j}}{\sqrt{\frac{2V_{ech}}{m}}}$, où $\overline{X_{ref}}$ désigne la performance moyenne du classifieur qui sert de référence. Sous $H0$, q ne suit plus une loi des écarts studentisés, mais les valeurs critiques sont également disponibles dans des tables usuelles (Zar, 1999).

5.4.2.3 Procédures d'ajustement pour le contrôle du taux d'erreurs global

Les tests de Tukey et de Dunnet sont spécifiques à l'utilisation de l'ANOVA. Les hypothèses sous-jacentes justifiant leur validité sont les mêmes que pour l'ANOVA. Afin de pouvoir être indépendant de l'ANOVA, de nombreuses procédures ont été développées pour pouvoir appliquer n'importe quel test statistique lors de la comparaison d'une paire de classifieurs, tout en garantissant le contrôle du taux d'erreurs global.

L'ajustement de Dunn-Bonferroni est probablement la plus simple et la plus répandue de ces procédures. Elle est décrite par Salzberg (1997) dans le cadre de l'apprentissage

automatique. Elle consiste à ajuster le taux d'erreurs de comparaison α^* de telle sorte que le taux d'erreurs global reste inférieur à un seuil fixé α. Il suffit pour cela de prendre $\alpha^* = \frac{\alpha}{c}$ où c est le nombre de comparaisons que l'on souhaite faire (en général $k-1$ ou $\frac{k(k-1)}{2}$). Cette procédure est très appréciée pour sa simplicité, mais elle est controversée car très conservative (Supattathum *et al.*, 1994; Demsar, 2006). Elle dégrade sensiblement la puissance du test utilisé. Une variante de cette procédure, appelée test de Sidak améliore quelque peu la puissance. Elle consiste à prendre $\alpha^* = 1-(1-\alpha)^{\frac{1}{c}}$. Rappelons qu'il s'agit de l'expression exacte de α^* en fonction α, dans le cas où les comparaisons sont indépendantes.

Afin d'améliorer la puissance du test utilisé de manière conséquente, les statisticiens ont introduit de nouvelles procédures. Contrairement aux deux précédentes, elles ne considèrent plus que α^* doit être constant. Elles le font varier en fonction du nombre d'hypothèses qui ont déjà été testées ou du nombre d'hypothèses qui ont déjà été rejetées. Toutes ces procédures commencent par calculer les p-valeurs associées à chacun des c tests à effectuer. Ces p-valeurs sont ensuite triées par ordre croissant. Nous indicerons alors ces c p-valeurs de la plus faible à la plus élevée : $p_1 < p_2 < ... < p_c$. Nous noterons H_i l'hypothèse nulle associée à p_i. H_i énonce que la différence entre les deux classifieurs comparés n'est pas significative. Rappelons que plus une p-valeur est petite, moins il est vraisemblable que la différence observée ne soit due qu'au hasard, et donc que l'hypothèse concurrente soit valide. On distingue alors deux grandes familles de procédures.

- **Procédures descendantes** : elles évaluent les hypothèses séquentiellement en partant de la moins vraisemblable H_1. Pour tester chaque hypothèse H_i, la p-valeur associée p_i sera comparée avec un seuil qui dépend de son rang : $\alpha(i)$. Soit j le plus petit indice tel que $p_j > \alpha(j)$. Les hypothèses $H_1, ..., H_{j-1}$ seront rejetées, tandis que $H_j, ..., H_c$ seront acceptées.
- **Procédures ascendantes** : le principe est similaire. Elles évaluent également les hypothèses séquentiellement en comparant p_i à $\alpha(i)$, mais en partant de l'hypothèse la plus vraisemblable H_c. Soit j le plus grand indice tel que $p_j \leq \alpha(j)$. Les hypothèses $H_1, ..., H_j$ seront rejetées, tandis que $H_{j+1}, ..., H_c$ seront acceptées.

Les différentes procédures de chaque famille diffèrent par le choix de $\alpha(i)$. Demsar (2006) donne quelques exemples de telles procédures et de leur application à la comparaison de classifieurs. Suite à une comparaison empirique de six d'entre elles, Supattathum *et al.* (1994) constatent que la procédure ascendante de Holland et Copenhaver est la plus satisfaisante. Elle permet en effet d'obtenir une bonne puissance (faible taux d'erreurs de type II) tout en assurant un contrôle efficace du taux d'erreurs global α. Elle se caractérise par le choix de $\alpha(i) = 1-(1-\alpha)^{\frac{1}{c-i+1}}$. Aussi avons-nous décidé d'y avoir recours par la suite.

5.5 k classifieurs évalués sur n bases de données

Nous avons vu à la section précédente comment l'ANOVA à mesures répétées pouvait être utilisée pour comparer plusieurs classifieurs évalués sur une seule base de données. Or, comme nous l'avons mentionné précédemment, il est fréquent que des classifieurs soient évalués sur différentes bases de données. Dans de tels cas, l'hypothèse de normalité faite par le test de Student a été vivement critiquée du fait de la non-commensurabilité des performances estimées sur des domaines distincts. Or l'ANOVA repose également sur l'hypothèse que les mesures de performance $X_1, X_2, ..., X_k$ des k classifieurs sont distribuées

selon des lois normales. Le problème est donc ici identique. L'ANOVA suppose de plus que les variances de ces k lois normales sont identiques. Cette hypothèse est également fortement contestable. Pour ces raisons, nous estimons qu'il est préférable de se tourner vers des solutions non paramétriques (Brazdil et Soares, 2000; Demsar, 2006).

5.5.1 Test de Friedman

L'équivalent non paramétrique de l'ANOVA à mesures répétées est le test de Friedman (Zar, 1999). Demsar (2006) en donne une description dans le cas précis de la comparaison de classifieurs. Il correspond à l'application de l'ANOVA non pas directement sur les performances des classifieurs à comparer, mais sur les rangs de ces performances. De même que le test de Student est un cas particulier de l'ANOVA à mesures répétées, le test de Wilcoxon, présenté à la section 5.4, est un cas particulier du test de Friedman que nous allons maintenant détailler.

Soit r_i^j le rang de X_{ji}. Nous noterons $R_j = \frac{1}{n}\sum_{i=1}^{n} r_i^j$ le rang moyen obtenu par le classifieur C_j. Le test de Friedman a pour objectif d'évaluer la vraisemblance de l'hypothèse nulle $H0 : R_1 = R_2 = ... = R_k$. Si celle-ci est suffisamment faible, il sera possible d'inférer que les performances des différents classifieurs ne sont pas équivalentes. Sous $H0$, il est possible de montrer que la statistique de Friedman $\chi_F^2 = \frac{12n}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right]$ suit une loi du χ^2 à $k-1$ degrés de liberté, lorsque le nombre de classifieurs k et le nombre de bases de données n est suffisamment grand.

On estime que cela vaut pour $k > 5$ et $n > 10$ (Demsar, 2006). La règle de décision est alors similaire à celle qui a été détaillée pour le test de McNemar à la section 5.2, qui lui aussi repose sur un test du χ^2. Pour des valeurs de k et n plus petites, les valeurs critiques exactes du test ont été établies (Zar, 1999).

À l'instar de ce qui a été fait pour le test de Wilcoxon à la section 5.3, lorsque plusieurs classifieurs obtiennent des performances identiques sur une base de données, les rangs qui leur sont effectivement attribués correspondent à la moyenne des rangs qu'ils auraient obtenus si ces performances avaient été légèrement différentes. Par exemple si l'on a 5 classifieurs dont les performances sur la i-ième base de données sont telles que l'on a $X_{1i} > X_{2i} = X_{3i} = X_{4i} > X_{5i}$, nous aurons $r_i^1 = 1$, $r_i^5 = 5$ et $r_i^2 = r_i^3 = r_i^4 = \frac{2+3+4}{3} = 3$. Plus ces égalités de rang seront nombreuses et plus χ_F^2 sera sous-estimée. Pour corriger ce biais on calcule $\frac{\chi_F^2}{A}$ (Zar, 1999; Brazdil et Soares, 2000). A est un facteur correctif, dépendant du nombre T d'égalités entre rangs observées sur les n bases de données, ainsi que des nombres t_i $(i = 1..T)$ de classifieurs impliqués dans chacune des égalités observées. On a $A = 1 - \frac{\sum_{i=1}^{T} t_i^3 - t_i}{n(k^3-k)}$.

Ce test a l'inconvénient d'être très conservatif. Il passe à côté de nombreuses différences statistiquement significatives. Iman et Davenport ont introduit une statistique basée sur celle de Friedman, qui est nettement moins conservative :

$$F_F = \frac{(n-1)\,\chi_F^2}{n\,(k-1) - \chi_F^2}$$

Sous $H0$, F_F suit une F-distribution, celle qui est suivie par la statistique de Fisher et que nous avons évoquée à propos de l'ANOVA à la section 5.4. La règle de décision à utiliser est très proche de celle qui avait alors été présentée. La seule différence concerne le nombre de degrés de liberté de cette F-distribution : $k-1$ pour le numérateur et $(k-1)\,(n-1)$ pour le dénominateur.

5.5.2 Tests *post-hoc* associés au test de Friedman

Le test de Friedman nous permet de savoir si les différences observées entre les différents classifieurs sont significatives ou non. Si tel est le cas, il faut ensuite procéder à des comparaisons multiples via des tests que l'on qualifie de *post-hoc*, pour savoir quels classifieurs diffèrent. Lorsque l'on souhaite comparer les $\frac{k(k-1)}{2}$ paires de classifieurs, l'équivalent non paramétrique du test de Tukey est le test de Nemenyi (Demsar, 2006). Pour toute paire de classifieurs (C_i, C_j), l'hypothèse nulle énonce que les rangs moyens de C_i et C_j sont identiques. On a $H0 : R_i = R_j$.

Sous $H0$, $q = \frac{R_i - R_j}{\sqrt{\frac{k(k+1)}{6n}}}$ suit une loi normale.

Contrôler le taux d'erreurs global sachant que l'on doit effectuer $\frac{k(k-1)}{2}$ comparaisons, impose d'utiliser des valeurs critiques liées à celles qui sont fournies par la loi des écarts studentisés. Le test de Nemenyi consiste alors à appliquer la règle de décision suivante.

$$|q| \geq \frac{q_{k,\infty,\alpha}}{\sqrt{2}} \quad \Rightarrow \quad \text{Rejeter H0}$$
$$|q| < \frac{q_{k,\infty,\alpha}}{\sqrt{2}} \quad \Rightarrow \quad \text{Accepter H0}$$

où $q_{k,\infty,\alpha}$ vérifie $P\left(x \sim q_{k,\infty} \notin [-q_{k,\infty,\alpha}, q_{k,\infty,\alpha}]\right) = \alpha$, $x \sim q_{k,\infty}$ indiquant que x suit la loi des écarts studentisés dont les paramètres sont le nombre k de classifieurs comparés et un nombre de degrés de liberté infini.

Ce test, tout comme sa version paramétrique (le test de Tukey), ajuste le taux d'erreurs de comparaison, pour prendre en compte le fait que $\frac{k(k-1)}{2}$ comparaisons seront effectués et ainsi contrôler le taux d'erreurs global. Il ne faut donc pas l'utiliser lorsqu'un des classifieurs sert de référence et que l'on souhaite comparer les $k-1$ autres à celui-ci.

Dans de tels cas, il vaut mieux utiliser un autre type de test et contrôler le taux d'erreurs global via l'une des procédures d'ajustement décrites à la fin de la section 5.4. Demsar suggère d'utiliser un z test basé sur la loi normale. En effet, sous $H0$, q suit une loi normale. Il est donc possible de calculer la p-valeur associée. Nous pouvons ensuite appliquer la procédure d'ajustement ascendante de Holland-Copenhaver, qui nous semble mieux adaptée que celles qui ont été proposées par Demsar (2006).

5.6 Conclusion

Nous avons vu dans ce chapitre quelles techniques pouvaient être mises à profit pour s'assurer que les différences observées expérimentalement entre différents classifieurs ne sont pas le fruit du hasard, introduit via la procédure d'échantillonnage des données. Selon le nombre de classifieurs à comparer, et le nombre de bases de données sur lesquelles ces classifieurs sont évalués, différentes solutions sont plus ou moins adéquates. Nous avons également vu, à propos de l'utilisation du test de Student, que la mesure de performance choisie avait son importance, ainsi que la procédure d'échantillonnage qui doit pouvoir garantir un minimum d'indépendance entre les échantillons, ce qui n'est pas aisé lorsqu'une seule base de données est à disposition.

Par la suite, nous serons amené à comparer plusieurs algorithmes de prétraitement : traitement des données manquantes et sélection d'attributs. Nous plaçant dans le contexte de la classification supervisée, ces algorithmes seront évalués par l'intermédiaire des classifieurs construits sur les bases de données qu'ils auront prétraitées. Nous aurons donc la plupart du temps k classifieurs à comparer, évalués sur n bases de données.

Au vu de ce que nous venons de présenter, nous mettrons donc en place un test de Friedman pour savoir si les classifieurs ont des performances équivalentes au vu des différences

que l'on peut observer. Lorsque ce test nous indiquera que ces différences sont significatives, nous aurons alors recours au test de Nemenyi pour identifier, parmi toutes les paires de classifieurs, celles qui associent deux classifieurs dont les performances sont distinctes. Lorsqu'une technique servira de référence à laquelle les $k-1$ autres techniques seront comparées, nous opterons pour un z test combiné à la procédure d'ajustement ascendante de Holland-Copenhaver.

Nous allons maintenant aborder plus précisément les questions liées au traitement des données manquantes et à la sélection d'attributs, au cours desquelles nous serons amené à mettre en œuvre la méthodologie comparative que nous venons de décrire. Mais avant cela, il nous paraît important de mettre en garde le lecteur contre des interprétations abusives des tests que nous venons de présenter.

Supposons qu'un test T soit appliqué à partir de données D pour choisir entre une hypothèse nulle $H0$ et l'hypothèse complémentaire $H1$ qui lui est associée. Pour un niveau de confiance donné $(1-\alpha)$, le fait que T rejette $H0$ indique qu'en considérant que $H1$ est vraie, nous pouvons garantir que nous avons moins de $\alpha\%$ de chances de nous tromper et que $H0$ soit vraie. En revanche lorsque T accepte $H0$, cela veut simplement dire que de D seulement, nous ne pouvons pas conclure que $H1$ est vraie en garantissant une probabilité d'erreur inférieure à α. Dans notre cas, lorsque les tests employés ne concluent pas sur l'existence de différences significatives entre les classifieurs comparés, cela signifie que les données recueillies ne suffisent pas à révéler l'existence de différences significatives. En l'état on ne peut écarter l'idée que les différences observées soient dues à l'aléatoire introduit par le processus d'échantillonnage, sans accepter de commettre une erreur avec une probabilité supérieure à α.

Chapitre 6

Traitement des données manquantes

La base de données à partir de laquelle nous avons construit un premier modèle d'évaluation des risques contient de nombreuses données manquantes : plus du quart. Notre algorithme d'apprentissage ayant besoin de travailler sur une base complète, nous avons remplacé les valeurs manquantes par la valeur moyenne de l'attribut correspondant. Ceci n'est guère satisfaisant. Dans ce chapitre, nous allons envisager l'utilisation d'autres techniques, plus avancées, afin d'améliorer la qualité de notre modèle. L'absence de certaines valeurs ou la présence de valeurs erronées est un problème récurrent que l'on retrouve dans de nombreux domaines, en particulier en analyse de risque, lorsque celle-ci se base sur l'analyse de données historiques, ce qui est notre cas. Aussi avons-nous décidé dans ce chapitre de traiter la question de façon aussi générique que possible.

6.1 Position du problème

La plupart des techniques actuelles d'analyse et de fouille de données sont fortement dépendantes de la qualité des données. Or, dans des applications réelles, il est fréquent que nombre de valeurs soient erronées, incohérentes, ou tout simplement manquantes. Pour mener à bien des analyses valides, il est donc important de tenir compte de ces problèmes. Si les statisticiens se préoccupent depuis longtemps de cette question (Little et Rubin, 2002), essentiellement pour traiter les problèmes de non-réponse dans les questionnaires, cela est plus récent pour les chercheurs travaillant sur l'ADN (Oba *et al.*, 2003), dans le domaine de l'ingénierie logicielle (Song et Shepperd, 2007) ou la fouille de données. En apprentissage, les données manquantes peuvent faire chuter les performances d'un classifieur (Acuna et Rodriguez, 2004), voire le rendre inutilisable.

Nous nous proposons ici de recenser les principales méthodes existantes de traitement des données manquantes en section 6.4, en présentant leurs points forts et leurs faiblesses, d'un point de vue théorique. Nous introduisons ensuite une nouvelle technique, basée sur des considérations issues de la théorie de l'information et qui correspond mieux à nos besoins. Nous l'avons développée en collaboration avec Thanh Ha Dang (Dang et Delavallade, 2006; Delavallade et Dang, 2007). Il n'existe pas de meilleure méthode dans l'absolu. Chacune est plus ou moins adaptée pour répondre à un objectif donné, en fonction du type de problème à traiter. Pour cette raison, nous comparerons les différentes méthodes dans un cadre expérimental bien normalisé. Ceci devrait nous permettre de caractériser le comportement des différentes techniques en fonction des particularités de la base de données considérée. Nous nous intéresserons uniquement aux données manquantes, une valeur erronée ou incohérente pouvant être considérée comme manquante[1].

[1] La difficulté avec ce type de données réside alors dans leur identification, ce qui est un problème

6.2 Mécanismes de génération des données manquantes

Avant de voir quelles techniques sont couramment employées pour remédier à l'absence de données, il est important de présenter les différentes hypothèses quant à la distribution des valeurs manquantes sur lesquelles reposent ces techniques. Comprendre pourquoi les données sont manquantes, ce qui revient à identifier le mécanisme de génération de ces données, peut en effet faciliter le choix d'un traitement adapté. Selon que l'absence d'une donnée sera due à la défaillance temporaire d'un capteur ou à la volonté délibérée de masquer une information, pour ne citer que quelques-unes des multiples causes possibles, les techniques de traitement seront vraisemblablement différentes. Little et Rubin (2002), et à leur suite tous les chercheurs du domaine, distinguent trois cas de figure. Afin d'illustrer ces différents mécanismes nous nous appuierons sur un exemple fictif. Q désignera la matrice indicatrice des valeurs manquantes. Si l'on note $Q = [q_{ij}]_{i=1..n,j=1..p}$, et si « ? » désigne l'absence d'une valeur, nous avons :

$$q_{ij} = \begin{cases} 1 & \text{si } v_{ij} \text{ est manquante : } v_{ij} = ?, \\ 0 & \text{sinon.} \end{cases}$$

Nous noterons respectivement V^o et V^m les parties observée et manquante de V.

Tab. 6.1 – Cette base de données complète sera utilisée pour illustrer les différents mécanismes de génération des données manquantes. Nous noterons V_1 le PIB par habitant et V_2 le nombre d'années en guerre civile.

Pays Id	PIB/habitant($)	Années en guerre civile
1	12330	0
2	16180	0
3	23200	0
4	2820	10
5	9300	10
6	4170	10

Les trois mécanismes de génération des données manquantes diffèrent par les hypothèses qui sont faites à propos de la distribution de ces données, c'est-à-dire $P(Q = 1|V)$.

- **MCAR** : les données manquantes sont supposées avoir été générées de manière complètement aléatoire (MCAR est l'acronyme de *Missing Completely At Random*). L'absence d'une valeur ne dépend d'aucune variable, que celle-ci soit observée ou non. La probabilité pour une donnée d'être manquante est constante. Avec nos notations cela se traduit par la simplification suivante : $P(Q|V) = P(Q)$. Le tableau 6.2 en donne une illustration.
- **MAR** : Le mécanisme de génération des valeurs manquantes est supposé être aléatoire (MAR est l'acronyme de *Missing At Random*). L'absence d'une donnée peut dépendre des valeurs observées sur les autres variables, mais pas de sa propre valeur, ce que l'on écrira $P(Q|V) = P(Q|V^o)$. Le tableau 6.3 donne un exemple dans lequel on peut observer ce mécanisme.

différent.

– **NMAR** : Cette fois on considère que l'absence d'une donnée peut être liée à n'importe quelle variable observable ou non (NMAR signifie *Not Missing At Random*[2]). Aucune simplification de $P(Q|V)$ n'est alors possible. Ce phénomène est le plus difficile à modéliser et est malheureusement assez fréquent. On l'observe chaque fois qu'un capteur n'est pas capable de mesurer certaines valeurs, celles qui sortent du champ de mesure pour lequel il a été calibré par exemple. Le tableau 6.4 en donne un exemple.

TAB. 6.2 – MCAR : Supposons que le PIB soit une variable difficile à calculer. On peut alors choisir de ne l'estimer que pour certains pays, pris aléatoirement. On a $P(V_1 = ?|V_2 = 0) = P(V_1 = ?|V_2 = 10)$ et $P(V_1 = ?|V_1) = P(V_1 = ?)$.

Pays Id	PIB/habitant($)	Années en guerre civile
1	12330	0
2	?	0
3	23200	0
4	2820	10
5	?	10
6	4170	10

TAB. 6.3 – MAR : Supposons qu'il soit très difficile de calculer le PIB de pays ayant connu 10 années de guerre civile. Le fait qu'une valeur soit manquante pour V_1 dépend alors de V_2. On a $\left(P(V_1 = ?|V_2 = 10) = \frac{2}{3}\right) \neq (P(V_1 = ?|V_2 = 0) = 0)$.

Pays Id	PIB/habitant($)	Années en guerre civile
1	12330	0
2	16180	0
3	23200	0
4	2820	10
5	?	10
6	?	10

Il est important de constater que le premier cas de figure est assez réducteur, l'hypothèse sous-jacente étant très forte, alors que dans le dernier cas de figure aucune hypothèse particulière n'est faite. Si cela semble plus satisfaisant sur le plan théorique, il faut être conscient du fait que dans ces conditions le problème sera très difficile à appréhender. Le deuxième cas de figure correspond quant à lui à un compromis, les hypothèses sont contraignantes mais suffisamment relâchées pour permettre de développer des modèles efficaces.

Notons enfin qu'il est quasiment impossible en pratique de déterminer lequel des trois mécanismes est à l'œuvre à partir des données (Schafer et Graham, 2002). Comme l'expliquent Schafer et Graham l'intérêt d'une telle classification des mécanismes est purement théorique, elle est utile pour apprécier le domaine de validité de telle ou telle technique de traitement des données manquantes. Pour conclure cette partie, nous reproduisons sur la

[2]On trouve également l'acronyme NI pour *Not Ignorable*, ou encore MNAR pour *Missing Not At Random*.

TAB. 6.4 – NMAR : Supposons que ce sont les pays qui transmettent à la communauté internationale la valeur de leur PIB. On peut imaginer que ceux qui ont un PIB trop faible préfèrent ne pas divulguer le chiffre. Dans ce cas, le fait qu'une valeur soit manquante pour V_1 dépend de sa propre valeur. On parle alors de données censurées. On a $P(V_1 = ?|V_1 \leq 5000) = 1$ et $P(V_1 = ?|V_1 > 5000) = 0$.

Pays Id	PIB/habitant($)	Années en guerre civile
1	12330	0
2	16180	0
3	23200	0
4	?	10
5	9300	10
6	?	10

figure 6.1 l'un de leurs schémas qui donne une représentation graphique unifiée de ces trois mécanismes.

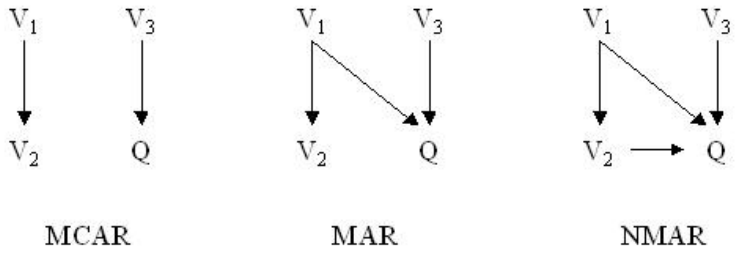

FIG. 6.1 – Une vue d'ensemble des mécanismes de génération des données manquantes pour une variable V_2. Q est la variable indicatrice des données manquantes relativement à V_2. V_1 désigne l'ensemble des variables disponibles autres que V_2. V_3 regroupe l'ensemble des causes de la génération des données manquantes qui sont indépendantes de V_1 et V_2.

6.3 Importance de la répartition des données manquantes

Si la distinction entre les différents mécanismes de génération des données manquantes s'avère utile essentiellement d'un point de vue théorique, la mise en évidence des différentes formes que peut prendre la répartition des données manquantes a un intérêt beaucoup plus pratique. En effet, ces différentes formes que nous appellerons motifs sont identifiables dans les applications réelles. Les statisticiens ont distingué trois principaux types de motifs (Little et Rubin, 2002; Schafer et Graham, 2002) :

- **univarié** : les données manquantes n'affectent qu'une seule variable ou alors un groupe de variables, mais dans ce cas les exemples pour lesquels les données manquent sont les mêmes pour toutes ces variables.
- **monotone** : on peut ranger les variables de telle sorte que si une variable n'est pas renseignée pour un exemple donnée, il en va de même pour toutes les variables suivantes.
- **quelconque** : Aucun réordonnancement de la matrice des données V ne peut faire apparaître l'un des motifs précédents.

À ces trois motifs, nous estimons utile d'en ajouter deux :

- **univarié étendu** : les données manquantes n'affectent qu'un groupe limité de variables. La répartition des données manquantes à l'intérieur de ce groupe peut être quelconque. Il s'agit donc d'une généralisation du motif univarié.
- **mono-instance étendu** : les données manquantes n'affectent qu'un groupe limité d'exemples. La répartition des données manquantes à l'intérieur de ce groupe peut être quelconque. Il s'agit de la transposition du motif précédent dans l'espace des exemples. Notons qu'il n'est pas besoin de définir la transposition dans l'espace des exemples du motif univarié strict, car il s'agit encore d'un motif univarié strict.

La figure 6.2 illustre ces 5 motifs en présentant l'allure générale de la matrice des données dans chacun des cas de figure.

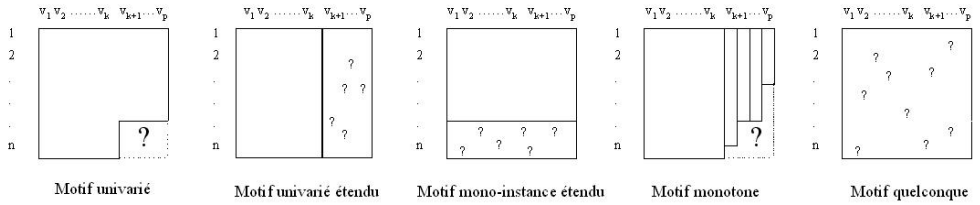

FIG. 6.2 – Motifs de répartition des données manquantes

Les motifs univariés stricts sont les plus simples à traiter, tandis qu'évidemment le motif quelconque sera le plus délicat à appréhender. Le motif monotone bien qu'il puisse paraître assez peu naturel est fréquemment rencontré en analyse de questionnaires, chaque fois que la réponse à un ensemble de questions est conditionnée par la réponse à une autre.

6.4 État de l'art sur le traitement des données manquantes

De nombreuses techniques de traitement des données manquantes ont été développées. Dans les années 90, Hu *et al.* (2000), sans prétendre être exhaustifs, en identifiaient déjà plus d'une vingtaine, pour la plupart issues des recherches en statistique. Depuis, les chercheurs en intelligence artificielle, bioinformatique et fouille de données entre autres, se sont mis à étudier la question et ont développé de nouvelles techniques. Recenser l'ensemble de ces techniques serait fastidieux. Aussi avons-nous opté pour une mise en évidence des principales caractéristiques des différentes méthodes. Ce travail nous permettra de dresser une taxinomie. Nous pourrons alors y placer les techniques les plus usitées et avoir ainsi une vue d'ensemble du domaine.

6.4.1 Vers une taxinomie des différentes méthodes

Lorsque l'on souhaite utiliser une base de données incomplète, trois stratégies sont possibles (Song et Shepperd, 2007).

1. Utiliser un algorithme qui permet intrinsèquement de travailler en présence de données manquantes, ou modifier un algorithme existant pour que cela devienne possible.

   Lorsque l'on fait de l'estimation de paramètres, l'algorithme *EM* (Expectation-Maximisation) introduit par Dempster *et al.* (1977) est une solution efficace, quoique coûteuse. Pour un problème de classification, des solutions ont été proposées pour adapter les arbres de décision, pour C4.5 ou CART par exemple (Feelders, 1999). Tous les algorithmes d'apprentissage basés sur des notions de distance ou de similarité (*k-moyennes*, *k plus proches voisins*) peuvent assez facilement s'adapter aux données manquantes. C'est ce que font Timm *et al.* avec l'algorithme des *c-moyennes floues* qui, à l'instar de l'algorithme *EM*, peut à la fois prendre en charge l'absence de certaines valeurs et leur trouver des valeurs de substitution.

   Timm *et al.* (2003) vont plus loin en proposant de considérer la distribution des données manquantes comme une donnée pertinente pour faire de la classification non supervisée. Il suffit de modifier la mesure de distance ou de similarité, par exemple en réduisant la dimension des vecteurs que l'on compare, pour n'intégrer à chaque fois que les composantes qui sont renseignées. Cependant on se retrouve à vouloir comparer des distances qui n'ont pas été mesurées sur les mêmes dimensions, ce qui peut poser problème. Certains travaux cherchent explicitement à intégrer le fait qu'une donnée soit manquante, en ajoutant une modalité supplémentaire pour chaque attribut incomplet.

2. Se ramener à une base de données complète par réduction de la dimension du problème. Pour cela tous les exemples de la base contenant des valeurs manquantes sont supprimés[3]. Cette technique, du fait de sa simplicité est fréquemment employée. Cependant elle présente deux inconvénients majeurs. D'une part, elle engendre de grosses pertes d'informations qui peuvent s'avérer dommageables, les techniques statistiques d'analyse des données ayant besoin d'un nombre suffisant d'échantillons pour que leurs inférences soient valides. Dans des cas qui ne sont pas rares, où la quasi-totalité des exemples possède des valeurs manquantes, elle devient même inutilisable. D'autre part, les statistiques, telles que la moyenne ou la variance, seront fortement

[3] On peut également choisir de supprimer toutes les variables dont certaines observations manquent, mais il faut être prudent car certaines peuvent être essentielles pour l'analyse.

biaisées, à moins que le mécanisme de génération des données ne soit complètement aléatoire (MCAR) (Magnani, 2003).

3. Se ramener à une base complète en trouvant un moyen adéquat pour remplacer les valeurs manquantes. On nomme ce procédé *imputation*, complétion ou substitution.

Avec certains algorithmes d'apprentissage il est possible d'adopter une quatrième stratégie qui consiste à considérer l'ensemble des valeurs observées et à ignorer l'ensemble des manquantes. Ceci suppose donc que les valeurs manquantes ne sont pas porteuses d'information et que le mécanisme de génération des valeurs manquantes est complètement aléatoire (MCAR). L'application de cette stratégie suppose que l'algorithme d'apprentissage est capable de traiter des exemples qui ne sont pas tous décrits par les mêmes variables et qui appartiennent donc à des espaces différents, de dimensions différentes. Ragel et Crémilleux (1998) ont montré l'intérêt d'une telle stratégie pour l'apprentissage de règles d'associations. L'application de cette méthode aux arbres de décision est cependant plus délicate et reste un problème ouvert.

La stratégie 1 n'étant pas toujours applicable, parce que l'on souhaite absolument utiliser un algorithme qui s'étend difficilement au cas des valeurs manquantes, et la stratégie 2 comportant des faiblesses rédhibitoires, celle-ci est la plus utilisée. C'est celle que nous adopterons pour notre problème. Nous aurions pu modifier *Salammbô*, à l'image de ce qui est fait dans CART ou C4.5. Mais comme le révèlent les études de Ragel et Crémilleux (1999), Feelders (1999) et Batista et Monard (2003), la substitution des valeurs manquantes appliquée en amont de la construction d'un arbre de décision est souvent plus efficace que le recours au traitement interne de ces valeurs par C4.5 ou CART.

Cette première analyse conduit à la typologie de la figure 6.3. Nous allons maintenant nous focaliser sur les techniques de substitution correspondant à la stratégie 3, en essayant de dégager les caractéristiques qui permettent de les différencier. La technique étiquetée CD, pour *Case Deletion* ou suppression de cas, correspond à la stratégie 2 dans laquelle on se ramène à une base de données complète par suppression de tous les exemples contenant au moins une valeur manquante.

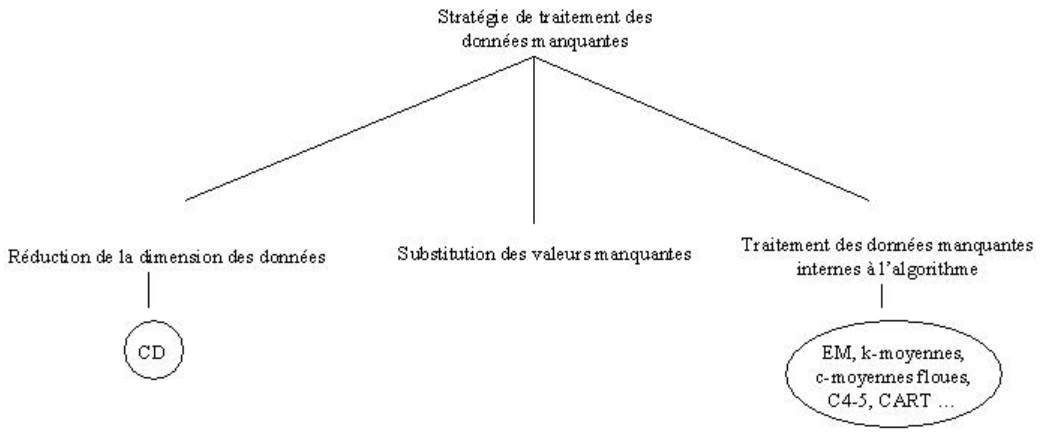

FIG. 6.3 – Les grandes catégories de méthodes pour le traitement des données manquantes

Hu *et al.* (2000) ont construit une typologie simple, reposant essentiellement sur deux alternatives découlant des questions suivantes

la méthode est-elle déterministe ou stochastique ?
Repose-t-elle sur la construction d'un modèle permettant de prédire les valeurs manquantes ou non ?
En s'inspirant de cette démarche nous proposons de rajouter de nouveaux critères afin de pouvoir catégoriser plus finement les différentes techniques. Hu *et al.* (2000) indiquent que les familles de méthodes qu'ils ont mises en évidence ne sont pas mutuellement exclusives et ne forment donc pas une partition. Cette remarque s'appliquera également à la taxinomie que nous allons introduire.

Nous notons $\mathcal{E}_i^o$ la partie observée de la variable v_i, alors que $\mathcal{E}_i^m$ en désigne la partie manquante. Il s'agit des ensembles d'exemples pour lesquels la valeur de v_i est observée ou manquante :

$$\begin{aligned} \mathcal{E}_i^o &= \{e_j \in \mathcal{E} / v_{ji} \neq ?\} \\ \mathcal{E}_i^m &= \{e_j \in \mathcal{E} / v_{ji} = ?\} \\ \mathcal{E} &= \mathcal{E}_i^o \cup \mathcal{E}_i^m \end{aligned}$$

n_i^o et n_i^m désigneront les cardinaux de ces ensembles. Dans la suite nous noterons $\hat{v}_{ij}$ l'estimation d'une valeur manquante v_{ij}. Supposons qu'une donnée v_{ij} soit manquante (valeur de v_j pour l'exemple e_i). Pour trouver une valeur de substitution diverses options s'offrent à nous :

1. *Considère-t-on le problème dans l'espace des variables ou dans celui des exemples ?*
   La substitution de v_{ij} peut se faire à partir des informations existantes à propos de l'exemple e_i contenues dans les autres variables $v_k \in \mathcal{V} \quad k \neq j$ (espace des variables). On peut préférer se focaliser sur les informations relatives à la variable v_j présentes dans les autres exemples $e_k \in \mathcal{E} \quad k \neq i$ (espace des exemples).

2. *Utilise-t-on l'information de classe y ?*
   Nous distinguerons alors les techniques de substitution supervisées qui utilisent cette information, des techniques non supervisées qui ne s'en servent pas. Cela aura son importance dans le contexte de la classification supervisée, lorsqu'il faudra spécifier le protocole expérimental. Nous y reviendrons à la section 6.6.3.

3. *A-t-on recours à un modèle de prédiction ?*
   Ceci correspond au second critère de Hu *et al.*, lorsque nous nous plaçons dans l'espace des variables. L'idée sous-jacente est d'essayer de tirer profit de la structure de corrélation qui peut exister entre les v_k ($k \neq j$) et v_j. La difficulté réside dans le choix du modèle, dans les hypothèses qui le sous-tendent et qui sont souvent invérifiables en pratique. De plus, lorsque peu de données sont disponibles, le modèle peut s'avérer statistiquement peu fiable.

4. *Si oui quel est le type de modèle utilisé ?*
   Classification, régression ou inférence bayésienne

5. *Le processus de substitution est-il déterministe ou stochastique ?*
   C'est le premier critère mentionné par Hu *et al.* Les méthodes stochastiques prennent en compte l'incertitude sous-jacente, liée au remplacement d'une valeur inconnue. Certaines approches déterministes peuvent également tenir compte de l'incertitude. Aussi aurait-il peut-être fallu dédoubler ce critère. C'est le cas des méthodes de substitution multiple déterministes (elles sont théoriquement envisageables, mais jamais utilisées) ou encore de l'approche par assignation de toutes les valeurs possibles (*APV* : *All Possible Values*) que nous décrirons plus loin.

6. *Prend-on en compte les informations au niveau local ou global ?*
   Autrement dit n'utilise-t-on que l'information de données proches de v_{ij} (au niveau des exemples ou des variables) ? Lorsqu'on s'intéresse à la proximité entre exemples, ce critère peut être regroupé avec le 2e. Il suffit pour cela de considérer que la variable de classe y permet d'identifier les exemples qui sont proches.

Ces critères en main, nous avons pu construire la taxinomie des méthodes de substitution des valeurs manquantes qui est décrite à la figure 6.4. Elle est représentée par un arbre, dans lequel chacun des nœuds correspond à un test binaire sur l'un de nos critères. Le fils de gauche rassemble les méthodes qui passent le test, alors que celles qui sont regroupées sous le fils de droite invalident ce test. Les différentes méthodes sont rangées dans les feuilles de cet arbre. Les abréviations et acronymes seront explicités, à la section suivante, dans laquelle nous détaillerons le principe et les caractéristiques des techniques correspondantes.

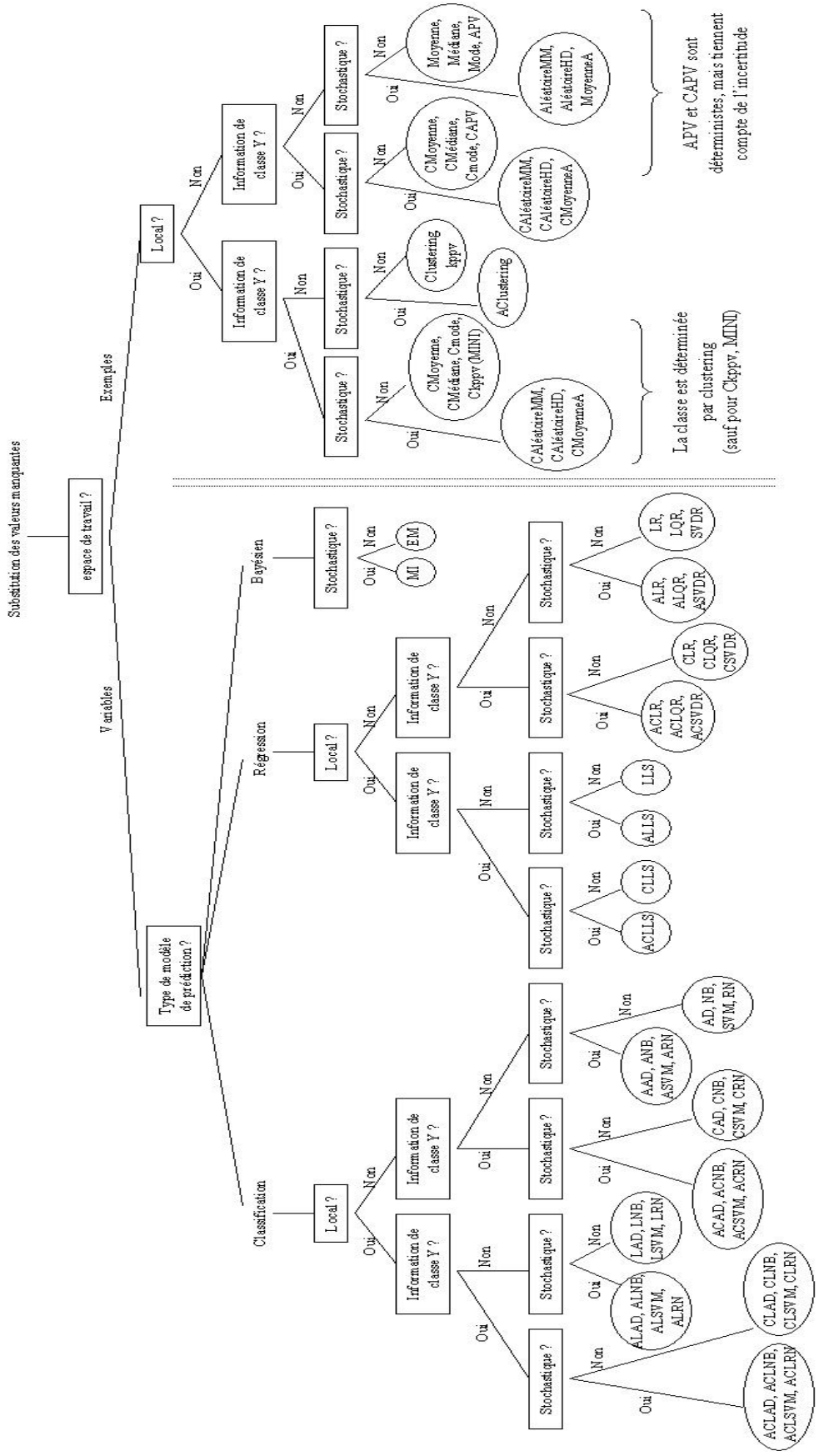

FIG. 6.4 – Taxinomie des différentes méthodes de traitement des données manquantes

6.4.2 Techniques de substitution des valeurs manquantes

Nous ne prétendons pas couvrir l'ensemble du domaine, mais nous évoquerons les méthodes les plus courantes, celles que nous avons incluses dans la taxinomie de la figure 6.4.

Toutes n'ont pas les mêmes propriétés, aussi est-il important de bien spécifier les objectifs que l'on s'assigne avant de choisir une méthode de substitution afin de pouvoir vérifier l'adéquation entre objectifs et propriétés de chaque méthode. Les principaux objectifs que l'on peut vouloir poursuivre sont les suivants.

- **Précision de la substitution** : la valeur de remplacement doit être aussi proche que possible de la *vraie* valeur[4].
- **Préservation de la distribution des données** : on s'intéresse plutôt aux paramètres de cette distribution : moyenne, variance de chaque variable, covariance entre les variables.
- **Précision de l'étape d'analyse** : dans notre contexte la phase d'analyse correspond à la construction d'un modèle de classification supervisée. Un des objectifs est alors de maximiser les performances du classifieur.
- **Complexité minimale**

Substitution dans l'espace des exemples

6.4.2.1 Substitution par la moyenne

Les valeurs manquantes de chaque variable sont remplacées par la moyenne de la variable considérée. Si cette méthode est simple et peu complexe, elle présente l'inconvénient de sous-estimer la variance et de biaiser la corrélation entre variables. La distribution des données est donc loin d'être préservée. Un autre problème provient du fait que l'estimateur de la moyenne est très sensible à la présence de valeurs aberrantes. Malgré cela, cette technique s'avère empiriquement plutôt satisfaisante (Acuna et Rodriguez, 2004). Pour pallier la sensibilité de la moyenne il peut être préférable d'utiliser la médiane qui est plus robuste. Dans le cas de données discrètes, on a recours au mode.

Dans un contexte de classification, comme celui que nous étudierons en section 6.6, il peut être intéressant d'estimer moyenne, médiane et mode relativement à chacune des classes et non pas sur la population toute entière. Les classes peuvent être connues à l'avance (supervisé) ou avoir été construites par des méthodes non supervisées (*EM*, *k-moyennes*, *nuées dynamiques*...).

> *Dans notre taxinomie, les méthodes correspondantes sont alors appelées :* CMoyenne, CMédiane *et* CMode*, le C indiquant que l'on tient compte d'une information de classe.*

Pour tenir compte de l'incertitude liée au processus de substitution on peut non pas considérer que l'on connaît avec certitude la valeur de substitution, mais tirer une valeur aléatoire centrée sur la médiane ou la moyenne. Ceci permet de rehausser la variance et donc de réduire le biais relatif à cette statistique. Généralement cette technique ne s'emploie qu'avec la moyenne, en supposant que la variable considérée suit une loi normale dont les paramètres sont estimés sur l'ensemble des données observables ou bien simplement sur les données de la même classe.

[4] Cet objectif est un peu utopique sachant que nous n'avons accès à la *vraie* valeur que sur des exemples jouets.

Les deux méthodes en question ont été nommées MoyenneA *et* CMoyenneA*, le A indiquant que l'on effectue un tirage aléatoire. Le tableau 6.5 récapitule les méthodes que nous venons de mentionner.*

TAB. 6.5 – Différentes techniques de substitution basées sur une mesure de tendance centrale

Moyenne	$\hat{v}_{ij} = \text{moyenne}\left(\{x\}\right)_{x \in \mathcal{E}_j^o}$
Médiane	$\hat{v}_{ij} = \text{médiane}\left(\{x\}\right)_{x \in \mathcal{E}_j^o}$
Mode	$\hat{v}_{ij} = \text{mode}\left(\{x\}\right)_{x \in \mathcal{E}_j^o}$
CMoyenne	$\hat{v}_{ij} = \text{moyenne}\left(\{x\}\right)_{x \in \mathcal{E}_j^o, \text{Classe}(x)=y_i}$
CMédiane	$\hat{v}_{ij} = \text{médiane}\left(\{x\}\right)_{x \in \mathcal{E}_j^o, \text{Classe}(x)=y_i}$
CMode	$\hat{v}_{ij} = \text{mode}\left(\{x\}\right)_{x \in \mathcal{E}_j^o, \text{Classe}(x)=y_i}$
MoyenneA	$\hat{v}_{ij} \sim \mathcal{N}\left(\text{moyenne}\left(\{x\}\right)_{x \in \mathcal{E}_j^o}, \text{écart-type}\left(\{x\}\right)_{x \in \mathcal{E}_j^o}\right)$
CMoyenneA	$\hat{v}_{ij} \sim \mathcal{N}\left(\text{CMoyenne}, \text{écart-type}\left(\{x\}\right)_{x \in \mathcal{E}_j^o, \text{Classe}(x)=y_i}\right)$

6.4.2.2 Substitution aléatoire

Une autre façon de traiter les valeurs manquantes d'une variable donnée v_j consiste à tirer aléatoirement une valeur dans le domaine de définition de v_j. Ceci revient à faire une hypothèse minimale sur les données, correspondant à la situation d'ignorance : toutes les valeurs sont équiprobables[5]. Ce domaine n'est pas connu *a priori*. On le détermine sur les échantillons observables $\mathcal{E}_j^o$. Nous avons noté cette méthode *AléatoireMM*, MM signifiant min-max, en référence à un domaine de définition d'une variable continue. Pour les variables discrètes, il s'agit simplement de l'ensemble des modalités prises par v_j et qui sont effectivement observées.

La méthode *AléatoireHD* fait partie de ce que l'on appelle les techniques *Hot Deck* qui visent à remplacer une valeur manquante v_{ij} sur v_j, en utilisant les valeurs prises par cette même variable sur d'autres exemples. La méthode *AléatoireHD* revient simplement à choisir au hasard (tirage uniforme), un exemple $l \neq i$ tel que v_{lj} soit observée. Cette valeur est alors utilisée pour remplacer v_{ij}.

Si nous intégrons une information de classe (mode supervisé ou suite à un clustering des différents exemples), nous pouvons raffiner l'identification du domaine de définition de v_j, classe par classe, avant de faire le tirage aléatoire. Ceci correspond à la méthode *CAléatoireMM*. *CAléatoireHD* consiste simplement à choisir aléatoirement un exemple qui donnera la valeur de substitution, non pas parmi tous les exemples, mais uniquement parmi ceux de la même classe que l'exemple à traiter. Le tableau 6.6 rappelle les définitions des méthodes aléatoires que nous venons d'introduire.

6.4.2.3 Substitution en utilisant l'ensemble des valeurs possibles

À l'instar des techniques de remplacement aléatoire, la méthode *APV* permet de tenir compte de l'incertitude. Elle s'abstient de faire la moindre hypothèse sur les données. L'idée est la suivante. Puisqu'on ne connaît pas la valeur manquante, le plus simple est encore d'envisager toutes les possibilités. Ainsi toutes les valeurs observées de v_j seront utilisées pour créer autant de nouveaux exemples, ne différant que par cette valeur. L'incertitude

[5] Le tirage est donc uniforme.

TAB. 6.6 – Techniques de substitution aléatoires

AléatoireMM	$\hat{v}_{ij} \sim \mathcal{U}\left(\min\left(\{x\}\right)_{x\in\mathcal{E}_j^o}, \max\left(\{x\}\right)_{x\in\mathcal{E}_j^o}\right)$
AléatoireHD	$\hat{v}_{ij} = v_{lj}/v_{lj} \in \mathcal{E}_j^o$
CAléatoireMM	$\hat{v}_{ij} \sim \mathcal{U}\left(\min\left(\{x\}\right)_{x\in\mathcal{E}_j^o,\text{Classe}(x)=y_i}, \max\left(\{x\}\right)_{x\in\mathcal{E}_j^o,\text{Classe}(x)=y_i}\right)$
CAléatoireHD	$\hat{v}_{ij} = v_{lj}/v_{lj} \in \mathcal{E}_j^o, y_l = y_i$

liée à la substitution est effectivement prise en compte. En revanche cela se fait de manière déterministe. Si v_{ij} est manquante on crée n_j nouveaux exemples à partir de e_i, où n_j est le nombre de valeurs distinctes de v_j. L'accroissement du nombre d'exemples dans la base est exponentiel en fonction du nombre de valeurs manquantes, ce qui peut vite devenir problématique.

Il existe également une autre version de cette technique, que nous avons appelée $CAPV$, dans laquelle une information de classe est prise en compte. La procédure de substitution est identique à celle de APV, à la différence près qu'on ne s'intéresse qu'aux valeurs de v_j pour les exemples appartenant à la classe de e_i. Grzymala-Busse et Hu (2001) notent que ces méthodes sont prometteuses, mais soulignent également les problèmes combinatoires qu'elles peuvent rencontrer.

6.4.2.4 k plus proches voisins

Pour chaque observation contenant des valeurs manquantes, on recherche ses k plus proches voisines. Dans le cas de variables continues, la valeur de remplacement correspond simplement à une moyenne pondérée des valeurs prises par ces k voisins pour la variable en question. Lorsque les variables sont discrètes, on procède à un vote majoritaire pour choisir la valeur la plus fréquente parmi les k qui ont été identifiées. Nous avons noté cette méthode, qui fait partie des techniques *Hot Deck*, *kppv*. La difficulté réside dans le choix du paramètre k et de la métrique utilisée, les distances les plus utilisées étant l'euclidienne, celle de Mahalanobis ou encore celle de Pearson. Ces distances sont également employées pour fixer les poids requis lors du calcul de la moyenne pondérée. L'avantage de cette méthode est de ne faire aucune supposition quant à la distribution des données, et de prendre en considération la corrélation entre variables. En revanche elle est assez gourmande en temps de calcul.

Lorsque de nombreuses données sont manquantes, la définition de la métrique est assez problématique. Pour calculer la distance entre deux exemples e_i et e_k, une technique simple consiste à projeter les vecteurs correspondants sur le sous-espace de dimension $q < p$ dans lequel ces deux exemples n'ont pas de valeurs manquantes. La distance entre les deux projetés est alors considérée comme la distance entre nos deux exemples. Prenons un exemple concret. Soient e_1 et e_2, 2 exemples décrits par 5 variables.

	v_1	v_2	v_3	v_4	v_5
e_1	2	4	?	3	?
e_2	?	1	3	7	?

On projette alors e_1 et e_2 sur (v_2, v_4) pour obtenir $pr\left(e_1\right)$ et $pr\left(e_2\right)$. En considérant la distance euclidienne, que l'on note d, on a :

$$d\left(e_1, e_2\right) = d\left(pr\left(e_1\right), pr\left(e_2\right)\right) = \sqrt{3^2 + 4^2} = 5$$

Ainsi la distance ne sera pas calculée sur des espaces de même dimension. Pour y parvenir, nous proposons d'utiliser une version itérative de l'algorithme *kppv*, que nous notons *kppvI*. Il s'agit de partir d'une substitution initiale (*Moyenne*, *Aléatoire* ou encore *kppv*) puis de calculer la distance entre exemples sur l'espace de départ (de dimension p), en considérant que les valeurs remplacées précédemment sont des valeurs observées. On procède alors à l'estimation de nouvelles valeurs de remplacement par l'utilisation classique de la méthode *kppv* et on recommence, jusqu'à satisfaire un certain critère d'arrêt. Ce peut être par exemple un nombre d'itérations maximal ou encore l'absence de modification des valeurs de substitution d'une itération sur l'autre.

Pour remplacer une valeur v_{ij} manquante, la technique *kppv* procède localement, en s'appuyant sur la valeur prise par v_j sur d'autres exemples, proches de e_i. On peut définir des versions *Ckppv* et *CkppvI* qui tiennent compte d'une information de classe. La méthode *Ckppv* est utilisée par Song et Shepperd (2007) sous le nom de *MINI*. Ils procèdent à une réduction amont de la dimension du problème en sélectionnant les variables clés *via* l'algorithme de sélection d'attributs d'ID3. Le calcul de leurs distances doit donc en être amélioré. Autre particularité, ils se placent dans un contexte d'apprentissage supervisé et disposent donc d'une variable *classe*. Mais ils n'utilisent qu'indirectement cette information, pour calculer la distance entre chaque exemple contenant une valeur manquante et chacune des classes. Les k plus proches voisins de l'exemple considéré, parmi ceux qui appartiennent à la classe dont il est le plus proche, sont alors utilisés pour déterminer la valeur de remplacement.

6.4.2.5 Classification non supervisée

Pour assigner une valeur de remplacement, il est également possible de procéder à un traitement amont de la base de données, de façon à regrouper les observations similaires. Pour cela on utilise une technique de classification non supervisée (ce qui correspond à la dénomination *clustering* dans la typologie de la figure 6.4). Par exemple on peut utiliser l'algorithme des *k-moyennes*, des *nuées dynamiques* ou encore les *c-moyennes floues* de façon à introduire plus de souplesse dans la classification.

Toutes ces méthodes vont regrouper les observations en k classes. Une fois ces regroupements effectués, l'assignation d'une valeur de remplacement consiste simplement à prendre la valeur correspondante du centre de gravité de la classe à laquelle appartient l'observation considérée. Dans le cas flou, on prend la moyenne des valeurs des centres de gravité, pondérée par le degré d'appartenance de l'observation à chacune des classes.

Les algorithmes de clustering procèdent de manière itérative. À chaque étape les centres de gravité sont estimés, puis chaque observation est affectée à une classe (ou des classes dans le cas flou), en fonction du centre de classe le plus proche. Ces techniques ne sont pas particulièrement rapides, comme tout algorithme itératif, mais elles permettent de réaliser une assignation itérative des valeurs manquantes. En partant d'une assignation initiale, *via* une autre technique telle que la *Moyenne*, l'algorithme standard peut être lancé. À chaque itération, l'ensemble des valeurs de chaque exemple[6] est utilisé afin de déterminer la nouvelle partition et les nouveaux centres de classe. Ensuite de nouvelles valeurs de remplacement sont estimées, en fonction de la classe à laquelle est temporairement affecté l'exemple. En fin d'algorithme, on dispose d'un regroupement de nos données, ce qui peut être utile pour l'analyse de la base, ainsi que d'une base complète.

Aucune hypothèse particulière sur les données n'est faite. Pour tenir compte de l'incertitude, on peut supprimer le déterminisme de la substitution en introduisant des pertur-

[6] Valeurs issues de la substitution à l'itération précédente et valeurs observées.

bations aléatoires autour des valeurs de remplacement. Par exemple, on peut envisager de tirer une valeur selon une loi de normale, dont la moyenne et l'écart-type sont estimés sur l'ensemble des exemples de la classe[7].

Une autre version plus rapide, consiste à ne calculer les similarités entre observations qu'à partir des valeurs observées (évaluation d'une métrique sur une dimension plus faible). Les valeurs manquantes ne sont alors remplacées qu'une fois que le regroupement final est obtenu. Cependant il s'avère empiriquement qu'il est plus judicieux, du point de vue de l'assignation des valeurs manquantes, de tenir compte de ces valeurs manquantes pour réaliser la classification (Timm *et al.*, 2003). Notons enfin que cette dernière version du clustering, si l'on omet la dernière phase de substitution, peut être utilisée uniquement pour créer une partition de l'ensemble des exemples $\mathcal{E}$, qui peut alors servir d'information de classe pour toutes les techniques de substitution à action locale telles que *CMoyenne* ou *Ckppv*.

Les techniques que nous avons vues jusqu'à présent se placent toutes dans l'espace des exemples (critère 1). Elles utilisent les valeurs prises par les autres exemples sur la variable manquante pour estimer la valeur de substitution. Parmi ces méthodes, celles qui s'appuient sur une mesure de tendance centrale, sur l'assignation de toutes les valeurs possibles ainsi que celles que nous avons nommées *Aléatoire* procèdent de manière globale. À l'inverse *kppv*, les méthodes dites de *clustering* ou toutes les méthodes utilisant l'information de classe y (celles qui sont préfixées par C) ne considèrent que les exemples les plus proches de celui pour lequel la substitution est envisagée. Elles agissent donc localement (critère 6).

Dans l'ensemble on peut dire que ces techniques cherchent à prédire toute valeur manquante d'un exemple à partir des valeurs prises par les autres exemples sur la même variable. Il est possible de renverser le point de vue (en tranposant la matrice des données V) et de s'intéresser non pas aux exemples mais aux variables. La tâche revient alors à prédire la valeur manquante d'une variable, à partir des valeurs prises par les autres variables pour cet exemple. La variable incomplète jouera alors le rôle de variable cible et sera qualifiée de variable dépendante ou de classe suivant que l'on se place dans le contexte de la régression ou celui de la classification. L'intérêt des méthodes prédictives est de pouvoir finement tirer parti des corrélations qui peuvent exister entre la variable cible et les autres. Leur faiblesse tient justement à cette caractéristique, lorsqu'il n'y pas véritablement de liens entre les variables. Autre point sensible, les hypothèses faites par le modèle prédictif sont souvent invérifiables.

Substitution dans l'espace des variables

6.4.2.6 Régression

Sous certaines hypothèses de linéarité et d'indépendance entre variables, il est possible de considérer qu'une variable, pour laquelle certaines observations manquent, peut être prédite par certaines autres, à l'aide d'une technique de régression. Les paramètres du modèle sont alors estimés de manière classique à partir des valeurs observées de la variable dépendante, par minimisation de l'erreur quadratique ou maximisation de la vraisemblance des données (Greene, 2003). Ces paramètres, que l'on nomme également coefficients de régression, sont ensuite utilisés pour prédire les valeurs manquantes en fonction des valeurs des variables explicatives.

[7] On se rapproche alors d'une version simplifiée de l'algorithme *EM* lorsque l'on suppose que les données sont générées par un mélange de gaussiennes.

Les diverses techniques s'apparentant à cette approche diffèrent en fonction du modèle utilisé et du choix des variables indépendantes. Dans le contexte des données manquantes, deux modèles sont en général employés : la régression linéaire et la régression logistique. Cette dernière est plutôt utilisée pour traiter les variables discrètes, alors que la régression linéaire est appliquée sur des variables continues. Dans ce cas-là, la valeur de substitution qui est ainsi identifiée correspond à l'espérance de la variable dépendante, conditionnellement au reste des données (Little et Rubin, 2002). Aussi trouve-t-on également dans la littérature l'appellation *moyenne conditionnelle* pour désigner cette méthode. Pour ce qui est du choix des variables explicatives, on distingue trois cas de figure.

- Toutes les variables disponibles (celles qui sont complètes) sont utilisées (approche globale). Suivant que la régression est linéaire ou logistique, nous avons nommé les deux méthodes correspondantes LR et LQR respectivement.
- Seules les k plus proches variables de celle que l'on cherche à modéliser sont retenues (régression locale). En référence à la méthode proposée par Kim *et al.* (2005) nous avons appelé cette méthode LLS, pour *Local Least Square*[8].
- Les variables de régression correspondent aux k axes principaux les plus porteurs d'information (vecteurs propres associés aux plus grandes valeurs propres). Ces axes s'obtiennent par décomposition en valeurs singulières de la matrice des données V (Oba *et al.*, 2003; Kim *et al.*, 2005). Nous appelons cette technique $SVDR$ (Régression par Décomposition en Valeurs Singulières).

Pour chacune de ces méthodes, il est d'une part possible de tenir compte de l'information de classe en n'utilisant que les exemples d'une même classe pour estimer les paramètres du modèle. D'autre part, pour refléter l'incertitude sous-jacente, on peut effectuer un tirage aléatoire, centré sur la valeur de remplacement originellement identifiée, selon une normale dont la variance correspond à celle des résidus ϵ_i (voir section 1.2.3.2). Cela revient à faire l'hypothèse d'homoscédasticité (tous les résidus ont même variance).

> *Lorsque l'information de classe est prise en compte, les méthodes sont préfixées par un C (Classe), par un A (Aléatoire) lorsque les méthodes ne sont pas déterministes.*

L'inconvénient de cette approche réside dans les hypothèses qui sont faites à propos de la distribution des données. Supposer une relation linéaire entre les différentes variables revient à faire une hypothèse qui est rarement vérifiée. L'indépendance des variables est, elle aussi, sujette à caution. En effet, lorsque l'on utilise la régression pour traiter une variable incomplète, il est supposé que les variables explicatives sont indépendantes et qu'il existe des corrélations entre celles-ci et la variable à prédire. Il n'est pas garanti que le modèle soit performant lorsque ces hypothèses ne sont pas vérifiées.

Lorsque la proportion de valeurs manquantes est importante, la régression ne peut être effectuée directement. C'est le cas lorsque toutes les variables contiennent des valeurs manquantes[9]. Dans ce cas, on a recours à un procédé itératif de remplacement des valeurs manquantes, à l'image de ce qui est fait avec les méthodes de classification non supervisée.

On procède à une substitution initiale, avec la méthode *Moyenne* par exemple, puis on applique la régression pour trouver les nouvelles valeurs de remplacement et on recommence jusqu'à satisfaire un certain critère d'arrêt. Souvent on s'arrête lorsque les valeurs de remplacement succéssives ne varient quasiment plus d'une itération sur l'autre.

[8] Comme le suggère la référence aux moindres carrés, c'est la régression linéaire classique qui est utilisée.
[9] Pour effectuer la décomposition en valeurs singulières, on est confronté au même problème.

Nous ferons référence à ces méthodes sous le vocable de régression itérée. *Toutes les techniques reposant sur ce principe seront suffixées d'un I pour symboliser le recours à un procédé itératif.*

6.4.2.7 Classification supervisée

Issue de la statistique classique, la régression correspond à une approche prédictive. Afin de généraliser cette approche il est possible de considérer que n'importe quel modèle prédictif peut être utilisé en lieu et place de la régression. Ce sont alors toutes les techniques de l'apprentissage automatique qui deviennent potentiellement applicables au traitement des données manquantes.

Pour des données catégorielles, les différents algorithmes de classification supervisée, comme par exemple les arbres de décision AD, le classifieur bayésien naïf NB, les machines à vecteurs supports SVM ou encore les réseaux de neurones RN, peuvent alors constituer des solutions efficaces et concurrentes de la régression logistique. Pour chaque variable v_j contenant une donnée manquante, un modèle de classification sera construit. La variable en question sera considérée comme la variable porteuse de l'information de classe. Les exemples pour lesquels v_j est observée, c'est-à-dire ceux qui appartiennent à $\mathcal{E}_j^o$, formeront la base d'apprentissage, tandis que ceux de $\mathcal{E}_j^m$ formeront la base de test.

Les algorithmes de classification supervisée présentent le grand avantage de ne faire que très peu d'hypothèses quant à la distribution des données. Conversano et Siciliano (2003) proposent par exemple une méthode utilisant les arbres de décision et Farhangfar *et al.* (2004) testent des approches basées sur *C4.5*, *NB* et *CLIP4* (classifieur à base de règles). Ils évoquent de façon générale le recours aux algorithmes d'apprentissage pour assigner de nouvelles valeurs aux données manquantes, mais ils ne précisent jamais comment sont traitées les variables continues. De plus, ils utilisent une terminologie qui, selon nous, est source de confusion. Ils parlent en effet d'algorithmes de substitution supervisée pour désigner ces techniques qui utilisent les valeurs observées de la variable incomplète pour construire un modèle permettant de prédire ses valeurs manquantes à partir d'autres variables. La confusion peut se produire lorsque la tâche de substitution correspond à un prétraitement d'une phase d'apprentissage supervisé, dans laquelle une des variables joue un rôle clé. Nous préférons réserver l'adjectif « supervisé » pour qualifier les techniques de substitution qui ont recours à cette variable clé pour trouver les valeurs de remplacement. La figure 6.5, schématise la façon dont les algorithmes d'apprentissage supervisé, classification et régression, sont utilisés.

Pour le cas des variables continues, ces techniques ne peuvent pas s'appliquer directement. Il faut procéder à une phase de discrétisation pour obtenir des variables symboliques, dont les modalités correspondent à des intervalles. Ces modalités peuvent alors être prédites avec les algorithmes de classification supervisée standards. Nous ne rentrerons pas ici dans les détails du processus de discrétisation, mais nous renvoyons le lecteur à la thèse de Marsala (1998) pour un aperçu des différentes méthodes de discrétisation supervisées. Pour les méthodes non supervisées on pourra se reporter à l'article de Dougherty *et al.* (1995).

Une fois qu'une des catégories de la variable symbolique a été prédite pour remplacer la valeur manquante, il faut pouvoir revenir à une valeur numérique. Pour ce faire, différentes techniques sont envisageables : prendre la moyenne ou la médiane des données qui appartiennent à l'intervalle correspondant à cette catégorie, ou encore prendre une des données de cet intervalle au hasard (tirage uniforme). On peut également considérer que les données de cette catégorie suivent une loi normale dont on peut estimer la moyenne et l'écart-type. La valeur de substitution est alors tirée aléatoirement suivant cette loi.

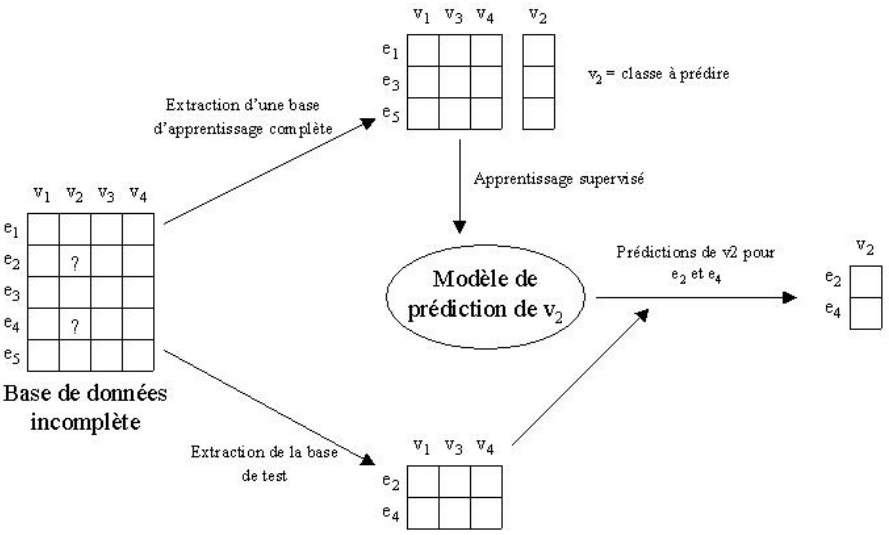

FIG. 6.5 – Substitution des valeurs manquantes à l'aide d'un algorithme d'apprentissage supervisé

Lorsque trop de données sont manquantes on retrouve le même problème qu'avec la régression : il n'est pas possible de réunir une base d'apprentissage complète. Ceci se produit lorsque tous les exemples et toutes les variables ont au moins une valeur manquante. Dans ce cas, nous sommes obligé de procéder itérativement en partant d'une substitution initiale, qui permet de construire un premier modèle. Celui-ci est alors utilisé pour raffiner l'estimation des valeurs de remplacement, ce qui permet de construire un nouveau modèle et ainsi de suite.

Ragel et Crémilleux (1999) propose cependant une méthode basée sur les règles d'association pour prédire les valeurs manquantes sans qu'il soit nécessaire de procéder de manière itérative. C'est là le grand avantage de leur méthode. Ils ont en effet proposé une méthode de construction de règles d'association qui permet d'ignorer les valeurs manquantes et dans le même temps de ne négliger aucune valeur observée (Ragel et Crémilleux, 1998).

Comme pour toutes les techniques, il est possible d'intégrer une information de classe pour construire le modèle prédictif. Enfin, à l'image de la régression sur les plus proches variables (LLS), il est possible de n'opérer la classification en utilisant que les variables les plus corrélées à celle que l'on essaie de prédire. Les méthodes reposant sur ce principe seront préfixées d'un L pour local.

6.4.2.8 EM

Une dernière approche assez fréquente consiste à utiliser l'algorithme *EM* pour estimer les valeurs manquantes (Dempster *et al.*, 1977; Ghahramani et Jordan, 1994; Little et Rubin, 2002). Il est généralement utilisé pour estimer les paramètres d'une densité de probabilité. Il peut être appliqué sur des bases de données incomplètes et présente l'avantage de procéder à l'estimation des valeurs manquantes en parallèle de l'estimation des paramètres.

On suppose l'existence d'un modèle de génération des données, par exemple un mélange de gaussiennes pour les variables continues. Les paramètres du modèle sont calculés suivant la méthode du maximum de vraisemblance, de manière itérative. Le principe est assez proche de celui des méthodes de substitution itératives par classification non supervisée comme les *c-moyennes floues*, ou encore de la régression itérative. La description de

Zou *et al.* (2005) de leur méthode de substitution basée sur une version simplifiée de *EM* correspond exactement à la méthode que nous avons appelée *régression itérée*.

À partir d'une estimation par défaut des valeurs manquantes, les paramètres du modèle sont réestimés, à chaque itération, à partir de la matrice complète, de manière à accroître la vraisemblance des données. Le modèle avec ses nouveaux paramètres est alors utilisé pour réestimer les valeurs manquantes. Puis on recommence jusqu'à ce que la convergence soit atteinte (ou considérée comme telle). À la fin de l'exécution de l'algorithme, on dispose non seulement des paramètres de notre modèle, mais également d'une matrice de données complétée.

Cette technique est très coûteuse en temps de calcul comme beaucoup d'approches itératives (Hu *et al.*, 2000; Magnani, 2003). De plus elle demande la spécification d'un modèle de génération des données. Cette tâche implique de faire un certain nombre d'hypothèses, ce qui est toujours délicat. Pour ces raisons, l'application de *EM* pour remplacer les données manquantes n'est pas toujours envisageable.

6.4.2.9 Assignation multiple

Lorsque l'on remplace une valeur manquante par l'une des techniques que nous venons de présenter, l'incertitude initiale qui caractérisait la base de données est totalement occultée. De plus un biais est introduit du fait de la déformation de la distribution initiale des données. Une méthode de traitement des données manquantes a été développée pour tenir compte de ces deux problèmes, l'assignation multiple[10]. Son principe est le suivant :

1. Assigner à chaque valeur manquante une valeur de remplacement selon un modèle de prédiction stochastique. Les modèles les plus couramment employés : Data Augmentation et Approximate Bayesian Bootstrap, sont rapidement abordés dans l'article de Grzymala-Busse et Hu (2001). Le lecteur désireux d'approfondir cette technique pourra se référer à (Little et Rubin, 2002; Schafer et Graham, 2002).
2. Recommencer M fois de façon à produire M bases complètes.
3. Effectuer l'analyse des données souhaitée sur chacune des M bases[11].
4. Agréger les résultats.

La répétition de la phase d'assignation permet de prendre en compte l'incertitude liée à la présence des données manquantes. Il faut également noter qu'intégrer dans le modèle final les erreurs faites sur chacune des M bases permet de réduire le biais du modèle global comparé à ce que l'on obtiendrait avec une assignation simple. Cependant l'assignation multiple, comme les techniques précédentes, n'a les propriétés escomptées que sous les hypothèses MCAR et MAR (Little et Rubin, 2002). Si cette méthode est théoriquement plus robuste que celles qui ont été vues jusqu'à présent, cela se paie par un accroissement évident de la complexité. Cela peut même s'avérer prohibitif lorsque la phase d'analyse qui doit être répétée M fois est elle-même coûteuse, ce qui est notre cas étant donné que l'induction d'arbres de décision est précédée d'une phase de sélection d'attributs.

On peut remarquer que nombre des techniques introduites supposent que l'on peut trouver, pour une valeur manquante d'une variable donnée, une valeur de remplacement à partir des variables observées. Elles se placent donc, pour les plus avancées, dans le cadre MAR. Aucune n'est véritablement adaptée, du moins théoriquement, aux données censurées (NMAR).

[10] Multiple Imputation : *MI*.

[11] Dans le cadre de notre travail, l'analyse qui nous intéresse correspond à une phase de classification supervisée *via Salammbô*.

6.5 Technique de substitution basée sur l'entropie

La plupart des méthodes existantes, en particulier les méthodes à base de modèles prédictifs, ont pour objectif de trouver des valeurs de substitution les plus proches possibles des valeurs originales. Elles sont également jugées sur leur capacité à préserver la distribution des données. Ce second critère est souvent invoqué par les statisticiens pour justifier leur préférence pour les méthodes stochastiques, qui ont tendance à moins sous-estimer la variance que les méthodes déterministes (Hu *et al.*, 2000; Little et Rubin, 2002).

Notre cadre méthodologique est celui de la classification supervisée. Nous cherchons surtout à obtenir un classifieur robuste et performant à partir de données incomplètes. Nos objectifs sont donc différents. Peu nous importe de compléter la matrice des données avec des valeurs aussi proches que possible de la réalité, à laquelle nous n'avons pas accès. Nous cherchons avant tout à trouver des valeurs qui nous permettront de construire un bon classifieur.

6.5.1 Description de notre méthode

Pour réaliser cet objectif nous avons choisi avec Thanh Ha Dang de compléter chaque attribut incomplet de façon à maximiser son pouvoir discriminant (Dang et Delavallade, 2006; Delavallade et Dang, 2007). L'idée sous-jacente est que l'absence de certaines valeurs détériore la capacité de discrimination d'un attribut. Nous essayons donc de la restaurer. Nous avons recours au gain d'information pour mesurer cette capacité de discrimination. Le gain d'information est en particulier utilisé dans l'algorithme ID3 lors du processus de sélection des attributs (Quinlan, 1986). Ce gain mesure l'écart entre l'entropie de la base d'exemples et l'entropie de cette même base, prise conditionnellement à l'attribut considéré.

Soit $v_i \in \mathcal{V}$. Notons I une mesure d'entropie. Le gain d'information apporté par v_i sur une base d'exemples $\mathcal{E}$, se note :

$$G(\mathcal{E}, v_i) = I(\mathcal{E}) - I(\mathcal{E}|v_i)$$

où I désigne l'entropie.

Notons $\mathcal{M}_i = \{m_{ij}\}_{j=1..n_i}$ l'ensemble fini des n_i modalités de v_i. Une substitution s_i relativement à la variable v_i est une application de $\mathcal{E}^n$ dans $(\mathcal{M}_i)^n$ qui associe à un vecteur de n exemples un vecteur contenant n valeurs prises parmi les modalités admissibles de v_i. Plus formellement on a :

$$\begin{array}{rcl} s_i : \mathcal{E}^n & \rightarrow & (\mathcal{M}_i)^n \\ (e_1, e_2, ..., e_n) & \mapsto & \left(s_i^{elt}(e_1), s_i^{elt}(e_2), ..., s_i^{elt}(e_n)\right) \end{array}$$

où s_i^{elt} correspond à une substitution élémentaire. C'est une application qui associe à un exemple une valeur de substitution de la manière suivante :

$$\begin{array}{rcl} s_i^{elt} : \mathcal{E} & \rightarrow & \mathcal{M}_i \\ e_j & \mapsto & \begin{cases} v_{ji} & \text{si } v_{ji} \neq ? \\ \hat{v}_{ji} & \text{sinon} \end{cases} \end{array}$$

La méthode que nous proposons consiste à identifier la fonction s_i qui permet de maximiser le gain d'information. Notons que le nombre de modalités n_i de v_i étant fini, l'ensemble S_i des substitutions possibles pour l'attribut v_i l'est également : $|S_i| = n_i^{|\mathcal{E}_i^m|}$. Le maximum que nous cherchons est donc atteint pour une substitution $s \in S_i$, ce que nous pouvons écrire :

$$s = \arg\max_{s_i \in S_i} \left(I(\mathcal{E}) - I(\mathcal{E}|s_i(e_1, ..., e_n))\right)$$

Le terme $I(\mathcal{E})$ ne dépendant pas de s_i, il est équivalent de minimiser l'entropie conditionnelle :

$$s = \arg\min_{s_i \in S_i} \left(I\left(\mathcal{E} | s_i\left(e_1, ..., e_n\right)\right)\right)$$

Cette quantité ne dépend que des valeurs prises par l'attribut pour chacune des observations et de leur classe. Cette méthode pourra donc être qualifiée de supervisée. Elle n'a pas recours aux variables autres que celle pour laquelle la substitution doit être effectuée et elle est déterministe. Ceci explique son positionnement dans la taxinomie, sous l'appellation *Entropie*.

Ne disposant que de l'information de classe pour trouver la valeur manquante correspondant à un élément, toutes les observations de même classe se verront attribuer la même valeur de remplacement pour l'attribut v_i, ce que Dang (2007) a montré de manière plus formelle. C'est là une des faiblesses potentielles de cette méthode, mais également sa force, car cela permet d'induire des modèles de classification qui auront tendance à mieux généraliser. Nous verrons mieux ce qu'il en est lors de l'étude expérimentale de la section 6.6. Une autre limitation vient du fait qu'elle soit spécifique aux données discrètes. Nous écartons cette remarque, comme nous l'avons fait précédemment à propos de la substitution basée sur des techniques de classification supervisée. Pour pouvoir traiter les données numériques il suffit en effet de les discrétiser.

6.5.2 Complexité et mise en œuvre algorithmique

Pour réaliser l'ensemble des substitutions de l'attribut v_i, la complexité de notre méthode est de l'ordre de $\mathcal{O}\left(n_i^{|\mathcal{E}_i^m|}\right) = \mathcal{O}\left(n_i^{n_i^m}\right)$, ce qui peut paraître rédhibitoire dès que le nombre de modalités n_i devient grand ou dès que v_i contient beaucoup de données manquantes. En tenant compte de la propriété relative à la substitution par une même valeur, des valeurs manquantes d'exemples de même classe, nous pouvons ramener la complexité à $\mathcal{O}\left(n_i^{\min\left(n_i^m, K\right)}\right)$, où K désigne le nombre de classes. Il s'agit cependant toujours d'une complexité exponentielle. Pour remédier à ce problème, nous nous sommes inspirés des techniques itératives telles que EM ou la régression itérée, pour construire une version itérative de notre méthode.

L'idée est de substituer, pour chaque variable, les valeurs manquantes une à une. À chaque itération, pour une valeur manquante v_{ji}, on calcule l'entropie conditionnnelle de $\mathcal{E}$ sachant v_i, en ne prenant en compte que les valeurs observées de v_i et en affectant temporairement à v_{ji} l'une des modalités m_{ik} de v_i. Ce calcul est effectué pour les n_i modalités de v_i. On affecte alors à v_{ji} la modalité associée à la plus petite valeur de l'entropie.

À la première itération, ne sont utilisés que les exemples dont les valeurs de v_i sont observées ($\mathcal{E}_i^o$). Pour toutes les itérations suivantes, on considère que les valeurs de substitution estimées à l'itération précédente sont des valeurs observées. Seront alors utilisés pour estimer v_{ji} tous les exemples à l'exception évidemment de e_j : $\mathcal{E} - \{e_j\}$.

Voyons plus formellement comment une valeur $v_{ji} \in \mathcal{E}_i^m$ est traitée. Nous supposons que $y_j = Classe\left(e_j\right) = c_k$. À la première itération, on considère l'attribut i de dimension $|\mathcal{E}_i^o| + 1 = n_i^o + 1$, c'est-à-dire v_i restreint aux éléments de $\mathcal{E}_i^o$ auxquels on ajoute e_j. La valeur de substitution $\hat{v}_{ji}$ correspond alors à la modalité permettant de minimiser l'entropie de $\mathcal{E}_i^o \cup \{e_j\}$ conditionnellement à cet attribut. Nous considérons par la suite uniquement l'entropie de Shannon (1948). Pour plus de détails sur les différentes entropies existantes, le lecteur pourra se reporter à la thèse de Dang (2007).

$$
\begin{aligned}
\hat{v}_{ji} &= m_{iz} \\
z &= \arg\min_{q=1..n_i} \left(I\left(\mathcal{E}_i^o \cup \{e_j\} \,|v_i, v_{ji} = m_{iq}\right)\right) \\
&= \arg\min_{q=1..n_i} -\frac{n_{iq}+1}{n_i^o+1}\left(\sum_{r=1,r\neq k}^{K} \frac{n_{iq}^{c_r}}{n_{iq}+1}\log_2\left(\frac{n_{iq}^{c_r}}{n_{iq}+1}\right) - \frac{n_{iq}^{c_k}+1}{n_{iq}+1}\log_2\left(\frac{n_{iq}^{c_k}+1}{n_{iq}+1}\right)\right) \\
&\quad - \sum_{l=1,l\neq q}^{n_i} \frac{n_{il}}{n_i^o+1}\sum_{r=1}^{K}\frac{n_{il}^{c_r}}{n_{il}}\log_2\left(\frac{n_{il}^{c_r}}{n_{il}}\right)
\end{aligned}
$$

Dans cette dernière équation n_{il} est le nombre d'exemples possédant la modalité m_{il} pour v_i et $n_{il}^{c_r}$ correspond au nombre d'exemples de la classe c_r qui prennent la modalité m_{il} de v_i.

Pour les itérations suivantes, le principe est le même, sauf qu'on ne considère pas uniquement les éléments de $\mathcal{E}_i^o$, mais également ceux de $\mathcal{E}_i^m$, en utilisant les valeurs de substitution trouvées à l'itération précédente. Nous mettons un terme à l'algorithme au bout d'un nombre prédéfini d'itérations ou lorsque l'entropie ne décroît plus. La complexité de l'algorithme est cette fois linéaire, en $\mathcal{O}\left(n_i \times n_i^m \times N\right)$ où N est le nombre d'itérations.

6.5.3 Exemple d'application

Le tableau 6.7 montre le comportement de notre méthode lors de l'initialisation, et le tableau 6.8 illustre sur le même exemple ce qui est fait lors de la première itération. Dans cet exemple nous considérons un problème avec une seule variable v_1, que nous noterons v, possédant 3 modalités : x, y et z. La matrice des données V est alors un vecteur de 10 éléments. Elle contient deux valeurs manquantes : v_6 et v_9. Ainsi nous avons

$$
\begin{aligned}
\mathcal{E} &= \{e_1, e_2, ..., e_{10}\} \\
\mathcal{E}^o &= \{e_1, e_2, e_3, e_4, e_5, e_7, e_8, e_{10}\} \\
\mathcal{E}^m &= \{e_6, e_9\}
\end{aligned}
$$

Lors de l'intialisation, c'est la modalité x qui minimise l'entropie conditionnelle, calculée sans prendre en compte v_9 et c'est z qui permet de minimiser cette entropie, en intégrant v_9, mais en excluant v_6. Lors de la seconde phase (première itération), aucune des deux valeurs n'a changé. Le processus est donc terminé. Notons au passage que les entropies ont cette fois été calculées en prenant en compte toutes les données. Pour la substitution de v_6, on a considéré que v_9 valait z, et pour la substitution de v_9, nous avons pris x comme valeur de v_6. Ce sont les deux valeurs que l'on avait trouvées à l'étape précédente.

TAB. 6.7 – Substitution des valeurs manquantes par minimisation de l'entropie : initialisation

$\mathcal{E}$	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}
v	x	y	x	z	x	?	y	y	?	x
y	1	1	2	2	1	1	2	2	2	1

$$
\begin{array}{lcl}
v_6\text{=x} \Rightarrow I\left(\mathcal{E}^o \cup e_6 | v \& v_6 = x\right) & = & -\frac{5}{9}\left(\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) \\
 & & -\frac{3}{9}\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) \\
 & & -\frac{1}{9}\left(\frac{1}{1}\log_2\frac{1}{1}\right) \\
 & = & \underline{0.707} \\
v_6\text{=y} \Rightarrow I\left(\mathcal{E}^o \cup e_6 | v \& v_6 = y\right) & = & 0.805 \\
v_6\text{=z} \Rightarrow I\left(\mathcal{E}^o \cup e_6 | v \& v_6 = z\right) & = & 0.888 \\
\hline
\Rightarrow \hat{x}_6 = x & & \\
 & & \\
v_9\text{=x} \Rightarrow I\left(\mathcal{E}^o \cup e_9 | v \& v_9 = x\right) & = & 0.846 \\
v_9\text{=y} \Rightarrow I\left(\mathcal{E}^o \cup e_9 | v \& v_9 = y\right) & = & 0.721 \\
v_9\text{=z} \Rightarrow I\left(\mathcal{E}^o \cup e_9 | v \& v_9 = z\right) & = & \underline{0.666} \\
\hline
\Rightarrow \hat{x}_9 = z & &
\end{array}
$$

TAB. 6.8 – Substitution des valeurs manquantes par minimisation de l'entropie : première et dernière itération. Les valeurs de $\hat{x}_6$ et $\hat{x}_9$ ne changent pas.

$\mathcal{E}$	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}
v	x	y	x	z	x	**x**	y	y	**z**	x
y	1	1	2	2	1	1	2	2	2	1

$$
\begin{array}{lcl}
v_6\text{=x} \Rightarrow I\left(\mathcal{E} | v \& v_6 = x\right) & = & -\frac{5}{10}\left(\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) \\
 & & -\frac{3}{10}\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) \\
 & & -\frac{2}{10}\left(\frac{2}{2}\log_2\frac{2}{2}\right) \\
 & = & \underline{0.636} \\
v_6\text{=y} \Rightarrow I\left(\mathcal{E} | v \& v_6 = y\right) & = & 0.724 \\
v_6\text{=z} \Rightarrow I\left(\mathcal{E} | v \& v_6 = z\right) & = & 0.875 \\
\hline
\Rightarrow \hat{x}_6 = x & & \\
 & & \\
v_9\text{=x} \Rightarrow I\left(\mathcal{E} | v \& v_9 = x\right) & = & 0.826 \\
v_9\text{=y} \Rightarrow I\left(\mathcal{E} | v \& v_9 = y\right) & = & 0.685 \\
v_9\text{=z} \Rightarrow I\left(\mathcal{E} | v \& v_9 = z\right) & = & \underline{0.636} \\
\hline
\Rightarrow \hat{x}_9 = z & &
\end{array}
$$

Il est possible de montrer que notre technique tend à privilégier la modalité m_j qui maximise la probabilité conditionnelle $P\left(c_i | m_j\right)$, où c_i désigne la classe de l'exemple auquel appartient la valeur manquante considérée (Dang, 2007). Il s'agit là d'une propriété asymptotique, vérifiée lorsque le nombre d'exemples de modalité m_j et de classe c_i tend vers l'infini. Cette propriété nous offre une autre solution pour la substitution initiale des valeurs manquantes qui est requise par notre technique. De plus elle permet de saisir les différences entre notre méthode et celle que nous avons nommée *CMode*. Celle-ci substitue en effet à une valeur manquante d'un exemple de classe c_i, la modalité m_j qui vérifie $P\left(m_j | c_i\right) \geq P\left(m_k | c_i\right), \forall k$.

6.6 Analyse comparative empirique

La caractérisation des différentes techniques de traitement des données manquantes que nous nous sommes efforcé de réaliser est un point important pour prendre la mesure de leurs atouts et faiblesses. Cependant cela n'est pas suffisant pour nous permettre de choisir l'une d'elles lorsque nous sommes confronté à un cas concret. Ce point est essentiel. En fouille de données, de nombreux algorithmes sont disponibles pour traiter tel ou tel problème. Nous savons de plus qu'aucune méthode ne sera jamais capable de surpasser toutes les autres sur l'ensemble des problèmes envisageables. Aussi est-il crucial de cerner les domaines spécifiques, les types de problème particuliers, pour lesquels il est possible d'identifier une, voire un petit groupe de méthodes dont les performances sont remarquables. L'objectif, *in fine*, est de pouvoir apporter un certain nombre de recommandations à un utilisateur confronté à un problème concret et qui souhaiterait savoir quelle méthode il doit mettre en œuvre.

6.6.1 Taxinomie du point de vue de l'utilisateur

En nous inspirant des travaux de Liu et Yu (2005) en sélection d'attributs, nous avons construit une taxinomie des méthodes de substitution des valeurs manquantes, en nous plaçant du point de vue de l'utilisateur cette fois. Elle est représentée par l'arbre de la figure 6.6.

L'intérêt de cette nouvelle taxinomie est de mettre en avant l'ensemble des critères dont un utilisateur dispose pour caractériser son problème et pour lequel il serait bon de pouvoir identifier la ou les méthodes les plus adaptées. Ces critères se rangent dans deux grandes catégories selon qu'ils caractérisent la base de données à disposition de l'utilisateur, ou les connaissances qu'il a sur le problème. Chacun des deux nœuds principaux se ramifie ensuite afin d'affiner la définition d'un critère. Les différents critères qu'un utilisateur se doit de considérer correspondent aux feuilles de cet arbre.

Pour un problème donné, l'utilisateur doit traiter un type de données particulier, en fonction de ses objectifs et contraintes. Nous n'avons mentionné que les contraintes de temps, car ce sont surtout elles qui peuvent influer sur le choix de la technique de substitution. Quant aux objectifs que se fixe l'utilisateur, leur impact sur la sélection de la bonne méthode est évidemment notable. D'un point de vue théorique, ce sont toujours les méthodes stochastiques qui respectent le mieux la distribution des données et parmi elles l'*assignation multiple* est sûrement la plus efficace. Notons que c'est en considérant l'objectif *optimisation des performances de la tâche finale* que nous avons développé la technique *Entropie*.

Les caractéristiques de la base de données que l'utilisateur doit traiter s'imposent à lui et influent également grandement sur la qualité des différentes techniques. Au regard de la tâche de substitution à effectuer, nous avons relevé cinq critères permettant de caractériser la nature des données.

- **Information de classe** : si elle n'est pas donnée, il faudra recourir à une phase de classification non supervisée afin de créer des classes. Lorsqu'elle est donnée il faut considérer la répartition des exemples dans les différentes classes. Il est en effet probable que les techniques soient plus ou moins adaptées en fonction du caractère plus ou moins équilibré de cette répartition. Les coûts associés aux erreurs sur les différentes classes dans un problème de classification supervisée peuvent également jouer un rôle. On peut cependant inclure ce critère dans celui qui est relatif à la mesure de performance lorsque l'objectif est d'optimiser les performances d'un classifieur.

- **Type d'attributs** : la méthode entropique que nous avons développée ne s'applique pas directement aux données continues. La régression linéaire ne traite quant à elles que les données continues.
- **La taille de la base de données** : le nombre d'exemples n et le nombre de variables p peuvent également jouer un rôle non négligeable, ainsi que peut-être le rapport p/n. Les bases de données génomiques ou textuelles contiennent fréquemment beaucoup plus de variables que d'exemples alors que la théorie statistique d'analyse des données a jusqu'ici plutôt envisagé les cas inverses.
- **La complexité de la base de données** : il existe différents travaux visant à caractériser la difficulté d'une base de données au regard d'une tâche particulière. Pour la classification supervisée on pourra par exemple se reporter à l'article de Dang *et al.* (2006).
- **Distribution des données manquantes** : selon le mécanisme de génération des données manquantes, le taux de valeurs manquantes et les motifs de distribution de ces données, les techniques à utiliser diffèrent. D'un point de vue théorique nous savons par exemple que la suppression de cas n'est envisageable que sous l'hypothèse MCAR. En pratique cette méthode ne sera jamais utilisable lorsque le motif est quelconque et que toutes les variables et tous les exemples sont affectés par l'absence de données.

Les expérimentations que nous avons menées ne sont évidemment, à elles seules, pas suffisantes pour qu'un ensemble complet de recommandations claires puisse être apporté. Nous espérons cependant contribuer à élaborer des bribes de recommandation en affinant notre compréhension des différentes méthodes. Nous avons ainsi procédé à une série d'expériences visant à analyser le comportement de ces méthodes sur des cas concrets couvrant une partie de l'ensemble des combinaisons de critères envisageables d'après notre taxinomie.

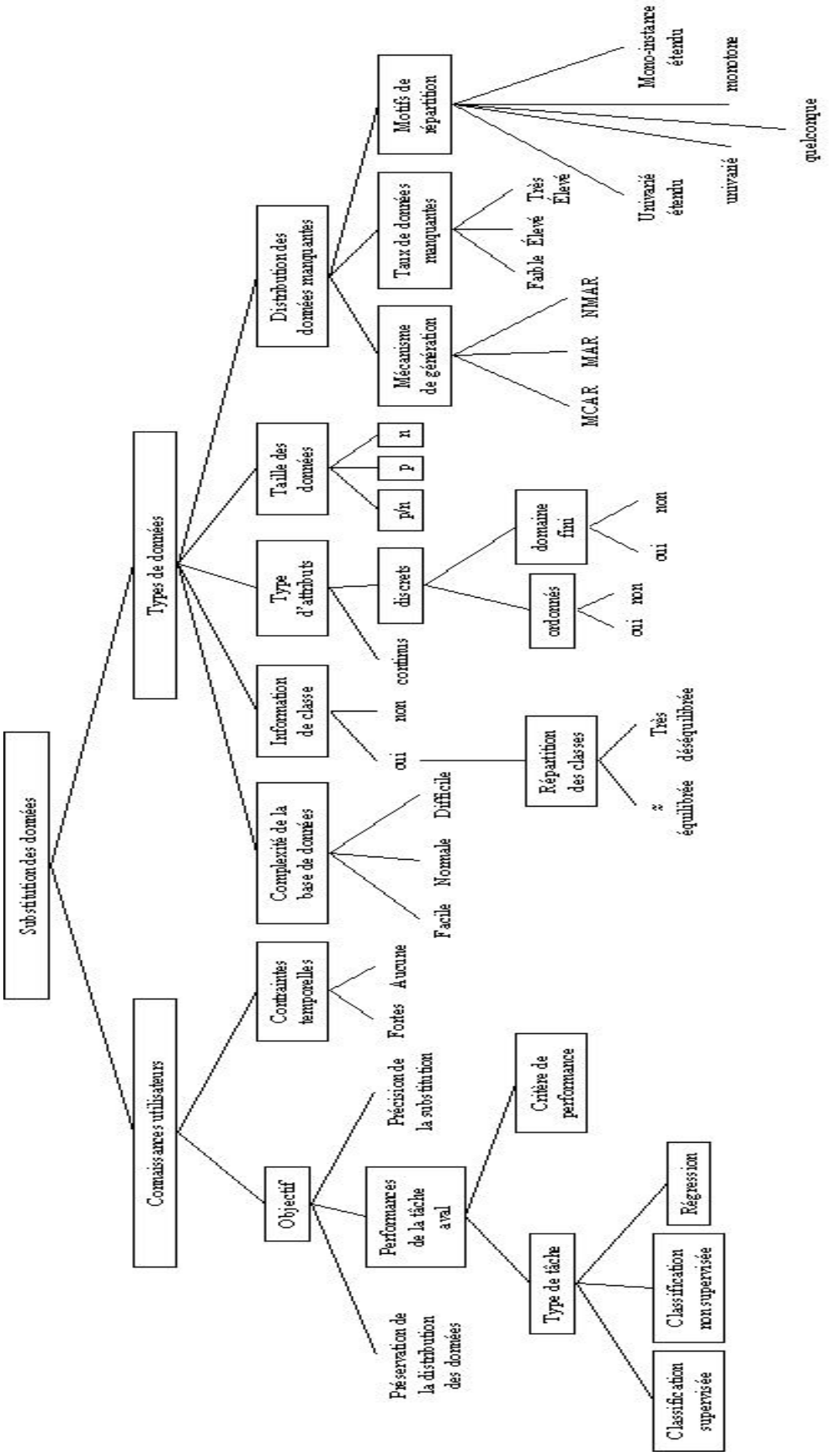

FIG. 6.6 – Taxinomie des techniques de substitution des valeurs manquantes, du point de vue de l'utilisateur

6.6.2 Objectifs

Nous souhaitons pouvoir comparer différentes techniques de substitution des valeurs manquantes dans un contexte de classification supervisée. Les performances de ces différentes techniques seront donc estimées à partir de celles d'un classifieur appris sur une base complétée. Notre hypothèse est la suivante : il n'existe pas de meilleure technique de remplacement des valeurs manquantes dans l'absolu, chacune est plus ou moins adaptée en fonction des caractéristiques de la tâche à réaliser. Nous souhaitons tester cette hypothèse en regardant l'effet des paramètres suivants, sur les performances de diverses techniques :

- taux de valeurs manquantes
- classifieur utilisé
- critère de performance

Avant de se lancer dans une série de tests comparatifs, il est important de fixer clairement le protocole expérimental qui sera utilisé, afin d'éviter toute erreur méthodologique qui pourrait biaiser l'interprétation des résultats.

6.6.3 Protocole expérimental

De nombreux travaux relatifs aux données manquantes consistent en l'évaluation empirique des performances de différentes techniques, à des fins de comparaison (Hu *et al.*, 2000; Grzymala-Busse et Hu, 2001; Farhangfar *et al.*, 2004; Acuna et Rodriguez, 2004; Batista et Monard, 2003). Les protocoles expérimentaux utilisés sont loin d'être identiques, même s'ils partagent certaines caractéristiques. Avant de présenter celui que nous adopterons, nous allons examiner les protocoles existants et essayer d'en dégager une typologie.

6.6.3.1 Objectif de la substitution

Le premier critère qui permet de distinguer les différents protocoles concerne l'objectif de la substitution. Rappelons trois des principaux objectifs que nous avons mentionnés section 6.4.2 :

1. proximité entre valeurs de remplacement et valeurs réelles
2. respect de la distribution de certaines statistiques
3. optimisation des performances de l'analyse de données subséquente

À chacun de ces points correspond un protocole expérimental. Pour se conformer aux objectifs 1 et 2 il faut disposer d'une base initialement complète, de laquelle des données seront enlevées. Une base complète sera générée par chacune des méthodes testées. Les performances seront ensuite évaluées en comparant la nouvelle base après substitution à la base d'origine. Ce sera alors la mesure de performance utilisée qui différenciera les différents protocoles.

L'étude menée par Hu *et al.* (2000) consiste à comparer des statistiques de base (moyenne, médiane, variance etc.) sur les données d'origine avec celles qui sont estimées sur les données après substitution, pour évaluer le biais introduit par chaque méthode, la meilleure étant évidemment celle qui est associée au biais le plus faible. L'objectif poursuivi est donc le 2e.

En bioinformatique, c'est l'écart entre les valeurs réelles et estimées qui permet de juger de la qualité d'une technique (objectif 1). Plus précisement, il s'agit de minimiser l'erreur

moyenne normalisée, au sens des moindres carrés, que l'on nomme *NRMSE* (Normalized Mean Root Square Error) (Kim *et al.*, 2005; Oba *et al.*, 2003) et que l'on calcule ainsi :

$$NRMSE = \sqrt{\frac{moyenne\left[(y_s - y_r)^2\right]}{variance\,(y_r)}}$$

où y_s est le vecteur contenant l'ensemble des valeurs substituées (sa dimension n^m est égal au nombre de valeurs manquantes) et y_r est le vecteur de même dimension, contenant l'ensemble des valeurs réelles qui ont été supprimées de la base d'origine. D'autres métriques sont envisageables pour mesurer la proximité entre les valeurs de substitution et les valeurs réelles. Ainsi Song et Shepperd (2007) évaluent les performances d'une méthode de substitution selon le critère *MMRE* (Mean Magnitude of Relative Error), qui est fréquemment employé dans le domaine du génie logiciel. Avec les mêmes notations que pour *NRMSE* on a

$$MMRE = \frac{100}{n^m}\sum_{i=1}^{n^m}|y_s\,(i) - y_r\,(i)|$$

Les protocoles correspondant aux objectifs 1 et 2 sont donc sans ambiguïté et plutôt simples à mettre en œuvre. Il en va autrement avec le 3e. Il s'agit d'opérer la substitution de valeurs manquantes sur une base de données et d'observer les performances d'un classifieur sur la base complète.

6.6.3.2 Indépendance des bases d'apprentissage et de test

Outre la question, en elle-même délicate, de la mesure des performances d'un classifieur, il faut s'attaquer au problème de la mise en place d'un protocole sans biais d'évaluation dudit classifieur. Nous avons vu l'importance qu'il y avait à distinguer la base d'apprentissage de la base de test, sur laquelle les performances du classifieur sont mesurées.

6.6.3.3 Rééchantillonnage

De plus, du fait du nombre limité d'échantillons, il est bon de multiplier les tests et de considérer la moyenne et la variance des performances sur un ensemble de couples (base d'apprentissage, base de test) pour ne pas sous-estimer la variance de ces performances. Ceci permet de rendre compte de la robustesse de l'algorithme d'apprentissage. Si l'on ne teste que sur un seul couple (apprentissage, test) il est possible que de bonnes ou mauvaises performances ne soient que le fruit du hasard, introduit par l'échantillonnage. La validation croisée permet de respecter ces deux contraintes : indépendance entre apprentissage et test et mesure des performances sur m corpus de données distincts. D'autres procédures de rééchantillonnage peuvent être employées, comme par exemple la méthode du *bootstrap*.

6.6.3.4 Biais méthodologique

Une fois cet élément du protocole fixé, il faut choisir le type de base que l'on va utiliser. Contrairement à ce qui se fait lorsque l'on veut mesurer un écart entre valeurs réelles et valeurs prédites, il n'est pas obligatoire de partir d'une base complète que l'on va ensuite « trouer ». En effet ce qui nous intéresse est de pouvoir comparer différentes méthodes de substitution. Pour cela la comparaison des performances d'un même classifieur induit sur une base complétée de différentes façons suffit. Par exemple Grzymala-Busse et Hu (2001) utilisent un ensemble de base de données UCI incomplètes[12].

[12] University of California Irvine http://www.ics.uci.edu/~mlearn/MLRepository.html

L'inconvénient de cette approche est que l'on ne maîtrise pas tous les paramètres relatifs aux données manquantes : leur proportion, le mécanisme les ayant générées (MCAR, MAR ou NMAR). Lorsque l'on souhaite pouvoir jouer sur ces paramètres, il faut partir d'une base complète sur laquelle nous allons agir pour supprimer certaines données selon l'effet que l'on souhaite mesurer. Cette approche est plus fréquemment retenue dans les analyses comparatives (Acuna et Rodriguez, 2004; Batista et Monard, 2003; Zou *et al.*, 2005).

Dans ce cas, un dernier choix permet de distinguer les différents protocoles : quelle(s) base(s) trouer ? En effet on peut imaginer trois options : seule la base de test est trouée, seule la base d'apprentissage est trouée, ou les deux le sont. Batista et Monard (2003), ainsi que Zou *et al.* (2005) n'enlèvent des données que sur la base d'apprentissage. Cette approche présente l'inconvénient de ne pas correspondre à un scénario réaliste. En pratique, les bases de données sur lesquelles un classifieur peut être appris contiennent des données manquantes, mais les futurs exemples qu'il faudra classer aussi. Mais ce protocole est tout de même séduisant, car il permet de comparer les algorithmes sur une base de test commune.

Acuña et Rodriguez procèdent différemment. Ils génèrent des valeurs manquantes dans une base de données qu'ils complètent *via* diverses méthodes. Pour chacune des bases complétées les performances d'un classifieur sont estimées par validation croisée sur cette base. Cela revient à trouer, et la base d'apprentissage, et la base de test. Cependant leurs résultats sont biaisés car ils estiment les valeurs de remplacement sur la base de test de la même façon que sur la base d'apprentissage, ce qui n'est absolument pas réaliste.

Considérons la méthode *CMoyenne* qui remplace une valeur manquante d'une variable et d'un exemple donné, par la moyenne de la variable considérée prise sur les exemples de la classe à laquelle appartient l'exemple incomplet. Utiliser cette technique sur la base globale revient à considérer connue la classe de tous les exemples. Ce problème se retrouve pour l'ensemble des techniques de substitution que nous avons qualifiées de supervisées. Même pour les techniques non supervisées, la méthode de remplacement n'est pas rigoureusement identique, selon que les exemples font partie de la base de test ou non.

Pour que le protocole soit réaliste, ces exemples doivent être pris un à un. Ainsi lorsque l'on utilise la méthode *Moyenne*, la valeur de substitution d'une valeur est la moyenne de la variable correspondante. Mais cette moyenne doit être estimée avec les exemples d'apprentissage uniquement, sans quoi on introduit un biais. Le protocole de Grzymala-Busse et Hu (2001) est biaisé pour les mêmes raisons. Lorsque test et apprentissage sont troués, le modèle permettant d'estimer des valeurs manquantes ne doit être construit qu'à partir des exemples de la base d'apprentissage.

6.6.3.5 Génération des valeurs manquantes

Rentrent également dans la spécification du protocole la façon dont sont supprimées certaines valeurs, ainsi que la proportion des valeurs qui doivent l'être lorsque l'on part d'une base initiale complète. Lorsque l'on multiplie les tests en utilisant plusieurs couples (apprentissage, test), il est possible de trouer la base de départ, avant d'en extraire ces couples. Mais il est préférable de commencer par extraire les couples avant de trouer chacune des bases, ce que font Batista et Monard (2003). Cela est indispensable si l'on veut disposer de bases de test complètes. Cela permet également de s'assurer que chacune des bases contiendra bien le taux de valeurs manquantes spécifié.

La plupart des études comparatives mentionnées jusqu'à présent ont toutes recours au mécanisme de génération MCAR, c'est-à-dire que chaque élément de la base de données a la même probabilité d'être manquant. Seuls les travaux de Hu *et al.* (2000) et Song et

Shepperd (2007) ont recours à des mécanismes autres que MCAR. Pour MAR par exemple, Song *et al.* choisissent une des variables de leur base, à partir de laquelle ils construisent une partition des exemples en quatre classes par discrétisation de la variable en quatre intervalles de même longueur. Un exemple a ensuite une certaine probabilité qu'une de ses valeurs soit manquante et cette probabilité dépend de la classe dans laquelle il se trouve.

Pour décrire complètement un protocole de comparaison des techniques de substitution des valeurs manquantes, il faut enfin préciser quelles variables seront amputées. Il est tout à fait envisageable de les considérer toutes (Farhangfar *et al.*, 2004; Hu *et al.*, 2000; Zou *et al.*, 2005). Cela permet de se rapprocher d'un contexte réel. Mais on peut également vouloir ne trouer que certaines variables, pour limiter la complexité de l'expérimentation. Ainsi Acuña et Rodriguez, Batista et Monard procèdent en premier lieu à l'identification des variables les plus pertinentes, grâce à un filtrage d'attributs (voir la section 7 relative à la sélection d'attributs). Des données seront alors supprimées uniquement sur ces variables.

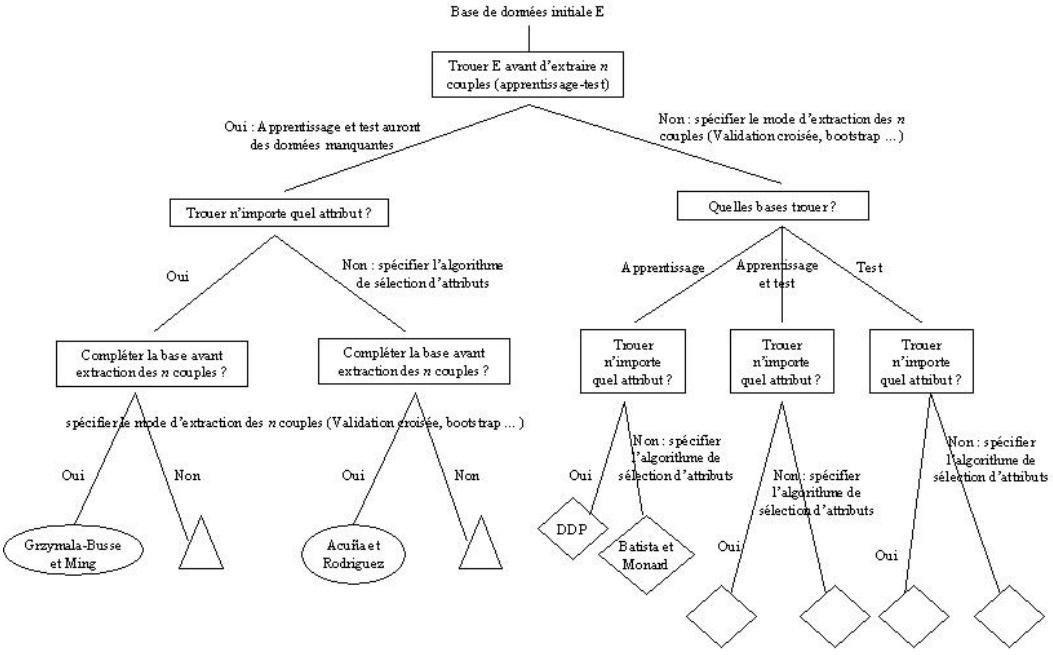

FIG. 6.7 – Taxinomie des protocoles d'évaluation des techniques de substitution des valeurs manquantes dans un contexte de classification supervisée

Pour synthétiser l'ensemble de ces remarques, nous avons dressé une taxinomie des différents protocoles. Seuls ceux qui mettent en œuvre un classifieur (objectif 3) et la génération de données manquantes artificielles présentent quelques difficultés. Aussi nous sommes-nous contenté d'illustrer sur la figure 6.7 la partie de la taxinomie qui leur est associée.

> *Les feuilles symbolisées par des ellipses correspondent à des protocoles incluant un biais, et que nous déconseillons vivement. Ceux qui appartiennent à des feuilles symbolisées par des triangles ne sont pas totalement satisfaisants. Seules les feuilles représentées par des losanges regroupent les protocoles que nous estimons satisfaisants.*

Notons que le protocole de Zou *et al.* (2005) n'y figure pas, car il ne considère que des tests simples sur un seul couple (apprentissage, test), et occulte ainsi les questions de robustesse. Outre les protocoles étudiés dans cette section, nous avons également placé dans cette taxinomie celui que nous avons adopté avec Thanh Ha Dang pour l'ensemble de nos expérimentations, noté *DDP*.

Contrairement à Batista et Monard (2003), nous préférons supprimer des valeurs de tous les attributs, car ce n'est que dans ces conditions que nous pouvons être sûr que toutes les variables importantes pour l'induction du modèle de classification seront affectées. Il n'est en effet aucunement garanti que celles qui sont jugées importantes par un filtre soit exactement celles dont ait besoin l'algorithme d'apprentissage. Supposons par exemple que plusieurs variables sont redondantes et que l'une d'entre elles est sélectionnée. Introduire des valeurs manquantes sur cette dernière ne perturbera pas outre mesure un classifieur puisqu'il pourra toujours utiliser l'une des variables qui lui était initialement corrélée et dont la distribution n'a pas été modifiée.

6.6.4 Résultats expérimentaux

Nous avons mené des tests comparatifs entre différentes techniques de substitution des valeurs manquantes avec un double objectif. D'une part, nous souhaitons voir comment se comporte la version itérative de la méthode *Entropie* que nous avons mise en place, ainsi que la version itérative des *k plus proches voisins* (*kppvI*) par rapport aux méthodes classiques. D'autre part, nous espérons mettre en évidence le fait que chaque méthode est plus ou moins adaptée aux caractéristiques du problème considéré. Ces caractéristiques sont celles que nous avons mises en évidence en établissant notre taxinomie des méthodes de substitution du point de vue de l'utilisateur (voir figure 6.6).

Les techniques comparées peuvent être regroupées dans les familles suivantes :

- moyenne : *Moyenne* et *Médiane*, ou *Mode* pour les données symboliques, ainsi que les versions supervisées *CMoyenne*, *CMédiane*, *CMode* et les versions stochastiques : *MoyenneA* et *CMoyenneA*
- aléatoire : *AleatoireMM*
- plus proches voisins : version classique *kppv* et itérative *kppvI*
- régression linéaire : version itérative, locale *LLSI* et la version stochastique associée *ALLSI*.
- classification supervisée : *J48*, *IB1* et *NB*, uniquement pour les données symboliques. Ces trois classifieurs correspondent respectivement aux implémentations dans Weka 3.4.7 de l'algorithme d'induction d'arbres de décision C4.5, du plus proche voisin et de l'algorithme naïve Bayes.
- entropie : pour les attributs numériques, trois discrétisations ont été envisagées :
  - *EW* (Equal Width) : segmentation en intervalles de longueurs égales.
  - *EF* (Equal Frequency) : segmentation en intervalles contenant tous la même proportion d'exemples.
  - *ID3* : discrétisation supervisée binaire, récursive. Elle génère une partition en deux sous-ensembles de manière à optimiser le gain d'information, puis elle recommence avec chacun des sous-ensembles, de manière récursive, jusqu'à vérifier un critère d'arrêt (nombre d'exemples minimum dans un intervalle par exemple).

Nous avons choisi d'utiliser le protocole *DDP*, dont le principe est schématisé figure 6.8. D'une part, ce protocole respecte bien les garde-fous mentionnés précédemment. La génération des données manquantes ainsi que leur substitution ne se fait qu'une fois que les bases d'apprentissage et de test ont été séparées. D'autre part, le fait que la base de test ne soit pas trouée nous place dans un contexte certes moins proche de la réalité, mais permet de mieux contrôler les paramètres de l'expérience, car toutes les chaînes d'apprentissage seront évaluées sur des bases de test identiques.

Nous avons appliqué ce protocole sur cinq bases de données symboliques et huit bases

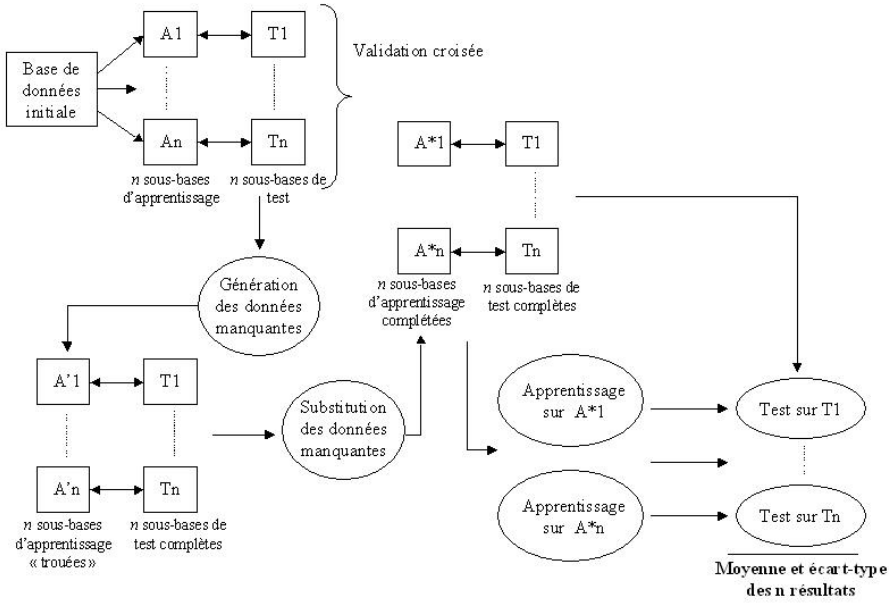

FIG. 6.8 – Protocole DDP pour l'évaluation des performances d'une technique de substitution des valeurs manquantes

numériques, toutes prises dans le *UCI repository*[13]. Ces 13 bases sont complètes, ce qui nous a permis de contrôler le mécanisme de génération des données manquantes ainsi que leur proportion. Les principales caractéristiques de chacune de ces bases symboliques sont résumées dans le tableau B.1 tandis que le tableau B.2 regroupe les informations concernant les bases numériques. Ces deux tableaux sont disponibles à l'annexe B

Pour chacune des 13 bases complètes nous avons construit 10 couples (apprentissage, test), et pour chacun de ces couples nous avons généré des données manquantes selon le mécanisme MCAR, avec 5 taux de valeurs manquantes différents (10%, 20%, 30%, 40% et 50%). Les performances de chacune des techniques ont été évaluées avec trois classifieurs : IB1 (plus proche voisin, implémenté dans Weka 3.4.7), J48 (C4.5, implémenté dans Weka 3.4.7) et NB (naïve Bayes également implémenté dans Weka-3.4.7) afin de voir si le classifieur de sortie influe sur les performances d'une technique de substitution.

Nous utilisons le taux de bonnes classifications, noté *Acc* dans la suite, pour comparer les différents classifieurs obtenus. Cette mesure est certes très discutable comme nous l'avons précédemment mentionné, mais c'est celle qui est utilisée dans les études empiriques sur les données manquantes auxquelles nous souhaitons pouvoir confronter notre travail.

Nous avons également eu recours à la moyenne des taux de reconnaissance de chaque classe (*Balanced Accuracy* pour les anglo-saxons), notée ci-après *BalAcc*. Reprenant les notations introduites à la section 2.3, nous notons $rappel\,(i)$ le taux de reconnaissance de la classe i, également appelé taux de rappel. Si K est le nombre total de classe, nous avons $BalAcc = \frac{1}{K}\sum_{i=1}^{K} rappel\,(i)$. Cette mesure est mieux adaptée pour les problèmes dans lesquels les observations sont inégalement réparties entre les différentes classes. En outre, inclure au moins deux mesures d'évaluation dans le protocole expérimental nous permet de juger de l'impact du choix de cette mesure sur le choix de la meilleure technique de substitution.

Lors de l'analyse des résultats nous allons être amené à comparer les classifieurs obtenus après substitution des valeurs manquantes par différentes techniques. Pour pouvoir conclure

[13]University of California Irvine http://www.ics.uci.edu/~mlearn/MLRepository.html

quant à la supériorité d'une technique sur une autre, il nous faut nous appuyer sur une méthode robuste permettant de juger du caractère significatif des différences observées. Nous aurons, aussi bien dans le cas numérique que symbolique, plus de deux classifieurs à comparer, chacun ayant été évalué sur plusieurs bases de données. Nous nous trouvons donc face à un problème de comparaison multiple.

Suivant les recommandations faites à la section 5.5, nous aurons recours au test non paramètrique de Friedman pour voir si l'on peut rejeter l'hypothèse selon laquelle toutes les techniques mises en balance ont les mêmes performances. Si tel est le cas, il nous faudra utiliser l'un des tests *post-hoc* associés au test de Friedman. Ayant introduit une nouvelle technique de substitution des valeurs manquantes, il nous importe essentiellement de voir comment celle-ci se comporte par rapport aux techniques existantes. Aussi nous tournerons-nous vers un z test combiné à la procédure ascendante de Holland-Copenhaver pour assurer le contrôle du taux d'erreurs global.

Les tests utilisés pour interpréter nos résultats expérimentaux sont basés non pas directement sur les performances de chaque technique, mais sur les rangs de ces performances. Afin de présenter des résultats synthétiques nous avons donc décidé de reporter la moyenne sur l'ensemble des bases de données, des rangs obtenus par chaque technique, et ce, pour chaque classifieur et chaque critère de performance. Précisons que nous avons adopté la convention suivante : meilleures sont les performances d'une méthode vis-à-vis des autres, et plus le rang de cette méthode est petit. Nous mettrons en gras le meilleur rang moyen pour chaque classifieur et chaque critère de performance. De plus, tous les rangs significativement différents de celui de notre technique Entropie *seront suivis d'une astérisque. Le niveau de confiance a été fixé à* 95% *pour toutes nos expériences.*

6.6.4.1 Données symboliques

Commençons par analyser les résultats obtenus sur les cinq bases symboliques. Ils sont synthétisés dans le tableau 6.9.

TAB. 6.9 – Comparaison statistique, sur des données symboliques, des techniques de substitution des valeurs manquantes[1]

Classifieurs / Méthodes	J48		IB1		NB	
	Acc	*BalAcc*	*Acc*	*BalAcc*	*Acc*[2]	*BalAcc*
AleatoireMM	4.8*	5.22*	5.54*	4.86*	**3.3**	3.48
Mode	4.82*	4.48*	3.4	3.52	4.52	4.04
CMode	**2.26**	**2.12**	**2.36**	**2.42**	3.64	**3.02**
J48-Classifieur	4.54*	4.66*	4.8*	5.18*	3.7	4.4
IB1-Classifieur	4.36	4.7*	4.34*	4.42*	3.98	4.3
NB-Classifieur	4.28	4.46*	4.72*	4.78*	4.56	5.2*
Entropie	2.94	2.36	2.84	2.82	4.3	3.56

[1] Ce tableau contient la moyenne des rangs, prise sur les 25 bases de données symboliques : 5 taux de valeurs manquantes pour chacune des 5 bases de données.

[2] Le test de Friedman considère que les différences observées entre tous les classifieurs ne sont pas significatives.

De ces résultats, il apparaît que la technique que nous avons proposée obtient des performances tout à fait satisfaisantes. Elle s'avère statistiquement supérieure à la quasi-totalité des autres techniques, pour au moins un classifieur et un critère de performance. Seul *CMode* obtient des rangs plus faibles, mais les écarts ne sont jamais significatifs. Il est à noter que le comportement de la méthode *Entropie* est beaucoup plus intéressant avec les classifieurs J48 et IB1, qu'avec NB. Cela n'est guère étonnant pour J48 puisque la fonction de discrimination utilisée par celui-ci est exactement celle que notre technique optimise. Notons que les différences observées ne sont significatives qu'avec J48 et IB1 et non avec NB. Avec ce dernier, on peut remarquer que le meilleur rang moyen est assez élevé, ce qui nous laisse supposer qu'aucune des méthodes envisagées n'est satisfaisante.

Ce dernier point conforte notre hypothèse selon laquelle l'algorithme d'induction utilisé au bout de la chaîne d'apprentissage influe sur la qualité de la méthode de substitution utilisée en amont. Le comportement de la technique *AléatoireMM* en offre un bon exemple. En effet, alors qu'elle obtient les rangs les plus élevés avec J48 et IB1, elle semble être la mieux adaptée avec NB. Même si les différences observées ne sont pas significatives, le changement observé entre J48 et IB1 d'un côté, et NB de l'autre, est assez net.

Avec J48 et IB1, les faibles performances de cette technique purement aléatoire ne sont guères étonnantes. Elle n'a été introduite dans l'étude comparative que pour servir de référence. En revanche, le revirement de situation avec NB est plus que surprenant, aucune technique ne semblant apporter une quelconque plus-value par rapport à cette méthode. Il serait bon d'approfondir ce point, par exemple en procédant à de nouveaux tests sur un plus grand nombre de bases afin de le confirmer ou de l'infirmer.

La mesure de performance semble également avoir un impact sur l'ordonnancement des méthodes de substitution. On observe en effet des différences entre les ordonnancements induits par les critères *Acc* et *BalAcc*, pour un même classifieur. C'est le cas pour IB1. Avec le critère *Acc* la méthode ayant le moins bon rang est *AléatoireMM*, alors qu'il s'agit de *J48-Classifieur* avec le critère *BalAcc*.

Il est également intéressant de constater que selon le critère, les résultats des tests statistiques diffèrent. Considérons le classifieur J48. Alors que les méthodes *IB1-Classifieur* et *J48-Classifieur* sont significativement inférieures à *Entropie* selon le critère *Acc*, tel n'est plus le cas avec *BalAcc*. Si l'on considère maintenant le classifieur NB, le test de Friedman conclut sur l'existence de différences significatives entre les différentes méthodes selon le critère *BalAcc*, mais pas selon le critère *Acc*.

L'utilisation d'algorithmes d'apprentissage supervisé, contrairement à nos attentes, ne se révèle pas très performante. Ainsi, aucun des classifieurs *IB1*, *J48*, *NB* ne semble être une solution recommandable, leurs performances étant quasiment systématiquement significativement inférieures à celle de la méthode *Entropie*. Elles sont de plus comparables à celles de l'aléatoire.

L'analyse des rangs moyens nous a permis de mettre en évidence des différences entre les comportements des différentes méthodes envisagées. Mais il s'agit d'une analyse globale, qu'il convient de raffiner si l'on veut être en mesure d'apprécier les points communs et les spécificités de ces techniques.

La figure 6.9 regroupe les graphiques donnant la moyenne des performances de l'ensemble des techniques en fonction du taux de valeurs manquantes pour chacune des bases de données de l'étude. Chacun des graphiques correspond à un classifieur et un critère d'évaluation particuliers. Une tendance émerge de ces graphiques : la moyenne des performances a tendance à décroître lorsque le taux de valeurs manquantes augmente. Ceci est bien en accord avec l'idée intuitive selon laquelle la dégradation de la qualité des données

s'accompagne d'une dégradation de la qualité des classifieurs construits à partir de ces données. Notons cependant que dans certains cas, les performances sont stables, ce qui est le signe d'une certaine robustesse.

La comparaison des différents graphiques d'une même colonne, correspondants aux résultats obtenus avec les trois classifieurs pour un même critère d'évaluation, permet de mettre en évidence l'influence du classifieur. Selon celui qui est considéré, la moyenne des performances sur une même base de données n'évoluera pas de la même façon en fonction du taux de valeurs manquantes. L'exemple le plus marquant est celui de la base *Car Evaluation*. Avec J48 et NB, quel que soit le critère d'évaluation, la moyenne des performances sur cette base décroît assez nettement alors qu'elle est stable avec IB1. Pour toutes les autres bases, les performances décroissent avec IB1. *Car Evaluation* est la base de données qui contient, de loin, le plus d'exemples. Rappelons qu'IB1 est un classifieur basé sur la recherche du plus proche voisin. Il est donc vraisemblable que si le nombre d'exemples est suffisant, ses performances soient relativement stables, même si la proportion de valeurs manquantes est grande.

Si l'impact du classifieur a pu de nouveau être mis en évidence à travers ces courbes, il n'en est pas de même pour le critère d'évaluation. Pour un classifieur donné, on peut en effet constater que les courbes obtenues pour les différentes bases de données ont la même allure quelle que soit la mesure de performance considérée.

Afin de caractériser les différentes méthodes de substitution nous avons tracé leurs performances sur les différentes bases de données, pour un taux de bonnes classifications donné. La figure 6.10 contient les graphiques correspondant aux trois classifieurs.

> *Sur la colonne de gauche le taux de valeurs manquantes est fixé à 10%, contre 50% pour la colonne de droite. Tous les graphiques de cette figure correspondent au critère* Acc[14].

Alors que toutes les méthodes semblent très proches lorsque seules 10% des données manquent, les méthodes *Entropie* et *CMode* se distinguent des autres par leurs bonnes performances lorsque la proportion des données manquantes atteint 50%. Ceci semble en accord avec l'analyse des rangs effectuée précédemment : ces deux méthodes sont les plus performantes en moyenne. Il est donc naturel d'observer que plus la proportion de valeurs manquantes est grande, mieux elles arrivent à se distinguer des autres techniques.

La situation est plus confuse avec le classifieur NB. Avec 50% de données manquantes, *Entropie* et *CMode* s'avèrent être à la fois les plus performantes sur certaines bases et les moins performantes sur d'autres. En tout état de cause, nous avons mis une nouvelle fois en évidence l'influence du classifieur. Cette influence est également notable sur la méthode *AléatoireMM*. Ses performances sont nettement moins bonnes que celles des autres méthodes, aussi bien avec 10% qu'avec 50% de données manquantes, pour les classifieurs J48 et IB1. En revanche, avec NB, ses performances sont au-dessus de la moyenne lorsque le taux de valeurs manquantes est de 50%.

Ces courbes permettent d'aller plus loin dans l'analyse comparative que le seul tableau des rangs. Parmi les faits saillants, on retrouve bien ceux qui ont été mis en exergue lors de l'analyse des rangs. Mais ces courbes ne sont pas suffisantes pour comprendre des résultats tels que ceux qui sont observés avec le classifieur NB. Il faudrait prendre en compte les caractéristiques intrinsèques des différentes bases de données pour aller plus loin. En revanche, gardons-nous de la tentation d'utiliser la moyenne des performances de chaque technique, prise sur l'ensemble des bases d'évaluation. Chacune de ces bases correspond en effet à un problème spécifique, avec un nombre de classes, une répartition des exemples

[14] Les résultats obtenus avec le critère *BalAcc* sont similaires.

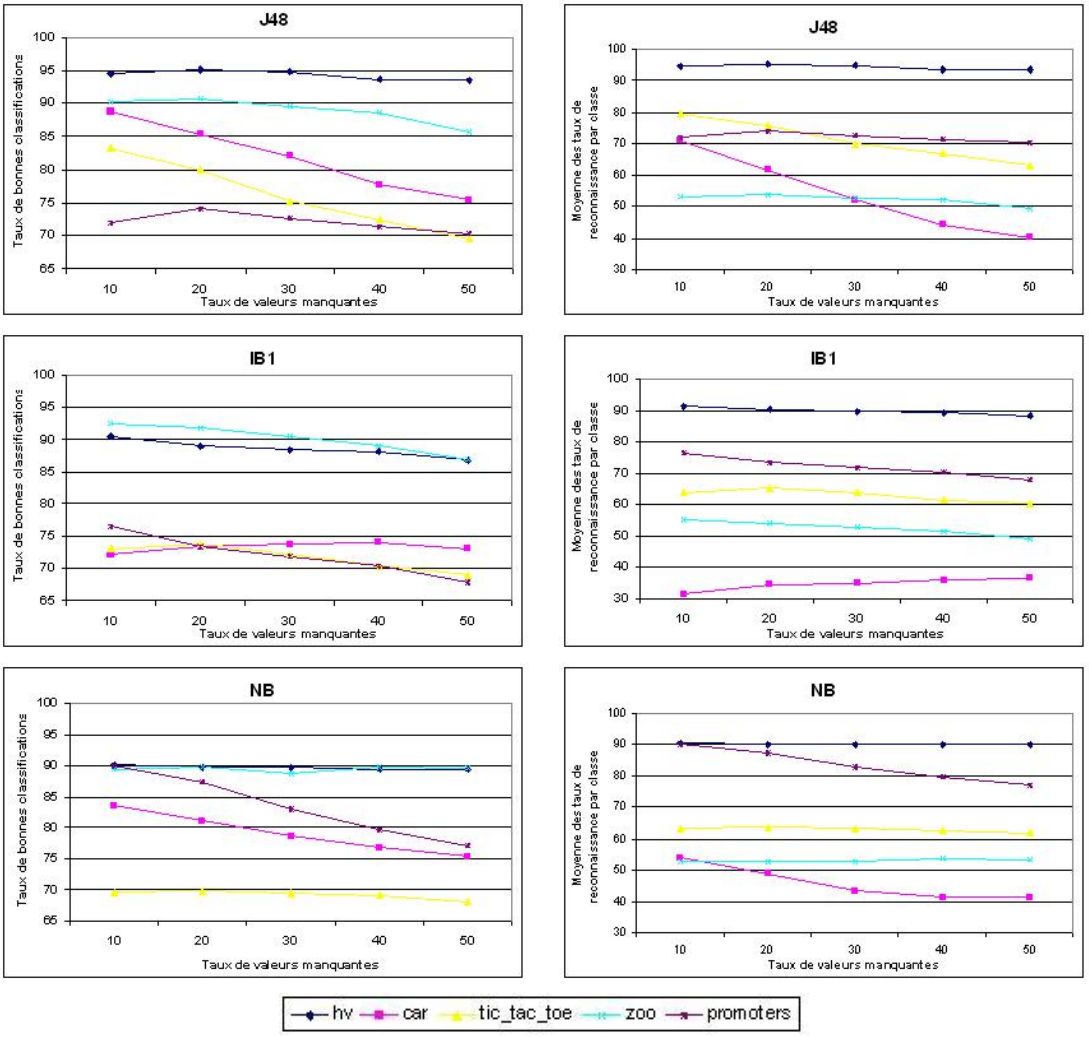

FIG. 6.9 – Performances moyennes des techniques de substitution en fonction du taux de valeurs manquantes pour chacune des bases de données étudiées

entre les classes, distincts. Aussi la moyenne des performances sur ces différents problèmes a-t-elle peu de sens. Il s'agit du problème d'incommensurabilité mentionné à la section 5.3.

6.6.4.2 Données numériques

Le tableau 6.10 fournit un récapitulatif des résultats obtenus sur les huit bases numériques. Comme dans le cas des données symboliques, elle ne contient pas les performances mais les rangs moyens des techniques de substitution entrant dans l'étude comparative. La moyenne est prise sur l'ensemble des 40 bases incomplètes testées : 5 taux de valeurs manquantes pour chacune des 8 bases. Dans ce tableau, les rangs typographiés en gras sont les meilleurs pour le classifieur et le critère d'évaluation considérés. Ceux qui sont suivis d'une astérisque sont significativement différents du rang de la méthode ID3-Entropie, *qui nous servira de référence. C'est en effet celle, parmi les trois techniques entropiques, qui semble avoir les meilleurs résultats.*

Dans l'ensemble les résultats du tableau 6.10 confirment ceux qui ont été obtenus sur les données symboliques. En effet, notre technique est ici aussi très prometteuse. Ses performances sont significativement supérieures à celles de toutes les autres techniques, exceptée *CMoyenneA*, pour au moins un classifieur et un critère d'évaluation. Elles ne sont de plus

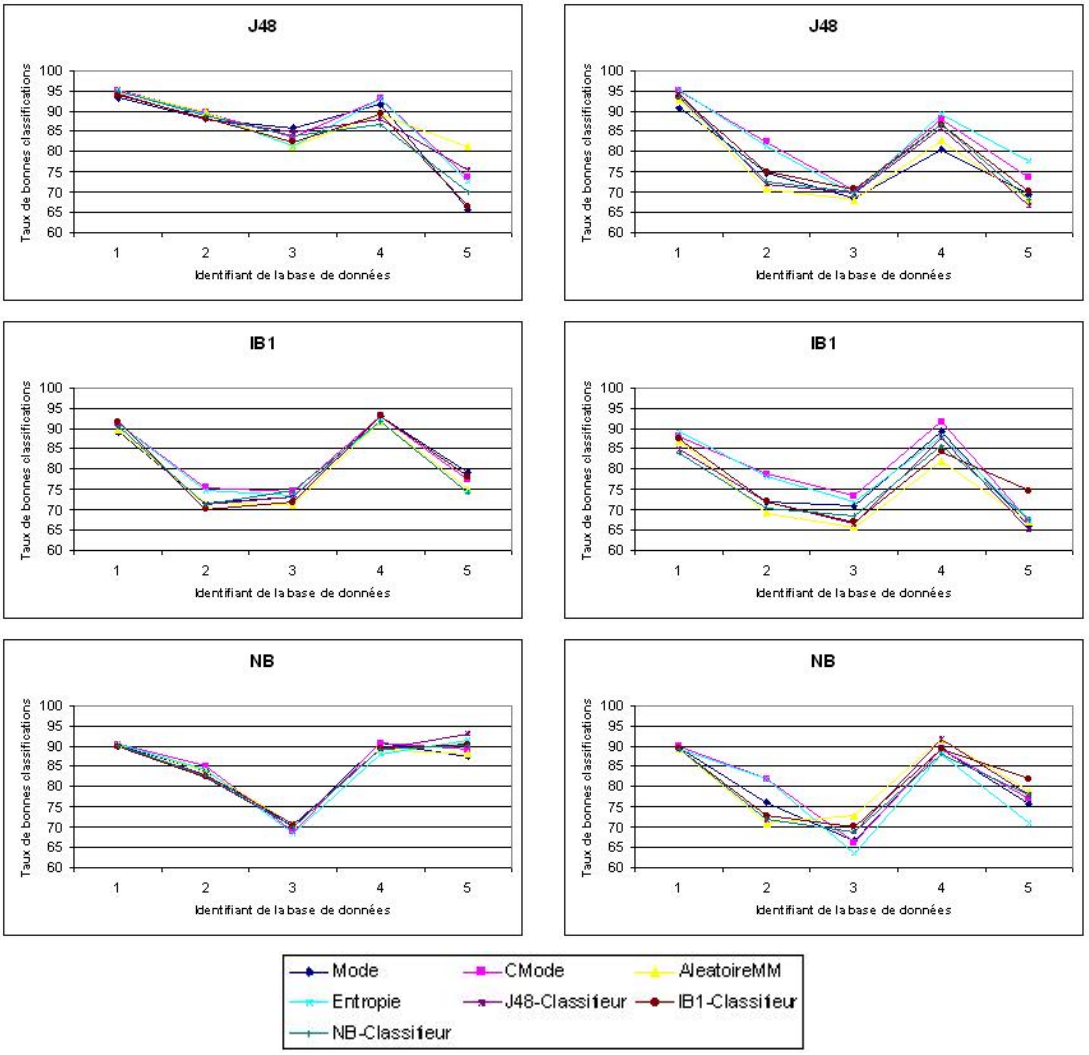

FIG. 6.10 – Performances moyennes des techniques de substitution en fonction des bases de données étudiées, pour un taux de valeurs manquantes fixé (10% à gauche et 50% à droite)

jamais significativement inférieures à celles d'une autre technique, y compris *CMoyenneA*. Ce sont donc les deux mêmes types de technique qui se distinguent. En effet, *ID3-Entropie* est une extension de notre technique entropique aux données numériques et *CMoyenneA* peut être considérée comme l'équivalent stochastique, pour les données numériques, de *CMode*.

En conformité avec ce qui avait été observé sur les données symboliques, nous pouvons également remarquer que la technique purement aléatoire ainsi que celle qui repose sur un modèle de prédiction utilisant un classifieur sont presque systématiquement statistiquement inférieures à notre technique. Nous déconseillons donc fortement leur usage.

Si cela n'est guère étonnant pour *AléatoireMM*, cela l'est plus pour *ID3-J48-Classifieur*. Cela pourrait s'expliquer par le fait qu'utiliser un classifieur pour prédire une donnée numérique est mal adaptée, et que ce sont surtout la discrétisation et la procédure de passage d'un intervalle à une valeur numérique qui dégradent les performances. Mais cette justification n'est pas valable. Les mêmes procédures sont en effet appliquées avec succès à *ID3-Entropie*. De plus, sur des données purement symboliques, nous avions déjà constaté

TAB. 6.10 – Comparaison statistique, sur des données quantitatives, des techniques de substitution des valeurs manquantes[1]

Classifieurs / Méthodes	J48		IB1		NB	
	Acc	*BalAcc*	*Acc*	*BalAcc*	*Acc*	*BalAcc*
AleatoireMM	7.68*	8.09*	8.6*	8.64*	7.39*	9.19*
Moyenne	5.39	4.79	4.75	4.44	7.03*	6.05*
CMoyenneA	5.25	**4.47**	**3.03**	**2.76**	3.73	**3.38**
ID3-J48-Classifieur	5.61	5.9	6.7	6.29*	5.56*	5.81*
5ppv	5.24	5.55	6	5.89	5.29	5.33*
1ppvI	5.54	5.35	6.81*	6.74*	4.98	5.36*
1LLSI	5.3	5.31	4.71	4.71	6.69*	5.7*
EW-Entropie	5.8	6.11	5.62	6.36*	5.64*	5.9*
EF-Entropie	4.69	4.62	4.55	4.92	5.09	4.9
ID3-Entropie	**4.51**	4.81	4.22	4.25	**3.62**	3.39

[1] Ce tableau contient la moyenne des rangs, prise sur les 40 bases de données numériques : 5 taux de valeurs manquantes pour chacune des 8 bases de données.

les faibles performances des classifieurs en tant qu'outils de substitution des valeurs manquantes.

L'influence du classifieur, mise en évidence sur les données symboliques, est ici aussi un fait saillant. Cette fois, c'est avec NB que l'on constate les plus gros écarts entre notre technique et les autres. Alors que les différences ne sont significatives qu'avec *AléatoireMM* pour J48, elles le sont avec 5 ou 7 des 9 autres techniques pour NB, selon la mesure de performance considérée. Ce dernier point suggère que le choix de cette mesure n'est pas anodin. On retrouve ici aussi l'une des conclusions que nous avions tirées des expérimentations sur données symboliques.

Quant aux autres techniques envisagées, les plus proches voisins, y compris la version itérative, ainsi que la régression linéaire itérée, leurs performances sont bien moins élevées qu'escomptées. Batista et Monard (2003) ainsi que Zou *et al.* (2005) avaient en effet relevé que les plus proches voisins était une technique de substitution très prometteuse. Rappelons cependant que le protocole expérimental de Zou et al. est loin d'être satisfaisant puisqu'il est très dépendant de l'échantillonnage. Les tests ne sont menées que sur une seule paire de bases (apprentissage,test). Quant à Batista et Monard, il faut préciser, d'une part, qu'ils utilisent une version particulière des plus proches voisins qui s'appuie sur l'identification de prototypes, et d'autre part, que seuls les attributs les plus discriminants contiennent des données manquantes. Les différences entre leurs constatations et les nôtres peuvent donc avoir deux explications, entre lesquelles nos expérimentations ne nous permettent pas de trancher.

- Leur version des plus proches voisins est nettement plus performante que la version classique.
- Laissant certains attributs intacts, leur mesure de distance est moins dégradée que la nôtre.

Notons aussi que les études précédentes utilisent la moyenne et non sa version supervisée qui intègre l'information de classe. Enfin, les comparaisons statistiques qui sont effectuées dans les études existantes sur le sujet reposent sur l'application répétée du test de Student, procédure peu recommandable dans notre cas, comme nous l'avons illustré à la section 5.4.

À l'instar de ce qui a été fait sur les données symboliques, nous allons maintenant essayer d'affiner notre analyse. L'évolution de la moyenne des performances en fonction du taux de valeurs manquantes est très proche de celle qui a été observée sur les données symboliques. Aussi ne tracerons-nous pas les graphiques correspondants. Ils ont la même allure que ceux qui ont été présentés dans la figure 6.9 : les performances décroissent avec l'augmentation du taux de valeurs manquantes.

La figure 6.11 permet de comparer les performances de chaque technique sur les différentes bases de données pour un taux de valeurs manquantes fixé à 10 et 50%. Nous ne donnons que les graphiques correspondant au classifieur IB1 avec le critère *Acc*. Les quatre autres paires de graphiques que nous aurions pu tracer ont la même allure et sont tout aussi délicats à interpréter. La situation est en effet plus confuse que dans le cas des données symboliques (voir figure 6.10). Les seuls enseignements que nous pouvons tirer de ces graphiques sont les suivants. Les techniques de substitution ont des performances voisines lorsque le taux de valeurs manquantes est faible : 10%. Lorsque ce taux est important, des écarts apparaissent, mais seule la technique *AléatoireMM* se distingue par ses piètres performances.

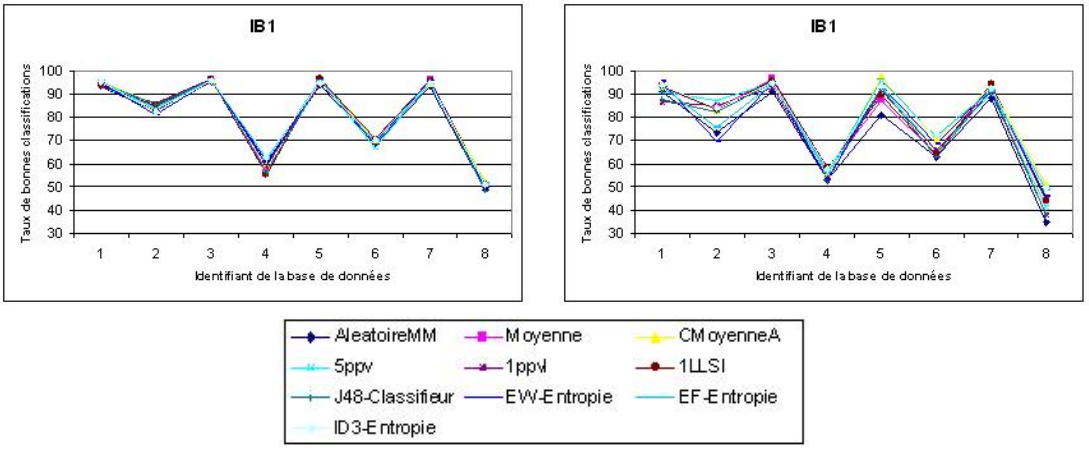

FIG. 6.11 – Performances moyennes des techniques de substitution en fonction des bases de données étudiées, pour un taux de valeurs manquantes fixé (10% à gauche et 50% à droite)

L'analyse des rangs des performances obtenues par les différentes méthodes substitution ne tient compte que de l'ordonnancement des différentes méthodes pour les différents problèmes considérés. De ce fait elle ne permet pas de se rendre compte de la magnitude des écarts entre les performances de ces méthodes. Étant donné que nous avons des résultats pour 3 classifieurs évalués sur 8 bases de données selon 2 critères de performance, visualiser l'ensemble des résultats bruts nécessiterait le tracé de 48 graphiques. Aussi avons-nous choisi de ne présenter sur la figure 6.12 que les performances relatives à une base de données : *Yeast*.

Notons que dans un souci de lisibilité, la disposition des graphiques respecte celle qui a été adoptée jusqu'ici : sur la colonne de gauche sont données les résultats relatifs au

critère Acc*, tandis que sur la colonne de droite il s'agit de* BalAcc*. À chaque ligne correspond un classifieur. De haut en bas nous avons J48, IB1 et NB.*

Nous avons retenu cette base car les performances observées y sont assez représentatives du comportement global des différentes méthodes de substitution. On retrouve en effet au travers de ces différentes courbes les caractéristiques mises en évidence jusqu'ici.

- Les performances sont décroissantes en fonction du taux de valeurs manquantes, et ce, quel que soit le classifieur ou le critère d'évaluation considéré.
- *ID3-Entropie* et *CmoyenneA* ont dans l'ensemble les meilleures performances et *AléatoireMM* les plus mauvaises.
- L'influence du classifieur est manifeste. Avec J48 et IB1 les méthodes se distinguent assez facilement. En revanche, avec NB, nous observons un groupe de techniques dont les performances sont très proches. Au regard du critère *Acc*, il est même très difficile de départager les 6 meilleures techniques.
- L'influence du critère d'évaluation est également notable, ne serait-ce que vis-à-vis des deux meilleures techniques. En effet *ID3-Entropie* semble dominer *CMoyenneA* lorsque l'on considère le taux de bonnes classifications, mais le rapport de forces s'inverse clairement lorsque la moyenne des taux de reconnaissance par classe est utilisée.

Ayant réalisé l'ensemble de nos expériences sur des bases complètes, il est également possible de juger de la qualité des prédictions qui sont effectuées. Cela ne correspond certes pas à l'objectif que nous nous sommes fixé, mais permet cependant de considérer les différentes méthodes sous un angle différent. Nous avons utilisé le critère NRMSE. Par souci de cohérence nous présentons les résultats obtenus sur la base *Yeast* qui nous a préalablement servi de support à l'analyse. Ces résultats sont illustrés à la figure 6.13.

Fait surprenant, contrairement à ce que nous observions à la figure 6.12, les courbes sur cette figure sont quasiment des constantes. Ceci signifie que les performances ne se dégradent pas ou du moins très peu lorsque le taux de valeurs manquantes augmente. De plus amples tests, sur d'autres types de bases de données, permettraient vraisemblablement d'approfondir ce point.

Les méthodes entropiques sont nettement moins performantes que les autres, hormis la technique purement aléatoire. Nous nous y attendions sachant qu'elles n'ont pas du tout été construites pour répondre à cet objectif particulier. Certes il ne s'agit que des résultats sur une seule base de données, mais sur les autres bases que nous avons considérées les comportements sont assez similaires, si ce n'est que la méthode *ID3-Entropie* est tout de même plus proche des autres, avec un taux d'erreurs moins élevé. *EW-Entropie* et *AléatoireMM* sont en revanche toujours aussi faibles. La discrétisation simpliste en intervalles de même longueur ne semble donc jamais adéquate, ce qui est plutôt rassurant.

Les autres méthodes sont assez proches et présentent des taux d'erreurs bien plus faibles. Nous voyons ainsi que les plus proches voisins, la régression linéaire locale itérée ou même le classifieur J48 sont à peu près équivalents à *CMoyenneA* qui obtient les taux d'erreurs les plus faibles. Le bon comportement des techniques de prédiction selon ce critère est assez intuitif. D'une part, cela est en accord avec les résultats de l'état de l'art (Kim *et al.*, 2005; Oba *et al.*, 2003). Rappelons d'autre part, que l'objectif sous-jacent de l'utilisation d'une technique de prédiction pour estimer les valeurs de substitution est la minimisation de l'erreur de prédiction. Cela est particulièrement clair pour la régression linéaire, dont les paramètres sont estimés par la méthode des moindres carrés.

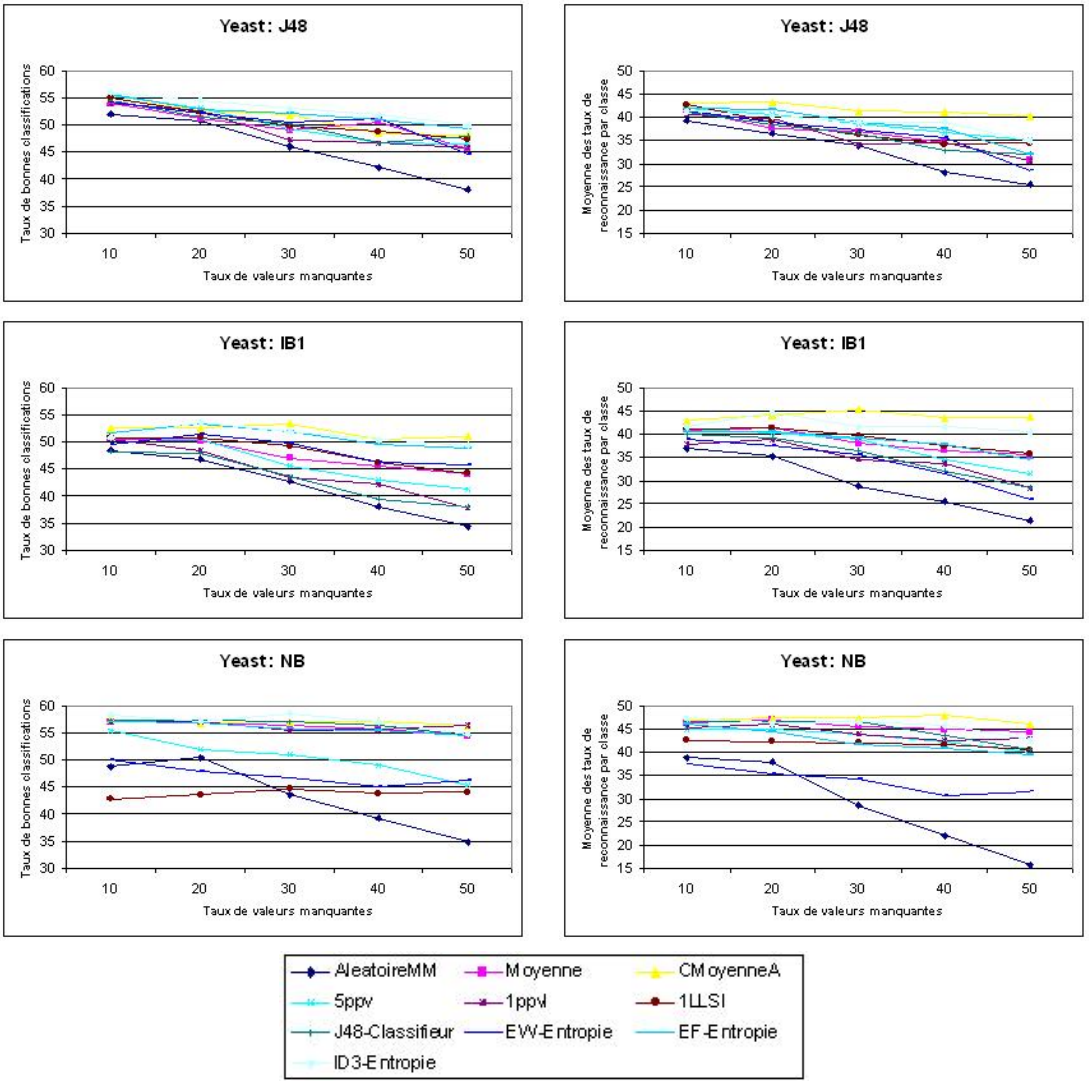

FIG. 6.12 – Performances des techniques de substitution des valeurs manquantes sur la base numérique *Yeast*

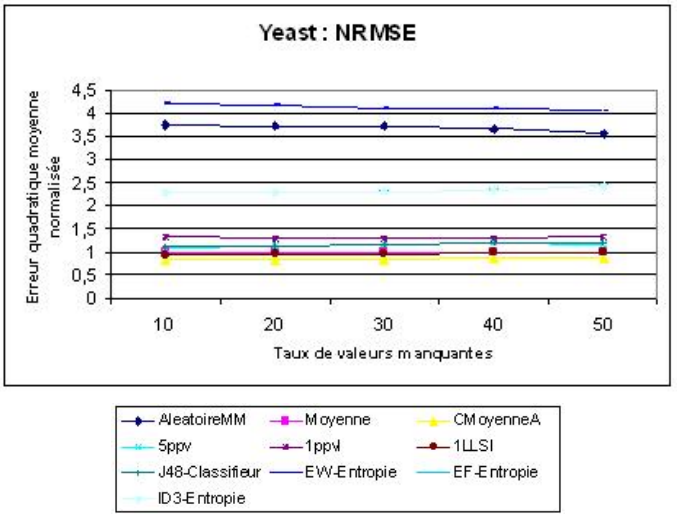

FIG. 6.13 – Erreur de prédiction des techniques de substitution des valeurs manquantes sur la base *Yeast*

6.7 Conclusion

Au vu de notre problème d'aide à la détection de crises et des spécificités des données afférentes, nous avons choisi de centrer notre étude des différentes techniques de traitement des données manquantes sur la tâche de substitution des valeurs manquantes. La mise en évidence des traits distinctifs des méthodes les plus répandues dans la littérature nous a permis de développer une taxinomie de ces méthodes. Nous avons également vu comment la plupart des méthodes de base pouvaient être étendues pour en créer de nouvelles, ce que nous avons appliqué aux *k plus proches voisins* pour en construire une version itérative.

Nous plaçant dans le contexte de l'apprentissage supervisé, nous avons pu changer de point de vue dans l'approche qui est classiquement adoptée dans le domaine. Nous avons ainsi développé une technique supervisée basée sur l'entropie, dans l'optique explicite de pouvoir améliorer la qualité d'un classifieur. Traditionnellement le problème est abordé sous l'angle de la minimisation de l'écart entre valeurs substituées et valeurs réelles.

Nous avons ensuite procédé à une comparaison empirique de diverses techniques afin de prendre la mesure, de manière pragmatique, de l'intérêt de notre nouvelle technique. Pour réaliser cette comparaison, nous nous sommes d'abord penché sur le problème du choix d'un protocole d'évaluation. Ce problème est souvent occulté dans la littérature. Or, en fonction de ce que l'on cherche à montrer, différents protocoles doivent être mis en place. À travers la construction d'une taxinomie de ces protocoles, nous avons pu mettre en relief les biais qui pouvaient être introduits par l'utilisation de protocoles inadéquats.

Les résultats obtenus sont plus que prometteurs. Notre méthode *Entropie* s'avère en effet très performante aussi bien sur des bases de données symboliques que numériques. L'utilisation de tests statistiques adaptés à notre cadre expérimental a mis en évidence l'existence de différences statistiquement significatives avec l'ensemble des autres techniques, hormis *CMode* et *CMoyenneA*.

Ces deux types de techniques peuvent toutes deux être qualifiées de supervisées dans la mesure où elles ont recours à l'information de classe. Il serait alors utile d'évaluer ces techniques sur d'autres protocoles, en particulier celui dans lequel les bases de test sont incomplètes, pour se rapprocher de cas d'applications réelles. Les techniques supervisées ne sont en effet pas applicables directement aux exemples de la base de test pour lesquels la classe est inconnue. Il faut utiliser, soit une autre méthode de substitution pour ces exemples, soit un classifieur, capable de traiter les données incomplètes, pour affecter temporairement une classe à chacun de ces exemples.

La technique de substitution que nous avons proposée repose sur la maximisation du gain d'information. En procédant ainsi, nous cherchons à restaurer la capacité de discrimination des différents attributs. Le gain d'information n'est qu'une mesure parmi d'autres de cette capacité de discrimination. Aussi serait-il intéressant de construire de nouvelles méthodes de substitution basées sur des mesures de discrimination aux propriétés différentes, afin d'apprécier l'impact que le choix de cette mesure peut avoir sur la qualité de la substitution. Nous reviendrons plus en détail sur ces mesures et leur application à la sélection d'attributs, à la section 7.3.1. Nous invitons le lecteur intéressé par une vue plus générale de ces mesures à se reporter aux travaux de Dang (2007).

L'hypothèse sous-jacente de notre approche est la suivante : *les valeurs manquantes dégradent la capacité de discrimination d'un attribut*. Or il est tout à fait envisageable que certains attributs ne soient que très peu, voire pas du tout, discriminants pour un problème de classification donné. Notre hypothèse ne sera donc pas valide pour ces attributs-là et il est possible que notre méthode génère alors des attributs dont le pouvoir de discrimination

soit totalement artificiel. Pour que cela ne se produise pas, il serait bon de procéder en amont de la substitution des valeurs manquantes à une phase de sélection d'attributs, de façon à supprimer ceux qui ne sont pas pertinents. Parmi les pistes d'amélioration de notre technique, deux autres points mériteraient également une attention particulière.

D'une part, l'algorithme d'optimisation itératif que nous avons développé pour minimiser l'entropie conditionnelle devrait être comparé à des algorithmes d'optimisation locale dont l'efficacité est reconnue, les algorithmes génétiques ou le recuit simulé par exemple. D'autre part, il faudrait faire en sorte que l'incertitude liée au processus de substitution puisse mieux être pris en compte. Sans recourir à des méthodes coûteuses de substitutions multiples, une solution consisterait à substituer une valeur manquante, non pas par une valeur précise mais par un sous-ensemble flou. Il faudrait alors cependant utiliser des algorithmes d'apprentissage qui soient capables de traiter directement des données floues.

En menant à bien notre étude empirique nous avions deux objectifs. D'une part, il s'agissait de mesurer l'intérêt pratique de notre méthode *Entropie*, ce que nous venons de souligner. D'autre part, nous souhaitions analyser le comportement des différentes méthodes dans des cas de figure particuliers afin de mettre en évidence l'influence de certains critères. L'objectif est de pouvoir aider un utilisateur à choisir une technique de subsitution. La figure 6.14 offre une synthèse de nos travaux relativement à cet objectif.

> *Ces critères sont entourés d'une ellipse. Nous avons encadré ceux dont nous n'avons pas évalué l'impact mais dont la valeur a été fixée de telle sorte qu'un domaine d'étude particulier a pu être circonscrit. Nous nous sommes par exemple placé dans le cadre de la classification supervisée. Nous disposons donc de l'information de classe, mais nous n'avons pas cherché à analyser le rôle que pouvait avoir la répartition des exemples dans les différentes classes.*

Des expériences que nous avons menées, force est de constater qu'il est difficile de tirer des lois générales qui nous permettraient de faire des recommandations précises à un utilisateur. Cependant, elles nous ont permis de mettre en évidence certaines tendances globales, relatives à l'ensemble des méthodes de substitution.

Lorsque les performances de ces méthodes sont estimées par le biais de celles d'un classifieur, elles ont tendance à décroître lorsque le taux de valeurs manquantes augmente. Il est intéressant de noter que le classifieur et la mesure de performance utilisée pour évaluer ce classifieur exercent une influence notable sur la qualité des techniques de substitution. Il n'y a pas de meilleure technique dans l'absolu. Selon le problème de classification auquel on s'attaque, les différentes techniques seront plus ou moins adaptées.

La qualité d'une méthode de substitution dépend de ce qu'on veut en faire, de l'objectif sous-jacent que l'on essaie d'atteindre. On observe ainsi que les techniques de prédiction qui se focalisent sur la proximité entre la valeur de substitution et une hypothétique valeur d'origine obtiennent de bonnes performances selon le critère NRMSE, alors que ce n'est pas le cas lorsqu'un classifieur est utilisé en aval de la substitution des valeurs manquantes. La situation est exactement contraire pour les techniques entropiques que nous avons développées expressément dans l'optique d'optimiser les performances d'un classifieur.

Comme l'illustre la figure 6.14, de nombreux critères restent à étudier comme par exemple l'influence du mécanisme de génération des données manquantes. Le champ d'investigation est encore large, même lorsqu'on ne considère que le domaine de la classification supervisée. Le plus important à notre avis est de bien sélectionner les bases de données utilisées pour réaliser de nouveaux tests, afin de pouvoir identifier les rôles joués par les caractéristiques intrinsèques des bases de données : nombre de variables, nombre d'exemples, répartition des exemples dans les différentes classes...

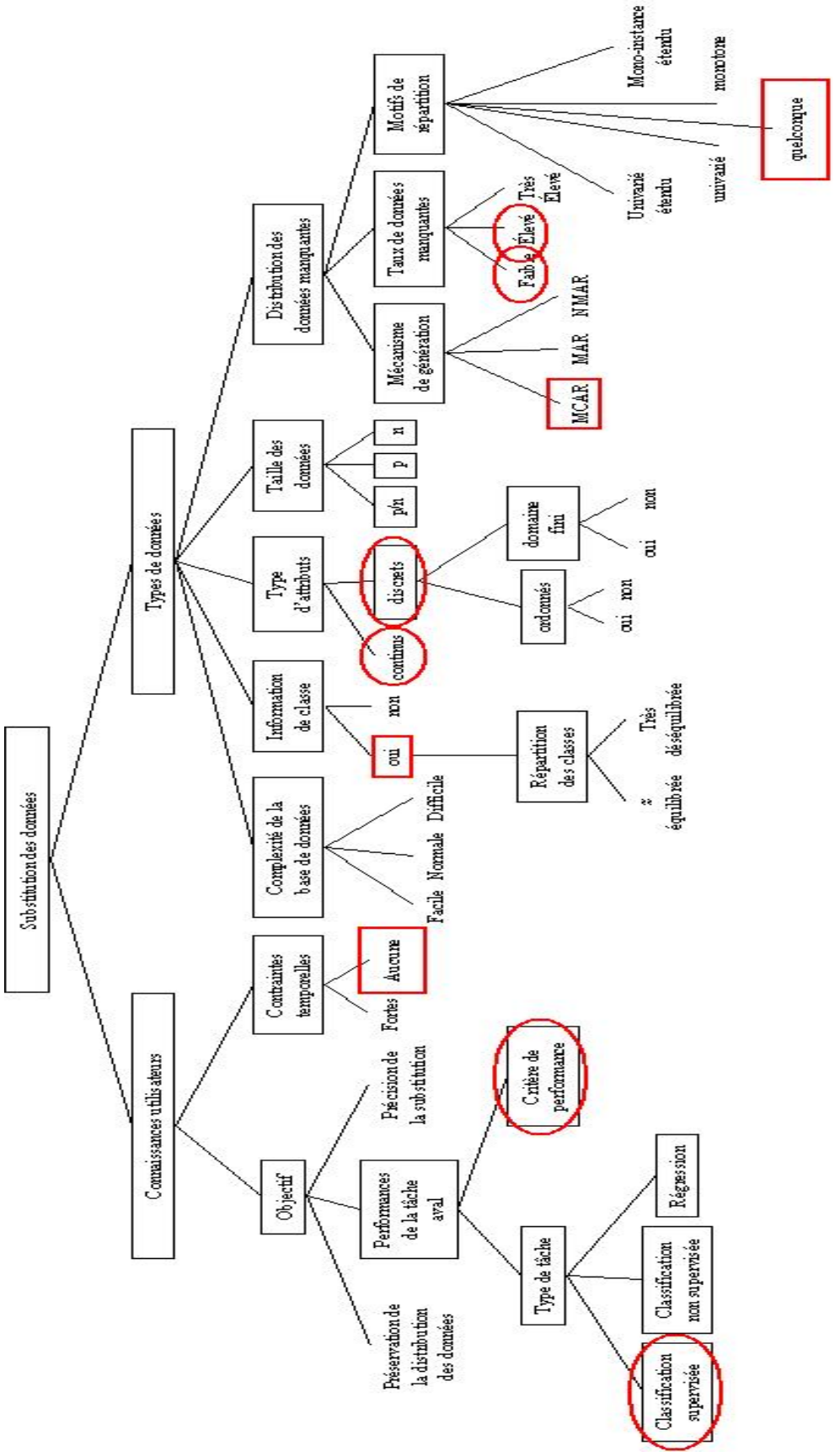

FIG. 6.14 – Taxinomie des techniques de substitution des valeurs manquantes, du point de vue de l'utilisateur : domaine couvert par nos expérimentations

Chapitre 7

Sélection d'attributs

Outre l'absence de nombreuses valeurs, la base de données qui a servi de support à nos premières expérimentations se caractérise par un nombre élevé d'attributs, nombre élevé en soi mais aussi relativement au nombre d'exemples disponibles. Nous avons vu à la section 2.3 au travers d'expérimentions menées sur cette base, que la réduction de la dimension du problème *via* une phase de sélection d'attributs constituait une étape importante voire indispensable pour construire un modèle performant.

Nous avons proposé d'utiliser un algorithme génétique pour chercher le sous-ensemble d'attributs maximisant les performances en classification de notre modèle, ce qui s'est avéré très efficace. Mais ce prétraitement est particulièrement coûteux et rallonge énormement l'apprentissage du modèle. Or nous avons pour objectif de construire une méthode générique qui puisse s'appliquer sur des bases de données beaucoup plus conséquentes. Nous avons par exemple construit de nouvelles bases sur les conflits qui contiennent près de cinq fois plus d'attributs que la base initiale.

Il nous faut donc envisager de nouvelles méthodes de sélection d'attributs, de moindre complexité. Ce besoin de réduction de la dimensionnalité n'est pas spécifique à la détection des conflits, mais est récurrent en fouille de données. Ne souhaitant pas restreindre notre champ d'application, nous aborderons cette question de façon aussi générique que possible, à l'image de ce que nous nous sommes efforcé de faire pour les données manquantes.

7.1 Position du problème

Les techniques d'analyse ou de fouille de données permettent d'apprendre un concept, d'extraire des informations pertinentes automatiquement à partir de données. Quelle que soit la technique mise en place, la qualité des données sur lesquelles se fait l'apprentissage joue un rôle fondamental. Outre le problème des données manquantes que nous avons abordé à la section 6, celui de la qualité des attributs utilisés pour décrire les exemples de la base de données est récurrent.

Une solution intuitive consiste à collecter autant d'attributs que possible en espérant que dans la collection finale il se trouvera suffisamment d'attributs de bonne qualité pour que l'algorithme d'apprentissage puisse apprendre un modèle performant. Incidemment, cela suppose que l'algorithme en question sera capable de repérer ces attributs.

7.1.1 Compromis biais-variance en apprentissage

Considérons un problème d'apprentissage supervisé. En reprenant les notations de l'annexe A, nous pouvons le formaliser de la façon suivante. Soit d_i le domaine de définition de v_i. Nous notons $D = d_1 \times d_2 \times ... \times d_p$ l'espace de dimension p correspondant au domaine de définition de $\mathcal{V}$. Soit D_y le domaine de définition de y. Chacun des exemples $e_i \in \mathcal{E}$ est décrit sur $\mathcal{V}$ et correspond à un point de $\mathcal{D}$. Nous supposons de plus que les e_i sont tirés aléatoirement, de façon indépendante, selon une loi de probabilité $\pi_{\mathcal{D}}$ sur $\mathcal{D}$. L'apprentissage supervisé consiste à chercher la fonction f qui approxime le mieux la fonction g inconnue qui relie les variables explicatives à la variable cible y.

$$\begin{array}{rcl} g : D & \rightarrow & D_y \\ e_i : (v_{i1}, ..., v_{ip}) & \mapsto & y_i = g(v_{i1}, ..., v_{ip}) \end{array}$$

L'objectif de l'apprentissage supervisé est de parvenir à approximer g à partir d'un ensemble de couples (e_i, y_i). On souhaite trouver une fonction f parmi l'ensemble des hypothèses H possibles, telle que l'erreur en généralisation $\epsilon_{\pi_{\mathcal{D}}}(f) = E_{\pi_D}[P(f(e) \neq g(e))]$ soit minimale.

Théoriquement, plus on intègre d'attributs dans le modèle et meilleures sont ses performances. L'erreur de Bayes optimale est en effet monotone, décroissante en fonction du nombre d'attributs (Kohavi et John, 1997). Mais ceci n'est vrai que si l'on dispose d'un nombre d'exemples infini. En pratique ceci n'est évidemment pas le cas. L'erreur de généralisation doit être estimée par l'erreur empirique :

$$\epsilon(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} dist(f(e_i), y_i)^2$$

avec

$$dist(f(e_i), y_i) = \begin{cases} |f(e_i) - y_i| & \text{en régression} \\ 0 & \text{en classification, si } f(e_i) = y_i \\ 1 & \text{en classification, si } f(e_i) \neq y_i \end{cases}$$

Sous ce formalisme, des éléments théoriques ont été avancés qui montrent que réduire la dimension du problème en supprimant certains attributs permet de réduire l'erreur empirique (Trunk, 1977; Ng, 1998). Plus le nombre d'attributs à prendre en compte pour déterminer l'hypothèse optimale f est grand et plus cette hypothèse sera complexe. Théoriquement cela devrait permettre d'approximer plus finement g, mais cela demande d'estimer beaucoup de paramètres.

En pratique, le nombre d'observations étant limité, il n'est pas possible d'estimer de manière robuste ces paramètres (variance élevée) et le risque est grand de faire du sur-apprentissage. Si trop peu d'attributs sont utilisés, les paramètres du modèle seront certes aisément estimés, mais le biais du modèle sera trop important. L'ensemble des hypothèses que l'algorithme pourra parcourir pour approximer g sera trop réduit et ne contiendra que des hypothèses bien trop éloignées de g. On retrouve le compromis biais-variance, classique en apprentissage. Pour plus une analyse approfondie du compromis biais-variance, en régression mais aussi en classification, le lecteur pourra se reporter à la thèse de Geurts (2002).

L'enjeu de la sélection d'attributs réside donc dans l'ajustement de ce compromis. Il s'agit de sélectionner les attributs de façon à guider l'algorithme d'apprentissage vers un sous-espace de l'espace des hypothèses, dans lequel une bonne approximation de g peut être trouvée. Ce sous-espace ne doit pas être trop complexe pour que les paramètres du

modèle puissent être estimés de façon robuste à partir des données à disposition, qui sont en nombre limité.

Outre ces considérations théoriques, de nombreuses études empiriques ont mis en évidence l'intérêt de la sélection d'attributs pour une tâche de classification supervisée. Les performances en généralisation peuvent être améliorées, parfois assez nettement, lorsqu'un sous-ensemble de l'ensemble initial d'attributs est retenu (voir par exemple (Doak, 1992; John *et al.*, 1994) ou plus récemment (Al-Shahib *et al.*, 2005; Cohen *et al.*, 2005)).

7.1.2 Objectifs de la sélection d'attributs

Dès que le nombre de variables est élevé et que le nombre d'exemples décrivant le domaine est limité, il semble donc important de procéder en amont de l'apprentissage à une phase de sélection d'attributs. Or, les cas d'application pour lesquels le nombre d'attributs varie entre une centaine et plusieurs dizaines de milliers alors que le nombre d'exemples est au mieux du même ordre de grandeur sont légion. Citons en particulier la fouille de texte (Forman, 2003), l'analyse des séquences ADN (Ding et Peng, 2003) ou encore la recherche d'images (Bins et Draper, 2001). La détection des crises comme le reflète la base de données que nous avons introduite à la section 2.2 est également concernée (Delavallade *et al.*, 2005).

Si l'amélioration des performances est l'un des principaux atouts de la sélection d'attributs, d'autres sont également importants suivant les caractéristiques du problème que l'on a à traiter.

- **Coût d'acquisition des attributs réduit**. Il peut en effet s'avérer coûteux de collecter certains attributs. S'en passer peut donc se révéler fort utile.
- Durée d'apprentissage réduite. Réduire le nombre d'attributs restreint l'espace de recherche de la fonction f. L'algorithme d'apprentissage est donc plus rapide. Souvent cet apprentissage est fait hors-ligne, les contraintes temporelles peuvent donc paraître secondaires. Cependant en très grande dimension, la complexité du problème peut être telle que l'apprentissage n'est pas possible. Langley et Iba (1993) ont par exemple montré que pour l'algorithme du plus proche voisin, le nombre d'exemples nécessaire pour atteindre une erreur en généralisation donnée croît de façon exponentielle par rapport au nombre d'attributs non pertinents.
- **Facilité d'interprétation des modèles**. Apprendre sur un espace de dimension plus faible permet de construire des modèles moins complexes qui seront, par voie de conséquence, plus simples à interpréter. Cela est flagrant avec les modèles de régression. Il sera toujours plus facile d'analyser un modèle contenant moins de variables explicatives. Des études ont également montré que la sélection d'attributs en amont de l'induction d'arbres de décision permettait non seulement d'améliorer les performances en classification (Perner, 2001), mais également de réduire la taille des arbres induits (Quinlan, 1993; Seban et Nock, 2001).
- **Mise en exergue des variables pertinentes pour la classification**. Cela contribue aussi à faciliter l'interprétation des résultats qui est une étape essentielle dans une optique telle que la nôtre, d'aide à la décision. L'un des enjeux est en effet de pouvoir focaliser l'attention des décideurs sur certaines caractéristiques du problème considéré.

Dans cette partie, nous nous focaliserons sur la sélection d'attributs dans le cadre de la classification supervisée, qui est celui dans lequel se place cette thèse. Les travaux correspondant dans le cadre de la régression sont très nombreux, les statisticiens ayant étudié le sujet de longue date. Pour plus d'informations sur ce sujet nous invitons le lecteur à se reporter aux ouvrages de Hastie *et al.* (2001) et Miller (2001). Quant à la sélection d'attributs pour la classification non supervisée, nous renvoyons le lecteur à l'article de Liu et Yu (2005).

Le reste de cette partie est organisée de la façon suivante. Nous commencerons par formaliser le problème de la sélection d'attributs à la section 7.2. Nous verrons ensuite quelles sont les principales techniques qui permettent de répondre à ce problème à la section 7.3. Au cours de cet état de l'art nous proposerons des généralisations de certains modèles existants, ce qui nous amènera à proposer au cours de la section 7.4 une nouvelle méthode de sélection des variables continues. Enfin, la section 7.5 sera pour nous l'occasion de faire un pont entre cette partie et la précédente. Nous avons mené des expérimentations afin d'étudier la chaîne d'apprentissage dans sa globalité c'est-à-dire en considérant les interactions entre le traitement des valeurs manquantes, la sélection d'attributs et l'apprentissage.

7.2 Définitions du problème

Ne retenir que certains attributs impose de faire des choix, de décider quels attributs conserver en fonction d'un certain critère. Quels que soient les objectifs exacts que l'on cherche à atteindre, il semble naturel de privilégier les attributs les plus pertinents au regard de la tâche finale à effectuer, à savoir la classification dans notre cas. L'étude du comportement de divers classifieurs sur des bases de données artificielles a par ailleurs permis de mettre en évidence la dégradation des performances de ces classifieurs en présence d'attributs non pertinents (Molina *et al.*, 2002; Ng, 1998; Kohavi et John, 1997).

Dans ces expérimentations, les attributs non pertinents sont des attributs générés aléatoirement et qui n'ont aucun lien avec la variable cible à modéliser. S'il est aisé de construire des attributs non pertinents, il est beaucoup plus délicat de les identifier dans des applications réelles, du moins tant que cette notion reste définie de façon aussi vague. La notion de pertinence joue donc un rôle central dans la sélection d'attributs. Aussi nous semble-t-il essentiel de commencer par une analyse plus poussée de cette notion avant de formaliser plus précisément la tâche que doit réaliser un algorithme de sélection d'attributs.

7.2.1 Pertinence d'un attribut

La notion de pertinence en tant que telle n'a aucun sens. Il s'agit d'une relation. Un objet pourra en effet être pertinent pour tel objet, mais pas pour tel autre. Il convient donc toujours de préciser l'objet cible pour lequel on cherche à savoir si tel objet est pertinent ou non. Notons que nous pouvons raisonnablement poser que cette relation est réflexive, mais rien ne permet d'affirmer *a priori* qu'elle doit être transitive ou même symétrique.

Dans notre contexte, il nous importe de trouver des attributs qui sont pertinents pour une tâche de classification supervisée. On s'intéressera donc à la pertinence d'un attribut v_i de $\mathcal{V}$ pour la classe y que l'on cherche à modéliser. Une définition satisfaisante de la pertinence dans ce cadre doit d'une part refléter le sens commun, et d'autre part pouvoir être utilisable concrètement ou au moins permettre une meilleure compréhension du problème.

Les travaux précurseurs de philosophes et logiciens tels que Gärdenfors (1978) ont exercé une influence notable sur les travaux liés à la pertinence en intelligence artificielle.

Nous nous focaliserons sur les retombées dans le domaine de l'apprentissage automatique[1]. Bell et Wang (2000) retracent la filiation entre les différentes définitions de la pertinence qui ont pu être proposées depuis les travaux de Gärdenfors. Pour Gärdenfors, la notion de pertinence est une relation ternaire et non binaire.

> Un objet o sera pertinent relativement à une hypothèse h dans le contexte c, si la vraisemblance de h, connaissant c, est affectée par la connaissance supplémentaire de o. Dans le cas contraire o sera considéré comme non pertinent.

Gärdenfors a développé une axiomatisation pour formaliser cette définition, mais nous ne retiendrons pour la suite que les idées principales sur lesquelles elle repose.

- La notion de contexte est fondamentale (voir à ce sujet les réflexions inspirées de la pragmatique de Ekbia et Maguitman (2001)).
- La pertinence se caractérise par la variation de la vraisemblance d'une hypothèse entre deux états de connaissance.

Nous allons maintenant voir comment ces idées ont été appliquées pour qualifier et parfois quantifier la pertinence d'une variable.

Les travaux de Pearl (1988) sur la notion d'indépendance conditionnelle, qui reflète la non-pertinence, sont à la base des réseaux de croyance aujourd'hui largement utilisés en intelligence artificielle. La construction de ces réseaux est fortement apparentée à la sélection d'attributs. Koller et Sahami (1996) ont d'ailleurs repris les idées de Pearl pour construire un algorithme de sélection d'attributs. Un attribut X sera considéré comme non pertinent relativement à un attribut Y dans un contexte Z, si et seulement si Y est indépendant de X conditionnellement à Z. Le contexte Z correspond ici à un ensemble d'attributs ne contenant ni X ni Y. Dans le cadre probabiliste utilisé par Pearl, cela nous amène à la première définition de la pertinence. Nous noterons $r(X, Y, Z)$ la relation qui exprime la pertinence de X relativement à Y dans le contexte Z et la négation de cette relation sera notée $\overline{r}(X, Y, Z)$.

Définition 1

$$\overline{r}(X, Y, Z) \Leftrightarrow P(Y|X, Z) = P(Y|Z)$$

Autrement dit la connaissance de X n'apporte, par rapport à Z, aucune information supplémentaire sur Y.

Dans le domaine de la sélection d'attributs en classification supervisée, de nombreuses définitions ont été posées. Molina *et al.* (2002) en offrent, à notre connaissance, la synthèse la plus récente et la plus complète. Nous ne mentionnerons que celles qui nous semblent utiles pour expliciter le fonctionnement des diverses techniques de sélection que nous présenterons à la section 7.3. Aussi invitons-nous le lecteur intéressé à se référer à cette synthèse. Comme nous centrons désormais notre propos sur la classification supervisée, nous considérerons toujours la pertinence au regard de cette tâche de classification. Nous nous permettrons par conséquent de parler d'attributs pertinents, sans préciser à chaque fois qu'il s'agit d'attributs pertinents relativement à la classe y que l'on cherche à modéliser.

Kohavi et John (1997) proposent d'affiner la classification binaire des attributs (pertinents,non pertinents) en segmentant la catégorie des attributs pertinents. Pour eux, il faut distinguer les attributs fortement pertinents des attributs faiblement pertinents. Les

[1] Les travaux en logique de révision de croyance et autres logiques non monotones qui ont joué un rôle majeur en intelligence artificielle dans les années 70 se sont également beaucoup intéressés à cette notion de pertinence (Ekbia et Maguitman, 2001; Delgrande et Pelletier, 1998).

premiers sont indispensables et doivent être intégrés dans tout modèle de y. Les seconds apportent de l'information sur la classe y dans certains contextes. En fonction du modèle que l'on souhaite construire, certains d'entre eux peuvent être importants. Nous noterons les deux relations correspondantes r_{forte} et r_{faible}. Pour tout ensemble $\mathcal{W} \subseteq \mathcal{V}$ de m attributs, nous noterons W la variable jointe des variables appartenant à cet ensemble. Ainsi pour $\mathcal{V} = \{v_1, ..., v_p\}$, nous avons $V = (v_1, ..., v_p)$, V désignant une variable aléatoire. Enfin $\mathcal{S}_i$ désignera l'ensemble $\mathcal{V} - \{v_i\}$ Avec ces notations les définitions de Kohavi et John s'expriment de la façon suivante :

Définition 2

$$r_{forte}(v_i, y) \Leftrightarrow \exists v_{ji}, s_{ji}, y_j \text{ pour lesquels } P(v_i = v_{ji}, S_i = s_{ji}) > 0, \text{ tels que } P(y = y_j | v_i = v_{ji}, S_i = s_{ji}) \neq P(y = y_j | S_i = s_{ji})$$

Définition 3

$$r_{faible}(v_i, y) \Leftrightarrow \overline{r_{forte}}(v_i, y), \exists \mathcal{S}'_i \subset \mathcal{S}_i, v_{ji}, s'_{ji}, y_j \text{ pour lesquels } P\left(v_i = v_{ji}, S'_i = s'_{ji}\right) > 0 \text{ tels que } P\left(y = y_j | v_i = v_{ji}, S'_i = s'_{ji}\right) \neq P\left(y = y_j | S'_i = s'_{ji}\right)$$

Contrairement à ce que nous avions fait précédemment, le contexte dans lequel la pertinence est considérée n'a pas été spécifié au niveau des relations r_{forte} et r_{faible}. Ces deux définitions ne sont que des applications de la notion d'indépendance conditionnelle pour lesquelles deux contextes différents sont envisagés. Pour faire apparaître plus clairement ce contexte, nous pouvons réécrire ces deux relations de la façon suivante :

$$\begin{array}{rcl} r_{forte}(v_i, y) & \Leftrightarrow & r(v_i, y, \mathcal{S}_i) \\ r_{faible}(v_i, y) & \Leftrightarrow & \overline{r_{forte}}(v_i, y) \text{ et } \exists \mathcal{S}'_i \subset \mathcal{S}_i \text{ tel que } r(v_i, y, \mathcal{S}'_i) \end{array}$$

Cette distinction entre attributs fortement et faiblement pertinents a joué un rôle important dans le développement de nouvelles techniques de sélection d'attributs. Blum et Langley (1997) en ont proposé une reformulation, à laquelle peuvent être rattachés de nombreux travaux du domaine. Les nouvelles définitions qui en découlent sont apparentées aux mesures de cohérence[2] développées par Almuallim et Dietterich (1994). L'idée sous-jacente est que l'on peut éviter d'estimer directement les probabilités conditionnelles. Il suffit de considérer les valeurs des différents attributs pour les exemples de $\mathcal{E}$.

Si deux exemples de classes différentes ne diffèrent que par la valeur de l'attribut v_i, alors cet attribut contient une information importante sur la classe. Ceci n'est évidemment valable que pour des attributs discrets. Dans ce cas-là nous sommes sûr que les probabilités conditionnelles $P(y|v_i, S_i)$ et $P(y|S_i)$ diffèrent. Les définitions que nous donnons ci-après sont donc des restrictions des deux définitions précédentes. Blum et Langley parlent de pertinence par rapport à l'échantillon. Leurs définitions dépendent en effet de l'échantillon $\mathcal{E}$ d'exemples disponibles.

Définition 4

$$r^{\mathcal{E}}_{forte}(v_i, y) \Leftrightarrow \exists e_k, e_l \in \mathcal{E} \text{ tels que } \begin{cases} v_{kj} = v_{lj} \forall v_j \in \mathcal{S}_i \\ v_{ki} \neq v_{li} \\ y_k \neq y_l \end{cases}$$

[2] Les anglo-saxons parlent de *consistency measures*.

Définition 5

$$r^{\mathcal{E}}_{faible}(v_i, y) \Leftrightarrow \overline{r_{forte}}(v_i, y) \text{ et } \exists \mathcal{S}'_i \subset \mathcal{S}_i, e_k, e_l \in \mathcal{E} \text{ tels que } \begin{cases} v_{kj} = v_{lj} \forall v_j \in \mathcal{S}'_i \\ v_{ki} \neq v_{li} \\ y_k \neq y_l \end{cases}$$

Jusqu'ici nous avons considéré la pertinence d'un attribut relativement à la classe. Ce qui nous importe *in fine* ce sont les performances de la phase d'apprentissage pour laquelle la sélection d'attributs n'est qu'un prétraitement. Aussi est-il tentant d'intégrer directement l'algorithme d'apprentissage dans une définition de la pertinence.

Les techniques nommées traditionnellement *wrappers* reposent sur cette idée (Kohavi et John, 1997), formalisée par la notion d'utilité incrémentale introduite par Caruana et Freitag (1994). Leur formalisation peut être considérée comme une application des idées de Gärdenfors dans laquelle le contexte correspond à un ensemble d'attributs $\mathcal{W}$. L'hypothèse qui nous intéresse est le modèle résultant de l'apprentissage. La variation de la vraisemblance de cette hypothèse est mesurée comme la différence entre les performances du modèle appris à partir de $\mathcal{W}$ et celles du modèle appris à partir de $\mathcal{W}$ et de la variable dont on cherche à évaluer la pertinence. En reprenant la catégorisation de John et Kohavi des attributs faiblement et fortement pertinents, nous proposons de définir les notions de pertinence forte et faible relativement à un algorithme d'apprentissage A et un critère de performance J à maximiser. Le modèle $A(\mathcal{W})$ construit par A à partir d'un ensemble de variables $\mathcal{W}$ dépend également de l'ensemble d'exemples $\mathcal{E}$, mais nous simplifierons les notations en supprimant la référence à cet ensemble.

Définition 6

$$r^{A,J}_{forte}(v_i, y) \Leftrightarrow J(A(\{v_i\} \cup \mathcal{S}_i)) > J(A(\mathcal{S}_i))$$

Définition 7

$$r^{A,J}_{faible}(v_i, y) \Leftrightarrow \overline{r_{forte}}(v_i, y) \text{ et } \exists \mathcal{S}'_i \subset \mathcal{S}_i \text{ tel que } J\left(A\left(\{v_i\} \cup \mathcal{S}'_i\right)\right) > J\left(A\left(\mathcal{S}'_i\right)\right)$$

Outre les probabilités, toute mesure d'incertitude peut théoriquement être utilisée pour évaluer la vraisemblance d'une hypothèse. Il est également possible d'employer une mesure entropique issue de la théorie de l'information (Shannon, 1948). L'entropie d'une variable peut en effet être interprétée comme une mesure d'incertitude. Si les applications de l'entropie, notée I, à la sélection d'attributs sont nombreuses comme nous le verrons à la section 7.3, ce sont les travaux de Wang et Bell (1999); Bell et Wang (2000) qui ont le mieux formalisé cette approche et mis en évidence l'intérêt de ses propriétés pour la sélection d'attributs.

L'entropie conditionnnelle est utilisée pour prendre en compte le contexte dont nous avons souligné l'importance. Ce qui importe étant la variation de la vraisemblance prise en contexte, la pertinence sera définie *via* le gain d'information sur la variable cible y apporté par la connaissance supplémentaire d'un attribut v_i par rapport à un contexte donné. De même que nous avons présenté l'utilité incrémentale au travers des notions de pertinence faible et forte, nous allons spécialiser la définition de Bell et Wang pour définir la pertinence entropique forte et faible.

Définition 8

$$r^{ent}_{forte}(v_i, y) \Leftrightarrow \frac{IM(v_i, y|S_i)}{I(y|S_i)} > 0$$

IM désigne l'information mutuelle, qui est équivalente au gain d'information.

Définition 9

$$r_{faible}^{ent}(v_i, y) \Leftrightarrow \overline{r_{forte}}(v_i, y) \text{ et } \exists \mathcal{S}_i' \subset \mathcal{S}_i \text{ tel que } \frac{IM(v_i, y|S_i')}{I(y|S_i')} > 0$$

L'intérêt de cette approche ainsi que de celle Caruana et Freitag est de fournir non seulement un critère permettant de juger si un attribut est pertinent ou non, mais également une mesure quantitative de cette pertinence.

Au travers des définitions précédentes nous avons vu qu'il était possible de distinguer des attributs fortement pertinents, faiblement pertinents et non pertinents. Les études empiriques sur des bases de données artificielles, citées précédemment, ainsi que des exemples donnés par Guyon et Elisseeff (2003), ont mis en avant les effets néfastes sur les performances d'un classifieur que pouvaient causer des attributs non pertinents, mais aussi des attributs qu'ils qualifient de redondants. Diverses définitions de la redondance ont été données dans la littérature, souvent basées sur le concept de corrélation entre attributs. Nous aurons l'occasion d'y revenir à la section 7.3. Nous ne donnons ici que la définition due à Yu et Liu (2004) qui nous servira pour la suite. Celle-ci s'appuie sur le concept de couverture de Markov introduit par Koller et Sahami (1996).

Définition 10 Soit $M_i \subset \mathcal{V}, v_i \notin M_i$. M_i forme une couverture de Markov pour l'attribut v_i si et seulement si :

$$P(\mathcal{V} - M_i - \{v_i\}, y|v_i, M_i) = P(\mathcal{V} - M_i - \{v_i\}, y|M_i)$$

Une couverture de Markov pour v_i regroupe donc un ensemble d'attributs qui apporte une information sur y mais également sur l'ensemble des autres attributs de $\mathcal{V}$, qui subsume celle qui est apportée par v_i.

La notion de redondance peut alors être définie de la façon suivante :

Définition 11 v_i est redondant (sous-entendu par rapport à l'ensemble d'attributs $\mathcal{V}$ et au regard de la tâche de classification), si et seulement si $r_{faible}(v_i, y)$ et $\exists M_i \subset \mathcal{V}$ tel que M_i forme une couverture de Markov pour v_i.

Définie ainsi, la notion de redondance permet donc d'affiner encore la catégorisation des attributs en spécialisant la catégorie des attributs faiblement pertinents. Afin de synthétiser les différents points de vue sur la notion de pertinence, nous reproduisons sur la figure 7.1 le schéma de Yu et Liu (2004) qui offre une représentation de cette catégorisation.

7.2.2 Formalisations de la sélection d'attributs

Au vu de la catégorisation des attributs que nous venons de présenter, il est assez naturel de présenter la sélection d'attributs comme la recherche des attributs fortement pertinents et de ceux qui sont faiblement pertinents mais non redondants. Cette conception du problème, pour importante qu'elle soit[3], ne fait pas apparaître de manière explicite les objectifs que l'on cherche à atteindre en procédant à la sélection d'attributs. Ces objectifs sont implicitement intégrés dans les choix des mesures de pertinence et de redondance.

[3] Nous verrons à la section 7.3 que de nombreuses techniques s'en inspirent.

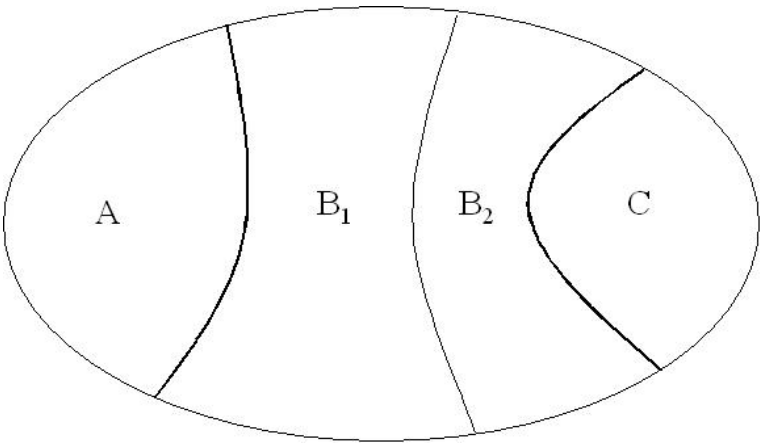

FIG. 7.1 – Une catégorisation possible des attributs : A={attributs non pertinents}, B_1={attributs faiblement pertinents et redondants}, B_2={attributs faiblement pertinents et non redondants}, C={attributs fortement pertinents}

7.2.2.1 Problème d'optimisation

Différents objectifs sont envisageables en fonction du domaine d'application considéré. Sans perte de généralité, nous pouvons supposer que les objectifs sont résumés par un critère J qu'il convient de maximiser. J est une application qui associe à tout ensemble d'attributs un score. Suivant les connaissances du domaine que l'on peut avoir, différentes formalisations sont envisageables.

- On fixe $d < p$ le nombre d'attributs à sélectionner. Trouver le sous-ensemble d'attributs $\mathcal{V}_{opt}$ de cardinal d qui maximise J.

$$\mathcal{V}_{opt} = \arg \max_{\mathcal{W} \subseteq \mathcal{V}, |\mathcal{W}|=d} J(\mathcal{W})$$

- On fixe J_o le seuil de performance acceptable. Trouver le sous-ensemble $\mathcal{V}_{opt}$ de cardinalité minimale dont la performance est supérieure à J_o.

$$\mathcal{V}_{opt} = \arg \min_{\mathcal{W} \subset \mathcal{V}, J(\mathcal{W}) \geq J_o} |\mathcal{W}|$$

- Les définitions précédentes imposent de fixer certains paramètres d ou J_o, et peuvent paraître restrictives. Une version plus générale est celle que nous avons adoptée à la section 3.1 lorsque nous avons mis en place un algorithme génétique pour sélectionner les attributs les plus pertinents sur la base des conflits intra-étatiques. Il s'agit simplement de trouver le sous-ensemble maximisant le critère J, sans imposer aucune contrainte sur la cardinalité du sous-ensemble en question.

$$\mathcal{V}_{opt} = \arg \max_{\mathcal{W} \subset \mathcal{V}} J(\mathcal{W})$$

  Notons que le maximum peut être atteint pour plusieurs sous-ensembles. Si tel est le cas, sera retenu celui de cardinalité minimale.

7.2.2.2 Préservation de caractéristiques essentielles pour la classification

Ces formalisations sont les plus couramment admises, mais certains auteurs envisagent le problème sous l'angle de la préservation de certaines caractéristiques qui peuvent être déteriorées par la suppression d'attributs. Bell et Wang (2000) s'intéressent à la quantité d'information portée sur la classe tandis que Koller et Sahami (1996) considèrent la distribution de la variable cible. On a ainsi les deux autres formalisations possibles suivantes.

- Trouver le sous-ensemble de plus faible entropie qui permet de préserver l'information utile pour la tâche de classification. Soit $\mathcal{X}$ l'ensemble des ensembles d'attributs qui préservent le gain d'information sur la classe y. On a

$$\mathcal{X} = \{\mathcal{W} \subset \mathcal{V}, IM(W, y) = IM(V, y)\}$$

  La tâche de sélection d'attributs consiste alors à trouver le sous-ensemble $\mathcal{V}_{opt}$ défini de la façon suivante :

$$\mathcal{V}_{opt} = \arg\min_{\mathcal{W} \in \mathcal{X}} I(W)$$

- Une modélisation bayésienne de la classification supervisée consiste à affecter à tout nouvel exemple la classe la plus vraisemblable au vu des données à disposition, c'est-à-dire celle qui maximise la probabilité conditionnelle de la classe sachant les données $P(y|V = v)$. Étant donné l'importance de cette distribution pour la classification, Koller et Sahami proposent de trouver le sous-ensemble d'attributs qui permette de la préserver au mieux. Si l'on note Δ une mesure de distance entre deux distributions de probabilités, l'objectif est alors de trouver un sous-ensemble $\mathcal{V}_{opt}$, de cardinalité minimale, tel que $\Delta(P(y|V), P(y|V_{opt}))$ soit suffisamment faible.
  Koller et Sahami ont proposé de définir Δ à partir de la divergence de Küllback-Leibler. Si l'on note v le vecteur des valeurs des variables de V pour un exemple e et $v|_{V_{opt}}$ la projection de v sur $\mathcal{V}_{opt}$ on a :

$$\Delta(P(y|V), P(y|V_{opt})) = \sum_{v} P(v) \sum_{i=1}^{K} P(y = c_i|V = v) \log_2 \left(\frac{P(y = c_i|V = v)}{P\left(y = c_i|V_{opt} = v|_{V_{opt}}\right)} \right)$$

7.2.2.3 Pondération d'attributs

Fortement apparentée à la sélection d'attributs, la *pondération d'attributs* consiste à affecter un poids à chaque variable, qui reflète l'importance de celle-ci vis-à-vis des autres variables. Il s'agit d'une généralisation de la sélection d'attributs, puisqu'il suffit de considérer des poids binaires (1 si la variable est sélectionnée et 0 sinon) pour se ramener à une tâche de sélection. Cette approche par pondération est essentielle pour l'ordonnancement des attributs ou pour l'utilisation de certains algorithmes d'apprentissage comme les plus proches voisins. Raymer *et al.* (2000) proposent par exemple d'apprendre ces poids *via* un algorithme génétique et de les réutiliser dans une moyenne pondérée pour effectuer le calcul des distances entre exemples. L'intérêt de la pondération d'attributs pour les plus proches voisins a été également mis en évidence par Kohavi *et al.* (1997). Les techniques de régression ou les réseaux de neurones opèrent intrinsèquement une recherche de la meilleure pondération possible.

Précisons cependant que la pondération d'attributs ne permet pas de réduire les coûts d'acquisition et de stockage de l'information. Il faut toujours autant d'attributs en entrée. Les modèles ne sont d'ailleurs pas plus simples, et l'apprentissage n'est pas non plus rapide puisqu'il y a tout autant d'attributs à prendre en compte. Aussi ne nous intéresserons-nous pas spécifiquement à ce domaine. Notons cependant que tout algorithme de pondération

peut facilement s'appliquer à la sélection d'attributs : il suffit de fixer un seuil sur les pondérations. Nous verrons quelques exemples de tels algorithmes à la section 7.3.

7.2.2.4 Extraction d'attributs

La sélection de variables telle que nous l'avons posée jusqu'à présent consiste à rechercher un sous-ensemble d'attributs qui optimise un certain critère sous certaines contraintes. Il s'agit d'un cas particulier de techniques de réduction des dimensions. D'autres techniques permettent également de réduire la dimension de l'espace des variables *via* la construction de nouvelles variables qui condensent l'information initiale.

Cette tâche, à laquelle il est souvent fait référence sous le vocable d'*extraction d'attributs*, est plus générale que la *sélection d'attributs*. Dans les deux cas il s'agit de trouver une transformation de l'espace d'entrée $\mathcal{V}$ vers un nouvel espace $\mathcal{V}_{opt}$. Alors que la sélection d'attributs n'envisage en guise de transformation que des projections, aucune contrainte n'est imposée dans le cas de l'extraction d'attributs. L'analyse en composantes principales (ACP), qui est probablement la technique d'extraction la plus répandue (Pechenizkiy *et al.*, 2003), opère une transformation linéaire de l'espace initial, la nouvelle base de l'espace réduit étant constitué des vecteurs propres de la matrice de variance-covariance des données (Saporta, 2006). D'autres techniques, utilisant la programmation génétique, construisent de nouvelles variables par application d'opérateurs prédéfinis entre les variables initiales. Envisager des transformations non linéaires de l'espace de départ est alors assez simple. Il suffit en effet d'intégrer des opérateurs non linéaires dans la liste de ceux qui sont considérés (Raymer *et al.*, 1996; Sherrah *et al.*, 1997; Smith et Bull, 2005; Guo *et al.*, 2005).

Contrairement à l'espace d'arrivée obtenu par sélection d'attributs, celui que l'on obtient par extraction d'attributs repose sur de nouvelles variables, combinaisons des variables de départ. Ces nouvelles variables, si elles permettent une amélioration des performances du modèle, n'en assurent plus en revanche l'interprétabilité. Or il s'agit là d'un point essentiel dans notre méthodologie d'évaluation des risques. Pour cette raison, nous écartons ce domaine de notre champ d'investigation pour nous concentrer sur la sélection d'attributs.

7.3 État de l'art sur les techniques de sélection d'attributs

Des différentes formalisations que nous venons de présenter, il ressort que la sélection d'attributs est avant tout une tâche d'optimisation sous contraintes. Il s'agit de rechercher le sous-ensemble d'attributs qui répond le mieux aux objectifs que l'on se fixe. L'espace de recherche à parcourir est un treillis. Si l'on représente chaque sous-ensemble d'attributs par un vecteur de dimension p (dimension d'origine), dans lequel la présence d'un attribut est signifié par un 1 et son absence par un 0, la figure 7.2 offre une schématisation de ce treillis. Par la suite, toute mention du « treillis », ou du « treillis de recherche », fera systématiquement référence à ce treillis.

Trouver le sous-ensemble qui répond exactement aux objectifs nécessiterait le parcours et l'évaluation de l'ensemble des 2^p sous-ensembles d'attributs possibles. Cette complexité exponentielle est vite prohibitive, même pour un nombre limité d'attributs. Hyafil et Rivest (1976); Blum et Rivest (1992) ont par ailleurs montré que ce problème d'optimisation était NP-difficile. En pratique il convient donc d'utiliser des heuristiques pour parcourir l'espace de recherche avec une complexité limitée.

Nous nous intéresserons dans un premier temps aux techniques qui se placent exactement dans le cadre de la recherche du sous-ensemble optimal au sein du treillis de recherche. Nous verrons ensuite comment d'autres formalisations ont été mises à profit pour réduire la

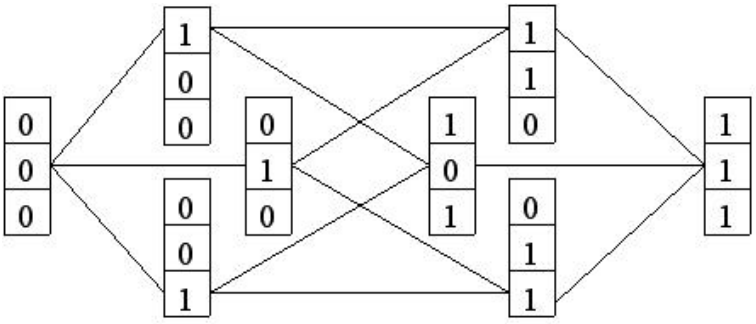

FIG. 7.2 – Espace de recherche pour la sélection de trois attributs : treillis représentant l'ensemble des sous-ensembles d'attributs. Chaque sous-ensemble est représenté par un vecteur contenant un 1 si l'attribut correspondant est sélectionné et un 0 sinon.

complexité de la tâche de sélection : la recherche de la pondération optimale et la recherche des attributs soit fortement pertinents soit faiblement pertinents et non redondants. Les deux types de techniques concernées peuvent très bien être décrits sous le formalisme générique de l'optimisation (Molina *et al.*, 2002; Liu et Yu, 2005), mais ils correspondent conceptuellement à des approches différentes du problème. C'est la raison pour laquelle nous avons choisi de scinder en trois grandes catégories les différentes méthodes. Cette distinction peut de plus se révéler fertile pour ce qui est de l'élaboration de nouveaux algorithmes de sélection d'attributs.

Nous nous concentrerons sur les principales familles de méthodes. Nous mettrons en évidence leurs principales caractéristiques de façon à identifier les critères de discrimination entre ces familles. Ceci nous permettra de construire une taxinomie qui nous donnera une vue d'ensemble du domaine. C'est dans cette optique que nous synthétisons sur la figure 7.3 les remarques précédentes sur les différentes approches conceptuelles de la sélection d'attributs.

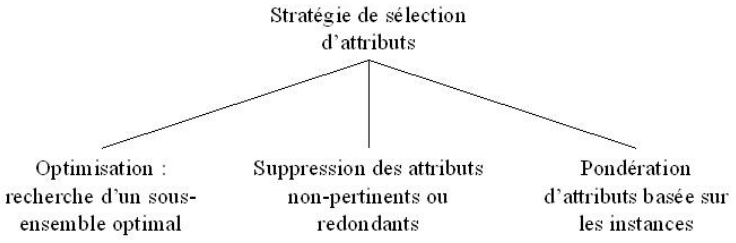

FIG. 7.3 – Une ébauche de taxinomie des méthodes de sélection d'attributs

Contrairement au traitement des données manquantes, pour lequel les caractérisations des différentes méthodes par le biais d'une taxinomie sont plutôt rares, voire inexistantes, divers auteurs ont proposé de telles caractérisations pour la sélection d'attributs. Un consensus semble avoir été trouvé car les différentes taxinomies existantes sont assez proches (Dash et Liu, 1997; Liu et Yu, 2005; Molina *et al.*, 2002). Elles reprennent et raffinent des catégorisations déjà proposées, par Blum et Langley (1997) par exemple. Nous allons maintenant raffiner cette ébauche de taxinomie en approfondissant chacune des trois branches de la figure 7.3.

7.3.1 Optimisation : recherche du sous-ensemble optimal

La sélection d'attributs peut donc s'interpréter comme la recherche parmi les sous-ensembles d'attributs de celui qui est optimal, relativement à un certain critère de performance. L'optimum n'étant pas forcément unique, il serait plus juste de dire que l'on en cherche un parmi ceux qui sont optimaux. Deux caractéristiques fondamentales se dégagent : l'organisation de la recherche et la fonction d'évaluation. La première spécifie la façon dont sera parcouru l'espace de recherche tandis que la seconde indique à quel point un sous-ensemble correspond à la solution souhaitée.

7.3.1.1 Organisation de la recherche

Nous avons vu que la recherche du sous-ensemble d'attributs était un problème NP-difficile et qu'une recherche exhaustive nécessitait le parcours de l'ensemble des 2^p états du treillis de recherche, ce qui est inenvisageable pour des applications réelles. Le choix de l'organisation de la recherche, qui correspond à la façon dont ce treillis sera parcouru, est donc une étape cruciale pour la construction de tout algorithme de sélection d'attributs. Commençons par identifier les critères sur lesquels des choix s'imposent, et qui une fois déterminés, permettent de définir complètement une organisation de recherche.

Dans un premier temps, il convient de fixer la stratégie de recherche que l'on souhaite adopter. Liu et Yu (2005) et Molina *et al.* (2002) distinguent trois grandes familles, sans pour autant s'accorder sur les dénominations de ces familles. Aussi expliciterons-nous pour chaque famille le choix de la terminologie employée.

- **Recherche optimale** : il s'agit de garantir l'optimalité de la solution trouvée. Si la recherche exhaustive en est un exemple, jamais utilisée en pratique, d'autres méthodes ont été proposées pour réduire la complexité de cette recherche. Par exemple l'algorithme FOCUS de Almuallim et Dietterich (1994) effectue un parcours en largeur du treillis à partir du sous-ensemble vide et s'arrête dès qu'il trouve un sous-ensemble cohérent. Le critère J d'évaluation d'un sous-ensemble correspond à une mesure booléenne de cohérence. L'optimalité est garantie du fait de la monotonie de ce critère. Nous y reviendrons plus en détail ultérieurement.
  L'algorithme *Branch & Bound* (B&B) de Narendra et Fukunaga (1977) permet également de trouver le sous-ensemble optimal pour un critère J. S'il n'est pas nécessaire que J corresponde au critère de cohérence de FOCUS, l'optimalité de la solution n'est garantie que pour des mesures J monotones par rapport à l'inclusion. Contrairement à FOCUS, B&B part de l'ensemble initial d'attributs (l'autre extrémité du treillis) et en supprime au fur et à mesure. Un seuil de performance minimale est fixé (*bound*) qui permet de réduire l'espace de recherche. Pour chaque état dont les performances sont inférieures à ce seuil, du fait de la monotonie de J, nous pouvons supprimer toutes les branches du treillis qui mènent à cet état.
  Comme nous avons pu le voir au travers de ces deux exemples, les stratégies optimales de recherche, quoique non exhaustives grâce l'utilisation astucieuse des propriétés de J, sont tout de même très coûteuses. Des variantes de complexité moindre ont été développées, aussi bien pour FOCUS que pour B&B. Ainsi AB&B, extension de B&B, réajuste automatiquement le seuil de performance minimum au cours de la recherche (Liu et Yu, 2002). Leur complexité reste exponentielle, de l'ordre de $\mathcal{O}(2^p)$. Aussi sont-elles parfois appelées stratégies « exponentielles » (Molina *et al.*, 2002). Liu et Yu (2005) préfèrent quant à eux le qualificatif « complètes ». Nous préférons cependant insister sur l'objectif qui sous-tend ce type de stratégies plutôt que sur ses propriétés.

- **Recherche sous-optimale déterministe** : pour réduire la complexité de la recherche, il est possible de ne chercher qu'un optimum local. Comme l'indiquent Liu et Yu (2005), les techniques adoptant cette stratégie sont généralement de complexité quadratique et utilisent des heuristiques pour guider un parcours en profondeur du treillis. C'est pour cette raison que ces auteurs qualifient d'« heuristiques » ces stratégies. Mais les méthodes stochastiques utilisent elles aussi diverses heuristiques. Ce n'est donc pas là un critère discriminant. Molina *et al.* (2002) préfèrent l'adjectif « séquentiel ». Si l'aspect séquentiel de la recherche opérée par de telles méthodes est effectivement discriminant, il est cependant trop restrictif puisqu'il exclut toutes les autres méthodes déterministes qui ne garantissent pas l'obtention d'un optimum global.
- **Recherche non déterministe** : cette famille rassemble les techniques parcourant le treillis de recherche de manière aléatoire. Molina *et al.* (2002) utilisent d'ailleurs ce terme pour qualifier ce type de recherche. Sans rentrer dans des débats philosophiques ou encore linguistiques, nous parlerons indifféremment de recherche non déterministe, aléatoire ou même parfois stochastique, simplement pour signifier qu'à partir d'un même sous-ensemble d'attributs de départ, le même algorithme de recherche, exécuté plusieurs fois, ne parcourra pas forcément les mêmes états. Le caractère déterministe de la recherche étant indépendant de l'objectif poursuivi, nous devrions donc distinguer quatre familles de stratégies et non trois. Nous écartons cependant les recherches aléatoires optimales qui n'ont pas d'intérêt. Le seul moyen de garantir que la solution trouvée de manière aléatoire est un optimum global consisterait à effectuer une recherche exhaustive. Peu importe alors de savoir dans quel ordre a été évalué chacun des états du treillis. Il n'y aura donc pas d'ambiguïté à parler de recherche stochastique, sans préciser que cette recherche est sous-optimale.

Ces différentes stratégies ne suffisent pas à caractériser complètement la façon dont est parcouru le treillis. Au vu des remarques précédentes, il apparaît que la recherche d'un sous-ensemble dans un treillis se caractérise par ces trois choix :

- le choix de l'état à partir duquel démarre la recherche,
- le choix de la fonction qui permet de passer d'un état à un autre,
- le choix d'un critère permettant de mettre fin à la recherche.

Le schéma de la figure 7.4 donne une vision d'ensemble de la sélection d'attributs prise comme recherche d'un sous-ensemble optimal au sein d'un treillis, inspirée de celle que Dash et Liu (1997) ont proposée.

L'initialisation de la recherche consiste à fixer l'état à partir duquel elle débutera. Les états aux extrémités du treillis sont souvent utilisés comme points de départ comme nous l'avons vu pour FOCUS et B&B. Si l'on part de l'ensemble vide, cela signifie que l'on privilégie la sélection des meilleurs attributs pour la suite de la recherche. Au contraire, une recherche initiée à partir de l'ensemble complet des attributs consistera à supprimer les attributs les moins prometteurs. Il y a donc un fort lien entre ce choix et celui de la méthode de génération des états successeurs.

Si ces deux états extrêmes correspondent à des choix très répandus, ce ne sont pas les seuls. Il est possible de commencer par n'importe quel état du treillis. Si nous disposons de connaissances particulières sur le domaine, que ce soit par expertise ou grâce à des expériences préalables, il est fortement conseillé de choisir un sous-ensemble d'attributs que l'on sait ou que l'on espère performant. Ceci permet d'accélerer la recherche. Dans l'optique d'une recherche stochastique, il est fréquent de tirer aléatoirement l'état de départ, qui ne

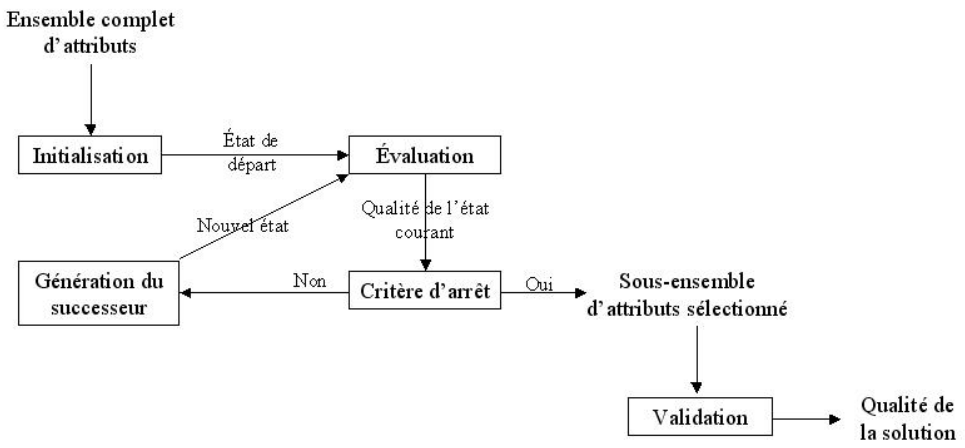

FIG. 7.4 – La sélection d'attributs : recherche dans l'espace des sous-ensembles d'attributs

sera donc vraisemblablement pas non plus l'un des états extrêmes du treillis.

Définir un critère d'arrêt est essentiel, sans quoi la recherche sera exhaustive ou sans fin suivant la méthode de génération des états successeurs. Le choix de ce critère dépend des objectifs que doit remplir la sélection d'attributs. Si l'on note $\mathcal{W}$ le sous-ensemble courant d'attributs sélectionnés, les critères d'arrêt les plus fréquents sont les suivants :

- Le nombre d'attributs sélectionnés correspond au cardinal d du sous-ensemble recherché : $|\mathcal{W}| = d$.
- Les performances du sous-ensemble courant sont suffisamment bonnes : $J(\mathcal{W}) \geq J_0$.
- Les performances du sous-ensemble courant sont moins bonnes que celles du sous-ensemble précédemment sélectionné : $J(\mathcal{W}) < J(\mathcal{W}_{prec})$.
- Le nombre d'états évalués a atteint une certaine limite fixée au départ.
- Le nombre d'états évalués sans que les performances ne s'améliorent a atteint une limite fixée au départ.

L'initialisation et le critère d'arrêt définissent les bornes de la recherche, mais le cœur de l'organisation de cette recherche est avant tout caractérisé par la façon dont seront générés les états que l'on a à évaluer, ce que nous avons appelé méthode de génération des successeurs. C'est essentiellement cette méthode qui permet de discriminer entre les différents algorithmes de recherche. Elle guide le choix de l'état de départ[4].

Recherche sous-optimale déterministe Les méthodes les plus simples et les plus répandues sont les techniques de recherche séquentielles (Aha et Bankert, 1996; Ng, 1998). Hall et Smith (1997) les utilisent par exemple dans l'algorithme CFS, réputé efficace et popularisé du fait de sa disponibilité dans la plate-forme de développement d'applications de fouille de données Weka (Witten et Frank, 2005). Selon la direction de la recherche on parlera de recherche séquentielle avant ou arrière, que nous désignerons par leurs sigles anglo-saxons : SFS (*Sequential Forward Search*) pour la recherche avant et SBS (*Sequential Backward Search*) pour la recherche arrière. SFS consiste à partir d'un ensemble vide d'attributs et à ajouter les attributs un à un, en prenant à chaque étape celui qui permet de maximiser J. Ce principe est illustré par l'algorithme 1. A désigne un critère d'arrêt.

[4]La réciproque est également vrai : le choix de l'état de départ restreint le choix de la méthode de génération des successeurs. En pratique, sauf cas exceptionnels, aucun des deux ne prévaut sur l'autre : ces deux choix sont concomitants.

De manière symétrique, SBS part de l'ensemble complet d'attributs, et les supprime un à un. À chaque étape est supprimé celui dont l'absence permet de maximiser J. La génération des successeurs qui est effectuée à chaque itération ne requiert que l'application d'un seul opérateur de base.

Algorithme 1 Algorithme de recherche séquentielle avant : SFS

Entrées: $J, A, \mathcal{V} = \{v_1, ..., v_p\}$
Sorties: $\mathcal{V}_{opt}$
Début
$\mathcal{V}_{opt} = \emptyset$
Tant que $\overline{A}$ **Faire**
$j = \arg\max\left(J\left(\mathcal{V}_{opt} \cup \{v_i\}\right), \forall v_i \notin \mathcal{V}_{opt}\right)$
$\mathcal{V}_{opt} = \mathcal{V}_{opt} \cup \{v_j\}$
Fin Tant que
Renvoyer $\mathcal{V}_{opt}$
Fin

La complexité au regard du nombre d'attributs, de SFS et SBS est théoriquement la même : $\mathcal{O}\left(p^2\right)$, puisqu'à l'étape k on calcule les performances de $p-k$ sous-ensembles (k attributs ayant déjà été sélectionnés ou supprimés) et que l'on a au maximum p itérations. Mais en pratique, les premières itérations de SBS sont très coûteuses. Il est en effet légitime de penser, sans que cela soit une hypothèse complètement irréaliste, que le coût de l'évaluation d'un sous-ensemble augmente avec sa cardinalité. Lorsque le nombre d'attributs attendu est plutôt faible, SFS s'avère nettement moins coûteuse. En revanche la relation s'inverse lorsque l'on s'attend à ce que peu d'attributs soient supprimés.

Notons cependant que la recherche arrière a un atout non négligeable. Contrairement à la recherche avant, elle permet d'évaluer l'influence de chaque attribut sur la classe en présence de l'ensemble des autres attributs. Le contexte, dont nous avons souligné l'importance précédemment, est donc bien mieux pris en compte avec SBS qu'avec SFS. Pour cette raison SBS est réputée plus performante, même si l'étude de Aha et Bankert (1996) tend à relativiser ce point.

Un des inconvénients des méthodes séquentielles simples, qui justifie cependant leur complexité modérée, est qu'elles ne remettent jamais en question les choix qu'elles ont effectués aux étapes précédentes. On les qualifie pour cette raison de méthodes gloutonnes. Ce comportement conduit vite à des optima locaux. Un attribut sélectionné (resp. supprimé pour SBS) à un moment donné ne pourra jamais être exclu (resp. sélectionné) par la suite alors que le contexte dans lequel a été évalué sa pertinence a changé[5].

Pour contourner cette difficulté, plusieurs approches sont envisageables, dont la plus simple consiste à considérer à chaque étape l'ensemble des états du treillis qui sont liés à l'état courant et pas seulement ceux qui contiennent un attribut de plus (ou de moins pour SBS). Pour chaque sous-ensemble on s'autorise donc à chercher dans les deux directions. Pour ne pas créer de cycles dans la recherche, il faut tenir à jour une liste d'états visités dont on empêche la réévaluation. Ceci correspond à une version plus générale de la méthode d'escalade, dite *hill climbing* dont SFS et SBS sont des cas particuliers.

L'hypothèse sous-jacente sur laquelle s'appuient SFS et SBS concerne la monotonie de la fonction d'évaluation : ajouter (resp. supprimer) un attribut à un sous-ensemble ne peut qu'accroître (resp. dégrader) ses performances. Certes la validité de ces méthodes ne repose pas explicitement sur cette hypothèse, mais c'est elle qui se cache derrière l'heuristique

[5] Les anglo-saxons parlent de *nesting effect*.

utilisée pour réduire l'espace de recherche. Or de nombreux critères d'évaluation ne sont pas monotones comme nous le verrons dans la section dédiée aux fonctions d'évaluation.

Pour pallier cette faiblesse Stearns (1976) a proposé de combiner à chaque itération de la recherche une suite de l opérations de recherche avant et une suite de r opérations de recherche arrière. Cette méthode, nommée *plus l-reprendre r* (*plus l-take away r* : PTA), peut s'appliquer pour corriger le biais aussi bien de SFS que de SBS. Pour une recherche arrière, on prendra $r > l$ et on commencera par supprimer des attributs. De manière symétrique, l'extension de la recherche avant consiste à prendre $l > r$ et à commencer par ajouter des attributs. Le problème de cette méthode est que le choix des valeurs l et r n'est pas aisé.

Pudil *et al.* (1994) ont proposé une extension de la méthode PTA dans laquelle il n'est plus besoin de fixer l et r. À chaque itération, c'est le critère d'évaluation J qui guide le choix de ces valeurs. Cette méthode adaptative, aussi appelée recherche séquentielle flottante, se décline elle aussi en recherche avant et arrière auxquelles nous nous référerons par les sigles anglo-saxons SFFS (*Sequential Floating Forward Search*) et SFBS (*Sequential Floating Backward Search*). L'algorithme 2 illustre le principe de SFFS. L'intérêt de cet algorithme est qu'il régule automatiquement le compromis entre recherche avant et arrière.

Algorithme 2 Algorithme de recherche séquentielle flottante avant : SFFS

Entrées: $J, A, \mathcal{V} = \{v_1, ..., v_p\}$
Sorties: $\mathcal{V}_{opt}$
  Début
  $\mathcal{V}_0 = \emptyset$
  $k = 0$
  Tant que $\overline{A}$ **Faire**
    $j = \arg\max\left(J\left(\mathcal{V}_k \cup \{v_i\}\right), \forall v_i \notin \mathcal{V}_k\right)$
    $\mathcal{V}_{k+1} = \mathcal{V}_k \cup \{v_j\}$
    $k = k + 1$
    $j = \arg\max\left(J\left(\mathcal{V}_k - \{v_i\}\right), \forall v_i \in \mathcal{V}_k\right)$
    Tant que $J\left(\mathcal{V}_k - \{v_j\} > J\left(\mathcal{V}_{k-1}\right)\right)$ **Faire**
      $\mathcal{V}_{k-1} = \mathcal{V}_k - \{v_j\}$
      $k = k - 1$
      $j = \arg\max\left(J\left(\mathcal{V}_k - \{v_i\}\right), \forall v_i \in \mathcal{V}_k\right)$
    Fin Tant que
  Fin Tant que
  Renvoyer $\mathcal{V}_{opt} = \mathcal{V}_k$
  Fin

Diverses études comparatives ont mis en évidence l'intérêt de cette technique (Jain et Zongker, 1997; Perner, 2001; Ferri *et al.*, 1994). Cependant plus récemment, une étude de Reunanen (2003) est venue remettre en cause ce constat. Utilisant un protocole clair d'évaluation des méthodes de sélection d'attributs, qui limite les risques de sur-apprentissage, il n'a pas observé de différence particulière entre les performances obtenues grâce à SFFS et grâce à SFS. Le coût de SFFS, nettement supérieur à celui de SFS, ne se justifie donc pas aisément. La nécessité d'adopter un protocole expérimental clair et non biaisé correspond à l'un des points sur lesquels nous avons insisté (voir section 6.6). Aussi y reviendrons-nous lors de notre étude expérimentale à la section 7.6.

L'atout principal de la recherche flottante est de permettre à tout moment la révision des choix effectués antérieurement, ce que les anglo-saxons appellent *back-tracking*.

La méthode dite *best first search* peut être considérée comme une généralisation de la recherche séquentielle simple qui intègre explicitement le *back-tracking*. Elle est également fréquemment employée (John *et al.*, 1994; Kotsiantis et Pintelas, 2004).

Au lieu de ne stocker à chaque itération que l'état qui parmi les successeurs permet de maximiser le critère de performance J, l'ensemble des successeurs est stocké et ordonné dans une pile par ordre décroissant de performance. Le treillis est toujours parcouru en profondeur en considérant à chaque itération le meilleur état successeur, mais lorsque les performances ne s'améliorent plus, le chemin cesse d'être considéré au profit du chemin correspondant au deuxième état le plus prometteur. On itère ainsi le processus jusqu'à épuisement des éléments de chaque pile. Notons que sans critère d'arrêt la recherche sera exhaustive. On désignera les extensions de SFS et SBS respectivement par BFFS (*Best First Forward Search*), décrite par l'algorithme 3, et BFBS (*Best First Backward Search*). SFS et SBS correspondent aux cas particuliers de BFFS et BFBS pour lesquels seul le premier élément de la pile est considéré.

Algorithme 3 Algorithme de recherche avant, BFFS

Entrées: $J, A, \mathcal{V} = \{v_1, ..., v_p\}$
Sorties: $\mathcal{V}_{opt}$
  Début
  $O = \{\emptyset\}$ // ensemble des états à évaluer
  $F = \emptyset$ // ensemble des états déjà évalués
  $\mathcal{V}_{opt} = \emptyset$
  Tant que $\overline{A}$ & $O \neq \emptyset$ **Faire**
    $\mathcal{W} = \arg\max\left(J\left(Z\right), Z \in O\right)$
    $O = O - \{\mathcal{W}\}$
    $F = F \cup \{\mathcal{W}\}$
    Si $J\left(\mathcal{W}\right) \geq J\left(\mathcal{V}_{opt}\right)$ **Alors**
      $\mathcal{V}_{opt} = \mathcal{W}$
    Fin Si
    Pour $v_i \notin \mathcal{W}$ **Faire**
      Si $\mathcal{W} \cup \{v_i\} \notin O$ & $\mathcal{W} \cup \{v_i\} \notin F$ **Alors**
        $O = O \cup \{\mathcal{W} \cup \{v_i\}\}$
      Fin Si
    Fin Pour
  Fin Tant que
  Renvoyer $\mathcal{V}_{opt}$
  Fin

Du fait des retours-arrières qu'elle autorise, cette technique permet de ne pas s'arrêter au premier optimum local trouvé. Elle est cependant nettement plus coûteuse que la recherche séquentielle simple et c'est par l'intermédiaire du critère d'arrêt que l'on peut établir le compromis entre complexité de la recherche et qualité des solutions trouvées. Par exemple, dans les versions ultérieures de CFS, Hall (2000) a choisi d'utiliser non plus SFS, mais BFFS en arrêtant la recherche si aucune amélioration de J n'est observée après avoir parcouru 5 chemins du treillis.

Une autre forme de généralisation de la recherche séquentielle, dite *beam search* a reçu l'attention de la communauté depuis les travaux de Aha et Bankert (1996) qui ont mis en évidence l'intérêt de cette technique, en particulier sur les problèmes contenant plus d'attributs que d'exemples. Il s'agit également d'une technique séquentielle, qui contrairement à SFS et SBS ne conserve pas une seule solution à chaque itération mais un ensemble de

solutions[6]. Cela lui permet d'incorporer implicitement le mécanisme de *back-tracking*. À chaque itération, les k états successeurs les plus prometteurs sont conservés dans une liste. Puis pour chacun de ces k états, les m états suivants sont générés et évalués et seuls les k meilleurs parmi les $k \times m$ états sont conservés et ainsi de suite. Ainsi plusieurs chemins sont parcourus simultanément, les moins prometteurs étant abandonnés au fur et à mesure.

Pour se convaincre qu'il s'agit bien d'une généralisation des recherches séquentielles simples, il suffit de considérer le cas où la liste des meilleurs états est de taille $k = 1$. Nous nommerons BFS (*Beam Forward Search*), décrite par l'algorithme 4, et BBS (*Beam Backward Search*), les extensions respectives de SFS et SBS. Si elle s'avère performante (Ng, 1998; Gupta *et al.*, 2002), cette technique est également relativement coûteuse. Lorsqu'aucune limite n'est imposée sur la taille de la liste $k = \infty$, elle parcourt le treillis de manière exhaustive. C'est donc la taille de la liste k, ainsi que le critère d'arrêt A qui vont permettre d'établir le compromis performance-complexité.

Algorithme 4 Algorithme de recherche avant, BFS

Entrées: $J, A, k, \mathcal{V} = \{v_1, ..., v_p\}$
Sorties: $\mathcal{V}_{opt}$
  Début
  $O = \{\emptyset\}$ // ensemble des k meilleurs états
  Tant que $\overline{A}$ **Faire**
    $G = \emptyset$ // liste temporaire des successeurs de tous les états de O
    Pour $\mathcal{W} \in O$ **Faire**
      Pour $v_i \notin \mathcal{W}$ **Faire**
        $G = G \cup \{\mathcal{W} \cup \{v_i\}\}$
      Fin Pour
    Fin Pour
    // H conserve les k meilleurs états
    Soit $H \subseteq G$ tel que $|H| = k$ & $\forall \mathcal{Y} \in H, \forall \mathcal{Z} \in G - H, J(\mathcal{Y}) \geq J(\mathcal{Z})$
    $O = H$
  Fin Tant que
  Renvoyer $\mathcal{V}_{opt} = \arg\max(J(\mathcal{W}), \mathcal{W} \in O)$
  Fin

Lorsque la fonction J à optimiser comporte de nombreux *extrema*, le risque est grand pour une méthode de recherche séquentielle de ne trouver qu'un optimum local. Nous avons vu que des solutions déterministes avaient été proposées pour surmonter cette difficulté au prix d'un accroissement de la complexité. Les méthodes stochastiques de recherche locale constituent un autre expédient répandu.

Recherche non déterministe Les méthodes de recherche précédentes ont toutes en commun une stratégie sous-optimale. Comme l'indique la terminologie choisie par Liu et Yu (2005), elles reposent toutes sur le choix d'une heuristique qui va guider le parcours du treillis de recherche. L'efficacité de ces caractéristiques dépend d'un certain nombre de paramètres, comme par exemple la monotonie du critère de performance J. Lorsque les hypothèses qui sous-tendent ces heuristiques ne sont pas vérifiées, les performances de ces méthodes peuvent se dégrader.

Pour remédier à ces difficultés, Liu et Setiono (1996c:b) proposent de se passer de toute heuristique en utilisant une technique purement aléatoire. Ils ont ainsi développé un

[6]Cela est également le cas pour la recherche flottante et *best first search*.

algorithme de Las Vegas, que l'on nommera LV, qui consiste simplement à choisir successivement, de manière aléatoire, un état dans le treillis et à évaluer ses performances. Comme tout algorithme de Las Vegas, une solution optimale est garantie, mais malheureusement dans un temps potentiellement infini. C'est la définition du critère d'arrêt (un nombre fini k de tirages aléatoires) qui permet de limiter la complexité de la recherche, faisant perdre par là même la garantie de la solution optimale. L'algorithme 5 en donne une description synthétique, dans laquelle *alea* est une fonction qui associe à un ensemble d'attributs $\mathcal{V}$ un sous-ensemble de $\mathcal{V}$, tiré de manière aléatoire.

Algorithme 5 Algorithme de recherche de Las Vegas, LV

Entrées: $J, A, k, \mathcal{V} = \{v_1, ..., v_p\}$
Sorties: $\mathcal{V}_{opt}$
  Début
  $\mathcal{V}_{opt} = \emptyset$
  Tant que $\overline{A}$ & $i < k$ **Faire**
    $\mathcal{W} = alea\,(\mathcal{V})$
    Si $J\,(\mathcal{W}) \geq J\,(\mathcal{V}_{opt})$ **Alors**
      $\mathcal{V}_{opt} = \mathcal{W}$
    Fin Si
    $i = i + 1$
  Fin Tant que
  Renvoyer $\mathcal{V}_{opt}$
  Fin

Dans la même veine, Skalak (1994) propose un algorithme de Monte Carlo pour sélectionner les exemples d'une base de données qui seront utilisés comme prototypes pour classer de nouveaux exemples selon le principe des k plus proches voisins. En transposant sa procédure dans l'espace des variables, nous obtenons exactement la même méthode de recherche que celle qui a été introduite par Liu et Setiono[7].

Skalak (1994) procède également à une recherche aléatoire pour identifier simultanément les meilleurs prototypes et les meilleurs attributs. La technique de recherche repose sur la version stochastique, dite à *mutation aléatoire*, de la recherche *hill climbing*. Les versions déterministes de cette méthode évaluent pour chaque état l'ensemble des successeurs avant de choisir le plus prometteur. La version stochastique, que l'on nommera SHC, choisit aléatoirement l'un des états successeurs, peu importe qu'il faille ajouter ou supprimer un attribut. Elle évalue le sous-ensemble résultant qui devient le nouvel état si ses performances sont meilleures que celles de l'état courant. L'algorithme 6 reprend ces éléments de manière plus formelle.

> *alea désignera une fonction qui tire aussi bien un sous-ensemble d'attributs au hasard, qu'un nombre au hasard entre* 1 *et* p*, lorsque* p *en est la variable.*

Méthode d'optimisation locale non déterministe, les algorithmes génétiques (AG) ont été largement employés en sélection d'attributs, avec un certain succès (Siedlecki et Sklansky, 1993; Vafaie et Imam, 1994; Yang et Hononvar, 1998; Cantu-Paz, 2004; Sepulveda-Sanchis *et al.*, 2002). Précisons, cependant, que les conclusions de diverses études comparant

[7] Un algorithme de Monte Carlo est également un algorithme probabiliste, qui à la différence d'un algorithme de Las Vegas, ne garantit pas l'obtention d'une réponse correcte au problème posé, mais assure qu'une solution approchée, avec une erreur limitée, sera trouvée dans un temps donné.

Algorithme 6 Algorithme de recherche stochastique, SHC

Entrées: $J, A, \mathcal{V} = \{v_1, ..., v_p\}$
Sorties: $\mathcal{V}_{opt}$
  Début
  $\mathcal{V}_{opt} = alea\,(\mathcal{V})$
  Tant que $\overline{A}$ **Faire**
    $j = alea(p)$
    Si $v_j \in \mathcal{V}_{opt}$ **Alors**
      $\mathcal{W} = \mathcal{V}_{opt} - \{v_j\}$
    Sinon
      $\mathcal{W} = \mathcal{V}_{opt} \cup \{v_j\}$
    Fin Si
    Si $J\,(\mathcal{W}) \geq J\,(\mathcal{V}_{opt})$ **Alors**
      $\mathcal{V}_{opt} = \mathcal{W}$
    Fin Si
  Fin Tant que
  Renvoyer $\mathcal{V}_{opt}$
  Fin

ces algorithmes aux méthodes de recherche séquentielles divergent. Ferri *et al.* (1994) observent ainsi que SFFS surpasse AG, tandis que Oh *et al.* (2004) arrivent à la conclusion inverse.

On peut attribuer ces divergences aux protocoles expérimentaux. Ceux-ci n'étant pas clairement définis, il est vraisemblable qu'ils diffèrent. L'autre explication peut venir des bases de données qui sont utilisées pour la comparaison. Chacune de ces méthodes est plus ou moins adaptée à un type de problème particulier qu'il serait bon de caractériser. Nous savons en effet depuis le théorème d'impossibilité de Wolpert et Macready (1997) que les méthodes d'optimisation sont toutes équivalentes, le théorème *No Free Lunch* indiquant que l'espérance des performances d'une méthode d'optimisation, sur l'ensemble des problèmes possibles, ne dépend pas de l'algorithme utilisé.

Les algorithmes génétiques ont été présenté à la section 3.1. Aussi nous contenterons-nous de rappeler leurs principales caractéristiques :

- initialisation de la population (composée de k individus),
- sélection des individus,
- opérateur de croisement (avec une probabilité de croisement P_c),
- opérateur de mutation (avec une probabilité de mutation P_m),
- recomposition de la population,
- critère d'arrêt.

L'algorithme 7 donne une description générique d'un algorithme génétique, pour une population P de k individus. Pour plus de détails, le lecteur intéressé pourra se reporter aux ouvrages de Michalewicz (1996); Man *et al.* (1999).

Les performances de SFFS et AG sont proches. Oh *et al.* (2004) ont mis en place une méthode hybride pour tirer parti des avantages et inconvénients des deux approches. À chaque itération, pour chaque nouvel individu, il est procédé à une phase d'optimisation locale selon la méthode SFFS. Les auteurs ont relevé expérimentalement que la complémentarité de ces deux approches était manifeste en grande dimension. Pour une valeur de p relativement faible, SFFS s'avère en revanche aussi efficace que la méthode hybride. L'apport des algorithmes génétiques est donc surtout notable en grande dimension.

Algorithme 7 Algorithme génétique, AG

Entrées: $J, A, k, P_m, P_c, \mathcal{V} = \{v_1, ..., v_p\}$
Sorties: $\mathcal{V}_{opt}$
Début
$P = Init\,(k, p)$ // Initialisation de la population
Tant que $\overline{A}$ **Faire**
$P_{nouv} = \emptyset$ // Population à la prochaine génération
Pour $i = 1..k/2$ **Faire**
$(Parent_1, Parent_2) = Selection\,(P, J)$
$(Enfant_1, Enfant_2) = Croisement\,(Parent_1, Parent_2, P_c)$
$Enfant_1 = Mutation\,(Enfant_1, P_m)$
$Enfant_2 = Mutation\,(Enfant_2, P_m)$
$P_{nouv} = P_{nouv} \cup \{Enfant_1, Enfant_2\}$
Fin Pour
$P = Recomposition\,(P, P_{nouv})$
Fin Tant que
Renvoyer $\mathcal{V}_{opt} = \arg\max\,(J\,(\mathcal{W})\,, \mathcal{W} \in P)$
Fin

D'autres types de méthodes apparentées aux algorithmes génétiques sont également efficaces en optimisation et peuvent être employés à bon escient pour la sélection d'attributs. Les stratégies évolutionnaires, quoique moins populaires que les algorithmes génétiques, en sont un bon exemple (Beyer et Schewefel, 2002; Back, 2004). À la différence des algorithmes génétiques seul l'opérateur de mutation est utilisé. Notons que la technique SHC peut être considérée comme un algorithme génétique dans lequel l'opérateur de croisement a été abandonné au profit de la seule mutation. Mais à la différence des stratégies évolutionnaires, une seule solution est maintenue à chaque itération.

Enfin, Inza *et al.* (2000) utilisent l'apprentissage incrémental à base de populations. Il s'agit d'un autre type de technique évolutionnaire, dans laquelle ni la mutation ni le croisement ne sont utilisés. Une population de solutions potentielles est maintenue, mais la régénération de la population se fait par rééchantillonnages successifs de la population globale, à partir d'une distribution de probabilité estimée sur la population courante en privilégiant les solutions les plus prometteuses.

À partir des différentes caractéristiques des méthodes de recherche que nous avons pu mettre en évidence, nous avons établi une taxinomie, schématisée à la figure 7.5. Les différentes méthodes abordées dans cette partie y sont situées au niveau des feuilles. Cette taxinomie met l'accent sur les différences principales qui permettent de distinguer ces méthodes.

Outre l'organisation de la recherche qui permet de fixer la manière dont le treillis de recherche sera parcouru, l'élément clé de tout algorithme d'optimisation est la fonction d'évaluation elle-même, celle qui doit être optimisée et que nous avons nommée J. Nous avons vu, en effet, que ses caractéristiques, telles que la monotonie ou la présence de nombreux *extrema*, pouvaient guider le choix de la fonction de recherche. Nous allons maintenant détailler les principales mesures qui sont utilisées dans le domaine de la sélection d'attributs.

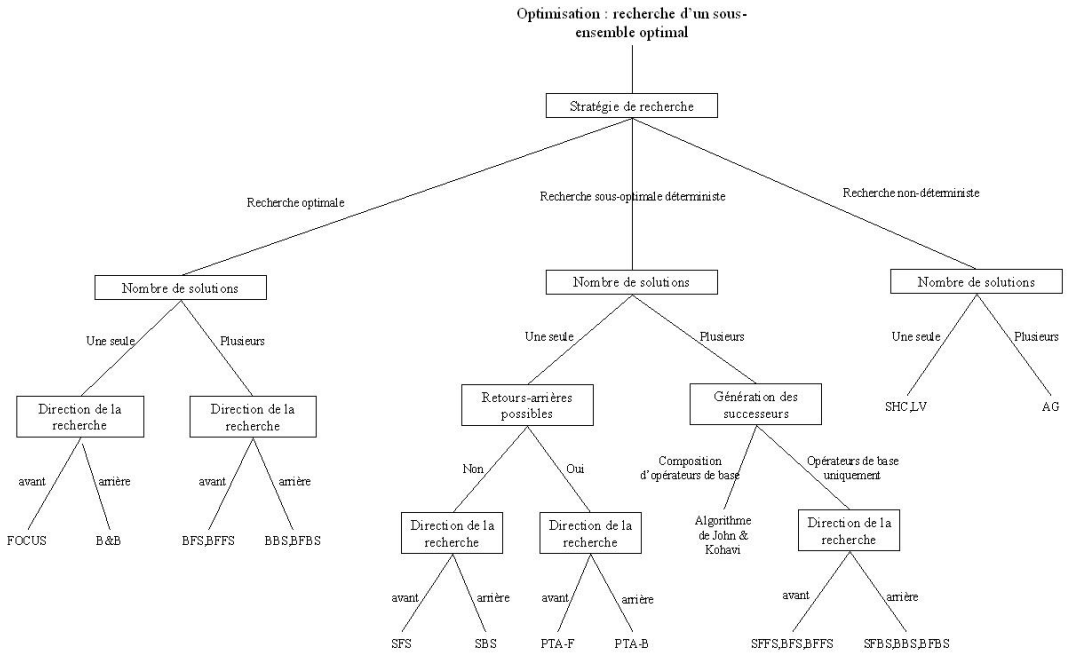

FIG. 7.5 – Taxinomie des méthodes de recherche utilisées pour la sélection d'attributs

7.3.1.2 Mesures d'évaluation

Lors du parcours du treillis de recherche, il nous faut être capable d'évaluer la qualité des différents états considérés. Autrement dit, il nous faut fixer la mesure d'évaluation J d'un sous-ensemble d'attributs qui sera utilisée pour guider la recherche.

***Wrappers* et filtres** Étant donné que nous envisageons la sélection d'attributs comme un prétraitement en vue d'une phase ultérieure de classification supervisée, il semble assez naturel de mesurer les performances de la sélection d'attributs par l'intermédiaire de celles du classifieur qui peut être appris à partir du sous-ensemble d'attributs sélectionné. Il a pourtant fallu attendre les travaux de Kohavi et John (1997) pour qu'apparaisse cette approche. Les algorithmes qui mettent en œuvre cette approche sont appelés *wrappers*, par opposition aux *filtres* qui utilisent une mesure de performance indépendante des traitements qui seront effectués ultérieurement.

Wrappers et filtres ont des propriétés complémentaires. Les premiers présentent l'avantage d'inclure le biais du classifieur qui sera utilisé *in fine* pour guider la recherche dans le treillis. Il s'agit là certainement de la meilleure heuristique à disposition (Blum et Langley, 1997). En revanche, ils sont par là même dépendants d'un algorithme d'induction particulier. Aussi faudra-t-il procéder à nouveau à une phase de sélection d'attributs chaque fois que l'on souhaitera employer un nouveau classifieur. Pour évaluer chacun des états du treillis, il faudra de plus relancer un processus d'apprentissage et tester le classifieur afin d'en déduire une mesure de performance de l'état considéré. Or la construction, voire le test, d'un classifieur sont des étapes coûteuses. Aussi l'inconvénient majeur de ces techniques réside-t-il dans leur complexité très élevée, ce qui les rend difficilement applicables sur des données de taille conséquente (grand nombre d'attributs ou grand nombre d'exemples).

À l'inverse les filtres sont indépendants du classifieur qui sera utilisé. Ils sont donc plus génériques et nettement moins coûteux. Ceci explique qu'ils soient préférés aux *wrappers* sur les problèmes en grande dimension, malgré leurs performances réputées inférieures

(Blum et Langley, 1997; Liu et Yu, 2005).

Pour les *wrappers*, le choix de la fonction d'évaluation d'un sous-ensemble d'attributs se ramène donc au choix d'une mesure de performance d'un classifieur. Nous en détaillerons quelques unes à la section 10.3.2. Forman (2003) en donne de nombreux exemples dans le domaine de la recherche d'information.

En ce qui concerne les filtres, diverses mesures ont été introduites, que l'on peut toutes rattacher, de près ou de loin, à l'une des définitions de la pertinence que nous avons mentionnées à la section 7.2.1. Notons à ce sujet qu'en prenant pour J une mesure des performances de l'algorithme d'induction, la sélection effectuée par un *wrapper* peut s'interpréter comme la recherche des attributs fortement et faiblement pertinents au sens des définitions 6 et 7.

Liens entre la notion de pertinence et les mesures de performance Derrière le choix du critère, se cache donc bien souvent, de façon plus ou moins explicite, la référence à une définition particulière de la pertinence. Cette formulation est cependant quelque peu trompeuse car elle pourrait laisser entendre que le choix de J dépend de celui d'une définition de la pertinence. Or il serait quelque peu illusoire de vouloir instaurer une relation de précédence entre les deux. S'il ne fait que peu de doute que les réflexions sur la pertinence ont motivé et guidé la réalisation d'algorithmes de sélection d'attributs, il semble également légitime de penser que les formalisations de la pertinence, telles que celles que nous avons proposées, ont elles-mêmes été guidées par des développements antérieurs de certaines mesures. Aussi nous contenterons-nous de souligner les liens qui existent entre ces deux choix, sans chercher à orienter ces liens.

Les définitions 2 et 3 mettent en avant l'importance du lien qui peut exister entre deux objets v et y. La variation de l'un de ces objets doit impacter le second, dans un certain contexte, pour que l'on puisse parler de pertinence. Les mesures de dépendance entre variables permettent d'évaluer l'importance de ce lien. Aussi ont-elles été largement utilisées dans la littérature. Nous les nommerons mesures de corrélation, le terme corrélation étant ici pris dans son acception première, sans référence particulière à la signification qu'elle peut avoir dans le domaine de la statistique. Ces mesures peuvent se répartir en deux sous-familles.

- Les mesures de *corrélation* statistique
- Les mesures de *divergence* entre distributions de probabilité

Les mesures d'*information* sont également fortement apparentées aux mesures de corrélation, mais elles peuvent également et plus directement être interprétées sous l'angle des définitions 8 et 9 de la pertinence.

Duch (2006); Dash et Liu (1997); Liu et Yu (2005); Molina *et al.* (2002) distinguent également une quatrième grande famille de mesures qui est fortement liée aux définitions 4 et 5 de la pertinence. Il s'agit des mesures de *cohérence*.

Au sein de chacune de ces familles, de nombreuses mesures ont été proposées. Avant d'en donner quelques exemples, rappelons que le critère J auquel elles correspondent est une fonction qui doit permettre d'évaluer la pertinence d'un sous-ensemble d'attributs relativement à la classe y. Nous commencerons par simplifier le problème en considérant dans un premier temps des mesures qui n'évaluent qu'un seul attribut. Nous verrons dans un second temps comment étendre les mesures précédentes à n'importe quel ensemble d'attributs.

Mesures de corrélation Elles ont pour objectif de mesurer le degré d'association entre deux attributs, que l'on nommera par la suite v et y. y correspond à la classe du problème d'apprentissage supervisé qui nous occupe, tandis que v désigne l'attribut dont on veut mesurer la pertinence relativement à y. Par abus de notation, y et v désigneront également les variables aléatoires correspondant à ces attributs.

Notons m_{cor} une mesure générique de corrélation. $m_{cor}(v, y)$ évalue le degré d'association entre v et y en mesurant à quel point ces variables sont statistiquement dépendantes l'une de l'autre. D'un point de vue statistique nous avons les équivalences suivantes :

$$\begin{aligned}
\text{Les variables } v \text{ et } y \text{ sont indépendantes} \quad &\Leftrightarrow \quad E[v, y] = E[v] \times E[y] && (7.1)\\
&\Leftrightarrow \quad P(v, y) = P(v) \times P(y) && (7.2)\\
&\Leftrightarrow \quad P(v|y) = P(v) && (7.3)\\
&\Leftrightarrow \quad P(y|v) = P(y) && (7.4)
\end{aligned}$$

Les différentes mesures de corrélation qui ont été proposées essaient toutes d'exploiter l'une ou l'autre des équivalences précédentes. De manière générique, si l'on note chacune des égalités précédentes sous la forme $r = s$, la fonction m_{cor} qui mesure l'écart par rapport à l'indépendance peut s'écrire de la manière suivante : $m_{cor}(v, y) = h(r, s)$ où h est une fonction croissante en r et décroissante en s ou inversement. Nous devons de plus avoir $|m_{cor}(v, y)| = 0$ lorsque v et y sont indépendantes et $|m_{cor}(v, y)| = 1$ lorsqu'il existe une dépendance fonctionnelle, entre v et y.

Le coefficient ρ de corrélation linéaire de Pearson est probablement la mesure de corrélation statistique la plus répandue :

$$\rho = \frac{E[v, y] - E[v]E[y]}{s_v \times s_y} = \frac{\sum_{i=1}^{n}(v_i - \overline{v})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(v_i - \overline{v})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Ce coefficient présente l'avantage d'avoir une interprétation géométrique simple. Il correspond au cosinus de l'angle formé par les vecteurs v et y, lorsque ceux-ci sont centrés (moyenne empirique nulle). En revanche il n'est adapté que pour détecter des dépendances linéaires entre v et y. L'utilisation de ce coefficient suppose de plus que ces attributs sont continus ou au moins discrets et ordonnés, sans quoi les notions de moyenne et de variance. perdent quelque peu de leur sens. Or dans notre contexte de classification supervisée, y est une variable symbolique. Hall (2000) décrit la façon d'adapter la définition de la corrélation dans de tels cas.

Pour des problèmes bi-classes, une mesure encore plus simple est définie par le critère *sep* de séparabilité des classes, qui évalue à quel point l'attribut v permet de séparer les classes c_1 et c_2. Nous noterons $v|_{c_i}$ la variable v restreinte à la classe c_i, composée des $n_{.i}$ observations dont la classe est c_i. Nous avons alors :

$$sep(v, y) = \frac{\overline{v|_{c_1}} - \overline{v|_{c_2}}}{s^2_{v|_{c_1}} + s^2_{v|_{c_2}}}$$

Le critère *sep* est la racine carrée du rapport de Fisher : rapport de la variance inter-classe sur la variance intra-classe. Des extensions existent pour les problèmes multi-classes. Xuan *et al.* (2004) en donnent un exemple d'application à la sélection d'attributs.

La statistique t de Student, qui ne diffère du critère précédent que par la pondération des variances intra-classes au dénominateur, est également fréquemment employée :

$$t(v, y) = \frac{\overline{v|_{c_1}} - \overline{v|_{c_2}}}{\frac{s^2_{v|_{c_1}}}{n_{.1}} + \frac{s^2_{v|_{c_2}}}{n_{.2}}}$$

Ces mesures peuvent être utilisées telles quelles pour ordonner les attributs. On peut alors choisir de ne conserver que les plus prometteurs : ceux pour lesquels le degré de corrélation est le plus élevé. Mais il est également envisageable, selon les besoins, de pouvoir décider pour chaque attribut s'il est ou non pertinent. En effet, à chacune de ces mesures correspond un test statistique qui nous permet de conclure, avec une probabilité d'erreur maximale donnée, quant à l'existence d'un lien effectif entre les deux variables v et y (Duch, 2006). Ces mesures de corrélation statistique mettent plus particulièrement en œuvre l'équivalence 7.1 caractérisant la relation d'indépendance statistique.

À l'inverse, les mesures de divergence reposent plutôt sur les caractérisations 7.2, 7.3 et 7.4 de l'indépendance. Elles sont de plus capables d'identifier des dépendances non linéaires. Elles quantifient le degré de divergence entre les deux distributions de probabilité impliquées dans chacune de ces caractérisations. Certains auteurs parlent de mesure de distance plutôt que de divergence. Mais ce terme est impropre car certaines de ces mesures ne sont pas symétriques. C'est le cas par exemple de la divergence de Küllback-Leibler δ_{KL}, le représentant le plus notoire de cette catégorie de mesures (Koller et Sahami, 1996; Cantu-Paz, 2004; Xing *et al.*, 2001).

Soient deux distributions de probabilité h et k définies sur l'univers X, cette divergence est définie de la façon suivante :

$$\delta_{KL}(h,k) = \int_{x \in X} h(x) \log_2 \left(\frac{h(x)}{k(x)} \right)$$

L'équation précédente met clairement en relief la non-symétrie de cette mesure. La distribution h joue un rôle particulier et est généralement appelée distribution de référence. Molina *et al.* (2002); Duch (2006) donnent de nombreux exemples d'autres mesures de divergence, nous nous contenterons de mentionner celles qui nous seront utiles par la suite : la distance de Kolmogorov-Smirnov et celle du χ^2.

Ces deux mesures sont habituellement utilisées dans des tests statistiques d'ajustement qui évaluent l'écart entre deux distributions de probabilité. Le test du χ^2 s'applique sur des variables discrètes[8]. Il consiste à chercher des éléments qui permettent d'infirmer l'hypothèse nulle d'indépendance entre v et y. Il se base pour cela sur l'équivalence 7.2 et compare la distribution empirique de $P(v,y)$, avec celle de $P(v)P(y)$. Cette dernière est estimée en supposant vraie la relation d'indépendance entre v et y.

Notons n le nombre total d'exemples, n_{ij} le nombre d'exemples de la classe j qui prennent la modalité i pour l'attribut v. $n_{i.}$ désigne le nombre total d'exemples prenant la modalité i pour v et $n_{.j}$ est le nombre total d'exemples de la classe j. Avec ces notations, la statistique du χ^2 s'écrit :

$$\chi^2 = \sum_{i,j} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

Cette statistique est souvent considérée comme une mesure de corrélation, bien que conceptuellement elle corresponde à une mesure de divergence entre distributions de probabilité. À l'image de ce que nous avons indiqué pour les mesures de corrélation, la statistique du χ^2 peut être utilisée telle quelle pour ordonner les attributs (plus la valeur de χ^2 sera élevée et plus v et y seront considérés comme dépendants). Cette mesure a été utilisée dans de nombreux domaines de l'apprentissage, tels que la discrétisation, la sélection d'attributs et l'induction d'arbres de décision[9] (Liu et Setiono, 1996a).

[8] Une phase de discrétisation est donc nécessaire pour les variables continues.

[9] Ceci n'est guère étonnant sachant que ces trois domaines sont étroitement liés.

Le test de Kolmogorov-Smirnov s'applique quant à lui à des données continues. Il compare deux distributions de probabilité sur la base de leur fonction de répartition empirique. La distance de Kolmogorov-Smirnov δ_{KS} est alors simplement égale à l'écart maximal entre les fonctions de répartition empiriques. Notons F_h et F_k les fonctions de répartition empiriques de deux distributions de probabilité h et k (définies sur X) que l'on souhaite comparer. Nous avons alors :

$$\delta_{KS}(h,k) = \max_{x \in X} |F_h(x) - F_k(x)|$$

Plus cette distance est grande et plus les deux distributions seront considérées comme distinctes. Moins répandue que la distance du χ^2, cette métrique a été également appliquée pour la discrétisation et la sélection d'attributs dans le cadre de l'induction d'arbres de décisions (Utgoff et Clouse, 1996). Un test statistique peut éventuellement être employé si l'on souhaite conclure sur l'existence d'une différence significative entre les deux distributions. Nous y reviendrons à la section 7.4.

À partir des équivalences 7.2, 7.3 et 7.4 caractérisant l'indépendance de deux variables, il est possible de choisir des fonctions f et g telles que $\delta_{KL}(f,g)$ et $\delta_{KS}(f,g)$ correspondent à des mesures de corrélation entre les variables v et y.

$$\begin{cases} f = P(v,y) & \text{et} \quad g = P(v)\,P(y) \\ f = P(v|y) & \text{et} \quad g = P(v) \\ f = P(y|v) & \text{et} \quad g = P(y) \end{cases}$$

Mesures d'information L'idée qui sous-tend l'utilisation des mesures issues de la théorie de l'information est assez proche de celle que nous avons évoquée à propos des mesures de corrélation. Aussi ces mesures sont-elles souvent considérées comme faisant partie de la classe étendue des mesures de corrélation. L'algorithme de sélection d'attributs de Hall et Smith (1997) est ainsi appelé filtre à base de corrélation, alors que la mesure d'évaluation J qui est utilisée est directement issue de la théorie de l'information. Pour une vue d'ensemble de cette classe étendue des mesures de corrélation dans le contexte de la sélection d'attributs, nous invitons le lecteur intéressé à se reporter à la thèse de Hall (1999).

Les approches conceptuelles de la pertinence diffèrent quelque peu entre ces deux types de mesures. C'est la raison pour laquelle nous les avons regroupées dans deux familles distinctes. L'objectif est ici de déterminer pour un attribut v, la quantité d'information qu'il apporte sur la classe y. Plus cette quantité est élevée et plus v sera considéré comme pertinent.

L'une des premières utilisations de ce type de mesures est due à Quinlan (1986) qui a intégré le gain d'entropie comme mesure de sélection d'attributs dans son système de construction d'arbres de décision ID3. Le gain d'information G correspond à la différence entre l'entropie de y et l'entropie de y lorsque v est supposé connu. G mesure donc la réduction de l'incertitude que l'on a sur y grâce à la connaissance de v. Si I désigne une mesure d'entropie, G s'exprime sous la forme suivante :

$$G(v,y) = I(y) - I(y|v)$$

En considérant l'entropie de Shannon, G peut se mettre sous la forme suivante :

$$G(v,y) = \sum_{v,y} P(v,y) \log_2 \left(\frac{P(v,y)}{P(v)\,P(y)} \right) \tag{7.5}$$

L'équation précédente suppose que v et y sont des variables discrètes. Dans le cas continu, il convient de remplacer la sommation par un calcul intégral. Mais l'estimation des densités

de probabilité pose cependant problème, du fait du nombre limité d'observations dont nous disposons en pratique (Renteria et Tanscheit, 2005). Si la méthode des fenêtres de Parzen est envisageable pour estimer ces densités (Peng *et al.*, 2005), l'approche la plus courante consiste à discrétiser les variables continues (Hall, 2000; Liu et Yu, 2003).

C'est exactement ce gain que nous essayons de maximiser dans la technique de substitution des valeurs manquantes que nous avons proposée à la section 6.5. Ce gain est également appelé *information mutuelle* car l'information apportée par v sur y est la même que celle apportée par y sur v. Il est assez aisé de montrer l'équivalence suivante :

$$\text{Les variables } v \text{ et } y \text{ sont indépendantes} \Leftrightarrow G(v,y) = 0$$

Ainsi, à l'instar des mesures de corrélation, les mesures d'information permettent de quantifier la dépendance statistique entre deux variables. Pour que le lien entre information mutuelle et mesures de corrélation soit encore plus net, signalons que lorsque l'on considère l'entropie de Shannon, l'équation 7.5 indique que l'information mutuelle correspond à la divergence de Küllback-Leibler. Nous avons en effet :

$$G(v,y) = \delta_{KL}\left(P(v,y), P(v)P(y)\right)$$

Le gain d'information, probablement grâce à la popularité de l'algorithme ID3, est devenue l'une des mesures d'évaluation les plus usitées dans le domaine de la sélection d'attributs. Ainsi Das (2001); Xing *et al.* (2001); Kotsiantis et Pintelas (2004) l'utilisent pour réaliser un filtrage initial des attributs, avant de procéder à une sélection plus poussée par d'autres méthodes. Mais il est également des exemples d'algorithmes dont le cœur du processus de sélection repose exclusivement sur cette mesure (Peng *et al.*, 2005; Renteria et Tanscheit, 2005).

L'inconvénient du gain d'information est qu'il comporte un biais favorisant les attributs v dont le nombre de modalités est grand (Kononenko, 1995). Afin de limiter ce biais, diverses solutions ont été proposées qui consistent en différentes normalisations du gain.

Quinlan (1993) a introduit la notion de *gain ratio*, GR, dans C4.5, l'extension d'ID3. Le gain G est alors normalisé par l'entropie de v. La mesure obtenue n'est alors plus symétrique. Pour conserver cette propriété Wu et Zhang (2004) ont quant à eux choisi d'utiliser le nombre de modalités de v comme facteur de normalisation. Ils nomment la nouvelle mesure *gain d'information équilibré* (*balanced information gain*), que l'on notera B_g. Hall (2000) utilise encore une autre méthode qui consiste à intégrer l'entropie de y dans le facteur de normalisation. Le critère correspondant est nommé coefficient d'incertitude symétrique (Press *et al.*, 2002) et sera noté SU. Ce coefficient a été utilisé dans d'autres travaux avec une terminologie différente. Wehenkel et Pavella (1991) le nomment par exemple « gain d'information normalisé ». L'emploi du terme « incertitude symétrique » étant plus usité, nous avons préféré conserver ce dernier. Les définitions formelles de ces trois nouvelles mesures sont les suivantes :

$$\begin{aligned}
GR(v,y) &= \frac{G(v,y)}{I(v)} \\
B_g(v,y) &= \frac{G(v,y)}{\log_2 k} \\
SU(v,y) &= 2\frac{G(v,y)}{I(v)+I(y)}
\end{aligned}$$

Pertinence d'un ensemble d'attributs Nous avons vu comment quantifier la pertinence d'un attribut, pris séparemment, vis-à-vis de la classe au moyen de mesures de corrélation statistique, de divergence et d'information. Les différentes mesures abordées peuvent être utilisées pour ordonner les différents attributs. Un algorithme de sélection basé sur un tel ordonnancement consiste simplement à retenir les d meilleurs attributs. Recourir à une telle solution implique d'une part que le nombre d'attributs à sélectionner est connu. D'autre part, seul le lien entre un attribut et la classe est pris en compte. Cela revient à supposer les attributs indépendants entre eux et à négliger les interactions potentielles entre ces attributs. Ces deux hypothèses sont très fortes. La première dépend de l'application, mais la seconde est beaucoup plus difficilement tenable.

Guyon et Elisseeff (2003) donnent quelques exemples mettant en exergue l'importance de la prise en compte de la redondance. Du fait des interactions avec d'autres attributs, un attribut v peut s'avérer non pertinent dans le contexte des attributs vis-à-vis desquels il est redondant, même s'il est corrélé à la classe. Pour cette raison, cette approche basée sur un ordonnancement obtenu uniquement à partir de mesures de corrélation entre un attribut et la classe n'est pas utilisée.

En évaluant la pertinence d'un attribut, indépendamment des autres, nous nous sommes contenté de définir la mesure J sur des singletons. Or pour pouvoir mettre en pratique les différentes techniques de recherche présentées précédemment, il nous faut pouvoir attribuer une mesure à un ensemble de plusieurs attributs. Théoriquement cela ne pose pas de problème particulier. Dans les formules données précédemment pour définir les différentes mesures, nous établissons la corrélation entre une variable v et la classe y. Nous avons supposé jusqu'à présent que v correspondait à l'une des p variables v_i de l'ensemble $\mathcal{V}$ de départ. Si l'on souhaite évaluer un sous-ensemble $\mathcal{W} = \{w_1, ...w_d\} \subseteq \mathcal{V}$, il suffit désormais de considérer que $v = (w_1, ..., w_d)$.

Cependant en pratique, le nombre d'exemples à disposition étant limité, il s'avère délicat, voire impossible d'estimer les différentes mesures de corrélation que nous avons envisagées, dès que d dépasse 2 ou 3. Peng *et al.* (2005) illustrent cette difficulté pour le calcul de l'information mutuelle. Pour contourner cette difficulté, l'approche la plus répandue consiste à construire une mesure d'ensemble à partir des mesures sur les singletons. Afin d'approcher la mesure d'ensemble théorique, il faut tenir compte aussi bien du lien qui existe entre les attributs de l'ensemble et la classe, que des interactions entre les attributs de l'ensemble.

L'évaluation du lien global entre un ensemble et la classe se fait par agrégation des mesures de pertinence de chacune des variables, selon une fonction d'agrégation Agg_p. Cette pertinence n'est pas contextuelle. Par la suite, sauf s'il y a ambiguïté, nous parlerons de *pertinence*, notée *pert*, sans préciser que le contexte n'est pas pris en compte, pour désigner ce lien entre un ensemble d'attributs et la classe. Pour un sous-ensemble $\mathcal{W} \subseteq \mathcal{V}$, nous avons :

$$pert\left(\mathcal{W}\right) = Agg_p\left(m_{cor}\left(v, y\right), \forall v \in \mathcal{W}\right)$$

Les interactions entre attributs d'un même ensemble, que l'on désignera de manière générique par *redondance* et qui sera notée *red*, seront évaluées par agrégation des mesures d'interactions élémentaires entre deux attributs de l'ensemble. Nous noterons Agg_r la fonction d'agrégation correspondante. Ces interactions élémentaires peuvent elles aussi être mesurées par le biais d'un degré d'association entre deux attributs. Aussi seront-elles quantifiées par les mesures de corrélation que nous avons décrites précédemment. Il suffit de remplacer dans les formules précédentes y par w, w désignant un attribut appartenant au même sous-ensemble que v. La redondance d'un sous-ensemble $\mathcal{W} \subseteq \mathcal{V}$ peut alors s'écrire :

$$red\left(\mathcal{W}\right) = Agg_r\left(m_{cor}\left(v, w\right), \forall v \neq w \in \mathcal{W}\right)$$

La mesure de redondance est nettement plus complexe à évaluer que la mesure de pertinence. Pour un sous-ensemble $\mathcal{W}$ de cardinal q, la pertinence nécessite l'évaluation et l'agrégation de q mesures élémentaires, tandis que ce nombre passe à $\frac{q(q-1)}{2} = \mathcal{O}\left(q^2\right)$ pour la redondance. Il faut en effet estimer la corrélation entre chacun des couples d'attributs du sous-ensemble.

Combiner ces deux mesures permet alors d'en construire une troisième qui reflétera la pertinence d'un sous-ensemble d'attributs vis-à-vis de la classe en tenant compte du contexte. Si l'on note Agg_J la fonction d'agrégation combinant ces deux mesures, on a :

$$\begin{aligned} J(\mathcal{W}) &= Agg_J\left(pert\left(\mathcal{W}\right), red\left(\mathcal{W}\right)\right) \\ &= Agg_J\left(Agg_p\left(m_{cor}\left(w,y\right), \forall w \in \mathcal{W}\right), Agg_r\left(m_{cor}\left(v,w\right), \forall v \neq w \in \mathcal{W}\right)\right) \end{aligned}$$

En sélection d'attributs, les sous-ensembles recherchés sont ceux qui contiennent des attributs fortement pertinents et faiblement pertinents mais non redondants. On privilégiera donc les sous-ensembles qui maximisent la pertinence et minimisent la redondance. Cela impose une contrainte sur le choix de l'opérateur Agg_J : il doit être croissant selon son premier argument (la pertinence) et décroissant selon le second (la redondance).

Les mesures d'évaluation qui ont été proposées dans la littérature diffèrent selon les mesures de corrélation utilisées pour l'évaluation de la pertinence et de la redondance, et selon les opérateurs d'agrégation qui sont choisis. Nous ne reviendrons pas sur les mesures de corrélation que nous venons de présenter. Précisons cependant que rien n'impose l'utilisation de la même mesure pour évaluer pertinence et redondance.

Pour ce qui est du choix des opérateurs d'agrégation, Hall (2000) propose d'utiliser la moyenne arithmétique pour agréger les scores de pertinence et de redondance élémentaires. L'opérateur Agg_J choisi est issu de la théorie de la mesure pour les sciences du comportement (Ghiselli, 1964) :

$$J(\mathcal{W}) = \frac{q \times pert\left(\mathcal{W}\right)}{\sqrt{q + q\left(q-1\right) red\left(\mathcal{W}\right)}}$$

D'autres types d'agrégateurs peuvent évidemment être retenus. Si Ding et Peng (2003) utilisent également la moyenne arithmétique pour Agg_p et Agg_r, ils envisagent en revanche deux autres opérateurs Agg_J : la différence et le rapport entre la pertinence et la redondance. Citons également les travaux de Renteria et Tanscheit (2005). Les auteurs ont recours aux mêmes opérateurs Agg_p et Agg_r, mais emploient pour Agg_J une différence pondérée :

$$J(\mathcal{W}) = pert\left(\mathcal{W}\right) - \beta \times red\left(\mathcal{W}\right)$$

Le paramètre β permet d'établir le compromis entre pertinence et redondance. Ce ne sont là que quelques exemples de choix d'opérateurs, révélateurs d'une certaine tendance : les opérateurs Agg_p et Agg_r sont très souvent des opérateurs de compromis, sans que ce choix soit justifié d'une quelconque façon.

Pour l'agrégation des scores de pertinence élémentaires, il peut être souhaitable de vouloir un sous-ensemble tel que tous les attributs soient suffisamment corrélés à la classe, sans que le bon comportement de l'un puisse compenser le mauvais comportement d'un autre. Si tel est le cas, les opérateurs de renforcement tels que les t-normes sont mieux appropriés. À l'inverse, il peut être souhaitable de vouloir rejeter un sous-ensemble, si une seule paire d'attributs s'avère trop redondante. On optera alors plutôt pour des opérateurs de renforcement tels que les t-conormes. Wu et Zhang (2004) par exemple ont choisi la plus petite des t-conormes : le maximum. En fonction du domaine, du degré de finesse de la sélection que l'on souhaite obtenir, divers choix sont possibles, qui ne se limitent pas à la moyenne arithmétique.

Mesures de cohérence Selon les définitions 4 et 5, un attribut v est pertinent s'il existe deux exemples de classes différentes qui prennent exactement les mêmes valeurs pour tous les attributs, à l'exception de v. Partant de cette notion de pertinence, Almuallim et Dietterich (1994) ont mis en place l'algorithme FOCUS qui se base sur la notion de cohérence. Il recherche le sous-ensemble minimal d'attributs qui est cohérent.

Deux exemples sont dits incohérents relativement à un ensemble d'attributs, s'ils ont exactement les mêmes valeurs d'attributs mais appartiennent à deux classes différentes. De manière symétrique nous dirons qu'un ensemble d'attributs est incohérent relativement à une base d'exemples, s'il existe dans cette base au moins une paire d'exemples incohérents relativement à cet ensemble. Soit $\mathcal{W} \subseteq \mathcal{V}$ un ensemble de q attributs. Les définitions précédentes se formalisent de la façon suivante :

$$e_i \text{ et } e_j \text{ sont incohérents relativement à } \mathcal{W} \Leftrightarrow \begin{cases} v_{ik} = v_{jk} & \forall k \in \{1, ..., q\} \\ y_i \neq y_j \end{cases}$$

L'incohérence, du point de vue des exemples est une relation ternaire, que l'on notera $Incons_{ex}(e_i, e_j, \mathcal{W})$. L'incohérence d'un ensemble d'attributs $\mathcal{W}$ relativement à la base d'exemples $\mathcal{E}$ est une relation binaire qui s'exprime de la manière suivante :

$$Incons_{att}(\mathcal{W}, \mathcal{E}) \Leftrightarrow \exists i \neq j \in \{1, ..., q\} \text{ tels que } Incons_{ex}(e_i, e_j, \mathcal{W})$$

La cohérence pour un couple d'exemples ou un ensemble d'attributs se définit simplement par la négation de l'incohérence. FOCUS repose sur les deux constats suivants.

- Tout ensemble d'attributs incohérent ne permettra pas de construire un classifieur qui puisse séparer tous les exemples, puisqu'il existe au moins un couple d'exemples incohérents pour cet ensemble d'attributs.
- Parmi les ensembles d'attributs cohérents, autant prendre celui de cardinalité la plus faible afin d'assurer une réduction de la dimension maximale[10].

La mesure d'évaluation utilisée par FOCUS est donc une mesure de cohérence binaire. On a :

$$J(\mathcal{W}) = \begin{cases} 1 & \text{si } Cons(\mathcal{W}, \mathcal{E}) \\ 0 & \text{sinon} \end{cases}$$

Cette mesure est sensible au bruit. Les valeurs de certains attributs, y compris la classe, peuvent être erronées ou manquantes. Il est certes possible, on l'a vu, de substituer de nouvelles valeurs en remplacement des manquantes. Mais l'incertitude liée à cette substitution est grande. Aussi faut-il pouvoir disposer d'une mesure plus fine de cohérence afin de tenir compte de ce bruit potentiel.

Pour pallier cette carence, Dash *et al.* (2000) ont introduit une mesure de cohérence plus robuste, non binaire, qui évalue le degré de cohérence d'un ensemble d'attributs, degré qui appartient à l'intervalle $[0, 1]$. L'évaluation de cette mesure pour un ensemble d'attributs $\mathcal{W}$ et une base d'exemples $\mathcal{E}$ repose sur le calcul du taux d'incohérence de $\mathcal{W}$ relativement à $\mathcal{E}$. Soit $\mathcal{N} = \{e'_1, ..., e'_i\} \subseteq \mathcal{E}$ l'ensemble des exemples de $\mathcal{E}$ qui sont incohérents relativement à $\mathcal{W}$. Le taux d'incohérence correspond à la différence entre le nombre total d'exemples incohérents et le nombre d'exemples de $\mathcal{N}$ qui appartiennent à la classe majoritaire parmi les classes représentées dans $\mathcal{N}$, normalisée par le nombre total d'exemples de la base $\mathcal{E}$.

$$\gamma(\mathcal{W}, \mathcal{E}) = \frac{|\mathcal{N}| - \max_{j=1..K} |\{e \in \mathcal{N}, Classe(e) = j\}|}{|\mathcal{E}|}$$

[10] On parle alors de *biais Min-Attributs*.

La mesure de cohérence est alors simplement définie comme le complément à 1 du taux d'incohérence :

$$J(\mathcal{W}) = 1 - \gamma(\mathcal{W}, \mathcal{E})$$

Liu et Setiono (1996b), en la couplant avec un algorithme de recherche aléatoire de type Las Vegas, ont mis en évidence expérimentalement l'intérêt du taux d'incohérence. Liu *et al.* (1998) ont en outre montré son intérêt d'un point de vue théorique. Il comporte deux atouts majeurs. D'une part, en utilisant un algorithme de hâchage, il est possible de calculer ce taux d'incohérence avec une complexité linéaire en fonction du nombre d'attributs. D'autre part, il est monotone par rapport à l'inclusion ensembliste. La mesure de cohérence qui en dérive l'est donc également. Il est en effet assez aisé de démontrer que l'on a la relation suivante :

$$\begin{aligned} \mathcal{W} \subseteq \mathcal{Z} \quad &\Rightarrow \quad \gamma(\mathcal{W}, \mathcal{E}) \leq \gamma(\mathcal{Z}, \mathcal{E}) \\ &\Rightarrow \quad J(\mathcal{W}) \geq J(\mathcal{Z}) \end{aligned}$$

Cette propriété est, on l'a vu, essentielle pour certains algorithmes de recherche tels que B&B. Ce n'est qu'en utilisant une fonction d'évaluation J monotone que l'on peut garantir l'optimalité du résultat fourni par cette méthode. Nous avons également souligné l'importance de cette propriété pour les algorithmes de recherche séquentielle simple, qui s'appuient de manière implicite sur l'hypothèse de monotonie.

Ajoutons qu'outre ces deux atouts, la mesure de cohérence, contrairement à celles qui sont basées sur une notion de corrélation, est définie directement et explicitement pour des ensembles d'attributs. Il n'est donc point besoin de recourir à un processus d'agrégation pour pouvoir évaluer des ensembles d'attributs autres que les singletons.

Notons cependant que chercher des exemples incohérents n'a de sens que relativement à un ensemble d'attributs discrets. Pour les attributs continus, il faut procéder antérieurement à une phase de discrétisation.

7.3.1.3 Combinaison de filtres et wrappers : méthodes hybrides

Les quatre familles de mesures que nous venons de présenter : mesures de corrélation statistique, de divergence, d'information et de cohérence, sont toutes indépendantes de la tâche finale pour laquelle la sélection est effectuée. Elles sont au cœur des algorithmes de filtrage, par opposition aux *wrappers* qui ont recours à l'algorithme *final* d'induction[11] pour guider la sélection.

Cette catégorisation des méthodes de sélection en fonction du type de mesure de performance a depuis été raffinée pour inclure une troisième classe regroupant les méthodes dites *hybrides*. Elles essaient de combiner les avantages des deux approches : la complexité modérée des filtres et la qualité des solutions fournies par les *wrappers*. Deux types de combinaisons émergent de la littérature.

La première consiste à utiliser l'algorithme final d'induction pour effectuer le réglage de certains paramètres utilisés par un filtre. On limite ainsi le nombre d'exécutions de l'algorithme d'induction et donc la complexité par rapport à un *wrapper* classique. Illustrons cet argument par un exemple.

Fixer un nombre prédéfini d d'attributs à sélectionner est un critère d'arrêt du parcours du treillis de recherche. Il est fréquemment employé dans le cadre des recherches séquentielles (Renteria et Tanscheit, 2005; Peng *et al.*, 2005), mais également aléatoires (Liu et

[11] Nous nommons algorithme *final* d'induction celui qui sera utilisé *in fine* pour apprendre un modèle à partir des attributs sélectionnés.

Setiono, 1996c; Oh *et al.*, 2004). Il s'avère cependant difficile de choisir ce paramètre. Une méthode hybride peut être mise en place pour prendre une décision quant au choix de d (Xing *et al.*, 2001; Das, 2001). Un filtre est utilisé pour identifier le meilleur sous-ensemble d'attributs pour une cardinalité k fixée. En répétant la sélection pour différentes valeurs de k, on obtient plusieurs ensembles d'attributs, chacun étant optimal parmi les ensembles de même cardinalité. Ces ensembles sont ensuite évalués et comparés par l'intermédiaire de l'algorithme final d'induction, afin de n'en retenir qu'un, celui qui permet d'induire le modèle le plus performant.

La seconde méthode de combinaison d'un filtre et d'un *wrapper* repose sur le constat suivant. Les *wrappers* obtiennent d'excellentes performances, généralement meilleures que celles des filtres, mais ils sont inapplicables en grande dimension à cause du coût prohibitif de la phase d'induction qui doit être répétée à chaque évaluation d'un nouvel ensemble d'attributs. Un compromis peut être trouvé en procédant à un filtrage préalable des attributs, afin de réduire de manière conséquente la dimension du problème. Ceci rend alors possible l'utilisation d'un *wrapper* pour obtenir une sélection plus fine des attributs (Kotsiantis et Pintelas, 2004; Cantu-Paz, 2004).

Nous présentons à la figure 7.6 une taxinomie des mesures d'évaluation utilisées pour la sélection d'attributs, qui reprend les caractéristiques principales des différentes mesures que nous venons d'aborder.

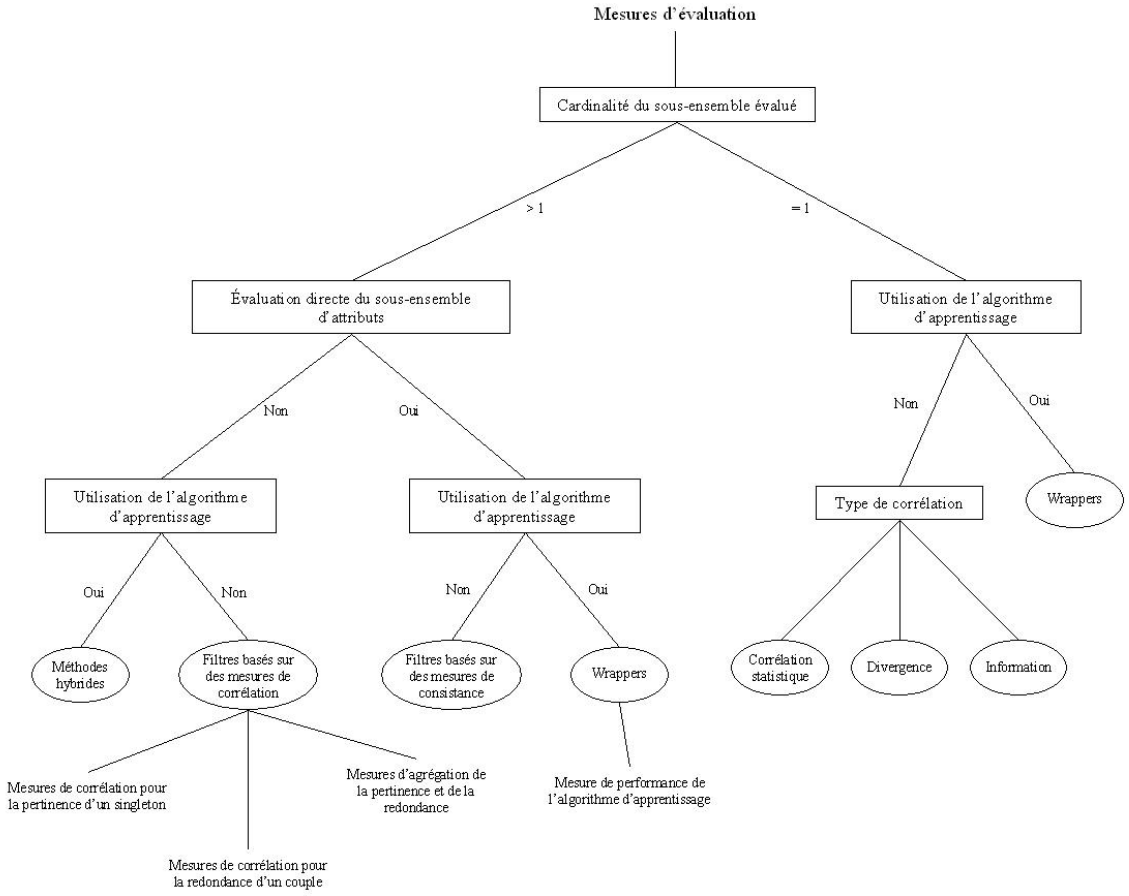

FIG. 7.6 – Une taxinomie des mesures d'évaluation pour la sélection d'attributs

7.3.2 Pondération d'attributs basée sur les instances

Parmi, les méthodes de sélection qui peuvent être construites en combinant l'une des méthodes de recherche avec l'une des mesures d'évaluation présentées précédemment, peu prennent effectivement en compte le contexte lorsqu'elles évaluent la pertinence d'un attribut. C'est le cas des *wrappers* et des algorithmes basés sur une mesure de cohérence, qui évaluent globalement un sous-ensemble d'attributs. Mais ils ne cherchent pas à évaluer directement la pertinence d'un attribut.

Pour tous les autres algorithmes, chaque attribut est évalué indépendamment des autres. Le contexte n'est envisagé que sous l'angle de la redondance. Les algorithmes de construction d'arbres de décision prennent certes en compte le contexte des attributs précédemment sélectionnés pour évaluer la pertinence de tout nouvel attribut, mais le problème persiste pour la sélection des premiers attributs, qui est une étape déterminante de l'induction. C'est la raison pour laquelle Kononenko et Hong (1997) qualifient de *myopes* ces mesures.

7.3.2.1 Objectifs

En classification supervisée, l'objectif, ou du moins l'une de ses interprétations, est de parvenir à discriminer du mieux possible les exemples de classes différentes. Aussi la pertinence d'un attribut peut-elle s'évaluer par l'intermédiaire de sa contribution à la délimitation de la frontière de décision. Cette frontière dépendant de l'ensemble des attributs, une mesure de pertinence fondée sur cette contribution ne sera pas handicapée par la myopie évoquée par Kononenko et Hong.

Pour mesurer cette contribution, il est possible de recourir à un algorithme d'induction, un réseau de neurones par exemple, ce qui sous-entend que l'on retrouve la distinction entre filtres et *wrappers*. Mais contrairement aux méthodes classiques de sélection d'attributs que nous avons présentées jusqu'ici, c'est l'espace des exemples et non celui des attributs qui est parcouru afin d'estimer de manière itérative les poids de chacun des attributs.

Cette différence explique que nous ayons choisi de distinguer cette approche de la sélection d'attributs classique basée sur le parcours du treillis de recherche. Il s'agit là d'un point de vue personnel, la plupart des états de l'art sur le sujet préférant regrouper ces deux familles dans un cadre unifié (Liu et Yu, 2005; Molina *et al.*, 2002). Nous pensons cependant qu'un tel cadre masque des divergences assez nettes, tant conceptuelles que formelles entre les deux approches.

7.3.2.2 Méthodes d'échantillonnage

Les méthodes de parcours de l'espace des instances sont bien moins nombreuses que celles qui ont été évoquées pour parcourir l'espace des attributs. L'objectif n'est pas de chercher le meilleur sous-ensemble d'exemples, mais d'en sélectionner un à partir duquel les poids des attributs seront évalués. Aussi parle-t-on plus volontiers de méthodes d'échantillonnage que de méthodes de recherche dans l'espace des exemples.

On distingue les méthodes exhaustives, qui conservent la totalité des exemples, des méthodes aléatoires qui n'en prennent qu'une partie. Le nombre d'exemples à sélectionner est alors un paramètre de la recherche. Les méthodes d'échantillonnage aléatoires diffèrent par le mode de tirage des exemples, avec ou sans remise, et par la distribution de probabilité à partir de laquelle sont tirés les exemples.

La méthode la plus répandue consiste à effectuer un tirage selon une loi uniforme et sans remise (Kira et Rendell, 1992; Skalak, 1994). Si l'on utilise une distribution uniforme mais avec remise, on retrouve la procédure d'échantillonnage avec auto-amorçage, plus connue

sous sa dénomination anglo-saxone *bootstrap*. Notons que les méthodes d'apprentissage basées sur le *boosting* ont recours à une procédure d'échantillonnage reposant sur un tirage aléatoire avec remise et selon une distribution qui n'est pas uniforme. Cette distribution varie au cours de l'apprentissage pour que les exemples mal classés à un moment donné aient une plus grande probabilité d'être sélectionnés.

7.3.2.3 Pondération à l'aide de *wrappers*

La pondération d'attributs à partir d'un *wrapper* repose sur le principe suivant. Chaque exemple retenu par la procédure d'échantillonnage est classé à l'aide d'un algorithme d'apprentissage. Si l'exemple est mal classé alors les poids des attributs ayant été utilisés pour classer l'exemple, sont diminués. S'il est bien classé, les poids sont augmentés. Le processus de rétro-propagation mis en place dans les réseaux de neurones pour estimer de manière itérative les poids des différentes connexions du réseau en est une illustration. Ce sont les poids des connexions partant de la couche d'entrée qui correspondent aux poids des différents attributs.

Lorsque le classifieur correspond à la méthode du plus proche voisin et que les données sont discrètes, une règle simple de remise à jour des poids, après classification de l'exemple e_j, s'exprime de la manière suivante (Payne et Edwards, 1998) :

$$w_i = \begin{cases} w_i\,(1+\mu) & \text{si } v_{ji} = v_{ki} \\ w_i\,(1-\mu) & \text{sinon} \end{cases} \tag{7.6}$$

où w_i désigne le poids de l'attribut i et e_k est le proche voisin de e_i à partir duquel est effectué la classification. μ est un paramètre réel, positif si e_j a été bien classé et négatif sinon.

7.3.2.4 RELIEF

L'algorithme RELIEF de pondération d'attributs a été développé par Kira et Rendell (1992) afin d'évaluer la pertinence des différents attributs. Son principe est proche de celui que nous venons de présenter. Il est généralement rangé dans la catégorie des filtres. Sa complexité étant suffisamment faible, comme nous le verrons par la suite, il est fréquemment employé pour sélectionner les attributs les plus pertinents, et ce indépendamment du classifieur qui sera mis en place *in fine*. Cette faible complexité et sa simplicité, ont rendu RELIEF très populaire. Il est de plus parfaitement représentatif de cette catégorie d'algorithmes de pondération. Aussi avons-nous décidé de le décrire plus en détail.

Soit $e_i \in \mathcal{E}$. Notons e_j le plus proche voisin de e_i appartenant à la même classe et e_k le plus proche voisin de e_i appartenant à une classe différente. Si l'on note $Dist$ une mesure de distance sur l'espace de dimension p, nous avons :

$$\begin{aligned} e_j &= \arg\min\left(Dist\left(e_i, e_h\right), e_h \neq e_i \in \mathcal{E}, y_i = y_h\right) \\ e_k &= \arg\min\left(Dist\left(e_i, e_h\right), e_h \neq e_i \in \mathcal{E}, y_i \neq y_h\right) \end{aligned}$$

$Dist$ peut être définie comme l'agrégation des distances élémentaires définies sur chacun des attributs. Notons $dist_l$ cette distance élémentaire pour l'attribut l. Par exemple, si $Dist$ est la distance euclidienne, on a $Dist = \sqrt{\sum_{l=1}^{p} dist_l^2}$.

Un attribut sera d'autant plus pertinent qu'il contribue à séparer les exemples de classes différentes et qu'il ne contribue pas à séparer des exemples de même classe. On s'intéressera donc à la contribution locale d'un attribut, pour deux exemples voisins de même classe e_i et e_j, et pour deux exemples de classes différentes e_i et e_k. Le poids de l'attribut v_l sera

croissant en fonction de $dist_l(e_i, e_k)$ et décroissant en fonction de $dist_l(e_i, e_j)$. On retrouve ici les caractéristiques mises en évidence par l'équation 7.6 de la remise à jour des poids dans le cadre de l'algorithme du plus proche voisin.

Toute mesure de distance élémentaire peut être adoptée. Elles n'ont pas besoin d'être identiques pour tous les attributs. Il ne pose donc aucun problème d'appliquer cette méthode sur des données hétérogènes, contenant attributs symboliques et numériques. Dans RELIEF par exemple elle est définie de la manière suivante :

$$dist_l(e_i, e_j) = \begin{cases} 0 & \text{si } v_l \text{ est symbolique et } v_{il} = v_{jl} \\ 1 & \text{si } v_l \text{ est symbolique et } v_{il} \neq v_{jl} \\ \min\left(\frac{|v_{il} - v_{jl}|}{dist_l^{max}/2}, 1\right) & \text{sinon} \end{cases}$$

où $dist_l^{max} = \max(dist_l(e, h), \forall e, h \in \mathcal{E})$.

RELIEF utilise une procédure d'échantillonnage aléatoire classique (sans remise et selon une distribution uniforme, notée $alea$) pour sélectionner un sous-ensemble de m exemples $\mathcal{H} \subseteq \mathcal{E}$. Pour chacun des exemples de $\mathcal{H}$, les poids des attributs sont remis à jour, selon le principe décrit dans l'algorithme 8. L'ensemble des poids est noté $\mathcal{W}$, celui qui correspond à l'attribut v_r étant noté w_r.

Algorithme 8 Algorithme RELIEF

Entrées: $m, \mathcal{V} = \{v_1, ..., v_p\}$
Sorties: $\mathcal{W} = \{w_1, ..., w_p\}$
  Début
  Pour $l = 1..p$ **Faire**
    $w_l = 0$
  Fin Pour
  Pour i=1..m **Faire**
    $e_i = alea(\mathcal{E})$
    $e_j = \arg\max(Dist(e_i, e_h), e_h \neq e_i \in \mathcal{E}, y_i = y_h)$
    $e_k = \arg\max(Dist(e_i, e_h), e_h \neq e_i \in \mathcal{E}, y_i \neq y_h)$
    Pour $l = 1..p$ **Faire**
      $w_l = w_l + \frac{dist_l(e_i, e_k)}{m} - \frac{dist_l(e_i, e_j)}{m}$
    Fin Pour
  Fin Pour
  Renvoyer $\mathcal{W}$
  Fin

RELIEF doit calculer, pour chaque e_i, $|\mathcal{E}| = n$ distances pour trouver les plus proches voisins e_j et e_k. Il faut alors remettre à jour les p poids. Ce processus étant répété m fois, la complexité de RELIEF est en $\mathcal{O}(m \times n \times p)$. Si l'on préfère choisir une recherche exhaustive et donc parcourir l'ensemble des exemples, la complexité devient $\mathcal{O}(n^2 \times p)$. Rappelons que la majorité des algorithmes de sélection que nous avons abordés à la section précédente ont une complexité quadratique en fonction du nombre d'attributs. Pour les bases de données contenant plus d'attributs que d'exemples, RELIEF s'avère donc particulièrement avantageux.

7.3.2.5 Extension de RELIEF

RELIEF ne permet pas de traiter les problèmes multi-classes. La version ReliefF de Kononenko (1994) en est une extension, qui permet de surmonter cette difficulté. Pour l'exemple e_i, au lieu de chercher le plus proche voisin e_k qui n'est pas de la même classe, ReliefF considère le plus proche voisin dans chacune des classes autres que celles de e_i. On considère ainsi l'ensemble des frontières de décision possibles. Kononenko a également fait en sorte de pouvoir prendre en compte $k > 0$ plus proches voisins. Dans cette version, la formule de remise à jour des poids s'écrit :

$$w_l = w_l - \sum_{e \in \mathcal{V}_k(e_i, y_i)} \frac{dist_l\left(e_i, e\right)}{k \times m} + \sum_{c \neq y_i} P\left(c\right) \sum_{f \in \mathcal{V}_k(e_i, c)} \frac{dist_l\left(e_i, f\right)}{k \times m}$$

où $\mathcal{V}_k\left(e_i, c\right)$ représente l'ensemble des k exemples les plus proches de e_i parmi ceux qui sont de classe c et $P\left(c\right)$ désigne la probabilité *a priori* de la classe c. L'extension proposée par Kononenko présente également l'avantage de pouvoir prendre en compte les valeurs manquantes, grâce à une modification de la mesure de distance élémentaire $dist_l$ (Kononenko, 1994).

7.3.2.6 De la pondération à la sélection d'attributs

En sortie de RELIEF ou ReliefF, nous disposons des poids associés à chacun des attributs. Ces poids reflètent leur pertinence vis-à-vis de la tâche de classification et permettent de réaliser un ordonnancement des attributs. Rappelons cependant que nous sommes avant tout intéressé par le problème de sélection et non d'ordonnancement des attributs. Il nous faut donc définir ensuite la façon dont sera effectuée la sélection.

Lorsque le nombre d'attributs d souhaité est connu, cela ne pose aucun problème : il suffit de ne conserver que les d attributs ayant les poids les plus élevés. Lorsque tel n'est pas le cas, ce qui est assez fréquent, il est possible d'utiliser l'algorithme final d'induction pour choisir le nombre d qui maximise les performances en classification. Cela revient à construire une méthode hybride. Il est aussi envisageable de fixer un seuil minimal de pertinence, souvent fixé à 0. Ne sont alors conservés que les attributs ayant un poids supérieur à ce seuil. Considérons la liste des poids obtenus par ReliefF, triés par ordre décroissant : $w_{\sigma(1)} > ... > w_{\sigma(n)}$, où σ désigne une permutation de {1,...,n}. Liu *et al.* (2002) ont proposé de définir d à partir des $n-1$ écarts entre $w_{\sigma(i)}$ et $w_{\sigma(i+1)}$.

$$d = \min_{j=1..n-1} \left(j \text{ tel que } \left(w_{\sigma(j)} - w_{\sigma(j+1)}\right) > \frac{1}{n-1} \sum_{i=1}^{n-1} w_{\sigma(i)} - w_{\sigma(i+1)} \right)$$

7.3.2.7 Méthodes voisines de RELIEF

RELIEF ou plutôt sa version étendue ReliefF fait partie des algorithmes standards de sélection d'attributs, mais comme nous l'avons suggéré, il ne s'agit que d'un représentant d'une classe de méthodes. D'autres approches ont été élaborées avec le même objectif. Raman et Ioerger (2002) ont ainsi recours à une notion de cohérence pour évaluer la contribution d'un attribut, tandis que Liu *et al.* (2002) envisagent l'utilisation d'une procédure d'échantillonnage actif, qui vise à choisir efficacement les m exemples à partir desquels la contribution de chaque attribut sera estimée. Ceci implique que la distribution de probabilité à partir de laquelle sont tirés les exemples n'est pas, contrairement à ReliefF, uniforme.

Hong (1997) a pour sa part introduit la notion de mérite contextuel (CM). L'idée sous-jacente est voisine de celle qui sous-tend ReliefF, mais les deux méthodes diffèrent sur deux

points. Pour estimer le mérite contextuel, ne sont utilisés que les exemples d'une classe autre que celle de l'exemple en cours d'analyse e_i. De plus, contrairement à ReliefF, au voisinage de e_i la contribution de chaque attribut à la tâche de classification est pondérée par une mesure de la difficulté de cette tâche à proximité de e_i. Cette difficulté est mesurée par une fonction décroissante de la distance $Dist$ qui sépare e_i des voisins appartenant aux autres classes. Ceci permet de relativiser la contribution de chaque attribut. Plus e_i sera proche de la frontière de décision et plus il sera important de considérer la contribution des attributs. Inversement, plus e_i sera éloigné de cette frontière et plus il sera facile de classer correctement e_i. Il conviendra alors d'accorder moins d'importance à la contribution respective de chacun des attributs. La remise à jour des poids de chaque attribut s'écrit, avec les mêmes notations que précédemment :

$$w_l = w_l + \sum_{c \neq y_i} \sum_{e \in \mathcal{V}_k(e_i, c)} \frac{dist_l\left(e_i, e\right)}{Dist\left(e_i, e\right)^2}$$

Pour conclure sur les méthodes de pondération utilisées en sélection d'attributs, nous décrivons dans la figure 7.7 leurs spécificités par le biais d'une taxinomie. Nous nous sommes attaché à construire cette taxinomie de telle sorte que ressortent les degrés de liberté sur lesquels il est possible de jouer lorsque l'on souhaite développer une méthode de pondération basée sur le parcours de l'espace des exemples.

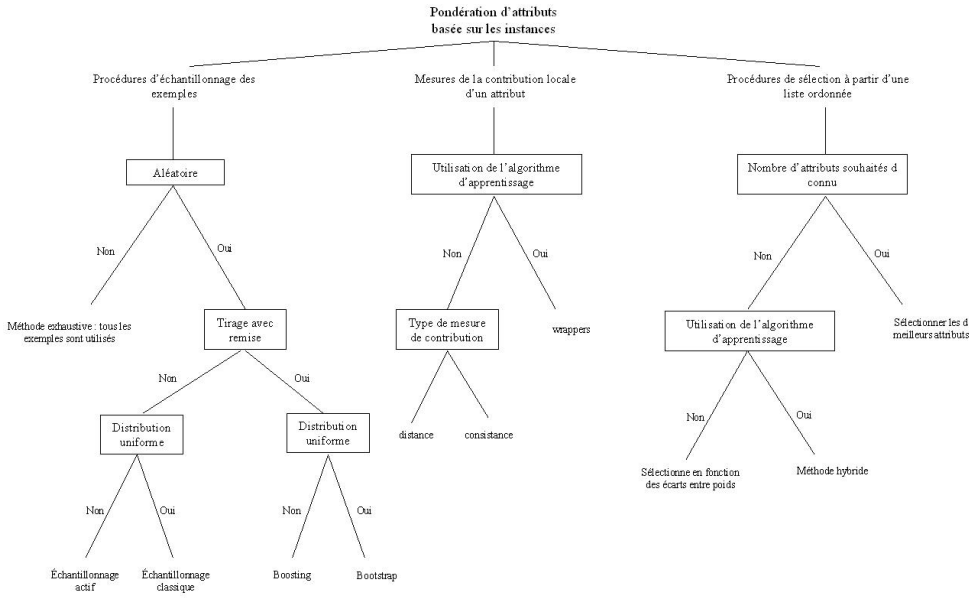

FIG. 7.7 – Une taxinomie des méthodes de pondération pour la sélection d'attributs : parcours de l'espace des exemples

7.3.3 Recherche des attributs pertinents non redondants

Grâce à la prise en compte du contexte, ReliefF et les méthodes qui lui sont apparentées sont particulièrement efficaces (faible complexité et bonnes performances) pour mesurer la pertinence des différents attributs. En revanche, elles ne prennent pas du tout en compte la redondance entre attributs, ce qui peut nuire aux performances de la tâche d'apprentissage (Bins et Draper, 2001). Ces remarques s'appuient sur les expériences de Molina *et al.* (2002), qui ont analysé le comportement de différents algorithmes de sélection d'attributs vis-à-vis de la pertinence et de la redondance.

Nous avons vu qu'à l'inverse, les méthodes de sélection qui parcourent le treillis de recherche dans l'espace des attributs évaluent les sous-ensembles d'attributs en prenant en compte la redondance. Elles ont cependant en général une complexité quadratique en fonction du nombre d'attributs. Il faut également préciser que dans l'évaluation de chaque nouvel ensemble d'attributs, c'est l'analyse de la redondance qui est la partie la plus coûteuse. Afin de limiter le coût de la sélection d'attributs tout en assurant le contrôle de la redondance du sous-ensemble final, diverses solutions ont été proposées que l'on peut toutes considérer comme des instanciations d'un troisième modèle de sélection d'attributs.

7.3.3.1 Découplage des analyses de la pertinence et de la redondance

Si nous revenons à l'analyse de la notion de pertinence que nous avons menée à la section 7.2.1, l'objectif de la sélection d'attributs est d'identifier les attributs fortement et faiblement pertinents. Les premiers doivent être conservés quoi qu'il arrive, tandis que nous devons faire un tri parmi les derniers afin de supprimer ceux qui sont redondants. Ceci a conduit Yu et Liu (2004) à définir un nouveau paradigme pour la sélection d'attributs. À notre connaissance, bien que des travaux antérieurs puissent être rattachés à ce paradigme (Bins et Draper, 2001; Xing *et al.*, 2001), ce sont les premiers à l'avoir formalisé. L'idée sous-jacente réside dans le découplage de l'analyse de la pertinence et de la redondance. Plutôt que d'intégrer dans une même mesure les analyses de la pertinence et de la redondance, afin de guider le parcours du treillis, il est possible de se passer d'un parcours effectif du treillis. Il est procédé dans un premier temps à l'analyse de la pertinence, afin de supprimer tous les attributs qui sont non pertinents, avant de réduire encore le sous-ensemble d'attributs finalement sélectionné par une analyse de la redondance.

Ce troisième modèle, dont l'objectif est de parvenir à identifier un sous-ensemble d'attributs pertinents et non redondants avec une complexité aussi limitée que possible, est essentiellement employé pour traiter les problèmes en grande dimension. Les différents algorithmes de sélection d'attributs que l'on ranger dans cette catégorie se différencient en fonction des techniques choisies pour effectuer l'analyse de la pertinence et de la redondance.

7.3.3.2 Analyse de la pertinence

Dans un souci de minimisation des temps de traitement, l'analyse de la pertinence se ramène à l'ordonnancement des différents attributs selon une mesure de pertinence donnée. Yu et Liu (2004) ont ainsi recours à l'incertitude symétrique $SU(v_i, y)$ entre un attribut v_i et la classe y, pour mesurer la pertinence de v_i au regard de la tâche de classification. N'importe quelle mesure de corrélation $m_{cor}(v_i, y)$ peut être employée en lieu et place de l'incertitude symétrique.

Par exemple Xing *et al.* (2001) ont opté pour le gain d'information ou information mutuelle, tandis que Wu et Zhang (2004) ont choisi le gain d'information équilibré que nous avions noté B_g. Bins et Draper (2001) adoptent pour leur part une solution quelque

peu différente puisqu'ils utilisent ReliefF pour réaliser l'ordonnancement des attributs, s'appuyant sur le fait qu'il s'agit d'une méthode rapide et performante pour identifier les attributs pertinents. La complexité de ReliefF est linéaire en fonction du nombre d'attributs, ce qui correspond à la complexité de l'analyse de la pertinence lorsqu'elle repose sur l'évaluation de la corrélation de chaque attribut avec la classe.

Une fois que les attributs ont été triés selon leur pertinence vis-à-vis de la tâche de classification, il faut alors choisir ceux qui seront supprimés. Nous avons déjà mentionné ce point à propos de ReliefF. Cela passe généralement par le choix d'un seuil de pertinence minimale, en deçà duquel les attributs sont considérés comme non pertinents.

7.3.3.3 Analyse de la redondance

Parmi les attributs retenus par l'analyse de la pertinence, il faut encore éliminer ceux qui sont redondants. Jusqu'à présent nous présentons la redondance comme s'il s'agissait d'une relation monadique. Or, la redondance n'est jamais la caractéristique d'un attribut pris indépendamment des autres. La redondance d'un attribut fait référence de manière implicite à la variable *classe*, mais aussi à un ensemble d'attributs. Il serait ainsi plus juste de parler de la redondance d'un attribut vis-à-vis d'un ensemble d'attributs en vue d'une tâche de classification. Mais pour alléger les notations, lorsqu'il n'y aura pas d'ambiguïté possible, nous omettrons de préciser ces deux points.

Le concept de couverture de Markov introduit par Koller et Sahami (1996) et dont la définition a été donnée à la section 7.2.1 (voir définition 10) permet de formaliser cette notion de redondance. Selon la définition 11, un attribut v_i est redondant par rapport à un ensemble $H_i \subset \mathcal{V} - \{v_i\}$ si $\exists M_i \subseteq H_i$ tel que v_i est indépendant de $(H_i - M_i - \{v_i\}, y)$ conditionnellement à M_i. Cette définition suggère l'utilisation d'un processus d'élimination des attributs redondants plutôt que de sélection des attributs non redondants.

La phase d'analyse de la redondance consiste à parcourir l'ensemble des attributs H retenus par l'analyse de la pertinence, afin de supprimer ceux pour lesquels on parvient à identifier une couverture de Markov. Cette méthode pourrait sembler fallacieuse dans la mesure où un attribut supprimé à un moment donné parce qu'une couverture de Markov le rendait redondant pourrait très bien ne plus avoir de couverture de Markov une fois que d'autres attributs ont été supprimés. Mais Koller et Sahami (1996) ont montré qu'il n'en était rien, justifiant ainsi la cohérence de cette procédure d'élimination.

Le problème de l'analyse de la redondance n'est pas résolu pour autant. Il est en effet très difficile, voire impossible de trouver une couverture de Markov pour un attribut donné. Aussi diverses approximations ont-elles été proposées afin de pouvoir procéder efficacement à la recherche d'une couverture de Markov. Koller et Sahami (1996); Xing *et al.* (2001) proposent ainsi de n'envisager pour un attribut v_i qu'un seule couverture de Markov, $\mathcal{M}_i$, correspondant à l'ensemble des k variables les plus corrélées à v_i. La mesure de corrélation utilisée par Koller et Sahami (1996) repose sur la divergence de Küllback-Leibler, tandis que Xing *et al.* (2001) utilisent la corrélation linéaire de Pearson. Supposant connu le nombre d'attributs souhaités, ils suppriment itérativement l'attribut v_i, pour lequel M_i minimise la divergence de Küllback-Leibler entre $P(y|M_i, v_i)$ et $P(y|M_i)$, l'idée étant que v_i est l'attribut pour lequel on dispose de la meilleure approximation d'une couverture de Markov. Cela revient à considérer que la divergence sus-mentionnée correspond à un degré de redondance et que l'objectif est de supprimer itérativement les attributs présentant le plus haut degré de redondance.

Liu et Yu (2003); Yu et Liu (2004) utilisent une approximation du concept de couverture de Markov qui repose également sur une mesure de corrélation mais issue cette fois-ci de

la théorie de l'information : l'incertitude symétrique[12]. Mais contrairement aux travaux précédents, dans un souci de réduction de la complexité, ils limitent les couvertures de Markov potentielles aux singletons. Ceci les amène à redéfinir l'approximation de ce concept de la manière suivante :

Définition 12 L'attribut v_j forme une couverture de Markov approchée de v_i si et seulement si $SU(v_j, y) \geq SU(v_i, y)$ et $SU(v_j, v_i) \geq \gamma$

où γ est un seuil qu'ils ont fixé de manière heuristique à $SU(v_i, y)$. L'idée sous-jacente est que si v_j et v_i sont suffisamment corrélés, la suppression de l'un des deux n'est pas dommageable, à condition évidemment de conserver l'autre. Sera alors supprimé celui qui apporte le moins d'information sur la classe.

7.3.3.4 Modèle générique de filtre rapide basé sur la corrélation

L'intérêt de cette nouvelle définition réside dans l'algorithme de suppression de la redondance qu'elle permet de mettre en œuvre. En effet, si les attributs sont triés par ordre décroissant de pertinence, il suffit de considérer ces attributs un à un et d'identifier pour chacun les attributs pour lesquels il forme une couverture de Markov approchée, c'est-à-dire ceux qui lui sont suffisamment corrélés (seconde condition de la définition 12) parmi ceux qui sont moins pertinents que lui (première condition de la définition 12). Ces attributs sont alors supprimés.

Pour tout attribut supprimé selon critère, il existera toujours une couverture de Markov approchée. Yu et Liu (2004) l'ont montré pour l'incertitude symétrique, mais il est immédiat de constater que la démonstration ne dépend pas de la mesure de corrélation choisie. Aussi peut-on généraliser leur procédure en considérant n'importe quelle mesure de corrélation m_{cor}. Nous obtenons alors l'algorithme 9 décrivant un filtre rapide basé sur la corrélation, où les analyses de la pertinence et de la redondance sont réalisées en séquence. Il s'agit d'une généralisation de l'algorithme de Liu et Yu (2003).

De manière générique, rien n'impose que les mesures de corrélation utilisées pour estimer la pertinence et la redondance soient identiques. γ_p et γ_r désignent respectivement les seuils utilisés par l'analyse de la pertinence et de la redondance.

Bins et Draper (2001) ont une optique quelque peu différente. Leur principale innovation réside dans l'utilisation d'une méthode de classification non supervisée pour identifier des groupes d'attributs similaires. Les attributs d'un même groupe sont alors qualifiés de redondants et seul l'un d'entre eux peut être conservé pour apporter l'ensemble de l'information contenue par les attributs du groupe. Sera retenu le meilleur représentant de chaque groupe au regard de la tâche de classification : l'attribut du groupe dont la pertinence est la plus forte.

Les auteurs utilisent l'algorithme des k moyennes pour réaliser la classification des attributs. Mais n'importe quelle méthode de regroupement est envisageable, comme par exemple les c moyennes floues.

Afin d'être le plus générique possible, nous supposerons que chaque attribut v_i appartient à tous les groupes avec un certain degré $\mu_h(v_i)$ où h est l'index du groupe. Dans le cadre probabiliste ce degré correspondra à une probabilité *a posteriori*, alors qu'il s'agira d'un degré d'appartenance dans le cadre de la logique floue. Lorsque le regroupement obtenu est une partition de l'ensemble des variables, le degré ne prendra que deux valeurs, 0 ou 1, suivant le groupe auquel l'attribut appartient effectivement.

[12] Wu et Zhang (2004) ont une approche voisine basée sur le gain d'information équilibré B_g.

Algorithme 9 Filtre rapide basé sur la corrélation

Entrées: $\gamma_r, \gamma_p, red, pert, \mathcal{V} = \{v_1, ..., v_p\}$
Sorties: $\mathcal{V}_{opt}$
1: **Début**
2: // par abus de notation $\mathcal{V}_{opt}$ désigne aussi bien un ensemble qu'une liste d'attributs
3: $\mathcal{V}_{opt} = \emptyset$
4: **Pour** $i = 1..p$ **Faire**
5: **Si** $pert\,(v_i, y) \geq \gamma_p$ **Alors**
6: $\mathcal{V}_{opt} = \mathcal{V}_{opt} \cup \{v_i\}$
7: **Fin Si**
8: **Fin Pour**
9: Trier $\mathcal{V}_{opt}$ par ordre décroissant de $pert$
10: **Pour** $i = 1..taille\,(\mathcal{V}_{opt}) - 1$ **Faire**
11: $F = \mathcal{V}_{opt}.getElmt(i)$
12: **Pour** $j = i + 1..taille\,(\mathcal{V}_{opt})$ **Faire**
13: $G = \mathcal{V}_{opt}.getElmt(j)$
14: **Si** $red\,(F, G) \geq \gamma_r$ **Alors**
15: $S_{best} = \mathcal{V}_{opt} - \{G\}$
16: $j = j - 1$
17: **Fin Si**
18: **Fin Pour**
19: **Fin Pour**
20: **Renvoyer** $\mathcal{V}_{opt}$
21: **Fin**

Il est possible de rattacher cette méthode au cadre analytique présenté précédemment. Considérons la mesure de corrélation red suivante :

$$red\,(v_i, v_j) = Agg_1\,(Agg_2\,(\mu_h\,(v_i)\,, \mu_h\,(v_j))\,, \forall h \in \{1, ..., k\})$$

où Agg_2 est un opérateur agrégeant les degrés d'appartenance au groupe h, de v_i et v_j. L'agrégation de ces scores sur les k groupes est effectuée par l'opérateur Agg_1. Lorsque le regroupement correspond à une partition de l'ensemble des attributs, cette mesure s'écrit simplement :

$$red\,(v_i, v_j) = \begin{cases} 1 & \text{si } v_i \text{ et } v_j \text{ appartiennent au même groupe} \\ 0 & \text{sinon} \end{cases}$$

Un moyen simple d'obtenir un tel comportement est de prendre l'opérateur *maximum* pour Agg_1 et l'opérateur *minimum* pour Agg_2.

Avec une telle mesure, la définition générique de la couverture de Markov approchée peut être conservée :

Définition 13 L'attribut v_j forme une couverture de Markov approchée de v_i si et seulement $pert\,(v_j, y) \geq pert\,(v_i, y)$ et $red\,(v_j, v_i) \geq \gamma$

Bins et Draper (2001) ont choisi d'utiliser ReliefF comme mesure de pertinence $pert$, mais là aussi n'importe quelle mesure peut être choisie. L'utilisation de la classification non supervisée permet de se passer d'une procédure itérative de suppression des attributs possédant une couverture de Markov. La complexité est transférée dans la réalisation du regroupement. Une fois que ce regroupement a été obtenu, l'analyse de la redondance est immédiate. Analysons justement la complexité de ces deux approches.

7.3.3.5 Complexité

Dans les deux cas l'analyse de la pertinence est linéaire en fonction du nombre d'attributs. Nous l'avions déjà mentionné pour ReliefF. C'est également la cas pour l'approche à base de corrélation. Il suffit en effet de calculer la corrélation entre chaque attribut et la classe. Dans les deux approches, c'est l'analyse de la redondance qui s'avère la plus coûteuse.

Le regroupement non supervisé du type k moyennes, nécessite le calcul de la corrélation entre chaque attribut et chaque centre de classe à chaque itération, soit une complexité de l'ordre de $\mathcal{O}(k \times q \times N)$ où N est le nombre d'itérations, q le nombre d'attributs issus de l'analyse de la pertinence et k le nombre de groupes. Lorsque ce nombre n'est pas connu, il est souhaitable de faire varier k, auquel cas la complexité est plutôt de l'ordre de $\mathcal{O}\left(q^2 \times N\right)$ (Bins et Draper, 2001).

En ce qui concerne l'approche à base de corrélation, pour chaque attribut issu de l'analyse de la pertinence, on évalue sa corrélation avec tous ceux qui ont un score de pertinence moins élevé que lui. Il y a donc au pire des cas (lorsqu'aucun attribut n'est estimé redondant) $\frac{q(q-1)}{2}$ corrélations à calculer. La complexité est donc quadratique en q[13]. Les complexités des deux approches sont donc du même ordre. Pour comparaison, les algorithmes reposant sur une recherche séquentielle dans l'espace des attributs évaluent dans le pire des cas $\mathcal{O}\left(p^2\right)$ états du treillis. Chaque évaluation nécessite q calculs de corrélation (la phase la plus coûteuse) pour estimer la redondance d'un nouvel attribut avec les q déjà sélectionnés.

Ce point sur la complexité permet de se rendre compte de l'intérêt des algorithmes effectuant les analyses de la pertinence et de la redondance en séquence au lieu de les combiner dans une mesure d'évaluation complexe qu'il faut calculer pour chaque nouvel état du treillis. Cet intérêt sera d'autant plus marqué que q est faible, c'est-à-dire que le nombre d'attributs sélectionnés par l'analyse de la pertinence sera faible. Il faut cependant prendre garde à ne pas en supprimer trop uniquement sur la base de la pertinence sous peine de dégrader la qualité du sous-ensemble final. En effet, il est fort possible que dans de tels cas, l'analyse de la pertinence conserve beaucoup d'attributs redondants, passant à côté d'attributs un peu moins pertinents mais qui ne sont pas redondants. Il faut donc trouver un compromis. Celui-ci se règle par l'intermédiaire du seuil γ_p.

7.3.3.6 Intérêt du test de Kolmogorov-Smirnov

L'inconvénient des méthodes que nous venons de décrire réside justement dans la difficulté qu'il y a à fixer de manière adéquate les seuils de pertinence et de redondance qui jouent un rôle fondamental dans la sélection[14]. Biesiada et Duch (2005) proposent d'utiliser la distance de Kolmogorov-Smirnov pour évaluer la redondance entre deux attributs. Cette distance permet de mesurer la divergence entre les distributions de probabilité des deux attributs concernés. Un test statistique peut lui être associé afin de juger avec un certain niveau de confiance, si les deux distributions sont identiques ou non. Nous reviendrons plus en détail sur ce test à la section 7.4. L'hypothèse nulle qui est testée lorsque l'on s'intéresse à deux attributs v_i et v_j est la suivante : $H0 : F_{v_i} = F_{v_j}$.

[13] Dans le meilleur des cas : lorsque l'attribut le plus pertinent forme une couverture de Markov approchée pour l'ensemble des autres attributs, seules $q - 1$ corrélations sont calculées.

[14] Pour la version reposant sur le regroupement, le seuil de redondance correspond au nombre de groupes que l'on souhaite créer, qui est tout aussi délicat à fixer. Des méthodes existent cependant qui ne nécessitent pas de connaître à l'avance ce nombre de groupes (Lemoine *et al.*, 2006).

F_v désigne ici la fonction de répartition de la distribution de probabilité de la variable v. Lorsque l'hypothèse nulle n'est pas rejetée les deux attributs sont considérés comme redondants.

L'intérêt de ce test réside dans la simplicité du choix du seuil de redondance. Il faut certes toujours en fixer un : le niveau de confiance souhaité, mais son interprétation est claire. Il correspond à la probabilité que la conclusion du test soit la bonne lorsque l'hypothèse nulle est rejetée. Les valeurs classiques de ce niveau de confiance sont 0.9, 0.95 ou 0.99 suivant la certitude avec laquelle on veut pouvoir conclure que deux attributs ne sont pas redondants. Biesiada et Duch (2005) ont en outre apporté des éléments empiriques qui montrent l'intérêt de cette modification de l'analyse de la redondance par rapport au modèle de Liu et Yu (2003).

Cependant il ne s'agit là que d'une réponse partielle aux problèmes soulevés précédemment : le seuil de pertinence est toujours aussi difficile à établir. Or ce seuil est crucial pour établir le compromis entre complexité et qualité du sous-ensemble sélectionné. Nous verrons à la section 7.4 quelle solution nous avons proposée pour surmonter cette difficulté. Mais avant cela nous présentons à la figure 7.8 les traits caractéristiques des méthodes entrant dans le cadre analytique que nous venons de présenter. Ceci nous permet de donner les éléments complétant la taxinomie générale des méthodes de sélection d'attributs que nous avions initiée à la figure 7.3.

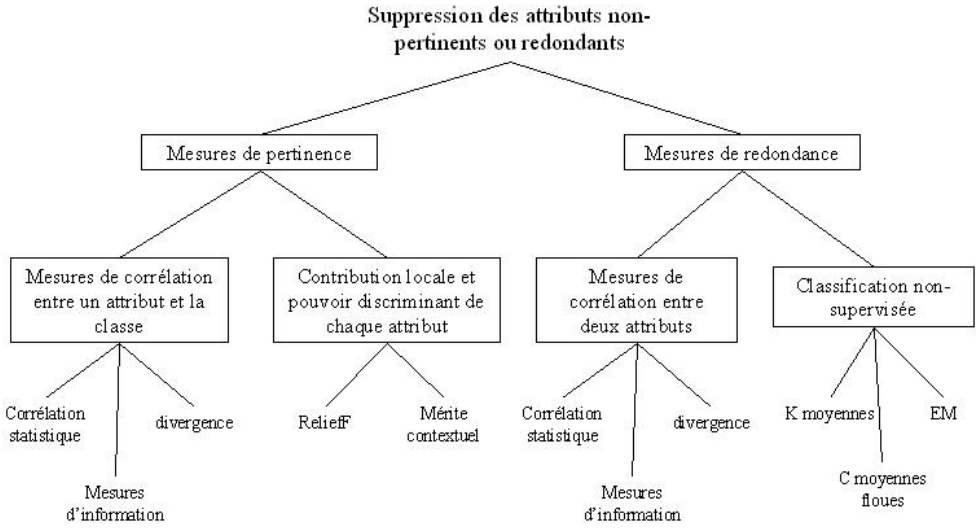

FIG. 7.8 – Taxinomie des techniques de sélection d'attributs basés sur l'analyse séquentielle de la pertinence et de la redondance

7.4 Filtrage basé sur le test de Kolmogorov-Smirnov

Les différentes approches de la sélection d'attributs, ainsi que leurs principales caractéristiques ayant été présentées de manière générique, il convient désormais de les replacer dans le contexte de notre travail : l'évaluation des risques. Rappelons à cet effet quels en sont les enjeux et plus particulièrement les contraintes que nous nous sommes imposées. Ceci nous permettra de statuer, d'un point de vue théorique, sur la méthode qui est la mieux adaptée à notre problème, parmi toutes celles que nous avons abordées.

7.4.1 Motivations

7.4.1.1 Choix d'une méthode de sélection

Au cours de la partie I et plus précisément de la section 2, nous avons contraint de la manière suivante le modèle d'évaluation des risques que nous souhaitions mettre en place.

- Le modèle doit être appris automatiquement en mode supervisé.
- Le modèle doit être transparent : les résultats de l'estimation doivent être interprétables facilement.
- Le modèle doit être le plus objectif possible : il doit intégrer le moins d'hypothèses possibles quant aux facteurs de risque potentiels.

Les deux premières contraintes ont d'ores et déjà été prises en compte. Nous avons en effet choisi de restreindre notre état de l'art aux méthodes de sélection d'attributs applicables à la classification supervisée. Nous avons de plus ignoré les méthodes d'extraction d'attributs qui créent de nouveaux attributs pouvant nuire à la compréhension du modèle final.

La troisième contrainte suggère que le nombre d'attributs décrivant les données doit être aussi grand que possible. Ne faisant aucune supposition sur la nature des facteurs de risque dont il convient de suivre l'évolution dans un processus d'évaluation des risques, il paraît en effet naturel d'essayer d'en collecter le plus grand nombre pour ensuite laisser l'apprentissage décider de ceux qui sont réellement pertinents. Pour que notre modèle soit aussi générique que possible, il nous faudra donc trouver une méthode de sélection qui soit capable de travailler en grande dimension.

Au vu de notre état de l'art, les *wrappers*, dont le coût est prohibitif en grande dimension sont donc à proscrire. Il nous faut ainsi trouver une autre approche que celle qui a été envisagée à la section 3.1 : un *wrapper*, dont la recherche dans le treillis est guidée par un algorithme génétique.

Parmi les filtres, la troisième et dernière approche que nous avons décrite semble la mieux appropriée. Contrairement aux méthodes apparentées à ReliefF, elle tient non seulement compte de la pertinence, mais également de la redondance. Elle le fait en outre avec une complexité qui est en moyenne moindre par rapport à celle des filtres utilisant une mesure d'ensemble pour guider la recherche dans le treillis.

Au vu du bon comportement empirique du filtre introduit par Yu et Liu (2004), tant du point de vue des performances en classification que des temps de calcul, nous avons décidé d'orienter nos recherches vers le modèle de filtre rapide basé sur la corrélation que nous avons proposé pour généraliser l'approche de Yu et Liu (voir algorithme 9). Ayant conservé la dénomination de Yu et Liu pour désigner l'algorithme générique, nous nous référerons à leur algorithme par le sigle FCBF (Fast Correlation-Based Filter).

7.4.1.2 Faiblesses du filtrage rapide basée sur l'incertitude symétrique

Les contraintes que nous venons de recenser correspondent à celles que nous nous sommes imposées. Il en est d'autres qui s'imposent à nous, du fait de la nature des données que nous avons à traiter pour réaliser l'évaluation des risques. Il nous faut également en tenir compte pour choisir la méthode de sélection adéquate. Les données relatives aux conflits armés intra-étatiques qui sont à la base de nos premières expérimentations (section 2.2) présentent quatre caractéristiques qui peuvent influer sur nos choix méthodologiques.

- Le nombre d'attributs est élevé. Ceci confirme la nécessité de se tourner vers des méthodes de sélection d'attributs de faible complexité.
- Nombreuses sont les valeurs d'attributs à ne pas être renseignées. Ce point nous a amené à nous intéresser à la substitution des valeurs manquantes (voir section 6). Nous y reviendrons à la section 7.5 lorsque nous envisagerons l'impact des données manquantes sur la sélection d'attributs.
- La répartition des classes (*crise* et *non-crise*) est fortement deséquilibrée. C'est pour cette raison que nous avons insisté sur l'importance de la mesure d'évaluation d'un classifieur, qui ne saurait être réduite au taux de bonnes classifications (voir section 2.3). Nous y reviendrons à la section 7.6 lorsque nous comparerons empiriquement différentes méthodes de sélection d'attributs.
- Les indicateurs utilisés sont des attributs continus, ou discrets et ordonnés. La capacité des méthodes de sélection d'attributs à pouvoir traiter des attributs continus devra donc également être considérée.

Si FCBF répond à nos exigences de complexité, nous avons souligné à la fin de la section précédente certaines de ses limites : le choix des seuils de pertinence et de redondance est problématique. Après avoir passé en revue les spécificités des données auxquelles la méthode de sélection doit pouvoir s'adapter, nous pouvons ajouter une autre limite : la prise en compte des attributs continus.

La mesure de corrélation utilisée par Yu et Liu pour évaluer la pertinence et la redondance des attributs est en effet une mesure issue de la théorie de l'information : l'incertitude symétrique. Elle ne s'applique pas directement aux attributs continus. Il faudrait pour cela estimer les densités de probabilité utilisées dans le calcul de l'entropie qui est à la base de l'incertitude symétrique. Mais les estimations ne sont généralement pas fiables lorsque le nombre d'exemples est réduit. Aussi Yu et Liu (2004); Hall (2000) ont-ils recours à une méthode de discrétisation pour convertir les attributs continus en variables discrètes sur lesquelles ils peuvent estimer facilement l'incertitude symétrique.

Cela pose problème car une étape supplémentaire doit être réalisée. Ceci accroît la complexité de la sélection d'attributs et augmente le nombre de paramètres du modèle puisque ceux qui sont inhérents à la phase de discrétisation doivent être inclus dans ce modèle. Avant de fixer ses paramètres il faut de plus savoir quelle méthode de discrétisation choisir. Ceci implique que plusieurs méthodes doivent être envisagées, leurs paramètres testés, avant de choisir celle qui convient.

Pouvoir travailler directement sur les données continues permettrait de se passer de cette étape supplémentaire et de conserver un modèle aussi simple que possible. Cette préférence pour des modèles simples correspond au principe du rasoir d'Occam (Domingos, 1999). Ce principe est fréquemment invoqué et appliqué dans les sciences expérimentales pour privilégier les solutions simples.

Nous avons vu en fin de section précédente comment Biesiada et Duch (2005), en remplaçant l'incertitude symétrique par la distance de Kolmogorov-Smirnov dans l'analyse de la redondance, résolvaient partiellement les problèmes sus-mentionnés de FCBF. À partir de cette distance il est en effet possible de construire un test statistique qui permet de conclure, pour un niveau de confiance donné, sur la redondance entre deux attributs. Le seuil de redondance est nettement plus simple à définir, son interprétation étant en outre immédiate. L'autre avantage de cette distance, que nous n'avions pas mentionné jusqu'alors, réside dans le fait qu'elle s'applique directement aux variables continues. Nous désignerons la méthode de Biesiada et Duch par le sigle KSCBF (Kolmogorov-Smirnov Correlation-Based Filter).

7.4.2 Description de la méthode

Nous avons choisi de construire un filtre rapide basé sur la corrélation, en nous appuyant exclusivement sur la distance de Kolmogorov-Smirnov, que nous nommerons par la suite KSF (Kolmogorov-Smirnov Filter). Autrement dit nous proposons d'effectuer non seulement l'analyse de la redondance mais également l'analyse de la pertinence à l'aide de cette distance et du test statistique qui lui est associé. Comme nous l'avons mentionné précédemment, cette distance, qui correspond à une mesure de divergence, peut être utilisée comme mesure de pertinence. Utgoff et Clouse (1996) l'emploient par exemple en lieu et place du gain d'information pour construire des arbres de décision. Voyons maintenant plus en détail comment ce test est appliqué dans chacune des deux étapes de l'analyse.

7.4.2.1 Analyse de la redondance

Pour l'analyse de la redondance entre v_i et v_j, nous avons indiqué précédemment que le test de Kolmogorov-Smirnov consistait à tester l'hypothèse nulle suivante $H0 : F_{v_i} = F_{v_j}$. Pour une variable v, F_v désigne sa fonction de répartition. On a :

$$\forall x \in \mathbb{R}, F_v(x) = \int_{t=-\infty}^{x} p_v(t)\, dt = P(v \leq x)$$

où P désigne la probabilité d'un événement, tandis que que p_v correspond à la densité de probabilité associée à la variable v.

Pour tester $H0$, on utilise la distance de Kolmogorov-Smirnov δ_{KS}, introduite à la section 7.3.1.2. Si $H0$ est correcte, δ_{KS} doit être nulle. Les densités de probabilité p_{v_i} et p_{v_j} sont *a priori* inconnues et très difficiles à estimer. Aussi utilise-t-on plutôt les fonctions de répartition empiriques que l'on peut construire à partir des valeurs de v_i et v_j prises par les n exemples dont nous disposons : $\mathcal{E} = \{e_1, ..., e_n\}$. Nous avons alors $F_{v_i}(x) = \frac{k}{n}$ où k correspond au nombre d'exemples pour lesquels la valeur de v_i est inférieure à x : $k = |\{e_h \in \mathcal{E}, v_{hi} \leq x\}|$.

Pour pouvoir calculer la distance $\delta_{KS}\left(F_{v_i}, F_{v_j}\right)$, l'étape la plus coûteuse correspond au tri des valeurs v_{hi} et v_{hj} qui a une complexité de l'ordre de $\mathcal{O}(n \times \log_2(n))$. Notons que la complexité du calcul de l'incertitude symétrique est du même ordre, puisqu'il faut également procéder à ce tri durant la phase de discrétisation.

Si $H0$ est vraie, il est possible de montrer que l'on a la relation suivante :

$$\forall t \geq 0 \lim_{x \to \infty} P\left(\beta_n \times \delta_{KS}\left(F_{v_i}, F_{v_j}\right) > t\right) = 2 \sum_{k=1}^{\infty} (-1)^{k+1} \exp\left(-2k^2 t^2\right) \tag{7.7}$$

où β_n est un facteur correctif qui tient compte du nombre d'exemples à partir desquels ont été estimées les fonctions de répartition empiriques. En pratique on prendra l'approximation suivante : $\beta_n = \sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}$ (Press *et al.*, 2002).

À partir de cette relation, la construction du test de Kolmogorov-Smirnov est immédiate. Soit $\hat{\delta_{KS}^{i,j}}$ la valeur observée de la distance de Kolmogorov-Smirnov entre les fonctions de répartition empiriques de v_i et v_j. Le membre droit de l'équation 7.7 calculé en prenant $t_{i,j} = \beta_n \times \hat{\delta_{KS}^{i,j}}$ donne en effet directement la p-valeur associée à ce test, notée $p_{i,j}$. Si l'on fixe le niveau de confiance à $1 - \alpha$, et si le nombre d'exemples n est suffisamment grand (en pratique $n > 30$), la probabilité que la valeur théorique de $\beta_n \times \delta_{KS}\left(F_{v_i}, F_{v_j}\right)$ soit au moins égale au $t_{i,j}$ estimé, est égale à $p_{i,j}$ sous $H0$. Si celle-ci est suffisamment faible il est peu vraisemblable que $H0$ soit vraie. Le test consiste donc à utiliser la règle de décision suivante :

$$\begin{array}{lcll} p_{i,j} \leq \alpha & \Rightarrow & \text{Rejeter H0} & v_i \text{ et } v_j \text{ ne sont pas redondants} \\ p_{i,j} > \alpha & \Rightarrow & \text{Accepter H0} & v_i \text{ et } v_j \text{ sont redondants} \end{array}$$

Cette méthode d'analyse de la redondance est exactement celle qui a été mise en place par Biesiada et Duch (2005) (avec $\alpha = 0.05$). Elle correspond au test effectué à la ligne 14 de l'algorithme 9. En observant plus finement cet algorithme, on s'aperçoit que ce test est effectué $(taille\,(\mathcal{V}_{opt}) - i)$ fois lorsque la i-ième variable la plus pertinente est considérée.

On retrouve donc les problèmes de comparaisons multiples évoqués à la section 5. Aussi avons-nous modifié l'analyse de la redondance telle que la présentent Biesiada et Duch (2005), afin de prendre en compte le fait que de multiples comparaisons sont réalisées à chaque étape. Reprenant les recommandations émises alors, nous avons décidé de mettre en place la procédure ascendante de Holland-Copenhaver pour ajuster le taux d'erreurs global.

7.4.2.2 Analyse de la pertinence

Pour l'analyse de la pertinence, nous proposons de considérer le problème sous l'angle suivant. Un attribut v_i sera d'autant plus pertinent vis-à-vis de la variable *classe* y, qu'il permet de discriminer les différences classes (modalités de y). Supposons pour le moment que nous n'avons que deux classes c_1 et c_2. Le pouvoir discriminant de v_i relativement à y, sera d'autant plus grand que la densité de probabilité de v_i sachant c_1 sera différente de la densité de probabilité de v_i sachant c_2. Une mesure de divergence peut alors être employée pour mesurer ce pouvoir discriminant et donc la pertinence d'un attribut.

Nous utilisons la distance $\delta_{KS}\left(F_{v_i|c_1}, F_{v_i|c_2}\right)$ comme mesure de divergence. Si l'on pose l'hypothèse nulle suivante $H0 : F_{v_i|c_1} = F_{v_i|c_2}$, cette distance peut être utilisée pour construire un test permettant de juger de la pertinence de v_i vis-à-vis de y. Lorsque $H0$ est rejetée, nous en concluons que v_i est pertinent. La différence observée entre les deux fonctions de répartition empiriques ne peut être due au seul hasard, du moins la probabilité que l'on se trompe est inférieure à α.

Lorsque nous avons $K > 2$ classes, la méthode n'est plus valable. Pour l'étendre il suffit de réaliser $\frac{K(K-1)}{2}$ tests pour juger du pouvoir discriminant de v_i sur toutes les paires de classes possibles. S'il existe au moins une paire pour laquelle v_i est jugé pertinent alors nous pourrons considérer que v_i peut être utile au processus de classification. Pour chaque attribut nous allons devoir mener en parallèle $\frac{K(K-1)}{2}$ tests. Il nous faudra donc ici aussi, ajuster le taux d'erreurs global. Comme précédemment nous mettrons en place la procédure ascendante de Holland-Copenhaver pour y parvenir.

7.4.2.3 Synthèse

Afin de mettre en évidence les différences entre les algorithmes FCBF, KSCBF et KSF, rappelons la filiation qui existe entre eux. FCBF qui fut proposé par Yu et Liu a servi de source d'inspiration à Biesiada et Duch qui ont développé KSCBF pour résoudre le problème du choix du seuil de redondance. Nous avons nous-même repris et étendu avec KSF ces deux algorithmes, afin d'éviter le recours à la discrétisation des attributs continus et de faciliter non seulement le choix du seuil de redondance mais également celui du seuil de pertinence.

Ces trois filtres sont assez proches et reposent sur des idées similaires. Ils font tous partie de la famille plus générique des filtres rapides à base de corrélation que nous avons décrite par l'algorithme 9. Au sein de ce formalisme commun, les trois filtres ne se différencient que par les choix des mesures de pertinence et de redondance, ainsi que des seuils correspondants γ_p et γ_r. Le tableau 7.1 récapitule ces différences.

TAB. 7.1 – Choix des mesures de pertinence et de redondance, des seuils de pertinence γ_p et de redondance γ_r

Filtre	$pert\,(v_i, y)$	γ_p	$red\,(v_i, v_j)$	γ_r
FCBF	$SU\,(v_i, y)$	$\frac{p}{\log_2 p}$ [1]	$SU\,(v_i, v_j)$	$SU\,(v_i, y)$ [2]
KSCBF	$SU\,(v_i, y)$	0 [3]	$KS\left(F_{v_i}, F_{v_j}\right)$	0.05 [4]
KSF	$\max\left(KS\left(F_{v_i\|c_k}, F_{v_i\|c_h}\right), k \neq h\right)$	$f\,(0.05)$ [5]	$KS\left(F_{v_i}, F_{v_j}\right)$	$f\,(0.05)$ [5]

[1] Il s'agit du nombre d'attributs que l'on souhaite conserver après analyse de la pertinence.
[2] On suppose que v_i est moins corrélé à y que v_j.
[3] Tous les attributs sont donc conservés après analyse de la pertinence.
[4] C'est la p-valeur associée à la mesure de corrélation qui est comparée à ce seuil.
[5] f désigne une fonction faisant varier le seuil original 0.05 afin d'ajuster le taux d'erreurs global.

En remplaçant la comparaison de l'incertitude symétrique à un seuil difficile à définir par un test de Kolmogorov-Smirnov pour effectuer les analyses de pertinence et de redondance, le filtre KSF que nous proposons permet de traiter directement les variables continues. De plus le choix des seuils est beaucoup plus simple puisqu'ils correspondent à des niveaux de confiance d'un test statistique. Ce sont là les deux lacunes du modèle original FCBF que nous cherchions à combler. Mais n'importe quelle statistique applicable directement aux données continues aurait aussi bien pu être envisagé.

Premièrement, les expériences de Biesiada et Duch (2005) laissent penser que la métrique de Kolmogorov-Smirnov peut remplacer l'incertitude symétrique sans dégrader les performances en classification. Pour être plus exact, elles ne permettent pas de montrer le contraire. Duch (2006), passant en revue les différentes mesures d'évaluation utilisées dans le filtrage, note au contraire que les tests de corrélation statistique tels que celui de Pearson ou de Student, sont assez mal adaptés lorsque le nombre d'échantillons est faible et suggère l'utilisation d'un test de permutation pour obtenir des estimations plus fiables des p-valeurs. Les expériences de Radivojac *et al.* (2004) corroborent ce point. Les tests de permutation s'accompagnent cependant d'un accroissement notable de la complexité, ce que nous voulons éviter.

Deuxièment, les tests de Pearson ou de Student appartiennent à la famille des tests paramétriques. Ils reposent sur un certain nombre d'hypothèses qui sont rarement vérifiables. Lorsque le nombre d'échantillons est faible, les statistiques sur lesquelles reposent ces tests ne peuvant plus être estimées de manière fiable. Le test de Kolmogorov-Smirnov fait quant à lui partie des tests non paramétriques et ses estimations sont moins sensibles au nombre d'échantillons.

7.4.3 Limites de la méthode

Nous n'avons jusqu'à présent détaillé que les raisons qui nous ont incité à développer une nouvelle méthode de sélection d'attributs ainsi que les forces de cette méthode. Pour parfaire cette description, il convient maintenant de présenter ses faiblesses.

Le principal inconvénient de notre approche concerne les attributs discrets. Si le test de Kolmogorov-Smirnov s'applique directement sur les attributs continus, ce que nous souhaitions, il nous faut préciser qu'il ne peut s'appliquer théoriquement que sur des données continues. Il s'appuie en effet sur les fonctions de répartition empiriques de variables continues. Contrairement à ce que laissent supposer les expériences sur des données discrètes de Biesiada et Duch (2005), la construction des fonctions de répartition empiriques pour des variables discrètes peut être vide de sens.

Pour préciser cette remarque, il est utile de faire un point sur les différents types d'attributs que nous considérons depuis le début de cette thèse. Une classification de ces différents types a été donnée de façon anodine à la figure 6.14, dans laquelle nous nous sommes efforcé de caractériser une base de données. Nous avons alors distingué les attributs continus des attributs discrets. Parmi les attributs discrets on peut encore distinguer ceux qui sont ordonnés de ceux que nous qualifions de symboliques. Une dernière distinction peut être faite en fonction du domaine de définition de ces attributs. Il peut en effet être fini ou non. Cette classification n'est évidemment pas exhaustive. Nous n'avons en effet pas pris en compte les attributs structurés, les variables linguistiques... Mais cette classification sera suffisante pour notre propos. Seul le caractère ordonné ou non des attributs discrets va nous intéresser.

Pour tout attribut discret et ordonné, étendre la notion de fonction de répartition empirique ne pose aucun problème. En effet, les attributs continus sont traités comme des attributs discrets ordonnés lorsque nous construisons cette fonction. À partir de l'ensemble fini des n exemples qui sont à notre disposition, pour une variable continue v_i, nous pouvons trouver une permutation σ telle que $v_{\sigma(1)i} < ... < v_{\sigma(n)i}$. La fonction de répartition empirique se calcule alors de la manière suivante : $\forall x$

$$F_{v_i}(x) = P(v_i \leq x) = \frac{k}{n}$$

où k est l'entier tel que $v_{\sigma(k)i} < x < v_{\sigma(k+1)i}$. Le point essentiel de ce calcul réside dans l'obtention de σ. Rechercher une telle permutation n'a de sens que si une relation d'ordre peut être construite sur le domaine de définition de v_i.[15].

Le test de Kolmogorov-Smirnov ne peut donc s'appliquer que pour comparer des fonctions de répartition de deux variables continues ou discrètes et ordonnées. Ajoutons cependant que dans le cas discret, il faut encore que la relation d'ordre utilisée pour trier les valeurs des deux variables soit la même, sans quoi la distance de Kolmogorov-Smirnov ne pourra être calculée.

[15] Ces réflexions sont issues de discussions avec J. Biesiada. Je tiens à l'en remercier.

Prenons un exemple. Soit v un attribut correspondant à la forme des yeux et w un attribut correspondant à la couleur des yeux d'une grenouille mexicaine hypnotique du Sud du Sri Lanka. Supposons que v ne puisse prendre que les formes ronde (R), triangulaire (T) et octogonale (O), tandis que w ne peut prendre que les couleurs verte (V), bleue (B) et jaune (J). Supposons de plus que nous disposons de 10 exemples de grenouilles mexicaines hypnotiques du Sud du Sri Lanka dont les caractéristiques sont les suivantes :

- v : 2 R, 3 T, 5 O
- w : 8 V, 1 B, 1 J

v et w sont des attributs discrets et non ordonnés. D'aucuns pourraient avancer que ces attributs sont ordonnés, par la longueur d'onde pour les couleurs et le nombre de côtés pour les formes. Outre le caractère fort discutable de tels ordonnancements, remarquons que la relation d'ordre sur v ne sera pas la même que celle que l'on pourrait trouver sur w.

L'analyse de la redondance entre v et w repose sur la distance de Kolmogorov-Smirnov entre leurs fonctions de répartition empiriques. Du fait de l'absence de relation d'ordre commune, plusieurs solutions sont possibles pour essayer d'approximer cette distance. Nous en donnons deux exemples à la figure 7.9. Ils correspondent aux deux ordonnancements suivants : $R < T < O < V < B < J$ et $R < V < T < O < B < J$. Pour le premier, la distance est de 1, tandis que pour le second la distance est de 0.6. Aucune de ces deux valeurs n'est la bonne, tout simplement parce qu'imposer une relation d'ordre commune à v et w est un acte dénué de sens. La seule conclusion valable est que le test de Kolmogorov-Smirnov ne s'applique pas sur de tels attributs.

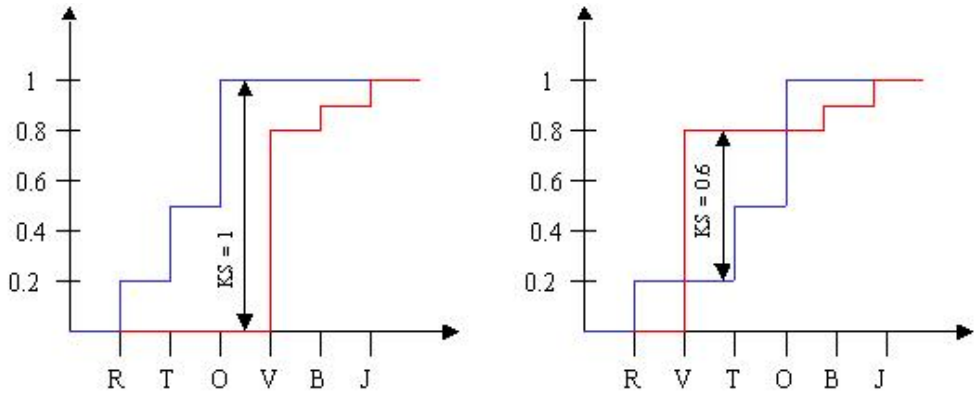

FIG. 7.9 – De l'inadéquation du test de Kolmogorov-Smirnov avec l'analyse de la redondance entre attributs discrets non ordonnés : distance de Kolmogorov-Smirnov pour deux ordres distincts choisis arbitrairement

Pour l'analyse de la pertinence, nous devons comparer les fonctions de répartition de $v|_{c_1}$ et $v|_{c_2}$. Les domaines de définition de ces deux variables étant les mêmes, si l'on trouve une relation d'ordre sur v, elle sera bien commune à $v|_{c_1}$ et $v|_{c_2}$ et donc il peut paraître sensé d'appliquer le test de Kolmogorov-Smirnov. Pour que cela soit le cas, encore faut-il que le choix de la relation d'ordre ne soit pas sujet à controverse. Car là aussi il est facile de trouver un exemple tel que le choix de la relation d'ordre influe sur la distance de Kolmogorov-Smirnov. Supposons que $v|_{c_1}$ et $v|_{c_2}$ ont les caractéristiques suivantes :

- $v|_{c_1}$: 2 R, 3 T, 5 O
- $v|_{c_2}$: 8 R, 1 T, 1 O

La figure 7.10 présente les différentes fonctions de répartition que l'on peut obtenir avec deux ordres différents : $R < T < O$ et $T < R < O$. Alors qu'avec le premier ordre on trouve une distance de 0.6, celle-ci n'est plus que de 0.4 avec le second ordre. Si l'on souhaite utiliser le test de Kolmogorov-Smirnov pour l'analyse de la pertinence des attributs discrets, et *a priori* non ordonnés, il faut donc construire une relation d'ordre pour ces attributs et prendre soin de la justifier, étant donné que tout autre relation d'ordre aurait pu conduire à des résultats différents.

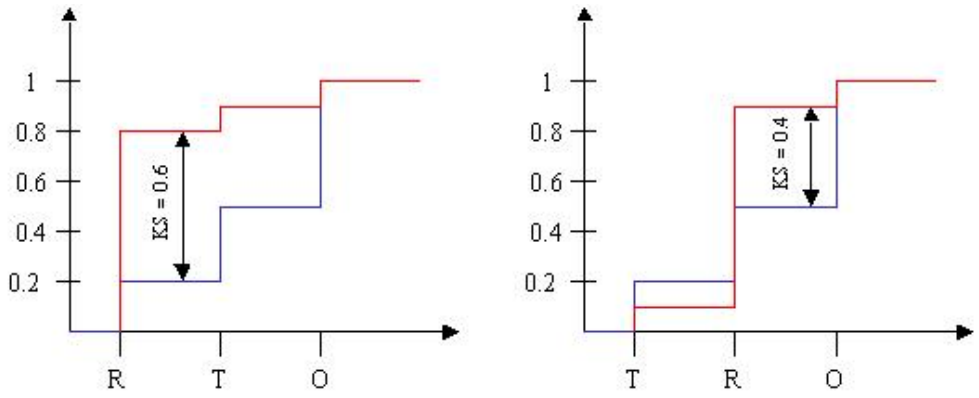

FIG. 7.10 – De l'inadéquation du test de Kolmogorov-Smirnov avec l'analyse de la pertinence d'attributs discrets non ordonnés : distance de Kolmogorov-Smirnov pour deux ordres distincts choisis arbitrairement

Utgoff et Clouse (1996) utilisent pourtant, avec un certain succès, ce test sur des données discrètes non ordonnées pour construire des arbres de décision. Ils n'ont pas besoin de recourir pour cela à la construction et à la justification d'une relation d'ordre particulière. La validité de leur approche réside dans le fait qu'ils ne considèrent que des arbres binaires. Un test portant sur un attribut v comportant k modalités $m_1, ..., m_k$ sera donc toujours de la forme $v = m_i$ ou $v \neq m_i$. Ceci revient à ne considérer que des attributs binaires. À l'attribut v en question, correspondent k attributs $v_1, ..., v_k$ ne prenant que deux valeurs : *vrai* ou *faux*, de telle sorte que l'on a :

$$v_i = \begin{cases} 1 & \text{si } v = m_i \\ 0 & \text{sinon} \end{cases}$$

Or pour de tels attributs, le test de Kolmogorov-Smirnov peut s'appliquer sans problème. En effet, quel que soit l'ordre que l'on utilise pour trier les modalités *vrai* et *faux*, la distance de Kolmogorov-Smirnov sera la même.

Pour synthétiser les remarques précédentes, disons que le test de Kolmogorov-Smirnov est applicable pour l'analyse de la pertinence sur tout type d'attribut, à condition de décomposer tout attribut discret non ordonné et comportant $k > 2$ modalités en k attributs binaires. Il en va de même pour l'analyse de la redondance entre deux attributs, à condition que leurs domaines de définition soient identiques ou du moins compatibles.

7.5 Substitution et filtrage

Nous avons jusqu'ici présenté différents algorithmes de sélection d'attributs en occultant le problème des valeurs manquantes. Or il s'agit d'un problème non négligeable qui affecte la plupart des données réelles. Nous avons certes déjà abordé ce sujet à la section 6, mais nous l'avons fait indépendamment du problème de sélection d'attributs. En pratique, nous allons

être amené à combiner traitement des données manquantes et sélection d'attributs afin de construire un classifieur aussi performant que possible. Aussi souhaitons-nous dans cette section étudier plus précisément la chaîne d'apprentissage dans son ensemble et analyser la façon dont les différents maillons de la chaîne influent les uns sur les autres.

7.5.1 Stratégies de combinaison

Deux solutions sont théoriquement envisageables pour construire la chaîne globale d'apprentissage, suivant lequel des deux prétraitements nous souhaitons placer en début de chaîne, l'algorithme d'apprentissage restant bien évidemment en fin de chaîne.

Il est possible de commencer par substituer les valeurs manquantes avant de réaliser la sélection d'attributs. Cela permet de ne pas modifier les méthodes de sélection d'attributs, puisqu'elles disposeront en entrée d'une base de données complète. Mais il est aussi possible de commencer par sélectionner un sous-ensemble d'attributs avant d'effectuer la substitution des valeurs manquantes. Cette approche présente l'avantage non négligeable de réduire les temps de traitement. La suppression d'attributs opérée durant le filtrage implique en effet une diminution du nombre de valeurs manquantes qu'il faudra substituer.

On peut de plus supposer que la suppression des attributs redondants et non pertinents permette d'améliorer la qualité de la substitution effectuée par les méthodes qui utilisent les autres variables pour construire un modèle prédictif d'une variable incomplète donnée. Il s'agit des méthodes dont l'espace de travail est celui des variables, par opposition à l'espace des exemples (voir à ce sujet notre taxinomie des méthodes de subtitution des valeurs manquantes, figure 6.4). Il serait donc intéressant de voir si l'on peut placer un filtre avant la substitution des valeurs manquantes dans la chaîne d'apprentissage, sans dégrader outre mesure les performances de cette chaîne.

7.5.2 Filtrage de données incomplètes

Réaliser le filtrage d'attributs en amont de la substitution suppose que les filtres utilisés sont capables de traiter des bases de données incomplètes. Or la plupart des techniques abordées jusqu'ici, y compris celle que nous avons proposée, ont été développées pour traiter des bases de données complètes. Seul l'algorithme ReliefF fait exception. Kononenko (1994) a en effet étendu RELIEF afin de rendre l'algorithme utilisable même lorsque certaines valeurs manquent. La mesure de distance, qui est centrale pour l'application de RELIEF, a été modifiée pour pouvoir comparer deux vecteurs incomplets.

7.5.2.1 Ignorer les valeurs manquantes

Le moyen le plus simple d'utiliser une méthode de sélection d'attributs, lorsque certaines valeurs manquent est encore d'ignorer ces valeurs. Mais cela revient à se priver d'informations potentiellement utiles pour déterminer la pertinence et la redondance des différents attributs. Timm *et al.* (2003) notent ainsi que la prise en compte des valeurs manquantes permet d'améliorer les performances d'un algorithme de classification non supervisée.

7.5.2.2 Prise en compte des valeurs manquantes

Lorsque les mesures d'évaluation (pertinence ou redondance) des méthodes de sélection d'attributs s'appliquent sur des attributs discrets, la prise en compte des valeurs manquantes peut se faire simplement en ajoutant une modalité pour chaque attribut, à laquelle correspondra toute valeur manquante. Les valeurs manquantes seront alors traitées comme n'importe laquelle des modalités existantes. C'est ce qui est proposé pour l'algorithme CFS

(Hall, 2000), qui repose sur l'incertitude symétrique (voir l'implémentation qui en est faite par Hall lui-même dans Weka 3.4.7 (Witten et Frank, 2005)).

7.5.2.3 Prise en compte des valeurs manquantes par KSF

Pour les méthodes qui traitent directement les attributs continus, sans passer par une phase de discrétisation, la prise en compte des valeurs manquantes n'est pas aussi immédiate. Le filtre de Kolmogorov-Smirnov que nous avons introduit précédemment est dans ce cas de figure, contrairement à la plupart des filtres que nous avons évoqués.

Rappelons que KSF repose sur la comparaison de fonctions de répartition empiriques. Ignorer les valeurs manquantes est possible et ne demande aucune modification de notre algorithme pour que des données incomplètes puissent être traitées. Mais cela peut s'avérer être une piètre stratégie. En effet, cela conduit à réduire le nombre d'échantillons à partir desquels les fonctions de répartition empiriques seront estimées. Si ce nombre est trop faible, les estimations ne seront pas fiables et on peut supposer que cela dégrade les performances de notre filtre. Il est donc essentiel de disposer d'une seconde stratégie, voisine de celle qui a été évoquée pour les données discrètes, qui permette de considérer une valeur manquante comme une modalité parmi les autres.

Pour mettre en œuvre une telle stratégie nous proposons de mesurer la corrélation entre deux variables (pour la pertinence et la redondance), non pas uniquement à partir de la distribution des valeurs observées mais également à partir de celle des valeurs manquantes. Ceci peut être effectué par l'adjonction d'un test du χ^2 au test de Kolmogorov-Smirnov, l'idée sous-jacente étant que deux attributs peuvent être considérés comme corrélés si les deux distributions des valeurs observées sont suffisamment proches **et** si les deux distributions des valeurs manquantes le sont également. Le test du χ^2, que l'on peut interpréter comme un équivalent du test de Kolmogorov-Smirnov pour les variables discrètes, est alors utilisé pour comparer les distributions des valeurs manquantes.

Soient G et H deux attributs que nous souhaitons comparer pour évaluer leur redondance ou la pertinence de l'un des deux (auquel cas l'autre doit correspondre à la variable *classe*). Notons H^o et G^o les parties observées de ces deux attributs, H^m et G^m les attributs binaires indiquant la présence ou l'absence des valeurs de H et G. Le test de Kolmogorov-Smirnov (KS) est employé pour estimer la corrélation entre H^o et G^o, tandis que le test du χ^2 est employé pour estimer la corrélation entre H^m et G^m. Les algorithmes 10 et 11 décrivent la procédure employée pour combiner ces deux tests durant les analyses de pertinence et de redondance.

7.5.3 Synthèse

Nous avons donc trois stratégies possibles pour construire une chaîne d'apprentissage globale :

- (A) substituer les valeurs manquantes et sélectionner les attributs ensuite
- (B) sélectionner les attributs à partir des valeurs observées uniquement, et substituer ensuite les valeurs manquantes
- (C) sélectionner les attributs en tenant compte des valeurs manquantes et substituer ensuite les valeurs manquantes

La nouvelle approche que nous avons introduite pour le filtrage d'attributs répond théoriquement à nos attentes. Grâce au test de Kolmogorov-Smirnov, elle permet de traiter directement les attributs continus, le choix des seuils de pertinence et de redondance étant

Algorithme 10 KSF : analyse de la pertinence pour bases de données incomplètes

Entrées: $y, G \in \mathcal{V} = \{v_1, ..., v_p\}$
Sorties: Pert // booléen indiquant la pertinence de G
Début
Pour $i \neq j \in \{1, ..., K\}$ // K classes **Faire**
$H0_{i,j}$: $F_{G^o|_{c_i}} = F_{G^o|_{c_j}}$
Si $H0_{i,j}$ est rejetée par le test KS // Ajustement ascendant de Holland-Copenhaver **Alors**
Renvoyer Pert=Vrai
Fin Si
Fin Pour
Pour $i \neq j \in \{1, ..., K\}$ **Faire**
$H0_{i,j}$: $F_{G^m|_{c_i}} = F_{G^m|_{c_j}}$
Si $H0_{i,j}$ est rejetée par le test du χ^2 // Ajustement ascendant de Holland-Copenhaver **Alors**
Renvoyer Pert=Vrai
Fin Si
Fin Pour
Renvoyer Pert=Faux
Fin

grandement facilité. L'adjonction d'un test du χ^2 rend en outre possible la prise en compte des données incomplètes. Il nous faut désormais confronter cette nouvelle technique à la réalité afin de voir si ces avantages théoriques se traduisent en avantages empiriques.

Algorithme 11 KSF : analyse de la redondance pour bases de données incomplètes

Entrées: $H, G \in \mathcal{V} = \{v_1, ..., v_p\}$
Sorties: Red // booléen indiquant la redondance entre H et G
Début
$H0$: $F_{G^o} = F_{H^o}$
Si $H0$ est rejetée par le test KS **Alors**
Renvoyer Red=Faux
Fin Si
$H0$: $F_{G^m} = F_{H^m}$
Si $H0$ est rejetée par le test du χ^2 **Alors**
Renvoyer Red=Faux
Fin Si
Renvoyer Red=Vrai
Fin

7.6 Analyse comparative empirique

À l'instar de ce qui a été fait sur les données manquantes, nous souhaitons approfondir l'analyse théorique des différentes méthodes de sélection d'attributs par une série d'expérimentations. Étant donné les contraintes que nous sommes imposées, nous nous focaliserons sur les filtres.

Depuis le théorème d'impossibilité de Wolpert et Macready (1997), nous savons pertinemment qu'aucun filtre ne surpassera les autres sur l'ensemble des problèmes. Il s'agit en effet d'algorithmes d'optimisation qui rentrent parfaitement dans le cadre défini par Wolpert et Macready. Afin d'aider un utilisateur à choisir tel ou tel filtre, il est donc important de mettre en évidence la catégorie de problèmes pour laquelle tel filtre est bien adapté.

Liu et Yu (2005) ont construit une taxinomie des méthodes de sélection d'attributs du point de vue de l'utilisateur. À la figure 7.11, nous en proposons une extension à la chaîne globale d'apprentissage, ce qui implique de tenir compte des caractéristiques des données relatives aux valeurs manquantes. Si nous ne prétendons pas couvrir l'ensemble des problèmes sur lesquels les filtres peuvent être appliqués, nous avons tout de même l'ambition de contribuer à la compréhension du comportement de différents filtres sur des données incomplètes. Une telle analyse comparative empirique n'a jamais été entreprise dans la littérature, du moins pas à notre connaissance.

Nous souhaitons de plus juger empiriquement l'intérêt qu'il peut y avoir à recourir au filtre que nous avons développé. Il comporte des avantages indéniables d'un point de vue théorique : absence de discrétisation et facilité de la détermination des seuils de pertinence et de redondance, mais encore faut-il qu'il ne conduise pas à une détérioration des performances en classification. Il nous faudra pour cela mener une étude comparative empirique. Cette étude sera également l'occasion de tester l'hypothèse suivante : la construction d'une chaîne globale d'apprentissage correspond à l'association de la méthode de substitution des valeurs manquantes optimale avec la méthode de sélection d'attributs optimale. Cette hypothèse qui est implicite à la plupart des travaux de la littérature, les deux maillons de la chaîne d'apprentissage étant systématiquement étudiés indépendamment l'un de l'autre, nous paraît hautement contestable. Elle mérite pour le moins un examen approfondi.

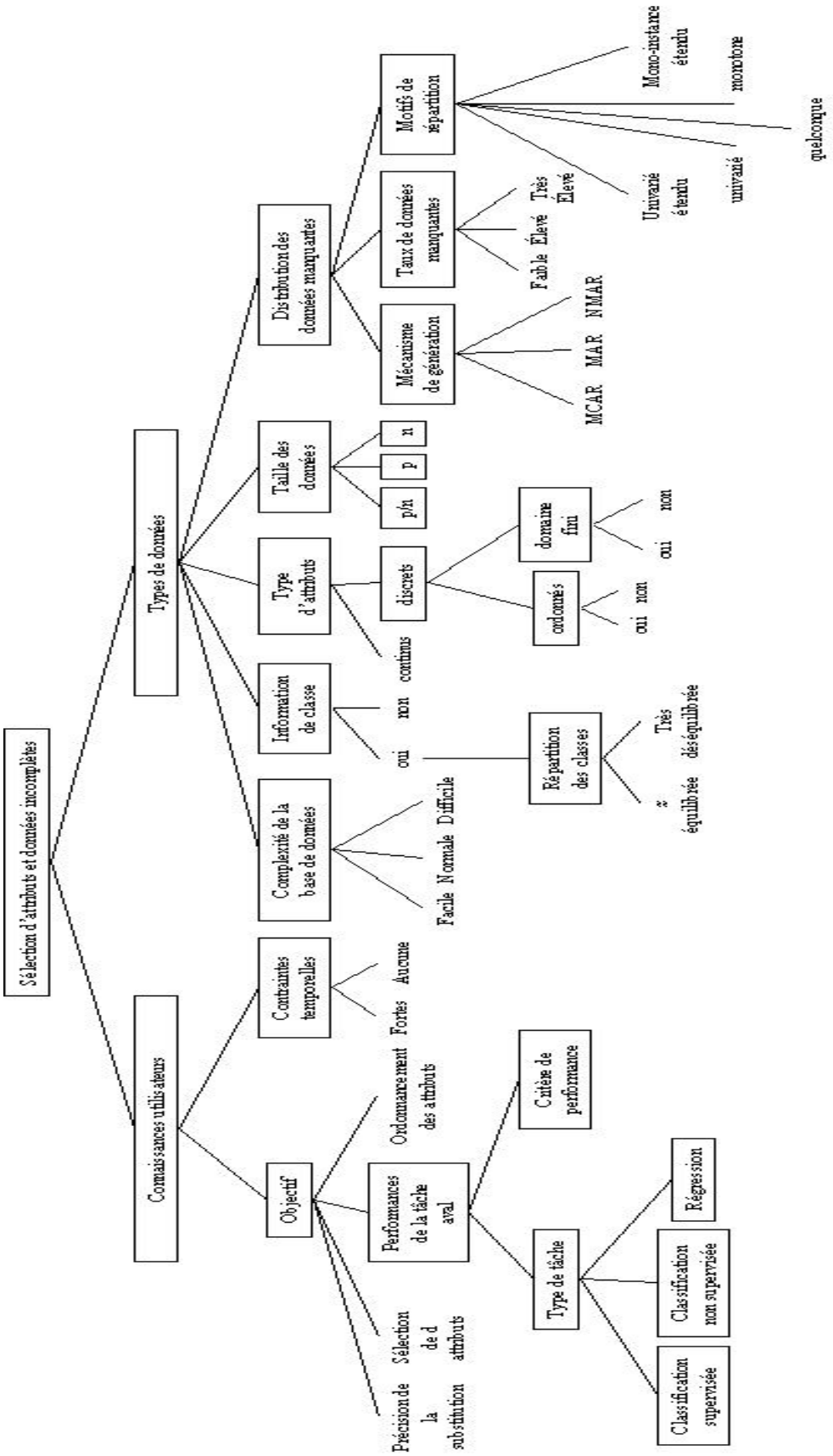

FIG. 7.11 – Taxinomie des techniques de sélection d'attributs sur données incomplètes, du point de vue de l'utilisateur

7.6.1 Protocole expérimental

Afin de pouvoir tirer des enseignements valides des expériences que l'on peut mener, un protocole non biaisé doit être mis en place. Il est également essentiel de rendre ce protocole explicite. Sans cela, il est impossible de reproduire les expériences et seule la bonne foi de l'expérimentateur fait office de preuve des résultats avancés. Quand bien même les résultats seraient avérés, des biais dans le protocole seraient de nature à modifier l'interprétation des résultats.

7.6.1.1 Biais méthodologique

Reunanen (2003) a mis en évidence le fait que la recherche flottante est généralement préférée à la recherche séquentielle simple dans le cas d'un *wrapper*, du fait de l'utilisation d'un protocole biaisé. De nombreuses études avaient en effet conclu à la supériorité de la recherche flottante en omettant de conserver un ensemble de test indépendant pour l'évaluation des performances en généralisation. Ainsi la validation croisée était effectuée sur le même ensemble d'exemples, aussi bien pour guider la recherche, que pour évaluer les performances, ce qui conduisait à un phénomène de sur-apprentissage. Il est donc essentiel de connaître le protocole utilisé pour pouvoir corriger l'interprétation le cas écheant.

Aussi ne pas révéler son protocole expérimental peut-il avoir des effets néfastes sur le développement du domaine : il a fallu attendre près de 10 ans entre les premiers résultats sur la recherche flottante et les travaux de Reunanen. On serait même tenté de taxer de désinformation (Capet, 2006) la non-divulgation d'un protocole expérimental, si l'intention de tromper était avérée, ce qui est heureusement rarement le cas.

Reunanen (2003) a insisté sur le fait que les échantillons de test, utilisés pour l'évaluation de l'erreur en généralisation, ne doivent jamais être employés durant la phase de sélection d'attributs. Lorsque l'on veut mesurer les performances d'un *wrapper*, cela impose de disposer de trois bases disjointes d'exemples. On les nomme généralement bases d'apprentissage, de validation et de test. Les deux premières sont utilisées pour la sélection d'attributs. Chaque fois qu'un sous-ensemble d'attributs doit être évalué, un classifieur est construit à partir de la base d'apprentissage restreinte à ce sous-ensemble. Il est ensuite évalué sur la base de validation. Ce n'est qu'une fois qu'un ensemble d'attributs a été sélectionné que le classifieur correspondant est construit sur l'union de la base d'apprentissage et de validation avant que ses performances ne soient évaluées sur la base de test.

Lorsque l'on ne dispose que de peu de données, il est possible de réduire la variance de l'estimation des performances en procédant à deux validations croisées emboîtées. La boucle intérieure permet de guider la recherche dans le treillis pour chacune des bases d'apprentissage, tandis que la boucle extérieure garantit que chacune des bases de test ne sera pas utilisée durant la sélection d'attributs.

Refaeilzadeh *et al.* (2007) ont tempéré l'argument de Reunanen en indiquant que le biais qu'il avait mis en évidence n'affectait pas l'ordonnancement des techniques de sélection d'attributs, mais seulement leurs performances absolues. Ainsi ce protocole peut être utilisé dans une étude comparative.

Faire en sorte que les exemples de test n'entrent jamais dans le processus de sélection d'attributs n'est pas suffisant pour garantir que le protocole ne sera pas biaisé. Ainsi Singhi et Liu (2006) notent qu'il faudrait également, d'un point de vue théorique, s'assurer qu'une fois la sélection d'attributs effectuée, il reste non pas une base d'exemples de test, mais deux bases, l'une pour apprendre le modèle et l'autre pour le tester. Ceci est très contraignant car le nombre d'exemples est souvent limité, ce qui rend très difficile l'application d'un tel protocole. Mais Singhi et Liu observent expérimentalement que le biais de sélection qui

résulte de l'apprentissage du modèle sur la base qui a servi à réaliser la sélection d'attributs est assez faible pour être ignoré dans les problèmes de classification supervisée, ce qui est notre cas.

Au travers d'expériences nous espérons approfondir notre compréhension du comportement de différents filtres sur des données complètes et incomplètes, dans le contexte de la classification supervisée. En supposant dans un premier temps que nous disposons de bases de données complètes, et en tenant compte des recommandations et autres mises en garde relatives aux protocoles expérimentaux que nous venons de formuler, nous pouvons nous contenter d'évaluer chaque filtre de la manière suivante.

7.6.1.2 Protocole retenu

La base de données est segmentée en dix sous-ensembles par le processus de validation croisée stratifiée, chaque sous-ensemble respecte la distribution des classes intiale. Pour chacun de ces sous-ensembles, le filtre est appliqué sur les neuf autres sous-ensembles. Un modèle de classification est alors appris sur ces mêmes neuf sous-ensembles de données, en ne considérant que les attributs sélectionnés par le filtre. Le classifieur résultant est alors testé sur le dixième sous-ensemble de données, ses performances sont conservées et on recommence avec chacun des autres sous-ensembles de données. La performance globale du filtre correspond à la moyenne des performances obtenues sur chacun des sous-ensembles.

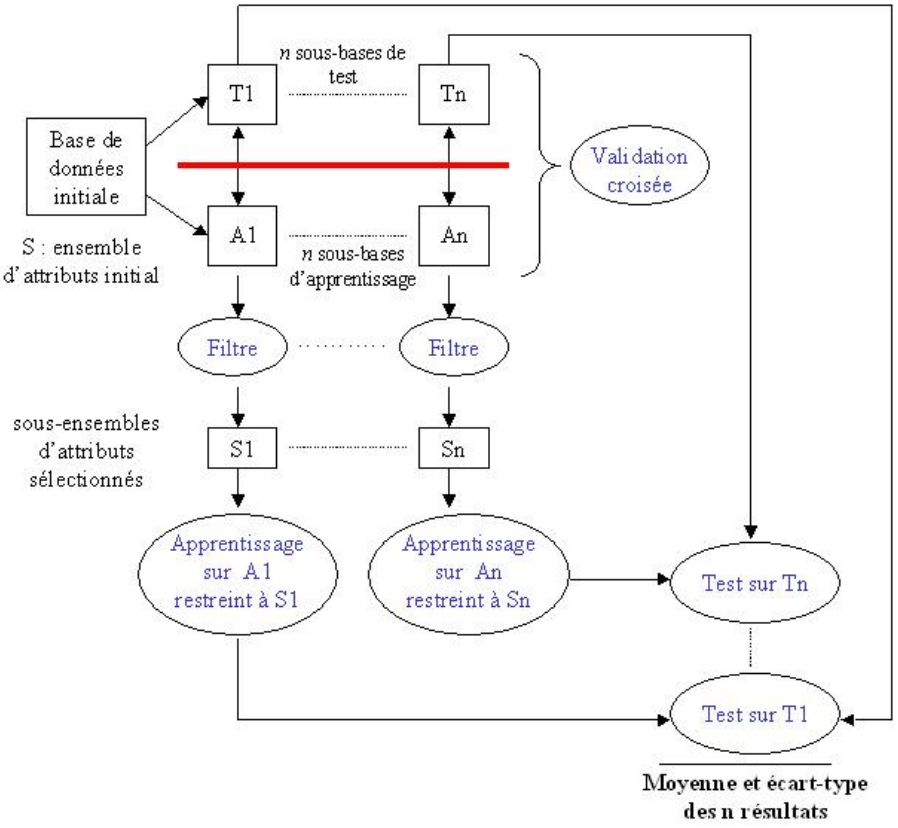

FIG. 7.12 – Protocole d'évaluation d'un filtre sur une base de données complète

Ce processus est décrit sur la figure 7.12. Les échantillons de test ne sont pas utilisés durant la sélection d'attributs, ce qui évite le biais décrit par Reunanen (2003). Quant au biais de sélection évoqué par Singhi et Liu (2006), il est bien présent puisque chaque

classifieur testé est construit sur la base d'exemples ayant servi à la sélection d'attributs. Mais, suivant leurs remarques, nous considérons que ce biais est négligeable.

L'étude de la chaîne d'apprentissage dans son ensemble, en incluant la substitution des valeurs manquantes impose de modifier quelque peu le protocole précédent. Nous en avons mis en place deux distincts. Le premier, illustré sur la figure 7.13, permet de mettre en œuvre la chaîne d'apprentissage dans laquelle les valeurs manquantes sont substituées avant que la sélection d'attributs n'ait lieu. Ceci correspond à la stratégie de combinaison notée (A). Le second protocole correspond quant à lui aux stratégies (B) et (C) : la sélection d'attributs est réalisée antérieurement à la substitution des valeurs manquantes. Il est représenté sur la figure 7.14.

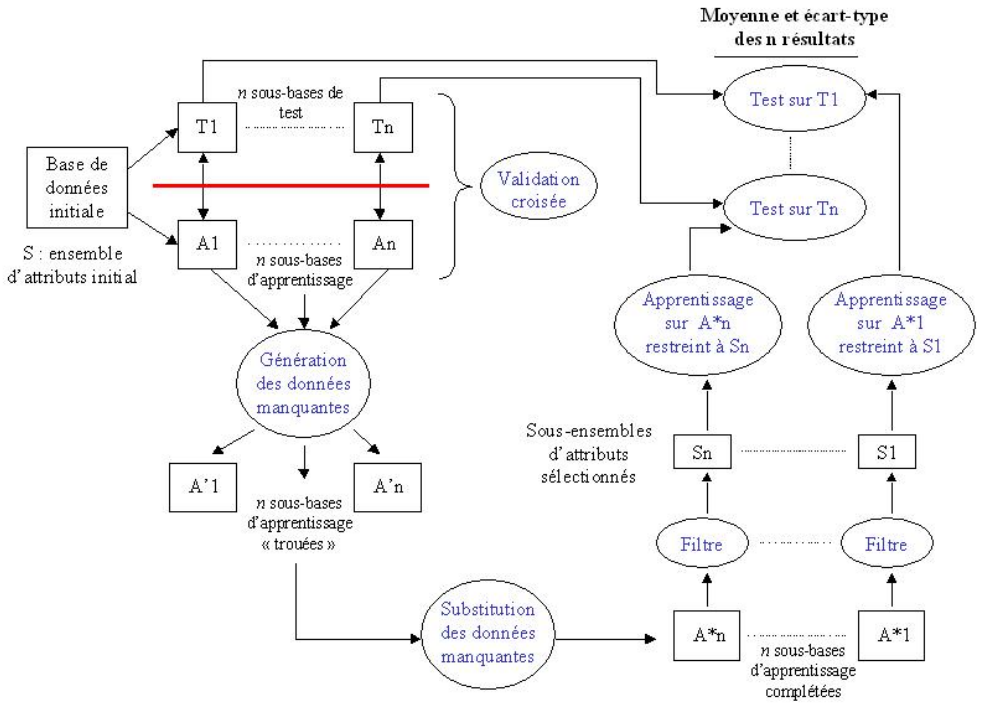

FIG. 7.13 – Protocole d'évaluation d'un filtre sur une base de données incomplète, stratégie (A) : substitution des valeurs manquantes puis sélection d'attributs

Les protocoles que nous avons décrits nous permettent d'obtenir des mesures de performance pour chacun des filtres considérés. Mais ces performances en elles-mêmes ne nous intéressent guère. Ce qui importe c'est de savoir quels filtres permettent d'obtenir de meilleures performances en classification. Nous allons donc être amené à les comparer. Chacun ayant été évalué sur différentes bases de données, nous nous retrouvons dans le cas de figure mentionné à la section 5.5 sur la comparaison de classifieurs. Aussi emploierons-nous le test de Friedman pour juger du caractère significatif ou non des différences observées.

Ayant introduit une nouvelle technique de filtrage que nous souhaitons confronter à celles de la littérature, nous disposons d'une méthode de référence à laquelle les autres seront comparées. Pour cette raison, les tests *post-hoc* seront effectués par l'intermédiaire de z tests, le taux d'erreurs global étant contrôlé par la procédure d'ajustement ascendante de Holland-Copenhaver.

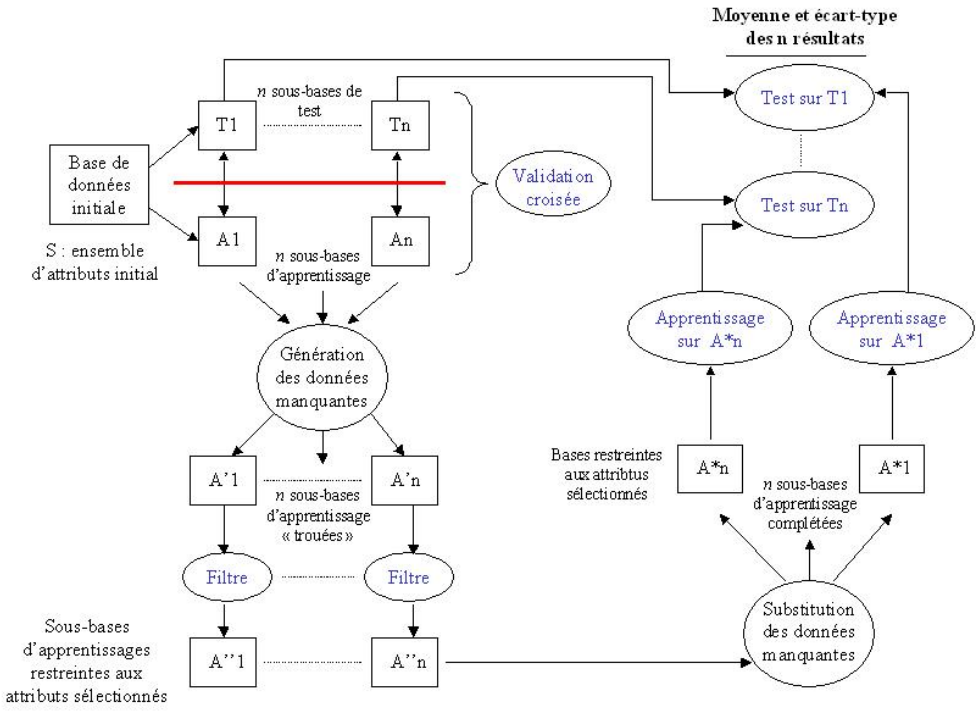

FIG. 7.14 – Protocole d'évaluation d'un filtre sur une base de données incomplète, stratégies (B) et (C) : sélection d'attributs puis substitution des valeurs manquantes

7.6.2 Résultats expérimentaux sur des bases de données complètes

L'objectif principal de nos expérimentations sur des données complètes et incomplètes est de mieux comprendre l'influence des valeurs manquantes sur le processus de sélection d'attributs. Nous avons ainsi inclus dans notre analyse comparative des filtres issus des trois grandes familles que nous avons mises en évidence au fil de notre état de l'art. Nous avons également intégré les deux filtres sur lesquels nous nous sommes appuyé pour construire notre nouvelle méthode KSF, afin de voir si l'extension que nous proposions présentait un intérêt. Cinq filtres entrent donc en ligne de compte dans notre étude empirique.

- CFS : nous avons pris la méthode recherche *best first*. Les attributs continus sont discrétisés selon la méthode de Fayyad et Irani (1993).
- ReliefF : suivant les recommandations de Yu et Liu (2004), seuls $m = 30$ exemples et les $k = 5$ plus proches voisins de chacun d'eux seront considérés durant l'estimation des poids. Pour passer de la pondération à la sélection d'attributs, nous avons décidé de ne retenir qu'un quart des attributs : ceux dont les poids sont les plus élevés. Ce seuil a été fixé de manière expérimentale.
- FCBF : comme pour CFS, les attributs continus sont discrétisés selon la méthode de Fayyad et Irani (1993).
- KSCBF
- KSF

Précisons que pour ReliefF et CFS, nous avons utilisé l'implémentation qui en est faite dans Weka 3.4.7 (Witten et Frank, 2005).

Nous considérons dans un premier temps l'application de chacun de ces filtres sur des données complètes. D'une part, nous souhaitons voir dans quelle mesure nos expériences peuvent corroborer les résultats expérimentaux de la littérature. D'autre part, nous voulons disposer d'une référence à laquelle il sera possible de comparer les résultats obtenus sur des données incomplètes afin de mettre en exergue l'impact des valeurs manquantes.

Le tableau B.3 de l'annexe B donne les noms et caractéristiques des cinq bases de données sur lesquelles les filtres ont été testés. Toutes sont issues de l'*UCI repository*[16] et ne comportent que des attributs continus. Nous avons en effet voulu comparer les différents filtres sur ce type de problèmes pour lequel nous avons spécialement construit l'algorithme KSF.

Étudiant la chaîne d'apprentissage dans sa globalité, nous avons utilisé différents algorithmes de classification supervisée aux propriétés bien distinctes afin de voir l'influence qu'ils pouvaient avoir. Nous avons repris ceux que nous avons utilisés dans la partie relative à l'étude des méthodes de substitution des valeurs manquantes : C4.5, le plus proche voisin et naïve Bayes. Ayant utilisé leur implémentation dans Weka 3.4.7, nous reprendrons par la suite les dénominations de Weka : J48, IB1 et NB respectivement. Afin de se faire idée de l'impact du choix de la mesure d'évaluation, nous en avons testé trois.

- Le taux de bonnes classifications noté *Acc*.
- La moyenne du taux de reconnaissance de chacune des classes noté *BalAcc*.
- L'aire sous la courbe ROC, notée AUC

Afin de donner une vision synthétique de nos résultats, nous avons choisi de reprendre le procédé de présentation des résultats de la section 6.6.4, relative à la comparaison empirique des méthodes de substitution des valeurs manquantes. Le tableau 7.2 donne ainsi la moyenne, estimée sur les cinq bases de données décrites précédemment, des

[16] University of California Irvine http://www.ics.uci.edu/~mlearn/MLRepository.html

rangs obtenus par chaque filtre, en fonction de l'algorithme d'apprentissage et de la mesure de performance considérés. Outre les cinq filtres cités précédemment, nous avons également évalué la méthode SansFiltre, qui comme son nom le suggère consiste à construire un classifieur à partir de l'ensemble des attributs, sans qu'aucune sélection n'ait été effectuée.

TAB. 7.2 – Comparaison statistique de filtres : moyenne des rangs de chaque technique, estimée sur les cinq bases de données complètes

	J48			IB1			NB		
	Acc	*BalAcc*	*AUC*	*Acc*	*BalAcc*	*AUC*	*Acc*	*BalAcc*	*AUC*
SansFiltre	**2.6**	**2.8**	4.1	**2.9**	3.1	3.1	4.5	4.5	2.7
CFS	2.8	3	**2.2**	4	3.8	4	**2.8**	**2.6**	**2**
ReliefF	3.8	3.6	4	3.4	3.6	3.4	4.4	4	4.8
FCBF	3.6	3.6	3	4.6	4.6	4.6	**2.8**	3.4	4.2
KSCBF	4.8	4.6	4.7	3.1	**2.9**	3.1	3.3	3.5	3.5
KSF	3.4	3.4	3	3	3	**2.8**	3.2	3	3.8

La principale information qu'il faut retenir du tableau 7.2 est qu'aucune des méthodes n'est statistiquement supérieure à une autre. Les rangs moyens de chacune d'elles sont très proches, si bien que le test de Friedman nous autorise à considérer que les différences observées ne sont pas significatives, et ce, quel que soit l'algorithme de classification, et quelle que soit la mesure de performance considérée. Ainsi contrairement à ce qu'affirment les auteurs des différents filtres comparés, nous n'observons pas que leur méthode surpasse les anciennes. Plusieurs raisons peuvent être invoquées pour expliquer cet écart entre nos résultats expérimentaux et ceux de la littérature.

Nous n'avons pas testé les méthodes sur les mêmes bases de données, aussi est-il difficile de comparer ces résultats à ceux de la littérature. Nous n'avons en effet utilisé que des bases de données contenant des attributs continus, contrairement à ce qui a été fait pour l'évaluation de CFS, FCBF, KSCBF où la majorité des problèmes traités contiennent des attributs discrets[17].

Yu et Liu (2004) ont également testé leur méthode FCBF, ainsi que CFS[18] sur la base musk2, qui fait partie de nos bases d'évaluation. Ils indiquent les performances de ces techniques, estimées par le taux de bonnes classifications (Acc) avec les algorithmes d'apprentissage J48 et NB. Nous pouvons donc comparer nos résultats bruts aux leurs. Pour FCBF couplé à J48 le taux de bonnes classifications qu'ils obtiennent est de 91.35%, quand nous obtenons 91.63%. Lorsque FCBF est couplé à NB, nous obtenons 83.64% de bonnes classifications au lieu des 84.59% obtenus par Yu et Liu[19]. Ces performances sont voisines, les différences pouvant facilement s'expliquer par le fait que nous n'avons pas échantillonné exactement de la même manière la base de données lorsque nous avons effectué la validation croisée (une part d'aléatoire est introduite dans l'échantillonnage).

Nos résultats corroborent donc ceux de Yu & Liu, du moins sur la seule base de données que nous avons en commun : musk2. Ce ne sont donc pas tant nos résultats qui diffèrent que

[17] Notons qu'au vu de nos remarques sur le test de Kolmogorov-Smirnov, l'utilisation de données discrètes faite par Biesiada et Duch (2005) pour évaluer KSCBF nous apparaît pour le moins inadaptée.

[18] Il s'agit d'une version antérieure de cet algorithme utilisant une recherche séquentielle simple et non la méthode *best first search* : CorrFS, dont les performances sont censées être similaires à celles de CFS.

[19] Pour CFS, nos performances sont également comparables.

l'interprétation que nous en faisons. La méthodologie utilisée par Yu et Liu (2004), mais aussi par Hall (2000) et Biesiada et Duch (2005), repose sur le test de Student appliqué à chaque paire de méthodes. Le nombre de fois où telle méthode s'avère statistiquement supérieure à telle autre selon ce test sert alors d'argument pour justifier son éventuelle supériorité. Nous avons précédemment insisté sur les défauts de cette approche, reprenant à notre compte les critiques de Salzberg (1997); Demsar (2006). Ce sont d'ailleurs ces critiques qui nous ont poussé à employer des tests de rangs, non paramétriques. Selon nous, les divergences que l'on observe entre notre analyse et celles de la littérature s'expliquent essentiellement par le choix de la méthodologie de comparaison des classifieurs.

Les performances de toutes les méthodes y compris SansFiltre étant statistiquement équivalentes, nous ne constatons pas une amélioration des performances lorsque la dimension du problème est réduite par sélection d'attributs. Mais nous ne constatons pas non plus de dégradation des performances. Ceci suffit à justifier l'intérêt des méthodes de filtrage. Pour certains types de problèmes, la sélection d'attributs réduit la durée de l'apprentissage et facilite l'interprétation, sans détériorer la qualité du modèle appris.

Concernant la nouvelle approche que nous avons proposée, ses résultats sont prometteurs. Certes elle n'apparaît pas meilleure que les autres, mais elle leur est équivalente. Or rappelons qu'elle permet de se passer de l'étape de discrétisation, ce qui réduit le nombre de paramètres à régler. Les choix des seuils de pertinence et de redondance sont, de plus, simplifiés.

Le tableau 7.2 semble confirmer notre hypothèse selon laquelle l'algorithme d'apprentissage influe sur la qualité d'un filtre. Autrement dit, il semble que pour choisir un filtre en vue d'une tâche de classification, il est important de savoir quel algorithme d'apprentissage sera utilisé. L'ordonnancement des méthodes diffère en effet selon le classifieur considéré. Avec J48, la méthode SansFiltre est celle qui obtient les rangs les plus bas, tandis qu'avec NB la première place échoit à CFS. Notre méthode KSF paraît enfin la plus prometteuse avec IB1.

Des remarques peuvent être faites à propos de la mesure d'évaluation d'un classifieur. Nous observons en effet que la méthode SansFiltre, avec le classifieur J48, obtient le rang moyen le plus bas lorsque l'on considère les mesures Acc et BalAcc. En revanche, lorsque les performances de J48 sont évaluées par l'aire sous la courbe ROC, cette méthode paraît nettement moins efficace, son rang moyen étant le deuxième plus élevé. L'impact du choix de la mesure de performance est également notable, pour cette même méthode, avec le classifieur NB. Mais dans ce cas, l'effet est inverse. SansFiltre obtient ainsi le plus mauvais rang avec Acc et BalAcc, tandis qu'avec AUC, son rang est le deuxième meilleur, derrière celui de CFS.

Nous allons maintenant aborder l'analyse de la chaîne d'apprentissage sur des données incomplètes, afin de voir dans quelle mesure sont corroborées les observations précédentes, constatées sur des bases de données complètes.

7.6.3 Résultats expérimentaux sur des bases de données incomplètes

Notre objectif n'est pas ici d'analyser finement les liens entre substitution des valeurs manquantes et sélection d'attributs, mais simplement de mettre en relief l'impact des valeurs manquantes sur le filtrage. Aussi avons-nous choisi de ne considérer qu'une seule méthode de substitution : la moyenne.

Pour contrôler les divers paramètres régissant la génération des données manquantes, nous sommes parti des mêmes bases complètes utilisées précédemment, desquelles nous avons ôté certaines valeurs selon un processus similaire à celui qui est décrit dans la section

6.6.3. Nous avons généré selon le mécanisme MCAR cinq bases de données incomplètes pour chacune des bases initiales, contenant respectivement 10%, 20%, 30%, 40% et 50% de valeurs manquantes. Pour chacune de ces bases, nous avons appliqué les cinq filtres précédents en considérant pour chacun d'eux les trois stratégies A, B et C, de combinaison avec la substitution par la moyenne. C'est-à-dire que nous avons construit trois chaînes d'apprentissage pour chaque filtre, à l'exception de ReliefF. Ce filtre a en effet été développé pour tenir compte des valeurs manquantes. La stratégie B, dans laquelle le filtrage est appliqué avant la substitution en ignorant les valeurs manquantes, n'a donc pas pu être testée pour ReliefF.

Nous avons ajouté une dernière chaîne d'apprentissage, nommée SansFiltre, dans laquelle seule la substitution des valeurs manquantes est effectuée, aucune réduction du nombre d'attributs n'étant opérée.

Comme précédemment nous présentons nos résultats par l'intermédiaire de la moyenne des rangs de chaque méthode, estimée sur les 25 bases de données incomplètes, pour un classifieur et une mesure de performance donnée. Ces résultats sont disponibles au tableau 7.3. Dans ce tableau, les chiffres suivis d'une étoile indiquent que les performances de la méthode correspondante sont statistiquement différentes de celle de la méthode de référence, le niveau de confiance ayant été fixé à 95%. La chaîne d'apprentissage A-KSF est celle des trois combinaisons, dans lesquelles intervient le nouveau filtre que nous avons proposé, qui semble la plus prometteuse. Aussi avons-nous choisi d'en faire notre méthode de référence.

TAB. 7.3 – Comparaison statistique de filtres : moyenne des rangs de chaque technique, estimée sur les 25 bases de données incomplètes (5 bases et 5 taux de valeurs manquantes par base)

	J48			IB1			NB		
	Acc	*BalAcc*	*AUC*	*Acc*	*BalAcc*	*AUC*	*Acc*	*BalAcc*	*AUC*
SansFiltre	7.9	6.9	9.1	6.7	7	7.2	9.3	8.4	7.7
A-CFS	5.2	6.4	6.6	6.1	6.5	6.5	**5.9**	**5.1**	2.72*
B-CFS	4.9	**5.2**	6.3	6.2	6.5	6.6	6.8	5.2	4.1
C-CFS	**4.8**	5.6	5.6	6.7	7.1	7.1	7.6	5.8	4.1
A-ReliefF	9	9.4*	7.2	8.2	8.2	7.9	8.3	8.3	8.9
C-ReliefF	9.6	10.3*	10	11.6*	11.8*	11.8*	10.7	10.5	10.8
A-FCBF	6.6	7.5	**4.1**	9*	8.8*	8.9*	6.7	8.1	7.7
B-FCBF	11.4*	11.7*	8.2	12.5*	12.2*	12.2*	9.1	12.5*	12*
C-FCBF	7.5	8.4	6.2	8.7*	8.4	8.4	7.4	8.5	8.2
A-KSCBF	7.5	6.5	8.9	6.3	6.7	7	9.1	8.1	7.8
B-KSCBF	10.1*	9.3*	10.3	8	8.1	8.3	8.8	9.1	9
C-KSCBF	10.5*	9.9*	10.8	8.5*	7.9	8	6.8	6.9	9.1
A-KSF	6.1	**5.2**	7.7	**4.9**	**5.2**	**5.2**	8.9	7.9	7.5
B-KSF	8.8	7.9	8.9	6.8	6.3	5.7	7.2	7	10
C-KSF	10*	9.9*	10.3	9.7*	9.3*	9*	7.4	8.6	10.5

Contrairement à ce que nous avons observé sur les données complètes, les résultats du tableau 7.3 font apparaître des différences significatives entre les méthodes. Deux techniques se distinguent assez nettement des autres : C-ReliefF et B-FCBF qui sont statistiquement inférieures à la nôtre dans presque toutes les configurations testées. Aussi déconseillons-nous vivement leur utilisation.

C-ReliefF correspond à l'application de ReliefF directement sur les données incomplètes, en amont de la phase de substitution. ReliefF ayant été spécialement étendu pour traiter les donnnées manquantes, il est assez étonnant de constater qu'il est plus efficace, du point de vue des performances, de remplacer les valeurs manquantes avant de l'utiliser. Ceci est à rapprocher des observations faites à propos de C4.5 dont les performances sont meilleures lorsqu'il est précédé d'une étape de substitution, alors qu'une méthode est intégrée dans C4.5 pour traiter les valeurs manquantes (Batista et Monard, 2003).

Les mauvais résultats de B-FCBF sont quant à eux moins surprenants, puisqu'ils indiquent que la prise en compte des valeurs manquantes est importante pour le filtrage. La méthode C-FCBF obtient en effet de meilleurs rangs et n'est pas statistiquement inférieure à A-KSF, avec IB1, lorsque celui-ci est évalué par l'intermédiaire du taux de bonnes classifications.

Cette dernière remarque suggère que les choix du classifieur et de la mesure de performance influent sur le choix de la méthode de prétraitement. Les expériences sur les données complètes le laissaient déjà supposer.

Concernant les classifieurs les tendances suivantes se dégagent : avec J48, la méthode C-CFS obtient les meilleurs rangs, tandis qu'A-KSF et A-CFS semblent respectivement les mieux adaptées aux classifieurs IB1 et NB. Avec J48, il nous faut cependant relativiser notre remarque car C-CFS n'obtient les meilleurs rangs qu'avec les critères de performance Acc et BalAcc. Lorsque l'on considère l'aire sous la courbe ROC, la méthode ayant le rang le plus faible est A-FCBF, ce qui tend à confirmer l'influence de la mesure de performance.

Parmi les trois stratégies de combinaison de la substitution des valeurs manquantes avec le filtrage d'attributs, il n'est pas aisé de conclure que l'une des trois domine les deux autres. Suivant le classifieur et le filtre utilisés nous observons des différences notables. Nous avons vu que pour FCBF, la prise en compte des valeurs manquantes était plus avantageuse lorsque ce filtre est appliqué directement sur les données incomplètes (stratégie C supérieure à la stratégie B). Mais il est difficile de départager la stratégie C et la stratégie A. Cette dernière semble tout de même préférable avec le classifieur J48. Pour les deux autres les rangs moyens sont très proches.

Pour ReliefF nous avons vu que la stratégie A était mieux adaptée que la stratégie C. Pour KSCBF, les stratégies B et C sont également dans ce cas, excepté dans le cas du classifieur NB, pour lequel A-KSF ne semble pas très performante. A-KSCBF est cependant toujours statistiquement équivalente à A-KSF, ce qui indique que pour ce filtre, la stratégie A est plus efficace.

Pour notre approche, nous avons vu que la stratégie était également plus efficace que les deux autres. Mais les performances de B-KSF sont équivalentes à celles de A-KSF, même si les rangs moyens sont toujours supérieurs avec la stratégie B. Nous serions donc tenté de dire que, pour KSF, il est préférable de compléter une base de données (même avec une technique aussi simple que la moyenne) avant de réaliser le filtrage. Nous espérions pouvoir mettre en évidence l'inverse, car réaliser le filtrage en amont de la substitution est moins coûteux, mais tel n'est pas le cas. Cela vient du fait, selon nous, que les méthodes de traitement des valeurs manquantes employées durant le filtrage ne sont pas efficaces.

Ignorer les valeurs manquantes n'est pas satisfaisant car cela revient à négliger une partie de l'information disponible. Quant au fait de considérer une valeur manquante comme n'importe laquelle des autres valeurs, il s'agit là d'une hypothèse peut-être trop forte.

L'hypothèse que nous faisons pour KSF est encore plus forte, puisque nous mettons sur un pied d'égalité la distribution globale des valeurs manquantes et celle des valeurs observées. Il s'agit là selon nous de la principale raison de l'échec de la méthode C-KSF, mais aussi de C-KSCBF dont l'analyse de la redondance repose également sur cette hypothèse.

Si l'on considère CFS, on se rend compte que le constat relatif aux stratégies de combinaison est moins tranché. À chaque classifieur, l'une des trois stratégies semble mieux adaptée : stratégie A pour NB, B pour IB1 et C pour J48. Mais l'écart entre les rangs obtenus par les trois types de chaîne d'apprentissage est tout de même assez faible. Il semble donc que CFS soit le plus robuste vis-à-vis des valeurs manquantes, puisque ses performances ne sont que faiblement affectées par le choix de la chaîne d'apprentissage. Mais rappelons qu'il s'agit tout de même de la technique la plus coûteuse parmi celles qui ont été envisagées. Contrairement aux autres, elle parcourt le treillis de recherche pour trouver le meilleur sous-ensemble.

La méthode SansFiltre qui nous sert de référence n'est jamais statistiquement différente de A-KSF. Il est donc possible de réaliser une réduction de la dimension sans dégrader les performances en classification. Ceci suffit, selon nous, à justifier l'intérêt des méthodes de sélection.

Le filtre KSF que nous avons développé obtenait des performances prometteuses sur les bases de données complètes. Les expériences sur les données incomplètes permettent également de mettre en évidence le bon comportement de ce filtre, du moins lorsque la substitution est effectuée en premier. A-KSF est en effet statistiquement supérieure à nombre de méthodes et les rangs moyens obtenus avec J48 et surtout IB1 sont parmi les meilleurs. Seul le filtre CFS, plus complexe, apparaît plus robuste. A-CFS est d'ailleurs la seule méthode à obtenir des performances statistiquement meilleures que celles de A-KSF, avec le classifieur NB et le critère AUC. Le rang moyen de A-CFS dans ce contexte est d'ailleurs étonnamment faible, nettement en deçà de tous les autres. Si l'on reprend les résultats du tableau 7.2, on se rend compte que CFS, s'il n'était pas statistiquement supérieur à KSF, obtenait déjà le meilleur rang avec NB et AUC. Les valeurs manquantes ont alors vraisemblablement contribué à la dégradation des performances des autres filtres, sans affecter outre mesure les trois versions de CFS, du fait de la robustesse de CFS évoquée précédemment.C'est là un biais de notre méthode de comparaison à base des rangs, ou plutôt de son application à des méthodes contenant les mêmes filtres.

7.7 Conclusion

La sélection d'attribut est une étape de prétraitement qui joue un rôle très important en apprentissage. Appliquée en classification supervisée, elle assure une réduction de la dimension du problème, ce qui permet de réduire la durée de l'apprentissage et de simplifier le modèle appris. Cette simplification facilite généralement l'interprétation de ce modèle. Autre avantage, la réduction de la dimension permet d'éviter le phénomène de sur-apprentissage, réduisant ainsi l'erreur de généralisation. Si l'accélération de la phase d'apprentissage n'est pas notre priorité, la qualité des performances et l'interprétabilité sont des caractéristiques essentielles du modèle d'évaluation des risques que nous essayons de construire. C'est la raison pour laquelle nous nous sommes penché sur l'étude des diverses méthodes de sélection.

Après avoir dressé un état de l'art des principales méthodes, nous nous sommes concentré sur celles qui correspondaient le mieux à notre besoin : les filtres. Contrairement aux *wrappers* ils n'ont pas recours à l'algorithme d'apprentissage pour guider leur recherche d'un sous-ensemble d'attributs optimal. Ils sont donc plus rapides. En grande dimension, ce point est essentiel puisque la grande complexité des *wrappers* les rend inapplicables. Parmi les filtres, nous avons focalisé notre attention sur la méthode proposée par Yu et Liu (2004), qui est de moindre complexité grâce au découplage des phases d'analyse de la pertinence et de la redondance.

En généralisant leur approche nous avons proposé un nouveau filtre, basé sur le test de Kolmogorov-Smirnov. L'avantage de notre méthode est double. D'une part, les attributs continus, qui constituent l'essentiel des attributs dont nous disposons dans notre contexte applicatif, peuvent être traités directement sans passer par une phase intermédiaire de discrétisation. D'autre part, les seuils de redondance et de pertinence sont beaucoup plus simples à choisir et à interpréter que dans la méthode originale de Yu et Liu.

Afin de juger de l'intérêt de notre approche d'un point de vue empirique, nous l'avons comparée aux filtres existants sur des données complètes et incomplètes. De ces expériences, CFS apparaît certes comme étant la méthode la plus robuste pour traiter les données incomplètes. Mais nos conclusions relatives à la méthode que nous avons développée sont plus qu'optimistes.

Par rapport aux deux filtres les plus réputés pour leur efficacité : FCBF et CFS, nous soutenons que KSF n'a pas à rougir de la comparaison. Ses performances sont en effet aussi bonnes voire meilleures sur les données incomplètes que celles de FCBF. Or nous avons vu que la complexité des deux filtres était la même. Quant à CFS, à l'exception du cas où NB et AUC sont utilisés, ses performances sont équivalentes, mais rappelons que la complexité de KSF est moindre.

Ces expériences ont également fait ressortir le gain de performance ou tout au moins la non-dégradation des performances, que pouvaient assurer certaines méthodes de réduction de dimension. Pour que cette réduction soit efficace, nous avons insisté sur deux points essentiels : il faut supprimer les attributs qui ne sont pas pertinents vis-à-vis de la classe ainsi que ceux qui sont redondants. L'évaluation de la pertinence est ici réduite à l'identification d'une simple relation de dépendance entre un attribut et la classe.

Or nous avons vu que la pertinence est une notion plus complexe qui doit être évaluée en contexte. Il faudrait donc théoriquement pousser plus loin notre analyse afin de chercher à identifier une relation de dépendance contextuelle entre un attribut et la classe. Le contexte correspond ici à un ensemble d'autres attributs. La redondance a été utilisée en partie pour cela, mais elle ne suffit pas. Pour quasiment tous les algorithmes que nous avons présentés, les interactions entre attributs ne sont pas pris en compte. Un attribut seul peut n'avoir aucune influence sur la classe, alors qu'en présence d'autres attributs, son influence est grande. La situation inverse peut également se produire.

Parmi les méthodes de recherche abordées durant notre état de l'art, celles qui reposent sur un parcours arrière du treillis sont bien adaptées à l'identification des interactions. Chaque attribut est, en effet, évalué en présence de tous les autres, ce qui n'est pas le cas pour les méthodes de recherche avant. Pour que les interactions puissent effectivement être identifiées, encore faut-il que la mesure d'évaluation choisie permette d'évaluer un ensemble d'attributs directement, sans pour autant n'être qu'une agrégation de mesures individuelles. Les *wrappers*, qui évaluent un sous-ensemble d'attributs par l'intermédiaire du classifieur construit à partir de ce sous-ensemble, répondent bien à ce besoin. Les mesures de cohérence sont également bien adaptées comme l'indique l'étude empirique de Zhao et Liu (2007).

Cohen *et al.* (2005) offrent quant à eux une illustration de l'intérêt des *wrappers* dans l'identification des interactions entre attributs. Ils se placent dans le cadre de la théorie des jeux et ont recours à la valeur de Shapley pour estimer la pertinence d'un attribut en tenant compte du contexte.

Nous souhaiterions nous inspirer des travaux de Jakulin et Bratko (2004: 2003) qui se basent sur la notion de gain d'interaction, pour parvenir à repérer les paires d'attributs qui interagissent. Pour chacune de ces paires, notre idée serait de construire un nouvel attribut résultant de la combinaison des deux qui interagissent. Cela permettrait de faire de l'extraction d'attributs et pas simplement de la sélection, avec une complexité moindre que celle des algorithmes de programmation génétique qui sont traditionnellement utilisés à cet effet (voir section 7.2.2). Le principal inconvénient d'une telle méthodologie réside dans la perte de compréhensibilité du modèle. En effet les attributs nouvellement créés seront difficiles à interpréter. Pour surmonter cette difficulté, des règles expertes pourraient être définies, indiquant les types de combinaison admissibles entre différents groupes de variables. Ce ne sont là que quelques pistes que nous pourrions explorer. Nous croyons fermement que les prochains développements dans le domaine du filtrage d'attributs, s'ils ne s'orienteront pas forcément vers la construction de nouveaux attributs, seront tous dirigés vers la prise en compte des interactions entre attributs.

Chapitre 8

Discussion

Nous avons abordé dans cette partie les deux principaux maillons de la partie amont de la chaîne d'apprentissage : le traitement des données manquantes et la sélection d'attributs. Nous les avons étudiés de manière aussi générique que possible, en essayant de tenir compte des contraintes sous-jacentes de notre projet d'évaluation des risques. Dans les deux cas, nous nous sommes efforcé de mettre en évidence les caractéristiques théoriques et empiriques des principales méthodes du domaine.

Sur le plan théorique, nous avons essayé de dégager de notre étude une taxinomie des différentes méthodes, afin d'apporter une vue synthétique et globale du domaine. Outre leur pouvoir descriptif, l'intérêt de ces taxinomies est de faire ressortir les critères qui distinguent les différentes méthodes, ce qui peut s'avérer fort utile pour la conception de nouvelles solutions, en combinant des critères de manière inédite. Ainsi, nous avons contruit une nouvelle méthode basée sur l'entropie en mettant l'accent sur la restauration du pouvoir discriminant des attributs alors que les méthodes usuelles cherchent avant tout des valeurs de subtitution aussi proches que possibles des valeurs d'origine.

Dans le domaine de la sélection d'attributs, nous avons généralisé la méthode de Liu et Yu (2003) qui repose sur le découplage de l'analyse de la redondance et de la pertinence, afin de construire un filtre basé sur le test de Kolmogorov-Smirnov, capable de s'appliquer directement sur les attributs continus et dont les seuils de pertinence et de redondance soient aisés à définir. Mais la généralisation que nous avons effectuée permet d'imaginer toutes sortes de combinaisons de méthodes pour réaliser les deux analyses de pertinence et de redondance. Au vu des performances de Bins et Draper (2001), nous serions tenté d'effectuer l'analyse de la redondance par classification non supervisée, en utilisant une technique n'imposant pas de fixer le nombre de groupes (Lemoine *et al.*, 2006).

Dans chacun des domaines, il est notoire, depuis les travaux de Wolpert et Macready (1997), qu'aucune technique n'est meilleure qu'une autre dans l'absolu. Aussi avons-nous essayé de contribuer à balayer certains types de problèmes afin d'identifier les méthodes les mieux adaptées pour chacun d'eux. Les méthodes que nous avons proposées, aussi bien pour la substitution des valeurs manquantes que pour la sélection d'attributs, se sont toutes deux avérées plus que prometteuses. Cependant, le travail est encore long avant de cerner un peu mieux les types de problèmes auxquels elles peuvent apporter des solutions mieux appropriées que les techniques existantes. Cette remarque vaut également pour les autres techniques. Selon nous, parvenir à une cartographie des types de problèmes et des solutions les mieux adaptées pour chacun d'eux, est l'un des principaux axes de recherche vers lequel la communauté doit s'engager.

Deux tâches principales doivent être réalisées : la caractérisation des différents types de problèmes d'une part, et l'analyse comparative empirique des méthodes existantes d'autre part. Nous avons mené nos études empiriques en nous efforçant de contrôler les para-

mètres relatifs aux valeurs manquantes. Afin de poursuivre la caractérisation des différents problèmes et des méthodes qui convient de leur appliquer, il serait bon de se tourner désormais vers la génération de bases de données artificielles. Nous pourrons ainsi assurer une meilleure maîtrise de l'ensemble des paramètres et pas uniquement de ceux qui ont trait à la distribution des valeurs manquantes : la proportion de variables non pertinentes, redondantes, la difficulté de la tâche de classification etc. Pour la première de ces deux tâches, nous avons proposé une batterie de critères, caractérisant les données, inspirée de celle que Liu et Yu (2005) ont fournie.

Notre principal apport concernant la seconde tâche est essentiellement d'ordre méthodologique. Nous avons vu qu'il était impératif de disposer non seulement d'un protocole d'évaluation clair et autant que possible non biaisé. Dans le domaine de la sélection d'attributs, qui a déjà fait l'objet de quantités de travaux, l'importance du protocole expérimental a déjà été souligné par divers auteurs. En revanche, tel n'est pas le cas dans le domaine de la substitution des valeurs manquantes appliquée à la classification supervisée. Ceci reflète, selon nous, une différence de maturité entre les deux domaines.

Le test de Student peut conduire à des conclusions fallacieuses lorsqu'il est appliqué à la comparaison de diverses paires de classifieurs. Or la quasi-totalité des études comparatives expérimentales dans les deux domaines qui nous préoccupent y ont recours. À partir des suggestions de Demsar (2006), pour la comparaison de plus de deux classifieurs, nous avons défendu l'utilisation du test non paramétrique de Friedman et des tests *post-hoc* qui lui sont associés.

Nous avons enfin mis à l'épreuve de données incomplètes diverses techniques de filtrage d'attributs. Cela nous a permis de constater l'importance de l'étude de la chaîne d'apprentissage dans son ensemble. À notre connaissance, la plupart des travaux n'abordent le traitement des valeurs manquantes et la sélection d'attributs que de manière indépendante, ce que nous avons d'ailleurs fait dans un premier temps.

Au travers d'expériences assez simples, ne mettant en œuvre qu'une seule technique de substitution, nous avons pu voir à quel point il était difficile de choisir la chaîne d'apprentissage appropriée. Il faut, en particulier, pouvoir décider de la stratégie à adopter pour combiner un filtre et une technique de substitution donnés. Or, il s'avère que la stratégie optimale est fonction des différents maillons de la chaîne que l'on souhaite combiner : méthode de substitution, de filtrage et de classification. Il est donc illusoire de penser qu'une analyse de chacun des maillons de la chaîne, pris indépendamment les uns des autres, est une solution suffisante pour construire l'ensemble de la chaîne.

Notons qu'en constatant l'interdépendance entre les seules étapes de filtrage et d'apprentissage nous allons contre l'idée reçue selon laquelle le grand avantage des filtres sur les *wrappers* réside dans leur indépendance vis-à-vis du choix de l'algorithme d'apprentissage. Ceci est vrai théoriquement. Une fois une base de données filtrée, n'importe quel algorithme d'induction peut être employé, mais les faits expérimentaux indiquent que les performances en classification seront plus ou moins élevées selon le classifieur qui est choisi.

Après cette étude générique des deux principales étapes de prétraitement utilisées en apprentissage supervisé, nous allons nous recentrer sur notre contexte applicatif : les conflits armés intra-étatiques. Nous avons vu à quel point la détermination de la chaîne d'apprentissage optimale était dépendante du type de problème à traiter. Aussi ne sont-ce pas tant les résultats empiriques de cette partie qui nous intéressent que la méthodologie mise en place pour y parvenir. Nous nous efforcerons, dans la partie suivante, de mettre en œuvre cette méthodologie pour comparer toutes les chaînes d'apprentissage qui, d'un point de vue théorique, semblent pouvoir correspondre à notre besoin. Outre l'intérêt que représente une telle analyse comparative pour notre cas d'usage, elle nous permettra d'approfondir notre travail d'analyse de la chaîne d'apprentissage dans sa globalité.

Troisième partie

Un nouveau modèle d'évaluation des risques

Dans la partie I nous avons commencé à élaborer un modèle d'évaluation des risques, basé sur l'induction d'arbres de décisions flous. Au travers une série d'expérimentations, nous avons mis en évidence les problèmes relatifs aux valeurs manquantes et à la sélection des variables pertinentes à partir desquelles est construit ledit modèle. Au cours de la partie II, nous avons alors focalisé notre attention sur ces deux points clés de l'apprentissage inductif. Nous avons désormais en main toutes les informations nécessaires pour construire un modèle global et générique d'évaluation des risques.

Nous pouvons d'ores et déjà en décrire l'architecture, ce que nous ferons au chapitre 9. À la section 9.1 nous commencerons par décrire le processus d'apprentissage qui permet de construire un modèle d'évaluation des risques pour un problème donné. Nous décrirons ensuite, au chapitre 9.2, la façon dont ce modèle est utilisé pour calculer un indice de risque pour chaque exemple à traiter.

Ainsi que nous l'avons suggéré à la fin de la partie précédente, le choix des différents modules du système global dépend du type de problème à traiter. Aussi allons-nous dans cette partie recentrer notre travail sur l'application qui sert de guide à cette thèse : l'évaluation des risques de conflits armés intra-étatiques, que nous aborderons plus en détail à la section 10. Ce sera pour nous l'occasion de voir comment instancier la méthodologie générique que nous avons mise en place et de la mettre à l'épreuve d'un cas pratique.

Nous commencerons par rappeler le contexte et les objectifs qui sous-tendent ce domaine applicatif afin d'en dégager les spécificités. La présentation de la base de données que nous avons mise en place pour traiter ce problème, décrite à la section 10.2, nous permettra d'approfondir la caractérisation du problème. Nous pourrons alors comparer diverses instanciations du modèle générique, afin de choisir celle qui est la mieux adaptée à notre application. Ce sera l'objet de la section 10.3.

Chapitre 9

Système global d'évaluation des risques

Au cours de la première partie, nous avons construit un premier modèle d'évaluation des risques, basé sur l'apprentissage inductif d'arbres de décisions flous. Nous nous sommes depuis efforcé de voir comment améliorer ce modèle tant du point de vue des performances en classification que de l'interprétabilité. N'oublions pas cependant que notre objectif est d'estimer un certain risque, et donc de pouvoir utiliser le modèle appris non seulement pour classer différents exemples en deux catégories, à risque ou non, mais également pour quantifier le risque en question. Nous distinguons ainsi clairemement l'apprentissage du modèle de son utilisation effective pour l'évaluation des risques. Nous allons maintenant détailler ces deux points afin de donner une vue globale de notre approche.

9.1 Apprentissage du modèle

Le cœur de notre système réside dans l'apprentissage d'arbres de décision flous. Cet apprentissage est réalisé par le logiciel *Salammbô* (Marsala, 1998). Nous nous sommes concentré précédemment sur la phase de prétraitement des données, en insistant sur les étapes de substitution des valeurs manquantes et de sélection d'attributs. Si *Salammbô* réalise l'apprentissage de l'arbre de décision flou, qui constituera le modèle final, son utilisation n'est rendue possible que par la substitution des valeurs manquantes. En effet, il ne peut pas traiter des bases de données incomplètes. Pour que cela soit possible, il eût été envisageable de modifier *Salammbô*. Mais les résultats empiriques de Feelders (1999); Batista et Monard (2003); Ragel et Crémilleux (1999) suggèrent qu'il est préférable de remplacer les valeurs manquantes avant de construire un arbre de décision, plutôt que de modifier l'algorithme de base pour tenir compte de ces valeurs manquantes. Quant à la sélection de variables, elle correspond parfaitement à nos besoins dans la mesure où elle facilite l'interprétation du modèle final sans en dégrader les performances. Dans de nombreux cas elle permet même de les améliorer. Outre ces trois modules principaux que nous venons de décrire, le système que nous avons mis en place contient deux autres modules de moindre importance.

Le premier se charge d'éliminer les attributs et exemples de la base dont le nombre de valeurs observées est trop faible. Ces attributs et exemples sont en effet jugés non fiables : ils ne contiennent pas assez d'information pour que des inférences « saines » puissent être faites. Nous sommes conscient que cela revient à occulter beaucoup d'information, mais nous estimons que pour que cette information puisse être mise en valeur, des mécanismes de gestion de l'incertitude et de l'imprécision doivent être mis en place. Sans ces mécanismes,

qui devraient être intégrés tant au niveau de l'apprentissage à proprement parlé que de la substitution des valeurs manquantes, le risque est grand que certaines règles apprises par *Salammbô* ne soient que purs artéfacts. Se pose alors la question des choix des proportions minimales de valeurs manquantes à partir desquelles un attribut ou un exemple sera jugé non fiable. Nous n'avons pas de réponse précise à apporter à cette question. Ces deux seuils sont deux paramètres de notre système qu'il convient de régler expérimentalement afin de trouver un compromis entre la perte d'information et la conservation d'attributs et d'exemples potentiellement non fiables.

Le second de ces modules annexes a pour objectif de normaliser les valeurs des différents attributs continus afin d'homogénéiser les domaines de définition. Cela n'est pas indispensable pour la construction d'arbres de décision, mais cela l'est pour nombre de méthodes de substitution des valeurs manquantes et de sélection d'attributs qui doivent comparer divers attributs entre eux. Nous avons envisagé deux types de normalisation. La première consiste à faire en sorte que les attributs aient mêmes moyenne et écart-type, tandis que la seconde ramène les domaines de définition de tous les attributs dans l'intervalle $[0; 1]$. Des expériences effectuées sur notre premier modèle indiquent que ces deux types de normalisation conduisent à des résultats similaires. Aussi ne nous étendrons-nous pas plus sur la question du choix de la méthode de normalisation.

L'inconvénient de la normalisation réside dans le fait que les règles induites lors de l'apprentissage ne sont plus aussi aisément interprétables qu'initialement. Mais la normalisation est une transformation réversible des données. Aussi appliquons-nous la transformation inverse aux données, une fois que substitution et sélection ont été effectuées, juste avant la construction de l'arbre de décision.

Nous avons maintenant fait le tour des différents modules composant notre système. Il nous reste à préciser les méthodes de substitution et de sélection que nous comptons mettre en place. Le choix de ces méthodes n'est pas simple. Nos expérimentations de la section 7.6, si elles n'ont pu mettre en évidence la prédominance de telles ou telles méthodes, nous ont tout de même appris que la méthode de sélection d'attributs et la méthode de substitution des valeurs manquantes ne doivent pas être choisies indépendamment l'une de l'autre. C'est une combinaison de deux éléments qu'il convient de choisir, et ce choix doit être fonction du type de données à traiter, de l'algorithme de classification utilisé ainsi que de la mesure d'évaluation de cet algorithme.

Afin de faciliter ce choix, nous proposons de mettre en place la méthodologie d'analyse comparative empirique décrite et utilisée dans la partie II. Autrement dit pour une base de données particulière à traiter, notre système global comparera différentes combinaisons de méthodes de substitution et de sélection, afin de choisir celle qui semble optimale pour une mesure de performance donnée. Afin de pouvoir traiter des bases de données de grandes dimensions, seuls des filtres seront considérés comme méthodes de sélection d'attributs.

Nous ne fixons *a priori* aucune mesure de performance, c'est un autre degré de liberté sur lequel peut jouer l'utilisateur. Nous reviendrons plus en détail sur ce point, mais précisons simplement qu'il nous semble essentiel dans un système d'aide à la décision de faire en sorte que l'utilisateur puisse guider l'apprentissage en fonction de ses attentes et de sa connaissance du domaine, ne serait-ce que pour fixer les coûts associés aux erreurs sur les différentes classes, même de manière approximative.

Pour ce qui est du choix des tests statistiques sur lesquels reposera l'analyse comparative, nous suivons les recommandations faites à la section 5. Aussi opterons-nous pour l'ANOVA lorsqu'une seule base de données sert de socle à l'apprentissage d'un modèle d'évaluation des risques. Mais lorsque plusieurs bases de données sont utilisées, nous nous

tournerons vers le test de Friedman. Contrairement aux expériences menées dans la partie précédente, aucune méthode ne sera considérée comme une référence. Aussi faudra-t-il procéder à la comparaison de toutes les paires de méthodes afin de voir comment elles se comportent les unes par rapport aux autres. Les tests *post-hoc* associés à l'ANOVA et au test de Friedman seront donc respectivement les tests de Tukey et de Nemenyi.

Pour synthétiser l'ensemble de ces remarques, nous donnons à la figure 9.1 la description de l'architecture globale de notre système d'apprentissage.

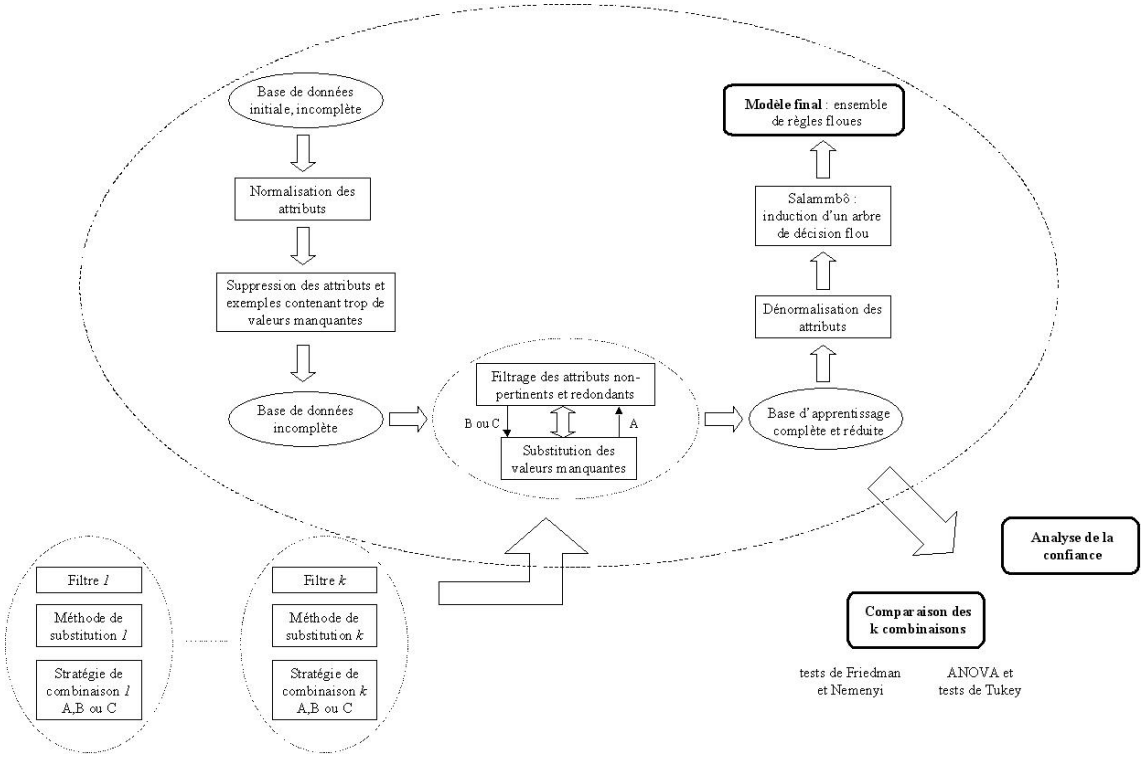

FIG. 9.1 – Architecture du système d'évaluation des risques

9.2 Utilisation du modèle

Le modèle résultant de l'apprentissage permet de classer de nouveaux exemples, c'est-à-dire de les affecter dans l'une des classes existantes. Dans notre application sur la détection des conflits armés intra-étatiques nous avons défini deux classes : *crise* et *non-crise*. Savoir à quelle classe appartient tel ou tel exemple constitue déjà une information importante qui permet de focaliser l'attention sur les exemples à risque. Mais il est également souhaitable de pouvoir quantifier l'incertitude sous-jacente, liée à la classification effectuée par le système. D'une part, ce point est essentiel pour le respect de notre contrainte de transparence : l'utilisateur doit savoir quelle confiance accorder aux décisions du système. D'autre part, dans la perspective de la mise en place de politiques de prévention ciblées, il importe tout autant de connaître la classe de chacun des exemples à traiter que de pouvoir les ordonner selon le degré de certitude que l'on a sur l'occurrence effective des événements sous surveillance. C'est en effet à partir d'un tel ordonnancement que des priorités pourront être fixées lors de l'établissement des politiques de prévention[1].

[1] La magnitude des dommages causés en cas de réalisation de l'événement sous surveillance devrait également être intégrée pour réaliser cet ordonnancement, mais nous avons précisé dans la partie I que

Salammbô construit un arbre de décision flou qui peut s'interpréter comme une base de règles floues (Marsala, 1998). Chaque chemin, composé de k tests portant sur les variables $v_{i_1}, ..., v_{i_k}$, correspond à un ensemble de K règles, où K est le nombre de classes. Pour la classe c_r la règle prend la forme suivante :

$$\text{Si } v_{i_1} = m_{i_1}, ..., v_{i_k} = m_{i_k} \text{ Alors } y = c_r \text{ avec le degré } P^* \left(y = c_r | v_{i_1} = m_{i_1}, ..., v_{i_k} = m_{i_k}\right)$$

Les m_{i_j} correspondent aux modalités floues susceptibles d'êtres prises par les attributs v_{i_j}. P^* désigne la probabilité d'événements flous. Ainsi $P^* \left(y = c_r | v_{i_1} = m_{i_1}, ..., v_{i_k} = m_{i_k}\right)$ correspond à la probabilité qu'un exemple appartenant à la feuille du chemin considéré soit de la classe c_r. Pour les arbres non flous, la pondération par cette probabilité conditionnelle n'intervient pas : chaque feuille n'est étiquetée que par la classe majoritaire.

Pour chaque nouvel exemple e à classer, l'application d'une règle permet de calculer les degrés d'appartenance de e à chacune des classes. Dans *Salammbô*, les calculs de ces degrés d'appartenance sont effectués à l'aide de mesures de satisfiabilité qui évaluent à quel point les caractéristiques de e satisfont les prémisses de la règle. Marsala (1998) a montré que ce calcul est parfaitement équivalent, à condition de choisir les mesures appropriées, à celui qui est effectué lors de l'application du modens ponens généralisé, la méthode traditionnelle utilisée dans les systèmes d'inférence floue. L'intérêt des arbres de décision pour l'évaluation des risques réside selon nous dans l'interprétation de ces arbres comme des ensembles de règles. Aussi présenterons-nous le calcul des degrés d'appartenance sous le formalisme du modus ponens généralisé.

Il s'agit d'une extension du modus ponens au raisonnement déductif sur des données floues. Le principe est le suivant[2]. On suppose connue une règle R de la forme suivante : $P \Rightarrow C$, où P désigne la prémisse, qui peut être complexe (conjonction de prémisses élémentaires), et C la conclusion. Lorsque l'on observe une prémisse P' on peut vouloir appliquer la règle R pour essayer de déduire un nouveau fait à partir de P'. Avec le modus ponens classique on doit avoir $P' = P$ ou $P' \Rightarrow P$ pour que R puisse s'appliquer. Mais dans le cas de données imprécises, on souhaite pouvoir utiliser R même lorsque P' ne correspond pas tout à fait à P. C'est ce que permet le modus ponens généralisé. Plus P' se *rapprochera* de P, et plus le fait C', inféré par l'application de la règle R à P', sera *proche* de la conclusion attendue C. Comme le fait remarquer Marsala (1998), l'avantage du raisonnement déductif *approché*, qui est à l'œuvre lors de l'application du modus ponens généralisé, est qu'il se conforme mieux au mode de raisonnement de l'esprit humain que le modus ponens classique.

Formellement, le modus ponens généralisé permet de calculer la fonction d'appartenance de C', à partir de celle de P, P' et C (Bouchon-Meunier, 2007). Si nous notons μ_A la fonction d'appartenance de A, nous avons :

$$\mu_{C'}(c) = \sup_p \top_{mpg} \left(\mu_{P'}(p), \mu_{imp}(p, c)\right) \tag{9.1}$$

où μ_{imp} désigne la fonction d'appartenance de l'implication choisie pour le modus ponens généralisé, et $\top_{mpg}$ est une t-norme associée à μ_{imp}.

Les observations dont nous disposons, pour un exemple $e_0 = (v_{i_1 0}, ..., v_{i_k 0})$ donné, sont précises. Cela permet de simplifier grandement l'expression de l'équation 9.1. Nous avons en effet :

$$\mu_{P'}(p) = \begin{cases} 1 & \text{si } p = e_0 = (v_{i_1 0}, ..., v_{i_k 0}) \\ 0 & \text{sinon} \end{cases}$$

nous ne nous en préoccupions pas dans cette thèse.

[2]Nous reprenons la description qui en est donnée par Marsala (1998).

L'équation 9.1 s'écrit alors :

$$\mu_{C'}(c) = \mu_{imp}(p = e_0, c) \tag{9.2}$$

L'application de règles floues à des données précises correspond à ce qui est fait en commande floue. Aussi avons-nous décidé de recourir à l'implication de Mamdani qui est particulièrement bien adaptée à ce domaine. L'expression du modus ponens généralisé devient :

$$\mu_{C'}(c) = \min(\mu_P(p = e_0), \mu_C(c)) \tag{9.3}$$

Rappelons que C correspond dans notre cas à l'une des K classes que peut prendre la variable y, et que P est la conjonction de prémisses élémentaires. Nous avons utilisé la t-norme de Zadeh pour exprimer le lien conjonctif qui unit l'ensemble des prémisses élémentaires de chaque règle. Pour chaque classe c_r, nous avons considéré une fonction d'appartenance discrète, nulle pour toute classe $c \neq c_r$ et qui prend la valeur $P^*(y = c_r | v_{i_1} = m_{i_1}, ..., v_{i_k} = m_{i_k})$ pour $c = c_r$. Il est assez simple de montrer que dans ce cas, les classes étant précises, nous avons :

$$\begin{aligned} P^*(y = c_r | v_{i_1} = m_{i_1}, ..., v_{i_k} = m_{i_k}) &= P(y = c_r | v_{i_1} = m_{i_1}, ..., v_{i_k} = m_{i_k}) \\ &= \frac{n^{c_r}_{v_{i_1}=m_{i_1},...,v_{i_k}=m_{i_k}}}{n_{v_{i_1}=m_{i_1},...,v_{i_k}=m_{i_k}}} \end{aligned}$$

où $n^{c_r}_{v_{i_1}=m_{i_1},...,v_{i_k}=m_{i_k}}$ est le nombre d'exemples qui appartiennent à la classe c_r parmi ceux qui appartiennent à la feuille correspondant à $v_{i_1} = m_{i_1}, ..., v_{i_k} = m_{i_k}$.
$n_{v_{i_1}=m_{i_1},...,v_{i_k}=m_{i_k}}$ est le nombre d'exemples de cette feuille.

L'application du modus ponens généralisé, pour la règle R, nous permet donc de conclure que le degré d'appartenance de l'exemple e_0 à chacune des K classes est de la forme suivante :

$$\forall r \in \{1, ..., K\}, \mu_{c_r}(e_0) = \min\left(\min\left(\mu_{v_{i_1}}(v_{i_1 0}), ..., \mu_{v_{i_k}}(v_{i_k 0})\right), \frac{n^{c_r}_{v_{i_1}=m_{i_1},...,v_{i_k}=m_{i_k}}}{n_{v_{i_1}=m_{i_1},...,v_{i_k}=m_{i_k}}}\right)$$

Pour chaque chemin, nous obtenons donc un degré d'appartenance de e_0 à chacune des classes. Dans le cas général, du fait de l'introduction du flou, tous les chemins sont activés lors du classement d'un exemple. Pour obtenir le degré d'appartenance global de e_0 à chacune des classes il nous faut donc agréger les degrés d'appartenance obtenus pour chacun des chemins. Marsala (1998) conseille d'utilise une t-conorme comme opérateur d'agrégation, l'idée étant que s'il existe au moins une règle qui attribue à un exemple un degré d'appartenance élevé à une certaine classe, le degré d'appartenance, pris sur l'ensemble de la base de règles, doit être au moins aussi élevé. C'est le maximum qui est utilisé dans *Salammbô*. La classification de e_0 par *Salammbô* consiste simplement à choisir la classe r dont le degré d'appartenance est le plus élevé. Ce degré peut donc s'interpréter comme un degré de confiance dans la décision prise par *Salammbô*.

Notons que si l'on avait pris l'implication de Larsen, à la place de celle de Mamdani (le produit remplace le minimum), le degré d'appartenance à une classe, étant donné un chemin de l'arbre, s'interpréterait directement comme la probabilité conditionnelle de cette classe sachant la feuille de ce chemin, pondérée par le degré de satisfaction de la prémisse :

$$\mu_{c_r}(e_0) = \mu_P(e_0) \times P(c_r | e_0)$$

Cette dernière expression met en évidence le lien qui existe entre l'approche floue et l'approche classique qu'elle généralise. En effet, dans le cas d'arbres de décisions non flous, seule la probabilité conditionnelle de la classe choisie est utilisée pour ordonner les différents exemples.

De nombreux travaux ont mis en avant les faiblesses de l'estimateur de $P(c_r|e_0)$ utilisé dans les arbres de décisions (Zadrozny et Elkan, 2001; Provost et Domingos, 2003; Alvarez *et al.*, 2007). Lorsque le nombre d'exemples d'une feuille est trop réduit, cet estimateur ne saurait être fiable. Pour pallier cette faiblesse, *Salammbô* n'utilise pas directement le degré d'appartenance à chaque classe pour prendre une décision : il le pondère par le rapport entre le nombre d'exemples de la classe appartenant à la feuille et le nombre d'exemples total n. Le score final s_{c_r}, relativement à chaque classe, s'exprime alors de la façon suivante :

$$\forall r \in \{1, ..., K\}, s_{c_r}(e_0) = \mu_{c_r}(e_0) \times \frac{n^{c_r}_{v_{i_1}=m_{i_1},...,v_{i_k}=m_{i_k}}}{n}$$

Contrairement aux travaux que nous venons de mentionner sur l'élaboration de probabilités conditionnelles fiables, le score s_{c_r} que *Salammbô* construit n'est pas destiné à fournir des probabilités bien calibrées, mais à quantifier la confiance que l'on peut avoir dans les décisions qui sont prises. C'est à partir de ce score que les différents exemples à traiter seront ordonnés.

Chapitre 10

Application aux conflits armés intra-étatiques

Nous avons insisté tout au long de cette thèse sur l'importance du caractère empirique de notre démarche. Nous allons appliquer dans cette section notre système d'évaluation des risques sur le cas concret qui a motivé ces travaux, à savoir l'anticipation des conflits armés intra-étatiques. Nous avons déjà mené une série d'expériences sur ce sujet au chapitre 2. Plutôt que de partir des mêmes données, nous avons préféré étendre le spectre de notre étude. Notre objectif ne se réduit pas simplement à la mise en évidence des améliorations apportées au système initial. Nous souhaitons d'une part pouvoir mettre en relief les forces et faiblesses de notre approche de manière générique. D'autre part, nous souhaitons que cette étude empirique réponde aux attentes des experts en veille géostratégique. Les données collectées lors de nos premières expérimentations, ainsi que le protocole suivi initialement, comportent bien trop de lacunes pour cela. Aussi a-t-il fallu les modifier. Ces modifications et leurs conséquences étant loin d'être mineures, nous les présentons à la section 10.2. Nous serons alors en mesure de décrire et analyser les résultats obtenus sur ces nouvelles données, ce que nous ferons aux sections 10.3 et 10.4.

10.1 Théories sur l'émergence des conflits

Sans prétendre couvrir l'ensemble des théories relatives au déclenchement des guerres civiles, nous allons introduire succintement les principaux paradigmes. Ceci nous permettra de mettre en évidence les variables clés identifiées par les chercheurs en sciences politiques ; variables que nous avons ajoutées à notre base de données afin d'être en mesure d'inclure les théories afférentes dans l'espace des modèles susceptibles d'être appris par notre système. Nous récapitulerons à la section 10.2.3 l'ensemble de ces variables de manière plus plus synthétique.

10.1.1 Modèle de Gurr

Parmi les premiers travaux sur les guerres civiles, ceux de Gurr (1971) ont particulièrement influencé le domaine. Il a proposé et testé empiriquement l'un des premiers modèles formels de déclenchement des conflits armés intra-étatiques (Gurr et Harff, 1998; Moore et Gurr, 1998). Son modèle[1] distingue les racines profondes d'un conflit ou causes structurelles, les facteurs accélérateurs et les événements déclencheurs. L'analyse structurelle a pour objectif d'identifier les États fragiles dans lesquels une crise est susceptible de se déclencher, tandis que l'analyse des accélérateurs et déclencheurs vise à repérer les séquences d'événements qui font basculer un État fragile dans la crise. Cette dernière se rapproche donc d'un suivi événementiel des conflits tel que celui qui a été proposé par Mouillet (2005). Aussi nous contenterons-nous ici d'évoquer son travail relatif aux facteurs structurels à l'origine des guerres civiles.

Selon Gurr, pour qu'un groupe d'individus se rebelle, trois conditions doivent être remplies. Les motivations ainsi que la capacité du groupe à mener une action collective doivent être suffisantes. Enfin les occasions favorisant le passage à l'action doivent être réunies. Divers éléments de ce modèle ont été fréquemment repris dans la littérature. Les principales différences interviennent dans le choix des variables destinées à tester l'influence de chacune de ces trois conditions. Les études empiriques nécessitent en effet le choix d'un certain nombre de variables explicatives pour modéliser les guerres civiles. Ce n'est que par l'intermédiaire de ces variables que les hypothèses théoriques sont testées. Comme nous le verrons par la suite l'interprétation des variables choisies pour modéliser tel ou tel phénomène est souvent sujette à discussion (Lichbach *et al.*, 2004). Aussi avons-nous fait en sorte d'inclure dans notre analyse un maximum de variables jugées importantes dans la littérature afin que les principales théories sur l'origine des conflits soient effectivement prises en compte. L'ensemble des variables que nous avons introduites sont décrites à l'annexe E. Les sources de ces variables sont, quant à elles, présentées à l'annexe F.

10.1.1.1 Sources de discorde

Pour Gurr les motivations jouent un rôle fondamental. Plus un groupe est défavorisé, discriminé par l'État vis-à-vis d'autres groupes, plus ses doléances[2] seront conséquentes et plus il sera incité à se rebeller. Les indicateurs ayant trait aux inégalités de revenus, au respect des libertés individuelles, au niveau de développement, sont essentiels pour tester l'hypothèse selon laquelle le ressentiment d'un groupe est un facteur primordial expliquant l'émergence des conflits. Des indicateurs issus de la Banque mondiale, le taux de mortalité infantile et autres indicateurs de développement, permettent de rendre compte de cette hypothèse (Goldstone *et al.*, 2000). Le nombre d'années durant lesquelles l'autonomie

[1] Le pluriel serait plus adéquat, étant donné que plusieurs variations autour du modèle que nous présentons ont été proposées par Gurr.

[2] Les anglo-saxons parlent de *grievance*.

d'un territoire a été supprimée fait également partie de cette famille d'indicateurs (Moore et Gurr, 1998). Nous avons ajouté l'intensité maximale des tremblements de terre. Les catastrophes naturelles peuvent en effet contribuer à léser des minorités et renforcer par là même leur ressentiment vis-à-vis du gouvernement en place.

Ambivalence de l'hétérogénéité de la population Les indicateurs reflétant l'hétérogénéité de la population sont fréquemment utilisés dans la littérature pour appuyer l'hypothèse précédente, à savoir que les doléances, les griefs d'un groupe à l'égard du gouvernement représentent un facteur de risque essentiel pour la stabilité d'un État. L'influence sur les guerres civiles des différences culturelles, généralement décomposées en différences ethno-linguistiques et religieuses, a été mise en avant par les travaux de Horowitz (1985) et est souvent considérée comme un fait établi dans l'opinion publique, principalement à cause du succès médiatique rencontré par la thèse du choc des civilisations de Huntington (1993).

Selon ces travaux, les différences entre communautés, ou plutôt la coexistence de communautés différentes, exacerbent l'inimitié inter-communautaire, ce qui peut conduire à l'usage de la violence. Les différences culturelles rendent en effet plus difficiles la communication et les échanges entre communautés, ce qui peut contribuer à accroître l'incompréhension dans le pays. Du fait de l'implantation transnationale de certaines ethnies, des tensions inter-étatiques peuvent de plus favoriser l'éclatement de conflits intra-étatiques. Notons que le lien de causalité peut parfois être inversé.

Les études théoriques et quantitatives sur le sujet sont cependant beaucoup plus nuancées et controversées que ne le suggèrent nos propos. La controverse porte sur trois sujets :

- Comment mesurer la diversité culturelle de la population d'un pays ?
- Au regard de l'histoire, peut-on effectivement conclure qu'il existe un lien entre diversité de la population et risque de guerre civile ?
- Si oui, quelle est la nature de ce lien et comment l'interprète-t-on ?

Ces trois sujets ne sont bien évidemment pas indépendants. La mesure de diversité est généralement choisie en fonction de ce que l'on souhaite montrer et donc de l'interprétation que l'on a du rôle joué par la diversité.

On distingue habituellement les mesures de fractionnement (Fearon, 2002) des mesures de polarisation (Garcia-Montalvo et Reynal-Querol, 2002), appliquées aux différences ethno-linguistiques ou religieuses. Les mesures de fractionnement correspondent à des mesures d'hétérogénéité. Elles mesurent la probabilité que deux individus tirés aléatoirement dans la population appartiennent à un même groupe. L'interprétation qui sous-tend l'utilisation de cette mesure est qu'une société sera plus susceptible de basculer dans la guerre civile si elle est hétérogène. De nombreuses études théoriques et empiriques observent cependant l'existence d'une relation de dépendance quadratique (en U inversé) plutôt que linéaire entre hétérogénéité et risque de conflit, le risque étant moindre pour les sociétés très homogènes et très hétérogènes (Hegre *et al.*, 2001; Collier *et al.*, 2006; de Soysa, 2004; Miguel *et al.*, 2004).

Dans une société fortement hétérogène aucun groupe ne peut prendre seul le pouvoir. Des coalitions se forment donc naturellement, réduisant ainsi le nombre de groupes potentiellement exclus des organes de décision. Ces groupes sont par ailleurs de taille insuffisante pour se rebeller. Une grande hétérogénéité est en effet marquée par la présence d'un grand nombre de groupes de faible taille. Les sociétés à risque sont donc telles qu'il existe au moins deux groupes de taille conséquente, l'un des deux étant exclu du pouvoir. On parle

alors de sociétés polarisées. Les mesures de polarisation visent à quantifier directement cette caractéristique et à remédier à certaines faiblesses des mesures de fractionnement (Garcia-Montalvo et Reynal-Querol, 2002).

Contrairement aux études précédemment citées, Fearon et Laitin (2003); Fearon (2005) n'observent aucune relation de dépendance, entre hétérogénéité et risque de conflit, alors que Collier et Hoeffler (1998: 2004) constatent au contraire qu'un accroissement de la diversité est associé à une diminution du risque de conflit. Précisons que dans une étude postérieure à celles que nous venons de citer, Collier *et al.* observent l'effet inverse. Schneider et Wiesehomeier (2006) observent quant à eux que les résultats de Garcia-Montalvo et Reynal-Querol (2002), mettant en évidence l'impact de la polarisation et du fractionnement ethnique et religieux sur les guerres civiles, ne sont valides que lorsque ces mesures sont utilisées pour expliquer l'occurrence d'un conflit et non son déclenchement. Ceci tend à montrer que la diversité de la population ne joue un rôle que sur la durée des conflits, qu'elle tend à rallonger.

Notons enfin que Sambanis (2004) arrive à des conclusions contrastées selon la définition de guerre civile utilisée. Ceci suggère que les conclusions des diverses études sont fragiles et peu robustes. Les divergences observées dans les résultats et les interprétations de Collier à ce sujet, sont révélatrices de cette fragilité. Comme le note Sambanis, la persistance de controverses relatives à l'interprétation des mesures de diversités ethnique et religieuse suffit à justifier la nécessité d'inclure ces mesures dans toute étude sur les guerres civiles.

Choix d'une mesure de l'hétérogénéité Nous avons intégré dans notre base de données deux indicateurs de fractionnement ethnique et religieux. Ils se calculent à l'aide de la formule de Herfindahl :

$$h = 1 - \sum_{i=1}^{k} \pi_i^2$$

où k est le nombre de groupes et π_i est le pourcentage de la population appartenant au groupe i.

La principale difficulté pour construire ces indicateurs réside dans l'obtention d'une liste des groupes ethniques ou religieux et de la part de la population qui les compose. La plupart des études sur le sujet utilisent comme indicateur de diversité ethnique l'indice de fractionnement ethno-linguistique basé sur la liste des groupes recensés par des ethnographes soviétiques en 1964 dans l'*Atlas Narodov Mira*, référencé en particulier dans l'article de Fearon (2002). Les décisions prises par les ethnographes soviétiques pour construire cette liste sont parfois discutables, certains pays étant couverts plus finement que d'autres. Fearon (2002) donne plusieurs exemples litigieux. Aussi a-t-il dressé, à partir de différentes sources, une nouvelle liste des groupes ethno-linguistiques plus fidèle à sa conception de l'identité ethnique.

L'intérêt du travail de Fearon réside selon nous dans la formalisation des principales règles employées pour établir la délimitation entre les différentes ethnies. Il a de plus clairement mis en évidence les décisions critiques qu'il a prises. Sans entrer dans les détails, précisons simplement que, reprenant à son compte certains des arguments des constructivistes, il a essayé de tenir compte de la perception qu'ont les individus de leur appartenance à un groupe. Nous employons le terme « essayé » car, ainsi qu'il le reconnaît lui-même, la meilleure méthode serait encore de demander directement aux individus ce qu'ils en pensent *via* des questionnaires.

Nous avons utilisé le résultat de ses travaux pour construire notre indicateur de fractionnement ethnique. Notre indicateur de fractionnement religieux a été construit à partir de la liste des groupes religieux fournie par le *World Factbook* de la CIA. Afin de tenir

compte des phénomènes de polarisation, nous avons également construit à partir de la liste des groupes ethniques et religieux quatre autres indicateurs de diversité. Il s'agit du nombre de groupes différents ainsi que du rapport entre la taille du groupe majoritaire et celle du deuxième groupe le plus important. Du fait des ressources limitées à notre disposition, nous avons considéré que ces six indicateurs sont invariants dans le temps. Il s'agit là d'une réduction contraire aux idées constructivistes énoncées précédemment, et qui de plus ne tient pas compte des phénomènes de migration, de fécondité ou de mortalité différentielle, voire des génocides qui modifient la structure ethnique d'un pays. Mais à l'instar de la plupart des chercheurs du domaine, nous supposons que les variations sur le plan mondial à l'échelle d'une trentaine d'années sont suffisamment faibles pour ne pas influer sur nos modèles.

10.1.1.2 Identité et cohésion des groupes rebelles

La capacité intrinsèque d'un groupe à agir est généralement occultée dans les études quantitatives sur les guerres civiles. Elle joue pourtant un rôle important dans le modèle de Gurr. Elle est d'autant plus forte que l'identité et la cohésion du groupe sont fortes. Il est important de noter que ces notions sont fortement infuencées par l'étendue des discriminations subies par le groupe considéré (Fearon, 2002). Aussi avons-nous supposé que les variables introduites précédemment suffisaient à rendre compte de ces phénomènes. Comme le notent Collier et Hoeffler (2004), les variables relatives à l'hétérogénéité de la population influent tant sur la force du ressentiment de certains groupes que sur leur cohésion. Plus la population est hétérogène et plus les différents groupes seront de petite taille ce qui affecte la cohésion et l'identité. Ils justifient de cette façon leur observation de l'effet stabilisant du fractionnement ethnique sur le déclenchement des conflits.

10.1.1.3 Occasions facilitant le déclenchement des guerres civiles

Gurr insiste sur l'importance des occasions qui rendent un groupe minoritaire plus fort, comme le soutien financier, armé de la diaspora ou de pays étrangers favorables à la rébellion. Il considèrent également les occasions qui rendent le gouvernement plus faible et donc moins susceptible de contenir une rébellion.

La puissance et les ressources du régime sont des exemples d'indicateurs reflétant cette capacité de l'État à user de moyens de coercition pour prévenir toute insurrection. Moore et Gurr (1998) incluent également dans cette catégorie la durée et la force de l'expérience démocratique. Ces indicateurs sont à double interprétation.

Ambivalence de la démocratie D'une part, les démocraties offrent un environnement plus favorable à la rébellion que les dictatures du fait de la répression moindre qu'elles mettent en place. D'autre part, plus respectueuses des libertés, elles génèrent moins de ressentiment dans la population et offrent un mode de contestation alternatif à la violence. Une révolte potentielle sera donc plus difficile à légitimer. Par conséquent il sera plus difficile de rallier à sa cause de nouvelles recrues. Le caractère démocratique d'un État est donc un facteur qui influe en sens contraires sur deux des trois dimensions identifiées par Gurr. Pour rendre compte de cette caractéristique étatique, Gurr et la majorité des chercheurs utilisent un indicateur agrégé issu du projet *Polity IV* qui fournit un score entre -10 et +10 correspondant au degré de démocratie d'un État. Les dictatures « pures » ont un score de -10 et les démocraties « pures » ont un score de +10. L'influence des institutions sur le déclenchement d'un conflit est loin d'être évident, les études empiriques sur le sujet ainsi que leurs interprétations divergeant assez nettement.

Les travaux de Hegre *et al.* (2001) ont mis en évidence l'existence d'un lien quadratique (dépendance en U inversé) entre degré de démocratie et risque de guerre civile, les démocraties et dictatures étant moins susceptibles d'être en crise que les États ayant un degré de démocratie proche de 0. Ces derniers sont qualifiés d'*anocraties* dans la littérature. Il s'agit d'États en transition qui présentent des éléments caractéristiques des deux types de régime.

L'interprétation de ce résultat est la suivante. Le risque de guerre civile est plus grand pour les *anocraties* car elles n'ont ni la capacité de répression des dictatures, ni les institutions démocratiques suffisantes pour que s'expriment pacifiquement les désaccords de la population. Selon le protocole expérimental utilisé, certaines études empiriques abondent dans ce sens (Hegre et Sambanis, 2006; Goldstone *et al.*, 2000), tandis que d'autres n'observent aucune influence notable des institutions sur le déclenchement des conflits (Collier *et al.*, 2006; Fearon et Laitin, 2003; de Soysa, 2004).

De la difficulté de mesurer le niveau de démocratie La principale critique de l'attention accordée aux *anocraties* a été formulée par Gandhi et Vreeland (2004). Les auteurs remettent en question l'existence même d'un indicateur de démocratie ainsi que la notion d'*anocratie*. Selon eux, ce concept est flou. La seule caractérisation sur laquelle s'accordent les différents chercheurs est liée à l'indicateur de démocratie : une *anocratie* correspond à un État dont le degré de démocratie se situe au milieu de l'échelle *Polity IV*. Gandhi et Vreeland, en revenant aux définitions utilisées dans le projet *Polity IV*, notent qu'un État se verra attribuer un degré de démocratie proche de 0 s'il est en proie à des conflits armés. Il y a donc un biais d'endogénéité fort lorsque l'indicateur de démocratie est utilisé pour expliquer le déclenchement d'un conflit. Observer un risque plus élevé pour les *anocraties* n'indique rien sur le rôle joué par les institutions. Du moins il est difficile de savoir si l'effet observé est dû à la nature des institutions ou à l'histoire des conflits dans le pays. À cause de ce biais et parce que nous avons également de sérieux doutes quant à l'existence d'un continuum entre démocratie et dictature, nous n'avons pas retenu cet indicateur.

Notons que suite aux travaux de Gandhi et Vreeland, Gleditsch *et al.* (2006) ont remis en question l'existence d'une relation de dépendance quadratique entre degré de démocratie et risque de conflit qu'ils avaient mise en avant dans une précédente étude (Hegre *et al.*, 2001). Ils ont en effet observé qu'une telle relation n'était valide qu'avec certains indicateurs de démocratie ce qui suggère un manque de robustesse des conclusions qu'ils avaient préalablement avancées.

10.1.2 Modèles centrés sur les occasions favorisant le déclenchement des conflits

10.1.2.1 Critiques du modèle de Gurr

Le modèle que nous venons de décrire a été vivement critiqué par différents chercheurs qui lui reprochent d'accorder une importance indue aux doléances ou *grievance* des rebelles. Ainsi Laitin (2004), s'appuyant sur une étude de cas, relève que les discriminations linguistiques dont peuvent être l'objet certaines communautés sont totalement décorrélées de l'occurrence de guerres civiles. Il observe en particulier que les concessions faites par le gouvernement pour reconnaître les langues de certaines communautés n'empêchent pas le moins du monde l'éclatement de la violence. Ceci va à l'encontre de l'idée selon laquelle redresser les torts subis par les minorités constitue une politique de prévention efficace. Rien ne permet de conclure qu'il ne s'agit pas d'une condition nécessaire, mais les faits contredisent en revanche l'idée selon laquelle de telles mesures préventives sont suffisantes.

Nous allons maintenant présenter les modèles concurrents qui ont pu être proposés dans la littérature. Leur pouvoir explicatif semble suffisamment fort pour que nous fassions en sorte d'introduire les variables à partir desquelles ils sont construits. Les articles les plus influents de ce courant de pensée sont à mettre à l'actif de Collier et Hoeffler (2004); Collier *et al.* (2006) d'une part et de Fearon et Laitin (2003); Fearon (2005) d'autre part.

Rejetant l'hypothèse selon laquelle les motivations d'un groupe d'individus jouent un rôle primordal dans leur décision de se rebeller, ils mettent l'accent sur l'importance des conditions rendant possibles la rébellion. Ils ne nient pas l'importance des griefs et du ressentiment dans l'éclatement des conflits, mais ils estiment que ce sont là des facteurs universellement répandus qui ne suffisent absolument pas à expliquer le déclenchement des conflits. Selon cette théorie, les griefs ressentis par la population sont suffisants dans tout pays pour expliquer que des groupes aient l'intention de se rebeller. Mais les rébellions n'éclatent que là où elles sont viables. Dans un premier temps l'article volontairement polémique de Collier et Hoeffler (2004) suggère, ne serait-ce que par son titre, *Greed and Grievance*, que la motivation principale des rebelles est l'appât du gain, l'avidité, et non le souci de réparer une injustice ou de réduire les inégalités. Mais leur résultat principal est le même que celui que nous venons d'évoquer : les occasions favorisant la rébellion expliquent bien mieux l'émergence des conflits que les motivations. Dans un article postérieur, ils adoptent d'ailleurs un ton plus consensuel, insistant sur la *faisabilité* de la guerre civile, plutôt que sur les motivations des différents acteurs (Collier *et al.*, 2006).

10.1.2.2 Méthodologie d'identification des variables pertinentes

L'influence de la théorie des jeux marque les travaux qui se situent dans ce courant de pensée. La guerre civile est considérée comme un jeu entre deux acteurs : le gouvernement et un groupe rebelle[3]. Les recherches théoriques s'attachent alors à identifier les conditions qui doivent être remplies pour qu'un groupe rebelle décide de prendre les armes contre le gouvernement en place. Des études empiriques sont ensuite menées pour valider ou invalider telle ou telle hypothèse théorique.

Les conditions favorisant l'éclosion d'une insurrection sont perçues comme des facteurs de risque. Leur identification est généralement réalisée *via* une analyse *coûts-revenus*[4]. Cette analyse consiste à faire la part des coûts imposés par le gouvernement aux rebelles potentiels (répression) des gains que ceux-ci peuvent espérer (reconnaissance personnelle, pillage, meilleur niveau de vie...). L'objectif est de repérer les facteurs qui offrent un avantage conccurentiel aux futurs rebelles sur le gouvernement et réciproquement. Les travaux de Fearon et Laitin ainsi que ceux de Collier et Hoeffler sont exactement dans cette lignée. Même si leurs interprétations diffèrent parfois sur certains points comme par exemple sur le rôle joué par l'exportation de matières premières (Fearon, 2005), ils développent des idées et un argumentaire voisins.

Contrairement à Fearon et Laitin qui sont des chercheurs en sciences politiques, Collier et Hoeffler sont des économistes de formation. Ceci peut expliquer que leurs interprétations divergent quelque peu sur certains points. Collier et Hoeffler considèrent que la guerre civile est une activité économique comme les autres et insistent sur les facteurs qui poussent les

[3] Certains considèrent $n+1$ joueurs : le gouvernement et n individus. L'objectif est alors de savoir si les individus acceptant d'entrer en rébellion sont en nombre suffisant pour que la rébellion puisse être viable (Epstein, 2002).

[4] Les anglo-saxons parlent de *cost-benefit* analysis, ce qui est fréquemment traduit par analyse *coûts-bénéfices*. Mais le terme bénéfice est impropre du fait de sa connotation en comptabilité. Il s'agit non pas des gains espérés, mais de la différence entre gains et pertes.

individus à se détourner d'une activité économique conventionnelle pour gagner les rangs de la rébellion.

10.1.2.3 Variables reflétant l'importance des occasions propices à la rébellion

Les principales variables introduites par ces auteurs et qui se sont révélées empiriquement importantes pour modéliser les guerres civiles sont à peu près les mêmes et sont listées ci-après. L'interprétation fournie pour expliquer le rôle joué par ces variables diffèrent cependant quelque peu selon les auteurs, comme nous allons le voir. La relation observée entre les différentes variables et le risque de conflit est toujours monotone. Aussi ne le préciserons-nous pas. Nous indiquerons simplement si la variable a un effet aggravant ou stabilisant sur ce risque selon qu'elle favorise ou défavorise le déclenchement des conflits. Nous avons introduit dans notre base de données la plupart de ces variables (voir annexe E). Lorsque tel n'est pas le cas nous le signalerons clairement.

PIB par habitants Plus il est élevé et plus le risque de conflit est faible. Il a donc un effet stabilisant sur le déclenchement des conflits. Pour Collier et Hoeffler ce phénomène correspond au fait que plus le niveau de vie est élevé et plus il est difficile de convaincre les individus de renoncer aux revenus issus de l'économie conventionnelle pour une activité à haut risque telle que la rébellion.

Pour Laitin et Fearon, le PIB par habitant est le reflet de la force des structures étatiques. Plus il est élevé et plus l'État est à même de dégager des revenus conséquents de la taxation. Il peut alors les mettre à profit pour consolider sa capacité de répression : police de taille conséquente et bien équipée, infrastructures développées et en bon état qui permet d'agir rapidement contre des foyers insurrectionnels... Cette capacité de répression suffit à dissuader d'éventuels rebelles de prendre les armes car les risques encourus et donc les coûts attendus sont alors trop élevés par rapport aux gains espérés.

Dans les deux cas, c'est le recrutement des rebelles qui est affecté par le niveau de vie. L'importance accordée aux conditions favorisant le recrutement des rebelles est cependant critiquable. Comme le notent Hendrix et Glaser (2005) en prenant l'exemple des enfants soldats en Sierra Leone ou au Liberia, le recrutement ne se fait pas toujours sur la base du volontariat. Une dernière interprétation tout aussi plausible est qu'un niveau de vie faible renforce les griefs de la population à l'égard du gouvernement. Cet argument corrobore la théorie évoquée initialement selon laquelle les motivations des rebelles priment. Mais Collier et Hoeffler le rejettent car les inégalités de revenus, qui devraient également jouer un rôle selon cette théorie, n'ont pas une importance significative empiriquement.

Taux de croissance du PIB par habitants Comme le PIB par habitant, il a un effet stabilisant. Les trois interprétations données précédemment sont à peu près les mêmes. Collier et Hoeffler ont soulevé les problèmes d'endogénéité qui pouvaient affecter cette variable, à savoir qu'une faible croissance peut n'être que le résultat de conflits précédents ou en cours. Il n'y aurait donc pas de lien causal mais une simple corrélation entre croissance et conflits.

Miguel *et al.* (2004) ont montré qu'il n'en était rien en instrumentant la croissance par les variations pluviométriques en Afrique. Sur ce continent, l'agriculture joue un rôle prépondérant dans l'économie et l'irrigation étant globalement faible, cette agriculture est fortement dépendante du climat. Pour l'Afrique, il est donc raisonnable de supposer que les variations pluviométriques, qui sont indépendantes des conflits, influent directement sur le taux de croissance. Ils ont observé que les chutes de pluie diminuaient le risque de conflit, tandis que ce risque était accru en période de sécheresse, ce qui confirme que les

facteurs économiques tels que le taux de croissance ont effectivement un impact sur les guerres civiles.

Exportations de matières premières Elles ont un effet aggravant. Selon Collier et Hoeffler, cet effet s'explique par le fait que les matières premières sont des ressources convoitées par les groupes rebelles pour financer leur armée. Ces ressources se trouvent pour la majorité d'entre elles dans des zones rurales qui occupent une superficie importante. Ceci facilite leur pillage puisqu'il est plus difficile pour le gouvernement de surveiller ces zones.

Mais l'exportation de matières premières constitue également une source importante de revenus pour le gouvernement. Collier et Hoeffler considèrent que lorsque ces exportations sont importantes, les gains qu'elles assurent au gouvernement sont supérieurs à ceux que les rebelles peuvent obtenir en en détournant une partie. Il peut alors investir de façon à disposer d'une capacité de répression suffisante pour contenir la capacité d'action que les rebelles sont susceptibles de développer. Ils estiment donc, et observent empiriquement, que le lien entre cette variable et le risque de conflits est quadratique. L'inclusion de cette variable dans le modèle de Collier et Hoeffler est destinée à rendre compte de conflits dans lesquels il est notoire que les rebelles exploitent certaines ressources afin de financer leur guerre. Ce phénomène a pu être observé, entre autres, en Colombie avec la coca, en Sierra Leone avec les diamants ou encore au Nigeria avec le pétrole. La création du processus Kimberley, dont l'objectif est de garantir que les diamants bruts écoulés ne sont pas originaires de zones de conflits, est assez significatif de l'importance accordée par la communauté internationale à l'influence des matières premières sur les conflits.

Exportations de pétrole Elles ont un effet aggravant. Fearon (2005) rejette l'existence d'une dépendance forte, qu'elle soit linéaire ou non, entre exportations de tout type de matières premières et risque de conflits. Il estime que toutes les matières premières n'ont pas la même influence. Selon lui, c'est la dépendance d'une économie vis-à-vis des exportations de pétrole qui importe.

Mais plutôt que de considérer que ce phénomène correspond à une occasion pour les rebelles de se financer *via* la taxation de ces exportations, il l'interprète comme le signe de structures étatiques faibles. Un État fortement dépendant de l'exportation d'hydrocarbures dispose d'une source de revenus conséquente sans qu'il soit nécessaire de mettre en place une bureaucratie, des infrastructures et un système de collecte d'impôts efficaces. De tels États auront donc tendance à être plus fragiles que les autres[5]. Cette fragilité, renforcée par la corruption, offre un avantage comparatif indéniable à d'éventuels rebelles.

Notons qu'il est également possible d'interpréter la fragilité et la corruption de l'État comme des facteurs contribuant au ressentiment de la population.

Nombre d'habitants Il a un effet aggravant. Collier et Hoeffler considèrent qu'un nombre élevé d'habitants favorise le recrutement de rebelles, tandis que Laitin et Fearon estiment qu'un nombre élevé d'habitants rend plus difficile la mise en place de politiques de répression.

On peut encore une fois interpréter cette influence du nombre d'habitants sur le risque de conflits sous l'angle de la théorie des motivations, en utilisant des arguments néo-malthusiens. En effet, pour un même niveau de ressources, il est plus difficile de subvenir aux besoins d'une population plus nombreuse, ce qui exacerbe les tensions et peut mener

[5] Cette argumentation peut paraître un peu légère car elle ne concerne que certains des exportateurs d'hydrocarbures. L'Irak avant l'invasion américaine de 2003, ou encore l'Iran et le Venezuela sont difficilement assimilables à des États dans lesquelles la bureaucratie et les infrastructures sont sous-développés.

à des conflits pour l'accès aux ressources. La densité de population joue donc un rôle très important selon cette dernière interprétation.

Nombre d'années écoulées depuis le dernier conflit Il a un effet stabilisant. Collier et Hoeffler parlent du piège de la guerre civile (*civil war trap*) pour évoquer le cercle vicieux dans lequel sont pris les États en crise : plus la dernière guerre civile que le pays a connue est récente et plus il probable qu'une nouvelle crise se déclenche. Cela peut s'expliquer par le fait qu'après une guerre civile, des stocks d'armes sont toujours dans le pays et que les hommes sont bien formés pour les utiliser.

Laitin et Fearon insistent plus, quant à eux, sur la désorganisation et la fragilité du gouvernement après une guerre civile. Mais on peut tout aussi bien considérer qu'une guerre civile accentue les ressentiments, l'esprit de vengeance étant d'autant plus marqué que la guerre est récente. Les motivations de rebelles potentiels sont donc renforcées.

Part des hommes âgés de 15 à 29 dans la population Elle a un effet aggravant, ce qui traduit le fait que le recrutement de rebelles est rendu plus facile par la présence d'un grand nombre d'hommes en âge de se battre.

Proportion de terrains montagneux Elle a également un effet aggravant, ce qui s'interprète par le fait qu'un terrain difficile offre un refuge naturel aux rebelles et rend plus difficile le déploiement et l'action des forces gouvernementales.

À la place de cette variable nous avons construit l'indicateur *Delta Altitude* qui correspond à la différence entre le point culminant et le point le plus bas d'un pays (Fearon et Laitin, 2003).

Nouvel État Il s'agit d'une variable binaire prenant la valeur 1 si l'État considéré est indépendant depuis moins de deux ans (borne incluse) et 0 sinon. Elle a un effet aggravant. Un État qui vient d'accéder à l'indépendance est en pleine structuration, donc fragile et mal organisé. Les risques de répression sont donc moindres pour les rebelles.

Une autre interprétation est que l'indépendance d'un État se fait toujours au détriment d'une certaine frange de la population dont le ressentiment est vif au lendemain de l'indépendance. Plutôt que cette variable binaire nous avons choisi de considérer le nombre d'années écoulées depuis l'indépendance.

Instabilité institutionnelle Elle a aussi un effet aggravant. L'interprétation de Laitin et Fearon est ici identique à celle de la variable précédente. Des institutions instables sont le signe d'un État fragile, incapable de mettre en place les structures préventives et répressives nécessaires. Cette variable binaire est construite à partir de l'indicateur de démocratie *Polity IV*. Elle prend la valeur 1 (instabilité) si le pays a connu un changement d'au moins trois points sur l'échelle de démocratie *Polity IV* dans les trois dernières années. Pour les raisons évoquées précédemment, qui nous font douter de l'objectivité de l'indicateur de démocratie, nous ne l'avons pas inclus.

10.1.3 Histoire des conflits et géographie

Nous avons décidé d'ajouter un certain nombre de variables afin que soient mieux représentés l'histoire des conflits et la géographie du pays. Ce sont des domaines de grande importance dans les théories sur l'origine des conflits, que peu d'indicateurs issus de la Banque mondiale décrivent.

10.1.3.1 Histoire des conflits

Pour mieux rendre compte de l'histoire des conflits nous avons avons intégré dans notre base de données 7 nouvelles variables dont les sources sont précisées à l'annexe D.

Nombre de morts directement liés aux conflits passés Ce nombre est estimé à partir de la base de données UCDP/PRIO mentionnée précédemment. Tous les conflits ayant fait au moins 25 morts ont été comptabilisés. Nous avons également inclus les actes unilatéraux de violence. Plus ce nombre est élevé et plus les conflits passés ont été violents et risquent d'avoir marqué les esprits. L'étendue des griefs et du mécontement de la population devrait donc être importante.

Il est cependant aussi possible de considérer que de violents conflits contribuent à déteriorer la base de recrutement des rebelles. Selon cette interprétation on devrait alors constater une influence stabilisatrice de cette variable sur le déclenchement des conflits et non pas aggravante comme le suggère la première interprétation.

Nombre d'années durant lesquelles le pays a connu un conflit Les sources et les interprétations possibles de cette variable sont les mêmes que pour la variable précédente.

Implication dans un conflit inter-étatique Cette variable a également été construite à partir de la base UCDP/PRIO. Pour certains, elle doit favoriser l'éclatement de la violence à l'intérieur du pays car celui-ci est plus faible, les forces gouvernementales étant accaparées par le conflit inter-étatique.

On retrouve ici l'argumentation de Fearon et Laitin ou encore de Buhaug dont l'analyse de la capacité des rebelles à se soulever relativement à la capacité du gouvernement à contenir une rébellion est tout à fait similaire à l'analyse *coûts-revenus* de Fearon et Laitin. Pour reprendre l'argumentation de Collier et Hoeffler, c'est avant tout la disponibilité d'armes et de soldats formés accompagnant un conflit inter-étatique qui accroît le risque de déclenchement d'un conflit intra-étatique.

Hegre *et al.* (2001) observent que cette variable ne joue pas un rôle significatif et proposent une autre interprétation : les effets aggravants que nous venons d'évoquer sont compensés par le fait qu'un conflit inter-étatique a tendance à raviver l'esprit patriotique et à effacer les divisions internes, du moins temporairement.

Nombre d'États voisins en guerre civile Les arguments de Collier et Hoeffler relatifs à la facilité de recruter et d'équiper une armée de rebelles sont également valables lorsque des États voisins sont en crise. Du fait des liens ethniques transnationaux dans certaines régions, il est également possible d'assister à un phénomène de diffusion des conflits. L'interdépendance des économies à un niveau régional peut également être un facteur important de diffusion. Les guerres civiles s'accompagnent en effet fréquemment d'une dégradation de l'économie nationale. Ces perturbations peuvent alors se répercuter sur les pays voisins, ce qui les affaiblit à leur tour, rendant plus facile l'éclosion d'une insurrection.

Même si les expérimentations de Hegre *et al.* (2001) ne permettent pas de constater l'existence d'un lien significatif entre cette variable et le risque de conflit, elle est l'une des plus pertinentes pour prédire le déclenchement des conflits dans les travaux de la *State Failure Task Force* (Goldstone *et al.*, 2000). Ce sont les travaux qui, rappelons-le, se rapprochent le plus des nôtres tant du point de vue de la méthodologie que des objectifs. Ward et Bakke (2005) ont par ailleurs comparé le pouvoir de prédiction des modèles de Collier et Hoeffler, de Laitin et Fearon et de la *State Failure Task Force*. Ils notent que seul ce dernier a un pouvoir prédictif significatif.

Nombre de personnes déplacées à l'intérieur des frontières Les variables relatives aux migrations sont assez peu utilisées dans les études économétriques, peut-être en partie à cause de problèmes d'endogénéité. Les migrations sont en effet souvent une conséquence des conflits. Mais ceci pose problème lorsque l'on veut identifier de manière fiable des liens de causalité. Notre objectif est l'anticipation et nous avons de bonnes raisons de penser que tenir compte des migrations peut nous aider dans cette tâche. D'une part, les migrations internes exacerbent le ressentiment de la population, d'autre part, les personnes déplacées n'ont que peu de choses à perdre en s'engageant dans une rébellion, ce sont donc des recrues potentielles que des rebelles peuvent facilement convaincre.

Nombre de réfugiés accueillis sur le territoire Les interprétations relatives à cette variable sont sensiblement les mêmes que pour le nombre de personnes déplacées.

Nombre de réfugiés originaires du pays Le nombre de personnes fuyant le territoire est symptomatique de troubles internes de grande intensité. Le chaos qui accompagne les mouvements de population aussi brusques que les flux de réfugiés constitue un environnement favorable à l'éclosion d'une insurrection. La légitimité du gouvernement devient en effet plus que contestable si ce n'était pas le cas avant et les arguments des rebelles ont toutes les chances de convaincre des recrues potentielles. Il est de plus difficile pour le gouvernement de surveiller efficacement le territoire lorsque des milliers de personnes sont sur les routes.

10.1.3.2 Géographie

La seule variable relative à la géographie du pays que nous ayons mentionnée est la proportion de terrains montagneux. Or de nombreux courants de pensée relèvent l'importance de cette dimension, à commencer par les néo-malthusiens. Des indicateurs comme la superficie ou la densité de population sont des indicateurs fournies par la Banque mondiale et font donc partie de notre base de données.

Superficie L'influence de la superficie sur les conflits est potentiellement double comme le révèle l'étude de Buhaug (2006). D'une part, les États disposant d'un territoire étendu sont plus susceptibles de connaître des mouvements de sécession. Il est en effet difficile pour des rebelles, dont l'implantation est locale, de renverser le gouvernement. Mais il est également difficile pour le gouvernement de contrôler efficacement l'ensemble du territoire. À l'inverse les États de faible superficie connaissent plus de conflits visant au renversement du pouvoir. Faire sécession dans un État de petite taille n'a pas beaucoup d'intérêt, la prise directe du pouvoir demandant peu d'efforts supplémentaires par rapport à la sécession.

Rivières Nous avons ajouté le nombre de rivières délimitant une frontière ainsi que le rapport entre la taille des frontières délimitées par des rivières et la taille totale des frontières. Les rivières frontalières peuvent en effet être l'objet de tensions inter-étatiques qui peuvent affaiblir l'État. D'autre part, les mouvements sécessionnistes qui sont responsables d'un nombre conséquent de guerres civiles ont pour objet le contrôle d'une partie du territoire, qui est presque toujours située à sa périphérie (Buhaug, 2006).

Étant donné l'importance de l'eau en tant que ressource naturelle, l'absence de cours d'eau dans ces régions est susceptible de dissuader toute velléité de sécession. L'accès à la mer joue aussi un rôle important dans une économie en favorisant les échanges. Nous supposons donc que des tentatives de sécession seront plus volontiers initiées dans des

régions ayant un accès à la mer. Aussi avons-nous ajouté une variable correspondant au rapport entre la taille des frontières maritimes et la taille totale des frontières.

Répartition de la population Collier et Hoeffler (2004) ont analysé l'influence de la répartition de la population sur le territoire en utilisant l'indicateur de concentration de Gini. L'introduction de cette variable repose sur l'idée selon laquelle une population plus dispersée rend plus difficile le contrôle du territoire. Mais on peut également considérer que les tensions inter-communautaires et le ressentiment de la population sont plus forts lorsque la concentration est plus importante, ne serait-ce qu'à cause d'une plus grande compétition entre les individus pour l'accès aux ressources.

Pour rendre compte plus finement de la dispersion géographique, nous avons choisi de focaliser notre attention sur les zones de faible et forte densités. Pour ce faire nous avons introduit dans notre base de données la proportion de la population habitant dans les zones de faible et forte densité ainsi que la proportion de la superficie occupée par ces zones. Nous avons également inclus la proportion de la population habitant dans des zones de haute altitude, ainsi que la proportion de la superficie des zones de haute altitude. Ces zones sont en effet supposées plus propices à l'éclosion des rébellions du fait de la difficulté pour le gouvernement d'y contenir une insurrection[6].

Nous avons recueilli ces données auprès du *Center for International Earth Science Information Network* (CIESIN) de l'Université de Columbia aux États Unis. L'inconvénient principal de ces données est qu'elles ne concernent que la répartition de la population en 1995. Dans nos expérimentations les variables que nous venons d'introduire sont donc invariantes dans le temps, ce qui ne reflète évidemment pas la réalité des phénomènes démographiques (migrations et accroissement naturel), même s'ils interviennent sur des périodes relativement longues. Aussi n'avons-nous introduit ces variables que pour décrire la géographie des États dans la période de l'après-Guerre froide.

10.1.4 Conclusion

Les théories de Collier et Hoeffler ou celles de Laitin et Fearon ne sont pas si éloignées qu'il pourrait y paraître des théories de Gurr. Tous reconnaissent l'importance des trois facteurs mis en évidence par Gurr : les raisons profondes qui motivent les individus à se rebeller, la cohésion d'un groupe sans laquelle la rébellion n'a aucune chance de voir le jour et enfin les occasions facilitant le passage à l'acte. Les divergences se manifestent sur l'importance qui est accordée à chacun de ces facteurs. Selon l'interprétation qui est retenue pour expliquer le déclenchement des guerres civiles, diverses variables seront employées pour modéliser ce phénomène. Nous n'avons recensé que les principales d'entre elles. Le lecteur intéressé pourra se reporter à Hegre et Sambanis (2006) pour plus de détails. Ils en dressent en effet une liste bien plus exhaustive.

[6] Il eût été plus judicieux de considérer les zones montagneuses et pas simplement de haute altitude, car des plateaux en altitude ne peuvent pas être considérés comme des zones offrant des refuges naturels aux insurgés. Notre choix a été contraint par la disponibilité des données.

10.2 Base de données sur les conflits armés intra-étatiques

Le but de notre application est d'apprendre un modèle de prédiction capable de repérer, à partir de leur contexte structurel, les pays dans lesquels un conflit armé est susceptible de se produire à un horizon de 1 ou 2 ans.

10.2.1 Définition de la classe *crise*

Plus formellement, nous disposons d'une variable cible ou classe, y, telle que :

$$y_i^{t,t+1} = \begin{cases} 1 & \text{si un conflit a eu lieu dans le pays } i \text{ durant les années } t \text{ ou } t+1 \\ 0 & \text{sinon} \end{cases}$$

10.2.1.1 Déclenchement et occurrence d'un conflit

Nous cherchons à prédire $y^{t,t+1}$ à partir d'un ensemble de variables $v_1, ..., v_p$ décrivant le contexte structurel de différents pays sur une période T antérieure à t. Autrement dit nous cherchons un modèle f de y, tel que $y^{t,t+1} = f\left(v_1^{T<t}, ..., v_p^{T<t}\right)$. Cette formalisation du problème correspond exactement à celle que nous avons mise en place lors de nos premières expériences à la section 2.3. Cependant elle ne répond pas pleinement à nos attentes.

Le principal problème vient du fait que la variable y indique l'occurrence d'un conflit et non son déclenchement, ce qui nous préoccupe avant tout. Utiliser ce formalisme peut prêter à confusion car l'étude porte alors aussi bien sur le déclenchement des conflits que sur leur durée, sans qu'il soit possible de distinguer les deux (Schneider et Wiesehomeier, 2006). Une solution consiste à modifier y de telle sorte que l'on ait :

$$y_i^{t,t+1} = \begin{cases} 1 & \text{si } y_i^{t-1} = 0 \text{ et un conflit a eu lieu dans le pays } i \text{ durant les années } t \text{ ou } t+1 \\ 0 & \text{sinon} \end{cases}$$

Si cette solution résout bien notre problème, elle en crée un nouveau. Cette nouvelle définition de y occulte en effet les déclenchements de conflits survenus moins d'un an après la fin d'un autre ou encore ceux qui surviennent alors qu'un autre conflit est engagé. C'est le cas par exemple de l'Angola. Entre 1975, année de son indépendance, et 2002, le pays a été plongé dans une guerre civile opposant trois groupes pour la prise du pouvoir : le Mouvement populaire pour la libération de l'Angola (MPLA), l'Union nationale pour l'indépendance de l'Angola (Unita) et le Front national de libération de l'Angola (FNLA). En 1991 un nouveau conflit a été déclenché par des groupes séparatistes de la province du Cabinda. Il s'agit bien d'un déclenchement de conflit à l'intérieur de la guerre civile angolaise. La solution proposée précédemment ne permet pas de prendre en compte de tels cas.

Pour y parvenir, nous avons utilisé la base de données des conflits armés UCDP/PRIO (version 4-2006) (Gleditsch *et al.*, 2002). Cette base recense pour chaque État de plus de 250000 habitants l'ensemble des conflits, aussi bien inter- qu'intra-étatiques, dans lesquels il a été impliqué depuis 1946. Nous n'avons retenu que les conflits intra-étatiques pour construire notre variable cible y.

10.2.1.2 Définition d'un conflit armé intra-étatique

La notion de conflit armé est centrale pour notre travail. Aussi est-il essentiel de préciser la définition que les auteurs de cette base de données en ont donné.

> Un *conflit armé intra-étatique* est un différend entre deux parties, dont l'une au moins est le gouvernement d'un État, relatif au gouvernement ou au territoire d'un État, et à propos duquel il est fait un usage des armes entraînant au moins 25 morts liés à des combats dans l'année.

Pour que cette définition soit complète, il conviendrait de préciser les notions de *partie*, *gouvernement*, *État*, *usage des armes*, *différend à propos du gouvernement* et *différend à propos du territoire*. Mais la signification exacte de ces termes n'influencera pas notre propos, aussi renvoyons-nous le lecteur aux définitions originales fournies par (Gleditsch *et al.*, 2002).

De la définition précédente, nous retiendrons les points suivants qui montrent l'intérêt et les limites de la base de données UCDP/PRIO.

- Seuls les conflits dans lesquels l'État est partie prenante sont recensés. Ainsi il n'est pas rendu compte des violences inter-ethniques qui peuvent toucher un État sans que ce dernier soit directement impliqué.
- Les États considérés doivent comporter au moins 250000 habitants. Certains États sortent *de facto* du champ de l'analyse.
- Les conflits qui n'ont pas pour objet la revendication d'un territoire ou le renversement du gouvernement sont également exclus. L'usage unilatéral de la violence par le gouvernement pour réprimer une minorité (génocide ou politicide) n'est pas non plus pris en compte.
- Seuls les conflits ayant entraîné 25 morts dans l'année sont comptabilisés. Les conflits de moindre intensité sont donc exclus, ainsi que ceux qui provoquent moins de 49 morts répartis entre la fin d'une année et le début d'une autre (24 morts en décembre et 24 morts en janvier par exemple). Notons cependant que ce seuil de 25 morts, pour arbitraire qu'il soit[7], est nettement moins élevé que le seuil fixé lors de nos premières expériences (1000 morts). Ceci nous permet d'inclure dans notre analyse des conflits de faible intensité.
- Seuls les morts liés à des combats sont comptabilisés. Les chiffres fournis ne tiennent donc pas compte des invalides et des morts, souvent plus nombreux que les victimes directes des combats, causés par les famines et maladies qui accompagnent les conflits. Ce sont pourtant des chiffres dont il conviendrait de disposer pour mener à bien une analyse de l'impact des conflits, partie intégrante de l'évaluation du risque. Ce point ne nous concerne cependant pas directement puisque nous avons préalablement indiqué que nous ne nous occupions que de l'estimation de l'incertitude liée au déclenchement d'un conflit et non de la quantification des conséquences d'un tel conflit.

10.2.1.3 Autres formes de violence intra-étatique

L'inconvénient majeur de cette base de données est qu'elle ne recense pas les conflits dans lesquels le gouvernement n'intervient pas directement, ni les génocides et politicides. Aussi avons-nous décidé d'ajouter à notre liste de conflits ceux au cours desquels il est fait un usage unilatéral de la force envers une minorité. Pour cela, nous avons utilisé la base de données *One-Sided Violence* constituée par l'université d'Uppsala dans le cadre du

[7] Est-il judicieux de rejeter les conflits ayant fait 20 morts dans un pays de 250000 habitants et de comptabiliser ceux pour lesquels 25 morts ont été dénombrés dans un pays d'un milliard d'habitants ? Voir (Sambanis, 2004) pour une analyse critique de ce seuil.

projet Uppsala Conflict Data Program (UCDP)[8]. Cette base de données couvre les usages unilatéraux de la violence dans les États de plus de 250000 habitants durant la période 1989-2005. Ici encore, il est nécessaire de s'arrêter sur la terminologie, pour que la nature des phénomènes considérés par les auteurs de la base de données soit clairement établie.

> Un *usage unilatéral de la violence* est une action armée engagée par le gouvernement d'un État ou par un groupe formellement organisé à l'encontre de civils, entraînant au moins 25 morts. Les meurtres commis dans le milieu carcéral ne sont pas comptabilisés.

Nous n'avons pas intégré les conflits intra-étatiques dans lesquels le gouvernement n'intervient pas car la seule base que nous ayons trouvée à ce sujet (UCDP *Non-State Conflict*) ne couvre que la période 2002-2005, ce qui ne correspond pas à la période que nous souhaitons traiter. Nous reviendrons sur le choix de la période étudiée un peu plus loin.

Notre approche se distingue de celles de la littérature dans la mesure où nous considérons des conflits de natures distinctes. Cette différence tient au fait que nous poursuivons des objectifs distincts de ceux de la majorité des études sur les guerres civiles. Celles-ci cherchent à identifier les mécanismes causaux qui expliquent le déclenchement des guerres civiles, alors que nous cherchons avant tout à anticiper l'éruption de la violence au sein d'un État, quelle que soit la forme qu'elle puisse prendre. Dans la mesure du possible nous voulons également identifier les facteurs de risque afin d'agir au plus tôt avant le déclenchement d'une crise. L'avantage de notre approche réside donc dans la plus grande couverture des phénomènes considérés. Le revers de la médaille est qu'en assimilant toutes les formes de violence à un même phénomène, il est possible que les explications que nous serons en mesure d'apporter sur l'origine de la violence soient empreintes de confusion.

Ces craintes se fondent en partie sur les résultats de l'étude de Buhaug (2006). En analysant l'influence de certaines variables sur le déclenchement de conflits intra-étatiques, en fonction de leur type, Buhaug a observé l'existence d'un lien entre l'hétérogénéité de la population et les mouvements sécessionnistes. En revanche il a constaté l'absence d'un tel lien avec les conflits ayant pour objet le renversement du gouvernement central. Il observe de même que l'influence des institutions sur le déclenchement d'un conflit dépend de la nature du conflit considéré. Si les démocraties sont peu susceptibles de connaître des révoltes visant le renversement du gouvernement, elles constituent cependant un environnement favorable à l'émergence de mouvements sécessionnistes. Cette étude suggère donc qu'il est important de désagréger la notion de conflit.

Il nous semble délicat, voire vain, de catégoriser les conflits, étant donné que bien souvent toutes les formes de violence s'entremêlent, comme l'illustre parfaitement le second conflit opposant les États-Unis et le Royaume-Uni à l'Irak. Initialement il s'agit d'un conflit inter-étatique. Mais peu après l'invasion anglo-américaine, il fut bien difficile de faire la part des événements propres à la guerre civile (opposition armée entre le gouvernement et un groupe formellement organisé), des actes de terrorisme ou encore des affrontements intercommunautaires. Les revendications sécessionnistes sont de plus mêlées à une tentative de renversement du gouvernement. C'est la raison pour laquelle nous n'avons pas tenu compte des recommandations de l'étude de Buhaug. Nous verrons lors de l'analyse des résultats expérimentaux si cette décision aura porté préjudice à notre analyse.

[8] C'est également dans le cadre de ce programme, en collaboration avec l'institut international de recherche sur la paix d'Oslo, que la base UCDP/PRIO sus-mentionnée a vu le jour. http://www.pcr.uu.se/research/UCDP/our_data1.htm

10.2.1.4 Étiquetage des données

Nous disposons d'une liste annuelle de conflits pour un ensemble assez large de pays (voir annexe D). Pour que l'interprétation de la variable y ne prête pas à confusion, il nous faut tout de même préciser la façon dont nous avons construit y à partir de cette liste. Nous avons utilisé les quatre règles suivantes :

1. Si un conflit armé est initié dans le pays i durant les années t ou $t+1$, alors $y_i^{t,t+1} = 1$.
2. S'il est fait un usage unilatéral de la violence durant les années t ou $t+1$ alors que le pays i n'est pas en conflit en $t-1$, alors $y_i^{t,t+1} = 1$.
3. Si durant cette période le pays est affecté par un conflit qui a débuté avant l'année t et qu'aucun nouveau conflit ne s'est déclenché ni en t, ni en $t+1$, alors ce pays est supprimé de notre base de données (du moins, l'observation correspondante aux années $t, t+1$).
4. Si aucun conflit n'est en cours ou initié durant les années $t, t+1$ et qu'il n'est pas fait usage unilatéral de la violence durant cette période, alors $y_i^{t,t+1} = 0$.

La règle 3 est controversée dans la littérature du domaine. Nous avons choisi l'approche de Collier et Hoeffler (2004) afin d'éviter de mêler l'analyse de la durée des conflits avec celle du déclenchement des conflits. Mais d'autres auteurs influents du domaine préfèrent conserver les observations correspondant à des pays dans lesquels un conflit est en cours pour ne pas perdre d'information. Ils considèrent alors que $y_i^{t,t+1} = 0$ pour ces pays (Fearon et Laitin, 2003).

10.2.2 Construction des observations

Nous venons de voir le changement assez conséquent de la classe y que nous avons opéré par rapport à nos premières expérimentations. Cependant ce changement ne suffit pas à lui seul à expliquer pourquoi nous n'avons pas conservé les données recueillies initialement, en modifiant simplement la classe y de chacune de nos observations. La raison principale réside dans notre souhait d'étendre la base de données initiale afin de tester un grand nombre d'hypothèses ayant trait aux conflits. Cette extension de la base concerne aussi bien le nombre d'observations que le nombre de variables explicatives.

10.2.2.1 Apprentissage de modèles spécifiques

Le nombre de pays dans le monde étant limité, augmenter le nombre d'observations n'est possible que si l'on considère une période plus longue de l'histoire[9]. Nous avons ainsi réuni des données de la période 1970-2002.

Il eût été envisageable d'inclure également les conflits ayant eu lieu avant 1970, comme le font de nombreuses études empiriques du domaine (Fearon et Laitin, 2003; Collier et Hoeffler, 2004). Mais les conflits ayant eu lieu avant 1970 sont essentiellement des guerres de décolonisation, qui ont des caractéristiques spécifiques. Il est en outre assez difficile de choisir l'État auquel les rattacher, l'empire colonial ou le futur État indépendant (Fearon et Laitin, 2003). Il semble difficilement acceptable de considérer qu'un État qui n'existe pas encore est en guerre civile.

Mais si l'on désigne la puissance coloniale comme l'État subissant la guerre civile, il convient alors de réajuster l'ensemble des variables explicatives pour faire en sorte qu'elles réfèrent à l'ensemble de l'empire colonial et non simplement à la métropole, ce qui pose

[9] Notre base de données initiale ne contenait que des observations de la période 1999-2000.

des problèmes non négligeables (Sambanis, 2004). Par ailleurs, avant 1970, de nombreuses valeurs font défaut au sein des variables explicatives utilisées pour modéliser les conflits.

Horizon d'alerte Nous procédons comme précédemment à l'évaluation du risque à un horizon de deux ans. Nous avons ainsi 16 observations potentielles pour chaque pays, pour les intervalles 1971-1972, 1973-1974 jusqu'à 2001-2002. Cette procédure est similaire à celle qui a été mise en place par Collier et Hoeffler (2004); Collier *et al.* (2006).

Comme eux, nous ne considérons pas des intervalles qui se chevauchent, de façon à ne pas compter plusieurs fois un même déclenchement de conflit. En revanche, nous n'essayons d'anticiper les crises qu'à un horizon de deux ans et non cinq comme dans les travaux que nous venons de mentionner. Le principal avantage réside dans le fait que nous disposons d'un plus grand nombre d'observations, ce qui permet d'induire des modèles plus robustes. De plus le choix de l'intervalle initial correspondant à la première observation de chaque pays est moins problématique. Nous n'avons en effet que deux, et non cinq, configurations possibles suivant que l'on commence par 1971-1972 ou 1972-1973. Ceci réduit donc la variabilité due à l'échantillonnage. Idéalement il faudrait tester chacune des configurations. C'est là un des reproches adressés par Marchal et Messiant (2002) à Collier.

De nombreux autres travaux considèrent que chaque année donne lieu à une nouvelle observation (Fearon, 2005; Fearon et Laitin, 2003; Lichbach *et al.*, 2004), ce qui règle complètement le problème précédent. Si cette approche convient bien à l'application de modèles économétriques visant à expliquer le déclenchement des crises, elle nous semble moins adaptée à la prédiction de ces déclenchements à partir de variables macro-structurelles. Ces variables évoluent pour la plupart lentement et ne permettent pas, selon nous, de prédire finement ces déclenchements. Aussi faire des prédictions sur un intervalle de plusieurs années plutôt que sur une seule année nous semble-t-il être un objectif plus raisonnable. Le choix d'un intervalle de deux ans est quelque peu arbitraire, mais il correspond à un compromis entre les deux approches que nous venons de présenter.

Périodes d'analyse Outre le simple fait d'accroître le nombre d'observations de notre base de données, considérer les différents pays sur la période 1970-2002 plutôt que 1999-2000 nous permet de construire des modèles différents selon l'époque considérée.

Une époque ou *période d'analyse* correspond à un intervalle de temps durant lequel toutes les observations dont nous disposons pour les différents États sont regroupées pour former une base de données. Nous avons considéré par exemple l'époque de l'après-Guerre froide. L'intérêt est que l'on ne présuppose pas que les contextes propices au déclenchement des conflits sont les mêmes en 1970 et 2002. Nous les supposons invariants dans le temps seulement durant l'époque considérée.

Le contexte géopolitique international a connu de profonds changements depuis la fin de la Seconde Guerre mondiale, si bien qu'il est loin d'être invraisemblable de penser qu'entre 1970 et 2002 la nature et les racines des différents conflits aient évolué. Il serait vain d'essayer de synthétiser en quelques lignes l'évolution des relations internationales et de la géopolitique mondiale depuis la fin de la Seconde Guerre mondiale. Tout au plus parviendrions-nous à reprendre quelques poncifs journalistiques sur la décolonisation, la Guerre froide ou encore la mondialisation et cela nous écarterait quelque peu de notre sujet.

Précisons simplement que pour beaucoup de chercheurs en sciences politiques la fin de la Guerre froide marque un tournant géopolitique essentiel et cela vaut en particulier pour les conflits intra-étatiques. Diverses études ont ainsi cherché à comprendre dans quelle mesure les mécanismes sous-tendant l'émergence de conflits ont été modifiés par l'effondrement de l'Union soviétique.

L'influence de cette rupture sur l'émergence de nouveaux conflits est très controversé. Certains jugent que la fin de la Guerre froide et de l'équilibre entre les deux blocs s'est accompagnée de la résurgence des nationalismes ethniques autrefois étouffés ou du moins contrôlés par les deux blocs (Huntington, 1993; Kaplan, 1994). Leurs idées ont été largement relayées par les médias. Mais des études empiriques plus récentes ne constatent aucun effet particulier lié à la fin de la Guerre froide (Hegre et Sambanis, 2006; Collier *et al.*, 2006), l'effet précédent étant compensé par la diminution de l'offre d'armements par les deux blocs et par la baisse de l'enjeu symbolique et stratégique attaché à chaque conflit, fût-il périphérique. Une recrudescence des guerres civiles a bien été observée au début des années 90, mais la tendance s'est vite inversée, si bien que les conflits armés intra-étatiques sont dans l'ensemble moins nombreux et moins violents depuis la fin de la Guerre froide (Gleditsch *et al.*, 2002; Fearon et Laitin, 2003).

Ces jugements contradictoires peuvent s'expliquer par un changement de la nature des conflits. Par exemple les actes de terrorisme, dont l'importance a crû depuis la fin de la Guerre froide[10], ne sont pas comptabilisés par les auteurs des études empiriques que nous venons de mentionner.

En considérant l'ensemble des guerres civiles entre 1970 et 2002, nous sommes à même de voir dans quelle mesure nos données corroborent les différentes thèses au sujet de l'influence du contexte géopolitique international sur l'émergence des conflits intra-étatiques. Pour ce faire nous avons découpé la période 1970-2002 en différentes sous-périodes que nous avons précédemment nommées époques ou *périodes d'analyse*. Les différences entre les modèles construits sur chacune de ces sous-périodes nous renseigneront sur une éventuelle évolution des facteurs de crise.

Groupes de pays L'idée d'une spécialisation des modèles par époque historique est séduisante car elle permet une meilleure prise en compte du contexte dans lequel évoluent les États que l'on étudie. Dans la même optique, nous avons construit une base de données par groupe de pays afin de spécialiser nos modèles aux particularités régionales. Il paraît en effet contre-intuitif de considérer que les facteurs de crise sont les mêmes dans toutes les régions du monde. La *State Failure Task Force*, si elle défend l'intérêt d'un modèle global tel que celui que nous avons construit lors de nos premières expérimentations, a tout de même développé un modèle spécifique pour les pays musulmans (Goldstone *et al.*, 2000).

Nous avons, pour notre part, choisi un découpage régional en sept groupes. Nous nous sommes inspiré du découpage effectué par la Banque mondiale à quelques exceptions près. Au lieu de considérer les groupes Amérique du Nord d'un côté et Europe et Asie centrale de l'autre, nous avons préféré construire le groupe des pays occidentaux et celui des pays d'Europe de l'Est et de l'Asie Centrale. Ces derniers correspondent aux anciennes républiques et États satellites de l'Union Soviétique. Sur la période 1970-2002, cela nous paraissait plus cohérent avec l'histoire commune de ces pays. Notons que du fait de l'intégration de la plupart des pays d'Europe de l'Est dans l'Union européenne, ce découpage mériterait d'être reconsidéré. La liste exacte des États composant les groupes suivants est donnée à l'annexe D.

- Afrique du Nord et Proche-Orient
- Afrique subsaharienne
- Amérique latine et Caraïbes
- Asie du Sud
- Asie du Sud-Est et Pacifique
- Europe de l'Est et Asie centrale

[10] À moins que ce ne soit simplement la couverture médiatique du terrorisme qui ait crû.

– pays occidentaux : outre l'Europe occidentale et l'Amérique du Nord, nous avons fait le choix, potentiellement sujet à discussion, d'inclure dans ce groupe l'Australie, la Nouvelle-Zélande ainsi que l'Afrique du Sud.

Afin de juger de l'intérêt de la construction de modèles spécifiques régionaux, nous avons également considéré un groupe, qualifié de *global* par la suite, contenant l'ensemble des États étudiés.

10.2.2.2 Variables descriptives

Lors de nos premières expérimentations, nous avons utilisé près de 150 variables explicatives quasiment toutes issues des indicateurs de développement de la Banque mondiale. Pour chacune d'elles nous avons construit deux attributs. L'un correspond à la valeur moyenne de cette variable, estimée sur une période d'une dizaine d'années, antérieure à 1999. L'autre correspond à la variation annuelle moyenne de cette variable estimée sur la même période.

L'idée était de prendre en compte des tendances statiques mais également dynamiques, reflétant l'évolution des indicateurs avant la période sur laquelle les prédictions sont effectuées. L'introduction de la dynamique des indicateurs est un moyen de prendre en compte le temps. La base de données ainsi constituée comporte trois principales faiblesses que nous nous sommes efforcé de pallier :

– La non-distinction entre la tendance d'évolution et la variabilité dans l'évolution d'un indicateur.
– L'hétérogénéité de la *période d'estimation* des moyennes et variations annuelles.
– La non-couverture de certains domaines pourtant jugés importants dans la littérature.

Prise en compte de la dynamique des indicateurs La variation annuelle moyenne d'une variable regroupe des informations portant aussi bien sur la tendance d'évolution de cette variable que sur la variabilité de cette évolution. Or ces informations sont de natures différentes. La première nous renseigne sur la croissance, décroissance ou stabilité de l'indicateur au cours du temps, tandis que la seconde porte sur les écarts que l'on peut observer sur une période donnée entre l'évolution de l'indicateur et la tendance générale.

À l'instar de ce qui a été fait par le CIFP (Ampleford *et al.*, 2001; Carment, 2001) il nous semble préférable de désagréger ces informations. Aussi avons-nous introduit pour chaque variable explicative trois et non plus deux attributs : la moyenne, la tendance et la variabilité, estimées sur une période donnée. Pour ce faire nous avons construit la droite de régression par moindres carrés de chacun des indicateurs, sur l'intervalle de temps correspondant à la période d'estimation des indicateurs.

Le score de tendance pour chaque variable et chaque période correspond à la pente de la droite de régression correspondante, tandis que le score de variabilité correspond à l'écart-type des résidus de la régression. Précisons que lorsqu'au moins la moitié des valeurs annuelles d'un indicateur sont manquantes, pour une période d'estimation donnée, la droite de régression n'est pas construite et la tendance et la variabilité sont étiquetées comme manquantes.

Périodes d'estimation Pour illustrer les différences entre les trois notions de période que nous considérons pour construire les observations de nos bases de données, nous avons représenté à la figure 10.1 les observations que l'on peut extraire à partir des données d'un même pays collectées sur la période 1970-2002. La période d'estimation considérée

s'étale sur 7 années, l'horizon de prédiction est de 2 ans et nous avons établi deux périodes d'analyse correspondant aux époques antérieure et postérieure à la fin de la Guerre froide. Avec de telles spécifications il est donc possible de dégager 13 observations par pays, les 7 premières et les 6 dernières étant respectivement regroupées dans deux bases de données, une par période d'analyse. Nous estimons donc pour chaque attribut, non pas une droite de régression, mais 13, une par observation.

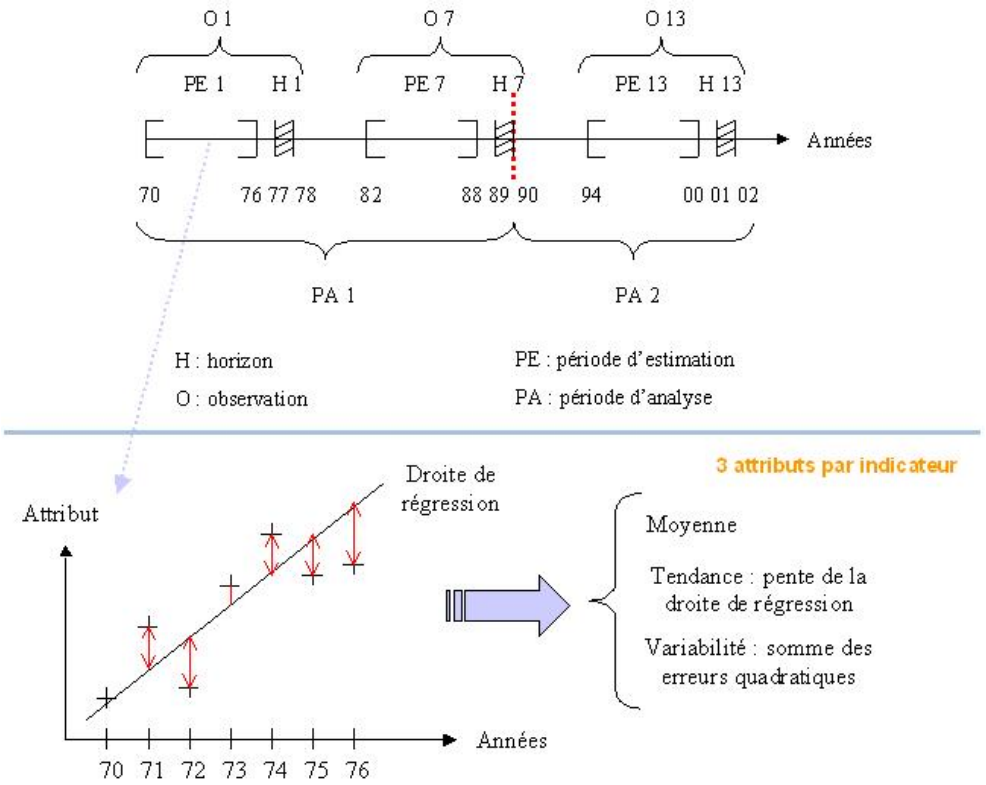

FIG. 10.1 – Différentes périodes considérées lors de la construction des observations pour un pays donné

Hormis le choix de la variation annuelle moyenne pour rendre compte de la dynamique des indicateurs, celui de la période d'estimation des moyennes et variations annuelles moyennes ne nous satisfait pas non plus. Pour construire notre base de données initiale nous avons en effet utilisé toutes les données disponibles jusqu'à 1998. Du fait de la proportion importante de valeurs manquantes, proportion qui diffère selon les indicateurs, les périodes d'estimation sont elles-mêmes différentes selon les indicateurs.

Nous avons souhaité homogénéiser ces périodes de façon à pouvoir analyser l'influence de l'histoire sur la qualité de nos modèles. Nous avons ainsi construit quatre base de données pour chaque groupe de pays et chaque période, chacune différant par le nombre d'années à partir desquelles moyenne, tendance et variabilité de chaque indicateur sont estimées. Nous avons ainsi considéré des périodes de 29, 15, 7 et 1 années.

À titre de comparaison, la plupart des études économétriques de la littérature relative aux conflits armés, qui sont celles qui se rapprochent le plus de nos travaux, utilisent uniquement la valeur des variables explicatives l'année précédant celle durant laquelle les conflits sont considérés. Ceci correspond à notre période d'1 an. Notons que sur une telle période, il est évidemment dénué de sens de chercher à estimer la tendance et la variabilité d'un indicateur. Ainsi nous aurons près de trois fois moins d'attributs pour les bases de données correspondant à cette période.

En faisant ainsi varier la fenêtre temporelle dans laquelle les variables seront estimées, notre objectif est de savoir jusqu'où il est utile de remonter dans le temps pour pouvoir réaliser des prédictions fiables.

10.2.2.3 Bases de données construites

En fonction des périodes d'estimation, le nombre d'observations pour chaque État diffère. En conséquence différentes périodes d'analyse ont été construites pour que le nombre d'observations de la base de données correspondante soit suffisamment important. Nous avons trois périodes d'analyse pour les périodes d'estimation de 1 et 7 ans, et respectivement deux et une périodes d'analyse pour les périodes d'estimation de 15 et 29 ans. Nous avons ainsi construit pour chacune des quatre périodes d'estimation, une base de données par groupe de pays et par période d'analyse. Certaines des bases de données résultantes contiennent un faible nombre d'observations et parfois certaines ne contiennent que trop peu de pays en crise pour qu'il soit intéressant d'essayer d'apprendre un modèle de classification. Nous les avons donc supprimées. Aussi n'apparaissent-elles pas à l'annexe C donnant les caractéristiques des différentes bases de données utilisées dans notre étude empirique. Ainsi au lieu des 72 bases potentielles, seules 53 ont été conservées.

Nous avons détaillé les décisions que nous avons prises à propos des observations de notre nouvelle base de données afin d'élargir le spectre de notre recherche. Pour compléter notre présentation il nous faut maintenant préciser quels indicateurs nous avons considérés.

10.2.3 Indicateurs macro-structurels

Suite à nos premières expérimentations nous avons mentionné la nécessité d'élargir le nombre de variables explicatives afin d'être à même de couvrir un plus grand nombre de domaines. L'intérêt est de pouvoir envisager des facteurs de risque de natures différentes. Pour ce faire nous avons employé la base des indicateurs de développement de la Banque mondiale (CD-ROM), considérant 432 indicateurs pour 208 pays[11] sous la forme de séries temporelles avec des estimations annuelles de 1960 à 2002. Nous n'avons utilisé que les données entre 1970 et 2002 ainsi que nous l'avons signalé précédemment.

10.2.3.1 Indicateurs retenus

Augmenter le nombre de variables à partir desquelles un modèle de prédiction sera construit n'est pas une fin en soi. Ayant choisi de faire le moins d'hypothèses théoriques possibles quant au déclenchement des conflits intra-étatiques, nous sommes enclin à utiliser autant d'informations que possible afin de ne négliger, autant que faire se peut, aucune théorie explicative potentielle.

Les indicateurs de la Banque mondiale permettent d'aborder un grand nombre de thématiques telles que la démographie, l'économie, la finance, l'environnement, l'énergie, les liens transnationaux... Certaines variables jugées importantes par les polémologues ont été également considérées. Nous avons justifié nos choix à la section 10.1. Nous ne faisons ici que rappeler les variables que nous avons incluses dans nos bases de données et qui sont issues de sources autres que la Banque mondiale .

[11] La base que nous avons construite ne contient des informations que pour 198 pays. Le différentiel s'explique par le fait que la Banque mondiale inclut dans sa liste des territoires qui ne sont pas des États indépendants, comme la Polynésie française ou les îles Vierges.

Hétérogénéité de la population

- Nombre de groupes ethno-linguistiques
- Rapport entre la taille du groupe ehtno-linguistique majoritaire et celle du deuxième groupe le plus important
- Indice de fractionnement ethno-linguistique
- Nombre de groupes religieux
- Rapport entre la taille du groupe religieux majoritaire et celle du deuxième groupe le plus important
- Indice de fractionnement religieux

Histoire des conflits

- Nombre de morts directement liés aux conflits passés
- Implication dans un conflit inter-étatique
- Nombre d'années en guerre civile
- Nombre d'années écoulées depuis le dernier conflit
- Nombre d'États voisins en guerre civile
- Nombre de réfugiés accueillis sur le territoire
- Nombre de réfugiés originaires du pays
- Nombre de personnes déplacées à l'intérieur des frontières

Histoire du pays

- Nombre d'années écoulées depuis l'indépendance
- Nombre d'années durant lesquelles l'autonomie d'un territoire a été supprimée

Géographie

- Proportion de la population habitant dans des zones de haute altitude
- Proportion de la population habitant dans les zones de faible densité
- Proportion de la population habitant dans les zones de forte densité
- Proportion de la superficie occupée par les zones de haute altitude
- Proportion de la superficie occupée par les zones de faible densité
- Proportion de la superficie occupée par les zones de forte densité
- Différence entre l'altitude la plus élevée et la l'altitude la plus basse
- Intensité maximale des tremblements de terre
- Nombre de rivières délimitant une frontière
- Rapport entre la taille des frontières délimitées par des rivières et la taille totale des frontières
- Rapport entre la taille des frontières maritimes et la taille totale des frontières

10.2.3.2 Indicateurs importants faisant défaut

Si les théories les plus importantes[12] sur l'origine des conflits ont été abordées, certaines variables manquent encore pour rendre compte d'un certain nombre d'éléments théoriques de la littérature. Nous font principalement défaut des indicateurs décrivant les institutions d'un État, l'étendue des libertés individuelles, la criminalité et l'escalade des conflits mineurs vers la guerre civile.

Nous avons refusé d'inclure dans notre base de données le niveau de démocratie que nous jugeons biaisé et trop subjectif. Des variables indiquant la tenue d'élections, ou l'existence d'une législature mériteraient d'être ajoutées en vue de futures expérimentations.

[12] L'importance est ici relative à l'impact que ces théories ont eu sur le développement des recherches sur les conflits.

Gandhi et Vreeland (2004) ont d'ailleurs observé que l'existence d'une législature, qu'elle soit indépendante et ait un réel pouvoir ou non, était un facteur contribuant à diminuer le risque de conflit.

L'étendue des libertés individuelles peut influer sur les motivations d'éventuels rebelles. Plus ces libertés seront restreintes et plus le mécontentement de la population risque d'être fort. Il eût donc été intéressant d'employer des indicateurs permettant de rendre compte de ces restrictions. Les données que l'on pourrait recueillir auprès d'organismes tels que *Freedom House*, ou *via* le projet *Minorities At Risk* dont Gurr est l'un des initiateurs, ne nous satisfont pas. Selon nous le problème réside dans le fait qu'il s'agit d'agrégats et non de mesures brutes, les règles d'agrégation étant largement discutables et variables au cours du temps. Nous préférerions employer des indicateurs recensant le nombre de journaux censurés, le nombre de prisonniers politiques...

Il serait également bon d'inclure diverses mesures du niveau de criminalité dans la liste des variables potentiellement pertinentes : le nombre d'incarcérations, le nombre de délits, de crimes. De tels indicateurs peuvent être considérés comme révélateurs du ressentiment de la population ou comme facteurs d'exacerbation de ce ressentiment. Liés à la criminalité, le nombre d'armes légères et lourdes en circulation ainsi que leur prix de vente[13], pourraient s'avérer être des facteurs pertinents, qui influent sur l'approvisionnement d'armées rebelles.

L'absence des variables que nous venons de citer n'est pas aussi dommageable qu'il pourrait y paraître. Les théories explicatives des conflits qu'elles permettraient d'appuyer sont en effet déjà bien représentées par les variables que nous avons précédemment incluses.

Mais il en est une autre, que Lichbach *et al.* (2004) promeuvent, pour la défense de laquelle nous ne disposons que de peu d'informations. Cette théorie s'appuie sur une interprétation clausewitzienne des conflits intra-étatiques. La guerre civile n'est considérée que comme la poursuite par d'autres moyens des interactions conventionnelles entre société et gouvernement. La société exprime ses désaccords et joue son rôle de contre-pouvoir par des moyens plus ou moins violents : critiques verbales du gouvernement, manifestations, insurrections... Le gouvernement peut adapter sa politique pour tenir compte des revendications de la population. Mais il peut également réprimer plus ou moins violemment l'opposition. La guerre civile serait donc un phénomène inhérent à la conduite de la politique intérieure.

Selon cette théorie, pour anticiper les guerres civiles il convient de surveiller en premier lieu l'escalade des conflits (pas forcément armés) de moindre intensité entre le gouvernement et la société. Si l'analyse événementielle de la situation d'un pays, telle que celle qui a été mise en place par Mouillet (2005), nous semble mieux adaptée pour observer l'escalade des conflits, il serait tout de même envisageable d'inclure dans un modèle d'évaluation des risques structurels des indicateurs reflétant l'intensité de la confrontation entre la société et le gouvernement. Lichbach *et al.* proposent d'utiliser à cet effet le nombre d'émeutes, de manifestations ainsi que les violations des droits et libertés.

10.2.4 Conclusion

Nous avons présenté l'ensemble des variables que nous avons adjointes aux indicateurs de développement de la Banque mondiale, ainsi que les raisons qui nous ont poussé à les inclure dans notre base de données. Nous avons parfois été confronté au problème de la multiplicité des interprétations des variables évoqué par Lichbach *et al.* (2004). Nous allons maintenant présenter les modèles qui ont pu être appris à partir des données que nous venons de présenter et voir dans quelle mesure certaines des hypothèses théoriques peuvent

[13] Suite à des discussions avec des chercheurs en sciences politiques de l'Institut Français des Relations Internationales (IFRI), nous souhaiterions disposer d'un indice Kalachnikov, calqué sur l'indice Big-Mac introduit par *The Economist*.

être réfutées ou non. Précisons que notre objectif n'est pas d'apporter une réponse ferme permettant de trancher entre diverses interprétations. Nous ne cherchons pas à expliquer les conflits, mais à les anticiper du mieux possible en donnant les moyens à un utilisateur, expert du domaine et du pays concerné, de construire sa propre interprétation au vu des faits. C'est pour cette raison que nous avons fait en sorte de n'exclure *a priori* aucune explication théorique possible.

10.3 Résultats expérimentaux

Afin d'illustrer l'intérêt du système générique d'aide à l'anticipation des crises que nous avons présenté à la section 9, nous allons le tester sur l'application qui a motivé ces travaux : la détection des conflits armés intra-étatiques. Rappelons que notre système n'est pas un modèle de détection mais une plate-forme d'évaluation et de sélection de modèles. Elle doit permettre d'évaluer et de comparer différents modèles afin de retenir celui qui est le mieux adapté au problème traité. Les résultats de nos expérimentations seront analysés d'un point de vue quantitatif et qualitatif afin d'évaluer les apports de nos travaux dans le domaine de la prévision des conflits armés intra-étatiques. Mais avant cela nous détaillerons les enseignements d'ordre méthodologique que l'on peut tirer de ces expérimentations en insistant sur l'importance de la méthode de sélection de modèles.

10.3.1 Protocole expérimental

Les principales caractéristiques des 53 bases de données, introduites à la section 10.2 et décrites à l'annexe C sont les suivantes :

- grande dimension (le nombre d'attributs varie entre 200 et 900)
- attributs continus ou discrets et ordonnés
- classes déséquilibrées (entre 4 et 25% d'observations de la classe *crise*)
- données manquantes réparties sur la quasi-totalité des attributs

Suite à notre travail sur les données manquantes et le filtrage d'attributs, durant la partie II, nous avons identifié un certain nombre de méthodes de prétraitement qui semblent bien adaptées à ce type de problème. Nous avons ainsi considéré 10 méthodes de substitution des valeurs manquantes et 4 filtres. Nous avons ainsi 40 associations possibles pour constituer une chaîne de prétraitement.

Nous avons mis en évidence à la section 7.5 trois stratégies de combinaison, nommées (A), (B) et (C), selon l'ordre dans lequel ces méthodes sont appliquées. Rappelons que la stratégie (A) consiste à appliquer en premier lieu la substitution avant de sélectionner les attributs tandis que le filtrage, sans prise en compte des valeurs manquantes, est réalisé avant la substitution dans la stratégie (B).

Enfin la stratégie (C) ne diffère de la stratégie (B) que par la prise en compte des valeurs manquantes lors du filtrage. Nous disposons donc de 120 méthodes de prétraitement. Lors de nos expérimentations nous en avons testé 129 après avoir ajouté 9 chaînes de prétraitement dans lesquelles il n'est procédé qu'à la substitution des valeurs manquantes. Dans la suite chacune de ces méthodes sera identifié par la stratégie employée suivie du nom de la méthode de substitution, lui-même suivi du sigle désignant la technique de filtrage.

Les méthodes de substitution des valeurs manquantes envisagées sont les suivantes. Nous reprenons les dénominations utilisées à la section 6.4.

1. mesure de tendance centrale : *Moyenne*, *CMoyenne*, *CMoyenneA*, *Médiane*.
2. aléatoire : *AléatoireMM*
3. plus proche voisin : *1ppv*, *5ppv*
4. régression linéaire locale itérée : *1LLSI*, *5LLSI*
5. entropie : *EF-Entropie*

Les quatre techniques de filtrage que nous avons considérées sont les suivantes :

1. CFS
2. FCBF
3. KSCBF
4. KSF

Lorsqu'aucun filtre n'est appliqué en sus de la substitution des valeurs manquantes, nous l'indiquerons par le terme *SansFiltre*. Comme indiqué précédemment, seules 9 des 10 méthodes de substitution ont été testées dans ce contexte. Ayant observé lors des expérimentations de la section 6.6.4 que les résultats obtenus avec les techniques *Moyenne* et *Médiane* sont très proches lorsqu'aucun filtrage n'est réalisé, nous avons choisi de ne considérer que l'une d'elles : la *Moyenne*.

Les 129 chaînes d'apprentissage ont été évaluées sur chacune des 53 bases de données à notre disposition selon le principe de la validation croisée stratifiée. Pour chacune des bases de données, nous avons tout d'abord procédé à leur segmentation en 10 sous-ensembles de même cardinalité respectant la distribution initiale des classes. 10 modèles ont ensuite été construits en prenant tour à tour chacun des 10 sous-ensembles comme base de test, l'union des 9 autres sous-ensembles faisant office de base d'apprentissage. Pour construire chacun des modèles nous avons tout d'abord appliqué les chaînes de prétraitement sur la base d'apprentissage courante. Les attributs sélectionnés durant cette étape ont été utilisés pour filtrer la base de test qui a ensuite été complétée par substitution des valeurs manquantes, en prenant soin de n'utiliser que la base d'apprentissage pour chacune des observations de test incomplètes. Une fois chacune des deux bases complétées et filtrées, un modèle de détection des crises a été appris par *Salammbô*[14] à partir de la base d'apprentissage, puis nous avons évalué ce modèle sur la base de test. Nous avons enfin pu obtenir les performances globales de chacune des méthodes en prenant la moyenne des performances obtenues sur chacune des 10 bases de test.

Le nombre d'exemples de certaines bases de données étant très faible, nous avons opté pour la version *leave one out* de la validation croisée. Sont alors construits autant de sous-ensembles que la base compte d'exemples. Un modèle différent est construit pour classer chacun des exemples. Nous avons appliqué cette procédure à toutes les bases de données contenant moins de 100 exemples, ainsi qu'à celles contenant moins de 10 exemples de la classe minoritaire.

Avant de présenter les résultats de nos expérimentations il est essentiel d'introduire les mesures de performance que nous avons considérées. S'il nous est possible de synthétiser ces résultats au moyen de matrices de confusion à l'image de ce que nous avons fait avec notre premier modèle à la section 2.3, il nous est cependant indispensable de recourir à des mesures de performance pour comparer les différentes méthodes et pour choisir le modèle qui est le mieux adapté à une base de données particulière.

10.3.2 Mesures de performances

Pour analyser les résultats des expérimentations de la partie II, nous avons utilisé le taux de bonnes classifications, la moyenne des taux de rappel de chacune des classes et l'aire sous la courbe ROC. Aucune de ces mesures n'est complètement satisfaisante pour notre problème.

[14] Le nombre d'exemples minimum que doit contenir un nœud pour pouvoir être partitionné a été fixé de manière empirique à $L = 5$

10.3.2.1 Critique des mesures de performance utilisées précédemment

Le taux de bonnes classifications n'est adapté que lorsque les observations sont équiréparties dans les différentes classes et lorsque les coûts associés aux erreurs de prédiction de chacune des classes sont identiques. Or il n'en est rien dans notre problème. Rappelons simplement que la classe *crise* regroupe entre 4 et 25% des observations selon les bases de données.

La moyenne des taux de rappel permet de compenser le problème de la répartition inégale des observations dans les différentes classes. Cependant l'utiliser revient à considérer que les erreurs de prédiction ont même coût quelle que soit la classe concernée. Dans notre contexte cette position est cependant difficilement tenable. Selon nous il importe surtout de ne pas passer à côté de crises potentielles. Le rappel de la classe *crise* est donc plus important que celui de la classe *non-crise*. Une moyenne pondérée des rappels de chacune des classes serait préférable et permettrait d'introduire un biais pénalisant les erreurs de prédiction des observations de la classe *crise*.

L'aire sous la courbe ROC permet de prendre en compte des différences dans la distribution des classes ainsi que dans la distribution des coûts d'erreur. Construire une courbe ROC nécessite cependant le calcul des probabilités *a posteriori* de chacune des classes. Ces courbes sont parfaitement adaptées pour des classifieurs probabilistes mais leur utilisation avec les arbres de décision est un peu plus problématique car la fiabilité de l'estimation des probabilités *a posteriori* dans les arbres de décision est douteuse. De nombreux travaux ont été réalisés dans ce domaine (Zadrozny et Elkan, 2001; Alvarez *et al.*, 2007; Provost et Domingos, 2003) mais leur application aux arbres de décision flous reste un problème ouvert.

10.3.2.2 F-mesure

Plutôt que de recourir à l'une de ces trois mesures dont aucune ne répond complètement à nos attentes, nous avons préféré utiliser une F-mesure. Cette mesure est en effet fréquemment employée pour des problèmes dont les caractéristiques sont voisines du nôtre : une classe largement minoritaire associée à un coût d'erreur nettement plus important que celui qui est associé aux autres classes (Lewis et Gale, 1994; Daskalaki *et al.*, 2006). Une F-mesure permet de combiner rappel et précision d'une même classe. Soit c la classe d'intérêt, elle s'exprime de la manière suivante :

$$F_m(c) = \frac{\left(\beta^2 + 1\right)(\text{précision}(c) \times \text{rappel}(c))}{\beta^2 \text{précision}(c) + \text{rappel}(c)}$$

β est un paramètre qui permet de biaiser la mesure en faveur du rappel ou de la précision. Lorsque $\beta < 1$ la précision aura plus d'importance, tandis que ce sera le rappel qui prédominera lorsque $\beta > 1$. Lorsque $\beta = 1$, qui est le cas le plus répandu dans la littérature, les deux ont même importance. La F-mesure est un opérateur d'agrégation du rappel et de la précision. Elle est en effet comprise entre 0 et 1 et est croissante aussi bien en fonction du rappel que de la précision. Lorsque rappel et précision ont la même valeur on a de plus la relation suivante : $F_m(c) = \text{rappel}(c) = \text{précision}(c)$.

L'objectif de nos travaux étant d'identifier les pays fragiles, susceptibles d'être le théâtre d'affrontements armés, nous avons appliqué cette mesure à la classe *crise*. Accordant plus d'importance au rappel qu'à la précision nous avons choisi de fixer le paramètre β à 2[15].

[15] Nous avons testé différentes valeurs de β supérieures à 1 avant d'arrêter notre choix de façon purement empirique.

Dans notre système la mesure de performance n'a d'intérêt que pour la comparaison et la sélection de modèles. Elle doit permettre de définir une relation d'ordre sur l'ensemble des modèles évalués afin qu'une décision quant au modèle à retenir puisse être prise. Elle ne doit être qu'un outil au service de l'utilisateur pour guider ses choix. Elle doit donc refléter dans la mesure du possible les préférences de cet utilisateur.

La F-mesure appliquée à la classe *crise* avec $\beta = 2$ répond assez bien à nos attentes. Nous souhaitons en effet privilégier les classifieurs qui parviennent à identifier un maximum de pays en crise tout en conservant un taux de fausses alertes raisonnable pour assurer la crédibilité du système.

Cependant avec une telle mesure il est très difficile d'exprimer l'indifférence. Des classifieurs associés des matrices de confusion voisines obtiendront bien souvent des scores de performance différents et seront donc considérés comme différents. Bien que cela puisse être souhaitable dans certains cas, laisser l'utilisateur final exprimer ses préférences nous semble être une solution mieux adaptée à notre tâche dont l'un des objectifs, rappelons-le, est d'apporter une aide à la décision. Une mesure de performance telle que la F-mesure est paramétrable et il est envisageable de laisser l'utilisateur fixer la valeur du paramètre β. Pour accroître les degrés de liberté de l'utilisateur il suffirait de mettre à sa disposition un large éventail de mesures paramétrables, mais les mesures de performance classiquement employées ont les mêmes difficultés que la F-mesure à exprimer l'indifférence.

10.3.2.3 Règles-mesure : prise en compte des préférences de l'utilisateur

Pour surmonter cette difficulté nous avons opté pour des mesures de performance à base de règles. Pour introduire un maximum de souplesse dans la définition de ces mesures nous avons considéré des règles floues. Le modus ponens généralisé présenté à la section 9.2 et illustré par l'équation 9.1 est le mode de raisonnement qui nous permet d'évaluer le score d'un classifieur.

Les règles floues qui définissent une mesure de performance portent sur des variables linguistiques qui correspondent à des mesures que l'on peut extraire d'une matrice de confusion, telles que le taux de bonnes classifications, le rappel, la précision ou la F-mesure de telle classe... Pour chacune des variables linguistiques choisies il faut alors définir les modalités qu'elle est susceptible de prendre ainsi que les sous-ensembles flous correspondants. Ceci vaut également pour la variable *score* qui correspond à la sortie de ce système d'inférence. Enfin les règles définissant la mesure de performance doivent être choisies.

Pour nos expérimentations nous avons utilisé la précision et le rappel de la classe *crise* comme variables d'entrée du système, l'idée étant de construire une mesure proche de la F-mesure décrite précédemment (avec $\beta = 2$), mais plus souple et moins précise de façon à transcrire l'indifférence entre matrices de confusion voisines. Nous avons pris trois modalités pour ces deux variables ainsi que pour la mesure de performance : faible, moyen et élevé. Les sous-ensembles flous correspondant sont donnés à la figure 10.2.

Les règles que nous avons employées sont les suivantes :

- **Si** le rappel ou la précision sont faibles **alors** le score est faible
- **Si** le rappel est moyen et **si** la précision n'est pas faible **alors** le score est moyen
- **Si** la précision est moyenne et **si** le rappel n'est pas faible **alors** le score est moyen
- **Si** le précision et le rappel sont élevés **alors** le score est élevé

Le mécanisme d'agrégation de la précision et du rappel que ces règles permettent de mettre en place est représenté sur la figure 10.3 dans laquelle nous n'avons pas tenu compte du flou.

Avec de telles règles de nombreux classifieurs peuvent être considérés comme équivalents

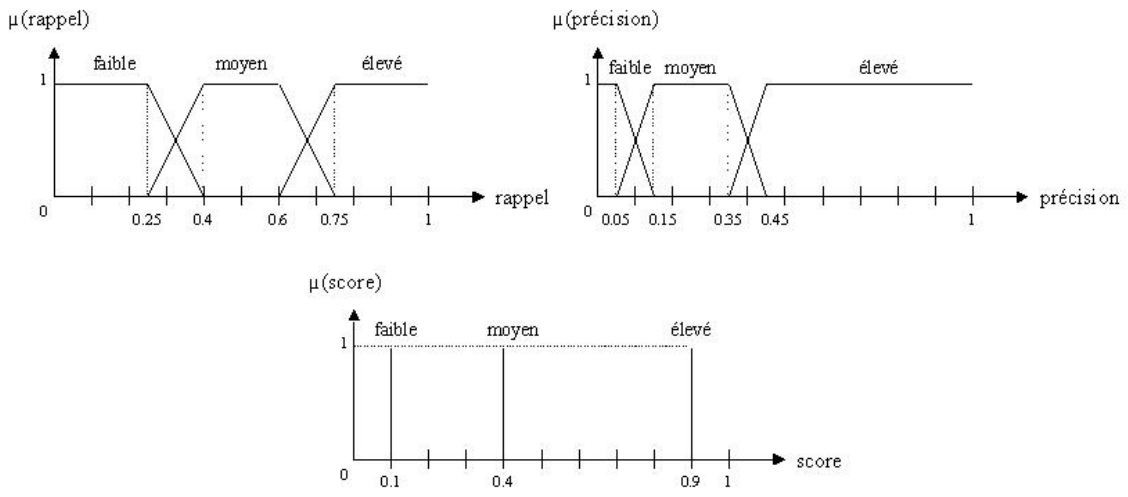

FIG. 10.2 – Sous-ensembles flous associés aux trois modalités des variables rappel, précision et score. La variable score correspond à la mesure de performance.

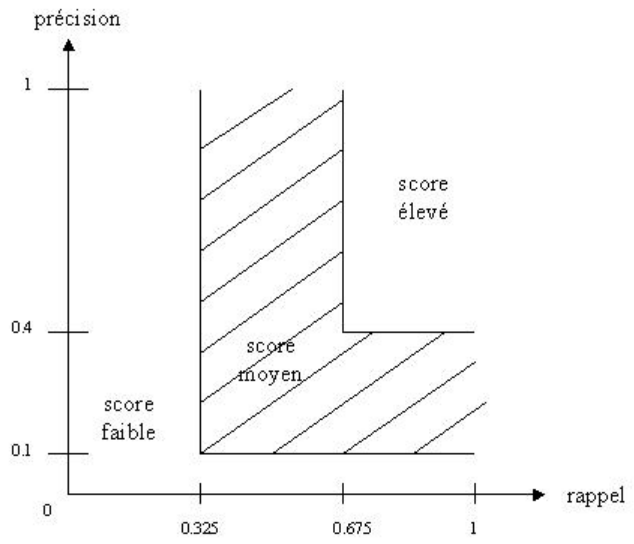

FIG. 10.3 – Processus d'agrégation du rappel et de la précision dans le cas non flou

bien que les matrices de confusion qu'ils permettent d'obtenir soient quelque peu différentes. Il est donc possible d'exprimer l'indifférence. À titre d'illustration nous donnons dans le tableau 10.1 quatre matrices de confusion qui sont toutes associées à un même score.

> *Rappelons que l'élément (i, j) d'une matrice de confusion correspond au nombre d'observations de la classe i auxquelles a été attribuée la classe j. La classe minoritaire, correspondant à la seconde ligne de chacune des matrices de confusion, est la classe crise, celle qui est prise en compte pour le calcul de la précision et du rappel.*

Dans l'analyse des résultats que nous allons détailler dans les sections suivantes nous avons utilisé la mesure de performance basée sur les règles floues que nous venons de présenter. Nous la nommerons *RèglesMesure*. Nous avons également considéré la F-mesure paramétrée par $\beta = 2$ afin de disposer d'une autre mesure répondant à nos besoins mais plus communément admise dans le domaine de l'apprentissage automatique. Nous utiliserons simplement le terme *F-mesure* pour la désigner.

TAB. 10.1 – Quatre matrices de confusion dont les performances sont jugées équivalentes par la mesure de performance à base de règles floues (score de 0.4)

109	5
7	9

111	3
8	8

111	3
9	7

110	4
7	9

10.3.3 Sélection d'un modèle de prédiction

Chacune de nos 53 bases de données correspond à un sous-problème pour lequel nous souhaitons construire un modèle de détection des conflits armés intra-étatiques qui soit aussi performant et fiable que possible. Disposant de 129 méthodes de prétraitement nous avons construit autant de modèles pour chacune des bases. Il nous faut donc être capable d'en sélectionner un parmi ces 129, et ce pour chacun de nos sous-problèmes. L'objectif de cette section est d'introduire la méthodologie que nous avons mise en place pour y parvenir.

10.3.3.1 Analyse de rangs

La sélection d'un modèle ne peut se faire sans analyse comparative préalable. En nous inspirant du travail réalisé à la partie II pour comparer les performances de différentes méthodes de sélection d'attributs et de substitution des valeurs manquantes, nous avons choisi de procéder à une analyse de rangs.

Sur chacune des 53 bases de données, les méthodes ont été ordonnées par performances décroissantes. Nous avons alors pu appliquer le test de Friedman pour estimer le caractère significatif des différences observées entre les différentes performances. Nous avons ensuite appliqué le test de Nemenyi, présenté à la section 5, à toutes les paires de méthodes afin de savoir lesquelles différaient.

Les 53 bases de données se référent toutes à un même problème et sont donc relativement homogènes. Il n'est donc pas aberrant de considérer la moyenne et l'écart-type des performances sur l'ensemble des bases de données à disposition. L'incommensurabilité des performances, évoquée à la section 5 pour justifier la faiblesse de l'ANOVA lors de la comparaison de classifieurs évaluées sur différentes tâches, ne pose donc pas de problème majeur. Si nous avons préféré le test de Friedman à l'ANOVA ce n'est pas donc pas pour cette raison mais plutôt à cause de la non-indépendance des bases de données. Certaines observations sont en effet présentes dans 2, 3 ou 4 bases de données, la seule différence étant la durée de la période d'estimation des différents attributs qui peut être de 1, 7, 15 ou 29 années (voir section 10.2). Cette dépendance entre les différentes bases de données influe sur la qualité de l'analyse de rangs mais plus encore sur l'ANOVA (Zar, 1999).

Méthodes statistiquement moins performantes L'analyse de rangs menée sur les 53 bases de données révèle que 30 méthodes sont statistiquement moins performantes que celle qui obtient le rang le plus faible. Ce résultat se vérifie avec les deux mesures de performance considérées : *F-mesure* et *Règles-mesure*. Ces méthodes, la moyenne et l'écart-type de leurs rangs ainsi que le rang minimum et maximum qu'elles obtiennent sont présentés au tableau 10.2. Les résultats étant similaires avec les deux mesures de performance nous ne donnons que ceux qui correspondent à *F-mesure*.

Le tableau 10.2 met en évidence la faiblesse de certaines méthodes de substitution. En effet seules quatre méthodes sur les dix testées sont représentées, tandis que tous les filtres sont présents. Sur notre problème il semble donc que la substitution des valeurs manquantes joue un rôle important dans la qualité des modèles appris. Les quatre méthodes

TAB. 10.2 – Liste des 30 méthodes statistiquement moins performantes que la meilleure des 129, ordonnées par rang moyen croissant. Les statistiques de rang associées à chacune des méthodes ont été estimées en prenant *F-mesure* comme mesure de performance.

Méthode	Moyenne	Écart-type	Minimum	Maximum
(C) EF-Entropie FCBF	78.84	32.62	12.5	126.5
(C) CMoyenne FCBF	79.13	34.71	7.5	126.5
(C) CMoyenneA KSCBF	80.68	29.7	14	127
(A) 1LLSI KSCBF	83.79	34.25	4	128.5
(A) CMoyenneA FCBF	84.77	36.02	2.5	127.5
(A) 1LLSI CFS	85.45	34.97	7.5	127.5
(A) CMoyenneA SansFiltre	85.9	36.98	2	129
(A) 1LLSI FCBF	86.06	33.21	3.5	129
(B) CMoyenneA KSF	88.59	34.69	11	123
(A) CMoyenneA CFS	88.75	35.48	2.5	127.5
(A) 1LLSI KSF	88.86	33.24	3.5	128
(A) CMoyenneA KSCBF	88.94	37.62	2.5	129
(C) CMoyenneA KSF	89.08	33.95	8	127
(A) EF-Entropie FCBF	89.33	42.64	1.5	129
(A) EF-Entropie SansFiltre	90.56	39.21	8	126
(C) EF-Entropie KSF	90.88	39.6	1.5	129
(A) CMoyenneA KSF	91.2	35.77	3	127.5
(B) EF-Entropie KSF	91.31	39.24	1	129
(C) CMoyenne KSCBF	93.25	32.52	2	128.5
(A) EF-Entropie KSCBF	93.56	37.38	12	128
(A) CMoyenne SansFiltre	94.17	36.01	7	129
(A) EF-Entropie CFS	94.25	38.68	8	129
(A) EF-Entropie KSF	94.54	36.88	13.5	129
(C) EF-Entropie KSCBF	94.68	28.48	22.5	128
(A) CMoyenne CFS	95.34	35.64	7	128.5
(A) CMoyenne FCBF	95.57	34.55	7	129
(C) CMoyenne KSF	95.77	35.92	13.5	127.5
(A) Cmoyenne KSF	96.2	36.06	4.5	129
(A) CMoyenne KSCBF	96.69	32.7	7	128.5
(B) CMoyenne KSF	97.6	32.09	13.5	129

mal adaptées à notre problème sont les suivantes *CMoyenne*, *CMoyenneA*, *EF-Entropie* et *1LLSI*.

Il est important de constater que trois d'entre elles sont des méthodes que nous avons nommées supervisées. Ce sont de plus les seules méthodes supervisées que nous ayons testées. Or ce sont celles qui semblaient les plus prometteuses lors de nos expériences sur les données manquantes (voir tableau 6.10). Plusieurs explications peuvent être avancées pour justifier un tel décalage.

Premièrement les caractéristiques des données traitées sont fort différentes. La distribution des observations dans les différentes classes est bien plus déséquilibrée dans le cas des conflits armés. Le nombre d'attributs est également bien plus élevé que lors de nos expérimentations sur les bases de données de l'UCI. Le mécanisme de génération des valeurs

manquantes est de plus ici inconnu. Il est peu vraisemblable que le mécanisme MCAR, celui que nous avons utilisé lors de nos précédentes expériences, soit à même de rendre compte de la distribution des valeurs manquantes dans ce cas réel.

Deuxièmement les mesures de performance considérées sont également fort différentes.

Troisièmement et il s'agit peut-être de l'explication la plus plausible, les tests sur les données de l'UCI ont été effectués en prenant soin de ne supprimer des valeurs que sur les bases d'apprentissage, les bases de test étant conservées intactes. Avec des bases de données réelles telles que celles que nous avons construites pour les conflits armés il n'est pas possible de procéder ainsi. Il n'est pas non plus possible d'employer les méthodes supervisées pour compléter les bases de test puisque la classe des observations de ces bases n'est pas connue.

C'est la raison pour laquelle nous avons utilisé la méthode *Moyenne* avec chacune de ces trois techniques supervisées pour traiter les bases de test. Il est vraisemblable que cette solution un peu simpliste ait contribué à dégrader les performances de ces méthodes. Il serait utile de chercher un moyen plus efficace de remplacer les valeurs manquantes sur les données de test.

Une solution pourrait être de classer les observations de test à l'aide d'un classifieur simple tel que le plus proche barycentre des différentes classes, afin de pouvoir utiliser la même méthode supervisée pour ces observations que pour celles de la base d'apprentissage. Pour la méthode *EF-Entropie* que nous avons proposée, il serait bon de tester d'autres méthodes de discrétisation plus sophistiquées afin d'analyser l'influence de la phase de discrétisation sur la qualité de la substitution.

Quant à la méthode *1LLSI*, il est intéressant de constater qu'elle n'apparaît dans le tableau 10.2 qu'au sein de chaînes de prétraitement pour lesquelles la stratégie (A) est employée. Avec cette stratégie la susbtitution est réalisée avant le filtrage. Lorsqu'aucun filtre n'est appliqué la chaîne d'apprentissage correspondante n'est pas statistiquement moins performante que les autres. Aussi la méthode *(A) 1LLSI SansFiltre* ne figure-t-elle pas dans le tableau 10.2.

Son rang moyen est cependant de 75.63, soit relativement proche du rang des 30 méthodes les plus faibles. En présence d'un grand nombre d'attributs substituer les valeurs manquantes en construisant un modèle de régression à partir d'une seule variable explicative ne semble donc pas être une bonne solution. Lorsque le nombre d'attributs est réduit après filtrage, cette méthode se comporte nettement mieux comme nous aurons l'occasion de le voir dans la suite.

En analysant le rang minimum obtenu par chacune des 30 méthodes les plus faibles il ressort que six d'entre elles se sont classées dans les 3 premières méthodes sur un problème donné. Elles obtiennent donc tout de même de bonnes performances sur certaines bases de données. Étant donné leurs piètres performances dans l'ensemble nous estimons cependant préférable de ne pas les considérer lors de la sélection des modèles à laquelle nous devons procéder pour chaque base de données, et ce y compris pour les quelques bases sur lesquelles elles s'avèrent efficaces.

L'analyse des rangs que nous venons de conduire nous permet de supprimer 30 méthodes mais elle est encore loin de résoudre notre problème de sélection de modèles. Il nous reste en effet 99 méthodes potentiellement éligibles. Ce grand nombre indique qu'il est difficile de départager les différentes méthodes. Nombreuses sont celles qui sont bien adaptées pour certains sous-problèmes. Ceci suffit selon nous à justifier l'intérêt de disposer d'une plateforme d'évaluation, de comparaison et de sélection de méthodes si l'on veut construire un modèle performant sur des problèmes spécifiques.

Méthodes obtenant les meilleures rangs avec *F-mesure* Bien que cela ne soit pas utile à notre tâche de sélection de modèles, nous allons nous attarder sur les 30 méthodes les plus performantes. Ceci nous permettra d'approfondir notre analyse des méthodes de prétraitement. Les résultats correspondant à *F-mesure* sont donnés au tableau 10.3. La terminologie est ici quelque peu tendancieuse dans la mesure où l'utilisation du superlatif laisse penser que ces méthodes ont des performances statistiquement supérieures à celles des autres méthodes. Or tel n'est le cas que vis-à-vis des 30 méthodes les plus faibles mentionnées précédemment. Nous ne faisons ici référence qu'aux techniques dont les rangs moyens sont les plus faibles.

TAB. 10.3 – Liste des 30 méthodes les plus performantes du point de vue des rangs moyens, ordonnées par rang moyen croissant. Les statistiques de rang associées à chacune des méthodes ont été estimées en prenant *F-mesure* comme mesure de performance.

Méthode	Moyenne	Écart-type	Minimum	Maximum
(B) 1LLSI KSCBF	47.42	29.74	4	120
(A) 1ppv SansFiltre	47.83	35.09	2	117
(A) Moyenne CFS	47.85	30.23	3	128
(B) 5ppv KSCBF	47.85	30.17	5.5	117.5
(B) Médiane KSCBF	48.01	29.83	1	117.5
(B) AléatoireMM CFS	48.22	30.34	2.5	127
(A) Moyenne KSCBF	48.55	26.66	3	110.5
(B) Moyenne KSCBF	48.73	31.38	10	119
(A) 1ppv FCBF	48.87	33.03	1	117
(B) 1LLSI FCBF	49.19	32.6	4.5	123
(B) 5LLSI KSCBF	49.26	31.56	1.5	120
(B) 5LLSI FCBF	49.44	32.78	4.5	123
(A) Médiane CFS	49.75	33.74	1	123.5
(A) Moyenne FCBF	50.08	29.75	2	128
(B) Moyenne CFS	50.35	28.97	3	114.5
(B) 5LLSI CFS	50.36	34.04	1	123
(B) 1LLSI CFS	50.61	31.45	2.5	126
(A) 5ppv SansFiltre	50.97	32.73	2	126.5
(B) Médiane KSF	51.11	39.07	1	127
(B) AléatoireMM KSCBF	51.55	30.27	8	121.5
(B) 1ppv KSCBF	51.6	29.94	2.5	117.5
(B) 5ppv CFS	51.65	30.58	2.5	114.5
(B) Médiane FCBF	51.75	32.78	4.5	123
(B) AléatoireMM FCBF	51.78	30.79	1	112.5
(B) 5ppv FCBF	51.92	32.98	4.5	123
(C) Médiane KSF	52.02	38.05	1	127
(B) Médiane CFS	52.21	31.9	5	114.5
(B) CMoyenneA KSCBF	52.5	34.14	1.5	122.5
(A) AléatoireMM CFS	52.53	33.11	1	126
(A) 1ppv CFS	52.76	31.77	2	120

Du tableau 10.3, il ressort que la méthode de substitution *1LLSI* intervient dans 3 des 30 chaînes de prétraitement les plus performantes avec la stratégie (B), c'est-à-dire lorsque le filtrage est réalisé en premier. Ceci confirme l'une de nos remarques précédentes, à savoir

que cette méthode n'est pas dénuée d'intérêt à condition de réduire le nombre d'attributs. La substitution basée sur les proches voisins ou sur la moyenne et la médiane s'avèrent également efficaces.

La présence dans le tableau 10.3 de méthodes employant la substitution aléatoire est surprenante. Nous avions déjà observé un tel phénomène à la section 6.6.4, mais les expériences portaient sur des données symboliques. En outre, sur des données continues, cette méthode était quasi systématiquement moins performante que les autres.

Ces remarques impliquent qu'aucune méthode de substitution ne semble réellement bien adaptée à notre problème. Le seul avantage de l'aléatoire est de préserver l'incertitude sous-jacente associée à la substitution des valeurs manquantes. Une technique comme les plus proches voisins qui est déterministe ne parvient pas à surpasser l'aléatoire. Nous n'avons malheureusement testé qu'une seule autre technique non déterministe *CMoyenneA* qui est supervisée et ne donne pas des résultats satisfaisants comme nous l'avons vu précédemment.

Une des perspectives ouvertes par cette remarque serait de tester d'autres méthodes non déterministes et non supervisées comme par exemple la version stochastique de la moyenne simple *MoyenneA*.

Autre point marquant dans ce tableau, la présence parmi les 30 méthodes de deux techniques n'ayant pas recours au filtrage (*(A) 1ppv SansFiltre* et *(A) 5ppv SansFiltre*). Dans les deux cas la substitution repose sur les plus proches voisins. Malgré le très grand nombre d'attributs, la sélection opérée en interne par *Salammbô* est donc suffisamment efficace pour que des modèles performants puissent être appris directement, sans recourir à une étape de filtrage préalable. Il serait bon de procéder à d'autres tests avec des classifieurs n'employant aucune méthode de sélection d'attributs en interne pour mieux rendre compte de l'influence du filtrage dans la chaîne de prétraitement.

Mais il est également possible de renverser l'argumentation pour se rendre compte de l'intérêt du filtrage. Malgré une forte réduction de la dimension opérée par la sélection d'attributs, il est possible d'apprendre beaucoup plus rapidement des modèles performants et simples.

En observant l'écart-type des rangs des différentes méthodes on observe de plus qu'il est possible d'apprendre des modèles plus stables en incluant le filtrage dans la phase de prétraitement. De nombreuses méthodes utilisant le filtrage ont en effet un écart-type plus faible que les deux techniques n'y ayant pas recours.

On peut constater que tous les filtres sont représentés parmi les 30 méthodes ayant les rangs les plus faibles, mais il est à noter que celui que nous avons proposé (KSF) ne fait partie des meilleures solutions que lorsqu'il est combiné à la méthode *Médiane*. L'écart-type élevé et les rang minimum et maximum obtenus par les méthodes *(B) Médiane KSF* et *(C) Médiane KSF* suggèrent que leurs performances ne sont pas stables. Elles sont très bien adaptées pour certaines bases de données puisque leur rang minimum est égal à 1, mais elles sont également totalement inadaptées à d'autres bases de données, leur rang maximum étant de 127. Rappelons que le rang le plus élevé possible est de 129.

Quant à la stratégie de combinaison de la substitution et du filtrage, elle semble influer sur la qualité des chaînes de prétraitement. Sur les 30 chaînes de ce tableau, 20 utilisent la stratégie (B) tandis que la stratégie (C) n'est employée que dans un seul cas. Il semble ainsi préférable de réaliser le filtrage en premier sans tenir compte des valeurs manquantes. Ceci peut en partie s'expliquer par la faiblesse des approches mises en place pour intégrer l'information véhiculée par la distribution des valeurs manquantes au niveau du filtrage.

Méthodes obtenant les meilleurs rangs avec *Règles-mesure* Afin de rendre compte des différences entre les deux mesures de performance nous présentons les résultats correspondant à *Règles-mesure* au tableau 10.4.

TAB. 10.4 – Liste des 30 méthodes les plus performantes du point de vue des rangs moyens, ordonnées par rang moyen croissant. Les statistiques de rang associées à chacune des méthodes ont été estimées en prenant *Règles-mesure* comme mesure de performance.

Méthode	Moyenne	Écart-type	Minimum	Maximum
(A) 1ppv SansFiltre	45.91	28.95	1.5	108
(A) 1ppv FCBF	50.58	29.15	3.5	108
(B) Médiane KSF	51.82	32.49	2	113.5
(B) 5LLSI FCBF	52.47	26.34	7.5	103.5
(B) 5ppv FCBF	53.11	25.87	7.5	106.5
(C) Médiane KSF	53.23	33.28	1.5	109
(B) 5LLSI CFS	53.24	25.55	8.5	109
(C) 5LLSI KSCBF	53.26	29.86	1.5	109
(B) 1LLSI FCBF	53.29	26.21	7.5	106.5
(B) Moyenne FCBF	54.12	25.9	3	106.5
(B) 1LLSI KSCBF	54.32	22.52	17.5	102
(B) Médiane FCBF	54.39	25.93	7.5	106.5
(A) 5ppv SansFiltre	54.65	30.81	1.5	104.5
(A) Moyenne KSCBF	54.68	23.87	7	104.5
(A) Médiane CFS	54.88	28.51	1.5	125.5
(A) Moyenne CFS	55.15	26.8	3.5	128
(B) 1ppv FCBF	55.31	25.44	7.5	103.5
(B) AléatoireMM FCBF	55.41	24.46	1	103.5
(A) 1ppv CFS	55.47	26.33	4.5	108
(B) Moyenne KSCBF	55.6	25.19	15.5	118.5
(B) AléatoireMM CFS	55.61	27.01	3.5	109
(B) Médiane KSCBF	55.68	24.24	15.5	106.5
(B) EF-Entropie FCBF	55.8	27.25	7.5	109
(B) CMoyenneA FCBF	55.87	26.73	3	109
(B) 5ppv KSCBF	56.12	25.45	13	106.5
(B) 1LLSI CFS	56.34	26.23	2	109
(A) 1ppv KSCBF	56.45	26.46	3.5	125
(A) Moyenne FCBF	56.99	25.4	9	128
(B) AléatoireMM KSCBF	57.08	23.5	19.5	109
(B) 5LLSI KSCBF	57.09	23.97	19.5	106.5

Les résultats obtenus avec la mesure de performance à base de règles sont assez proches de ceux observés avec *F-mesure*. Des 30 méthodes ayant les rangs les plus faibles, 25 sont communes aux deux, l'ordre de ces 25 méthodes différant quelque peu selon la mesure de performance considérée. Ceci est compréhensible dans la mesure où nous avons fait en sorte que la mesure à base de règles soit une version plus souple de *F-mesure*.

Si l'on excepte la méthode *(A) 1ppv SansFiltre* dont le rang moyen est nettement en dessous des autres dans le tableau 10.4, on peut observer que les rangs moyens obtenus par les différentes méthodes sont légèrement supérieurs à ceux que l'on peut observer avec *F-mesure*. Cette légère hausse peut s'expliquer par le fait que nous avons autorisé

des performances voisines à être considérées comme équivalentes afin de transcrire une certaine indifférence entre de telles performances. Cela entraîne de nombreuses égalités dans l'estimation des rangs, ce qui a pour conséquence de rehausser les rangs faibles et de diminuer les rangs élevés.

Pour conforter cette remarque nous pouvons préciser que les 30 méthodes les plus faibles selon le critère *Règles-mesure* ont des rangs moyens compris entre 74.58 et 93.28, alors que les résultats donnés au tableau 10.2 indiquent qu'avec le critère *F-mesure* ces rangs sont compris entre 78.84 et 97.6.

Quant à la méthode *(A) 1ppv SansFiltre*, la différence que l'on observe par rapport aux résultats du tableau 10.3 s'explique également par la souplesse introduite par les règles floues dans l'évaluation des classifieurs. Ses performances jugées bonnes avec le critère *F-mesure* sont nettement au-dessus des performances des autres méthodes, tandis que pour les bases de données sur lesquelles elle était moins bien classée selon *F-mesure* ses performances ne sont que légèrement inférieures à celles des autres méthodes, si bien que selon le critère *Règles-mesure* elle est jugée équivalente à ces méthodes, ce qui contribue à abaisser son rang. À un degré moindre ces mêmes remarques peuvent s'appliquer à *(B) Médiane KSF*. Elle obtient en effet le 3e meilleur rang moyen avec *Règles-mesure* et seulement le 19e avec *F-mesure*.

Afin d'illustrer les différences entre les deux mesures de performance que nous venons de mentionner, nous avons tracé sur les graphiques de la figure 10.4 les distributions des rangs de 7 méthodes de prétraitement en fonction de la base de données considérée.

L'ordre dans lequel les 53 bases de données sont considérées est propre à chaque méthode. Nous avons en effet fait en sorte de trier les rangs de chaque méthode par ordre croissant. Les identifiants que nous avons mis en abscisse réfèrent donc à des bases de données différentes selon la méthode concernée. Les quatre premières méthodes envisagées font partie des 30 méthodes ayant les rangs les plus bas et ce pour les 2 critères de performance. Les 3 dernières en revanche font partie des méthodes que nous avons choisi de rejeter. Leurs performances sont statistiquement inférieures à celles des quatre premières méthodes.

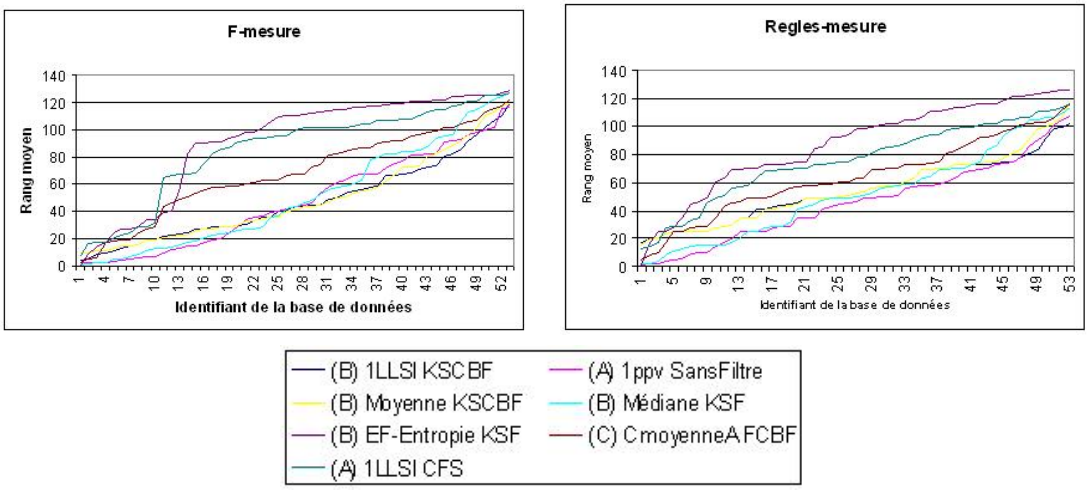

FIG. 10.4 – Distribution des rangs moyens obtenus par 7 méthodes sur l'ensemble des bases de données, avec les critères *F-mesure* et *Règles-mesure*. Pour chaque méthode, les bases de données ont été ordonnées de telle sorte que les rangs moyens soient triés par ordre croissant.

Ainsi que nous l'avons remarqué précédemment nous observons une atténuation des

différences entre les méthodes avec le critère *Règles-mesure*. Les courbes des trois plus mauvaises méthodes sont en effet plus proches de celles des quatre meilleures.

On constate également que la technique *(B) Médiane KSF* obtient de très bonnes performances (faibles rangs) sur de nombreuses bases. Mais ses performances sont de qualité bien moindre (partie droite de la courbe) sur certaines méthodes dont le nombre est également grand. Ces observations sont valables quel que soit le critère considéré, mais les différences entre rangs faibles et élevés sont nettement plus marquées avec *F-mesure*. Ceci explique que cette méthode soit bien mieux classée avec le critère *Règles-mesure*. Le comportement de la méthode *(B) 1LLSI KSCBF* se distingue de celui que nous venons d'évoquer par sa plus grande stabilité. Ses performances ne sont jamais aussi bonnes que celles de de *(B) Médiane KSF*, mais elles ne sont également jamais aussi mauvaises. La variabilité de ses performances est donc bien moindre.

Pour conclure notre analyse de rangs, nous pouvons constater que si les maxima des rangs des 30 meilleures méthodes sont inférieurs à ceux des 30 méthodes les plus faibles, la différence étant plus nette avec le critère *Règles-mesure*, ils sont cependant tous supérieurs à 100. Ceci indique qu'aucune des 129 méthodes n'est bien adaptée pour les 53 bases de données dont nous disposons et confirme si besoin était qu'une analyse plus fine et plus spécifique de chacune des 53 sous-tâches est nécessaire.

10.3.3.2 Analyse de la confiance

Abordons à présent la méthodologie que nous avons employée pour départager les 99 méthodes restantes sur chacune des sous-tâches, les 30 méthodes statistiquement moins performantes que les autres ayant été supprimées suite à l'analyse de rangs.

Une solution simple consiste à retenir pour chacune des bases de données la méthode maximisant le critère de performance choisi. Il nous faut cependant introduire de nouveaux critères afin de pouvoir distinguer les différentes méthodes obtenant des performances maximales. Rien ne garantit en effet que les performances maximales ne soient obtenues que par une seule méthode. Pour l'instant nous nous sommes contenté de comparer les différentes méthodes en fonction de leur capacité à prédire correctement ou non l'occurrence d'une crise. Estimer le pouvoir prédictif uniquement à partir de la matrice de confusion est cependant réducteur. De nombreuses informations pourtant fort utiles sont perdues, comme par exemple l'incertitude sous-jacente associée au classement de chaque exemple. *Salammbô* ne fournit certes pas des probabilités *a posteriori* valides, mais les décisions qu'il prend s'appuient sur les degrés d'appartenance aux différentes classes de chaque exemple.

Nous avons choisi d'intégrer cette information dans le processus d'évaluation d'un modèle afin de rendre compte de la confiance que l'on peut lui accorder. Nous disposons ainsi d'un nouveau critère d'évaluation sur lequel on pourra s'appuyer pour effectuer la sélection de modèles. Nous avons envisagé deux façons différentes de tenir compte de la confiance d'un modèle.

Seuil de rejet Dans un premier temps nous avons décidé d'introduire un seuil de rejet visant à exclure de l'analyse toutes les observations pour lesquelles les degrés de reconnaissance de chacune des classes sont trop proches. L'incertitude qui sous-tend les décisions prises par *Salammbô* à propos de ces observations est trop importante pour que ces décisions soient jugées fiables. Pour certaines observations, les degrés de reconnaissance des deux classes sont parfois identiques si bien que *Salammbô* leur attribue une classe de manière purement aléatoire. L'intérêt du seuil de rejet est donc de prévenir l'occurrence des cas de figure pour lesquels le comportement de *Salammbô* est proche de l'aléatoire. Nos deux

mesures de performance *Règles-mesure* et *F-mesure* sont alors appliquées sur les matrices de confusion générées après avoir supprimé toutes les observations litigieuses.

L'inconvénient de cette approche réside dans la non-prise en compte du nombre d'exemples ambigus. Considérons par exemple deux méthodes ayant même matrice de confusion sur une base de 100 exemples. Supposons qu'après suppression des décisions hautement incertaines nous obtenons les matrices de confusion du tableau 10.5 :

TAB. 10.5 – Deux matrices de confusion initialement identiques, après suppression des observations litigieuses

30	0
0	5

75	5
2	16

Nos deux mesures de performance nous inciteront à privilégier la première méthode qui ne commet aucune erreur une fois que les observations litigieuses ont été supprimées. Or il serait peut-être préférable d'accorder plus de crédit à la seconde. Si elle se trompe plus souvent que la première, elle est cependant capable de classer près de trois fois plus d'observations avec une marge d'erreur raisonnable.

Score de Brier En nous inspirant des travaux de Zadrozny et Elkan (2001); Alvarez *et al.* (2007) sur la calibration des probabilités des arbres de décision, nous avons construit une mesure proche de l'erreur quadratique moyenne de prédiction ou *score de Brier*. Cette mesure quantifie l'écart moyen pour l'ensemble des observations entre les probabilités *a posteriori* réelles et celles qui sont fournies par un classifieur. En notant c_k la k-ième classe et e_i le i-ième exemple, cette mesure s peut être exprimée par la formule suivante :

$$s = \sum_{i=1}^{n} \sum_{k=1}^{K} \left(\hat{P}\left(c_k|e_i\right) - P\left(c_k|e_i\right) \right)^2$$

$\hat{P}$ correspond à la probabilité *a posteriori* fournie par le classifieur tandis que P est la probabilité *a posteriori* réelle. Nous avons fixé $P\left(c_k|e_i\right)$ à 1 lorsque c_k est effectivement la classe de e_i et à 0 sinon. Un écart important est révélateur d'un manque de fiabilité du classifieur. Étant donné que nous sommes dans un problème à deux classes, nous avons $P\left(c_1|e_i\right) = 1 - P\left(c_2|e_i\right)$ et de même $\hat{P}\left(c_1|e_i\right) = 1 - \hat{P}\left(c_2|e_i\right)$. Nous avons ainsi :

$$s = 2 \sum_{i=1}^{n} \left(\hat{P}\left(c_k|e_i\right) - P\left(c_k|e_i\right) \right)^2$$

Le facteur 2 n'influant pas lorsque nous comparons les mesures de confiance de deux classifieurs, nous avons choisi de ne pas en tenir compte.

Avec cette mesure il n'est plus besoin de supprimer les observations associées à une incertitude élevée puisqu'elles pénalisent directement la mesure de confiance, mais l'application de cette mesure n'en est pas moins délicate. Rappelons que *Salammbô* utilise les degrés de reconnaissance de chaque classe pour prendre ses décisions, ces degrés n'étant pas des probabilités. Afin de pouvoir les interpréter comme des probabilités, nous avons utilisé les degrés normalisés fournis par *Salammbô* qui sont tels que leur somme est égale à 1.

Mais cette solution n'est pas complètement satisfaisante dans la mesure où une partie de l'information est perdue. Lorsque nous disposons des degrés de reconnaissance des deux classes deux paramètres importent, la différence entre les deux ainsi que la valeur du degré

le plus élevé qui est celui à partir duquel *Salammbô* prendra une décision. Plus cette valeur est élevée et plus grande est la confiance que l'on peut avoir dans la décision à condition que l'écart par rapport au degré de reconnaissance de l'autre classe soit suffisamment important. Avec notre processus de normalisation il n'est pas toujours possible de conserver l'information véhiculée par ces deux paramètres.

Supposons par exemple que nous avons deux observations e_1 et e_2 dont les degrés de reconnaissance de la classe 1 sont respectivement 0.1 et 0.8 tandis que les degrés de reconnaissance de la classe 2 sont tous deux nuls. Après normalisation nous avons $\hat{P}(c_1|e_1) = \hat{P}(c_1|e_1) = 1$. La confiance que l'on peut accorder au classement de ces deux observations est pourtant loin d'être identique. En raison de ces faiblesses nous avons choisi de limiter l'emploi de cette mesure de confiance[16].

Nous ne l'avons utilisée que lorsque deux modèles ou plus n'ont pu être départagés directement par la mesure de performance évaluée après suppression des observations litigieuses. Voyons en détail deux exemples issus de nos expérimentations qui nous permettront d'illustrer la méthodologie que nous avons employée pour tenir compte de la confiance.

Notre premier exemple se rapporte à la base *1 Asie du Sud Est, Pacifique 2*. Elle regroupe les observations du groupe *Asie du Sud-Est, Pacifique* de la troisième période d'analyse (les identifiants commencent à 0), autrement dit la période suivant la fin de la Guerre froide[17]. Pour chacun des attributs de cette base, la valeur de chaque observation a été construite uniquement à partir de l'année précédant celle où la classe de l'observation est estimée. Sur cette base trois méthodes obtiennent des scores voisins selon le critère *F-mesure*. Le tableau 10.6 donne les trois matrices de confusion correspondantes.

Tab. 10.6 – Matrices de confusion obtenues par les trois meilleures méthodes sur la base *1 Asie du Sud Est, Pacifique 2*

(A) Moyenne KSF	
80	5
3	12
F-mesure=0.78	

(A) 1LLSI SansFiltre	
81	4
3	12
F-mesure=0.79	

(B) 1LLSI KSF	
80	5
3	12
F-mesure=0.78	

Deux des méthodes sont strictement identiques tandis *(A) 1LLSI SansFiltre* obtient un score légèrement plus élevé du fait d'une précision légèrement supérieure. Cependant, lorsque l'on construit les matrices de confusion en supprimant les observations litigieuses, l'analyse est bien différente. Ces matrices sont données dans le tableau 10.7. C'est la méthode *(A) Moyenne KSF* qui domine les deux autres, *(A) 1LLSI SansFiltre* apparaissant comme la moins fiable des trois. Précisons que le seuil de rejet a été fixé à 10% du degré le plus élevé.

Tab. 10.7 – Matrices de confusion obtenues par les trois meilleures méthodes, après introduction d'un seuil de rejet, sur la base *1 Asie du Sud Est, Pacifique 2*

(A) Moyenne KSF	
80	2
3	11
F-mesure=0.8	

(A) 1LLSI SansFiltre	
81	4
3	9
F-mesure=0.74	

(B) 1LLSI KSF	
80	3
3	9
F-mesure=0.75	

[16] Nous devrions plutôt parler de mesure de non-fiabilité puisque le score de Brier est une mesure d'erreur.
[17] Les identifiants des périodes pour chaque groupe de pays sont disponibles à l'annexe C.

Considérons maintenant la base *15 Asie Du Sud 1*. Les matrices de confusion des trois meilleures méthodes sont identiques. Avec notre seuil de rejet, elles restent identiques. Pour différencier les méthodes nous avons alors eu recours au score de Brier. La méthode *(A) 5ppv CFS* semble la plus fiable, l'erreur quadratique moyenne commise étant de 0.086 contre 0.096 pour les deux autres. C'est donc elle qui a été retenue.

10.3.3.3 Résultats de la sélection de modèles

Nous avons procédé de la sorte pour l'ensemble de la phase de sélection de modèles. Les résultats sont présentés à l'annexe G.

> *Chacune des bases de données y est identifiée par le nom du groupe de pays auquel elle réfère. Ce nom est précédé du nombre d'années utilisées pour estimer les différents attributs : 1, 7, 15 ou 29. Enfin l'identifiant de la période d'analyse (0,1 ou 2) suit le nom du groupe de pays. Nous avons synthétisé l'ensemble de ces résultats dans le tableau 10.8. Nous y avons représenté l'ensemble des 53 chaînes de prétraitement sélectionnées. La cellule (i, j) de ce tableau contient le nombre de chaînes de prétraitement faisant intervenir la i-ième méthode de substitution associée au j-ième filtre. L'identifiant de la stratégie employée pour combiner ces deux méthodes est également indiqué dans la cellule.*

TAB. 10.8 – Synthèse des méthodes sélectionnées : nombre de fois où chaque filtre et chaque méthode de substitution des valeurs manquantes ont été sélectionnés

	SansFiltre	CFS	FCBF	KSCBF	KSF	$\sum$
AléatoireMM	1(A)		1(B)			2
CMoyenne		1(B)		1(C)		2
CmoyenneA		1(B)	1(A)			2
Moyenne		1(B)		1(A)	1(A),1(B)	4
Médiane		1(A)	2(B)	1(B),1(C)	1(A),3(B),4(C)	13
1ppv	2(A)	1(A)	3(A),2(C)	1(C)		9
5ppv	1(A)	1(A),1(B)		1(C)	3(A),3(B)	10
1LLSI	1(A)				1(A),1(C)	3
5LLSI		1(B)		1(C)		2
EF-Entropie		1(B)	2(B)	3(B)		6
$\sum$	5	9	11	10	18	53

Nous pouvons constater dans ce tableau que la méthode *EF-Entropie* qui apparaissait assez faible au vu de l'analyse de rangs intervient tout de même dans 6 des chaînes d'apprentissage sélectionnées. Les autres méthodes supervisées interviennent également quoique dans une moindre mesure (2 chaînes d'apprentissage uniquement).

Parmi les méthodes de substitution, les méthodes basées sur les plus proches voisins (*1ppv* et *5-ppv*) semblent très efficaces puisqu'elles interviennent dans 19 chaînes de prétraitement. Le remplacement par la médiane dont la complexité est moindre s'avère également très performante (13 chaînes d'apprentissage).

Quant aux méthodes de sélection d'attributs, toutes interviennent dans des proportions équivalentes à l'exception du filtre *KSF* que nous avons proposé et qui est présent dans 18 des 53 chaînes d'apprentissage, contre 11 seulement pour *FCBF* qui est la deuxième méthode la plus représentée.

Ces résultats contrastent quelque peu avec notre analyse de rangs qui n'avait pas fait ressortir de la sorte notre filtre. Cela tient au fait que les performances de KSF sont très

bonnes sur certaines bases de données et médiocres sur les autres ce qui a tendance à niveller le rang des méthodes y ayant recours. La remarque opposée peut être faite à propos de la méthode *SansFiltre* qui n'intervient que dans deux chaînes d'apprentissage alors que l'analyse de rangs avait mis en évidence le bon comportement de *(A) 1ppv SansFiltre*. Il semblerait que celle-ci ait globalement de bonnes performances sans pour autant être remarquable sur un grand nombre de bases de données. Notons que dans près de la moitié des cas la méthode de substitution associée à *KSF* est la médiane. Sur notre problème ces deux techniques semblent donc bien adaptées l'une à l'autre. Ce point avait déjà été mis en valeur lors de l'analyse de rangs.

Sur ce tableau de synthèse une dernière remarque mérite d'être faite. Les 10 méthodes de substitution, les 5 méthodes de sélection d'attributs ainsi que les 3 stratégies de combinaison sont toutes employées dans au moins une des chaînes de prétraitement sélectionnées. Ceci confirme l'intérêt de la plate-forme d'évaluation, de comparaison et de sélection de méthodes que nous avons mise en place.

Nous allons désormais faire abstraction des différentes méthodes de prétraitement et nous recentrer sur la détection des conflits armés intra-étatiques. Nous allons présenter les résultats que nous avons obtenus d'un point de vue tant quantitatif que qualitatif en ne considérant que les méthodes issues du processus de sélection que nous venons de décrire.

10.4 Analyse et interprétation des résultats

10.4.1 Analyse quantitative

10.4.1.1 Performances globales

Afin de fournir une vision globale des performances obtenues, nous avons estimé les performances moyennes, minimales et maximales ainsi que l'écart-type de quatre mesures de performance. Nous avons considéré les deux critères introduits précédemment sous les dénominations F-mesure et Règles-mesure ainsi que le rappel et la précision de la classe crise. Ce sont en effet les deux mesures à partir desquelles nos deux critères de performance sont estimés. Nous avons également ajouté la profondeur des arbres appris ce qui permet de donner une idée de la complexité des modèles de prédiction sélectionnés. Bien que nous ayons critiqué à la section 5 l'emploi de la moyenne et de l'écart-type pour synthétiser des performances évaluées sur des bases de données distinctes, nous avons décidé d'y avoir recours ici car toutes les bases de données en question traitent d'un même problème. Il y a donc une certaine homogénéité entre les différentes bases. Les résulats sont donnés dans le tableau 10.9.

TAB. 10.9 – Synthèse des performances obtenues par les méthodes sélectionnées sur l'ensemble des bases de données

	moyenne	écart-type	min	max
rappel	0.61	0.18	0.27	1
précision	0.73	0.18	0.4	1
F-mesure	0.62	0.16	0.3	0.93
Règles-mesure	0.56	0.24	0.13	0.9
Profondeur	2	1.3	1	6

Les performances sont moins élevées que celles que nous avions obtenues lors de nos premières expérimentations, mais rappelons qu'alors nous ne disposions que d'une seule base de données de taille modeste et qu'il était donc difficile d'estimer finement la confiance à accorder à nos résultats. Ajoutons également qu'identifier un déclenchement de crise est bien plus délicat qu'identifier l'occurrence d'une crise, tâche qui était la nôtre lors de ces premières expérimentations. L'occurrence d'une crise concerne aussi bien le déclenchement que la poursuite de la crise et rassemble donc plus d'observations.

Si l'on s'en réfère à la valeur prédictive des principaux modèles de la littérature sur les conflits armés intra-étatiques (Ward et Bakke, 2005), ces performances sont tout de même plus que satisfaisantes. En effet, d'après cette étude, seuls les modèles de la *State Failure Task Force* ont des performances intéressantes en prédiction avec un taux de rappel de la classe *crise* situé entre 60 et 70% avec un taux de bonnes classifications compris entre 70 et 80% (Goldstone *et al.*, 2000). Pour comparaison nous obtenons un taux de rappel de 61%, mais notre taux de bonnes classifications, non mentionné dans le tableau 10.9, est de 91%. Nos performances sont donc plus qu'encourageantes. Étant donné l'importance que nous accordons au rappel de la classe *crise*, notre marge de progression est encore significative comme le souligne l'existence d'un taux de rappel minimum de 27%.

Ce taux est obtenu sur la base de données *15 Afrique Subsaharienne 0* qui regroupe les observations antérieures à la fin de la Guerre froide des pays d'Afrique subsaharienne pour lesquels une période d'estimation de 15 ans a été considérée. Dans les années 1970, époque durant laquelle sont estimés les attributs de cette base de données, de nombreux pays de cette zone viennent d'accéder à l'indépendance et nombre d'entre eux ont connu au moins une guerre civile durant cette époque. La qualité des données les concernant est donc plus

que douteuse ce qui peut expliquer ces mauvais résultats. Ajoutons que le groupe des pays d'Afrique subsaharienne est celui qui contient le plus de pays. Aussi peut-on se demander si le découpage effectué en zones géographiques a été assez fin, l'histoire, la géographie, la culture des pays d'Afrique de l'Ouest étant par exemple assez différentes de celles des pays d'Afrique de l'Est. Peut-être un découpage en fonction des anciennes puissances coloniales eût-il été mieux approprié.

La profondeur des arbres construit est très faible. Ils contiennent en moyenne deux attributs seulement. Pour poursuivre notre comparaison avec la *State Failure Task Force*, précisons que les modèles de régression qui ont été développés par ce groupe de chercheurs ne comportent qu'entre trois et cinq variables explicatives, soit le même ordre de grandeur que les nôtres (entre 1 et 6 variables).

Cette vision synthétique des performances permet d'illustrer l'intérêt de notre approche de manière globale. Elle combine cependant les résultats de modèles spécifiques qu'il serait utile de désagréger afin de pouvoir analyser plus finement le comportement de ces modèles et de confirmer ou infirmer certaines hypothèses que nous avons avancées pour justifier la construction de nos bases de données à la section 10.2.

10.4.1.2 Importance de la période d'estimation

Intéressons-nous d'abord à la période d'estimation. Pour un pays et une période de prédiction donnés, nous avons construit quatre observations distinctes selon le nombre d'années qui ont été utilisées pour estimer les différents attributs. Ce nombre d'années correspond à la durée de la période d'estimation.

Rappelons que nous avons considéré des périodes de 1, 7, 15 et 29 ans. L'objectif implicite d'une telle démarche réside dans l'identification de la durée la mieux à même de synthétiser l'information nécessaire et suffisante pour réaliser la tâche de prédiction. Estimer les attributs sur des périodes plus courtes a pour conséquence de négliger la portée de certains phénomènes historiques. À l'inverse, estimer les attributs sur des périodes plus longues revient à accorder trop d'importance à l'histoire. Au-delà d'un certain laps de temps, les liens entre le contexte structurel et le déclenchement des crises sont trop diffus pour pouvoir être identifiés automatiquement à partir de nos données.

Pour tester cette hypothèse nous avons comparé les matrices de confusion obtenues par les modèles construits à partir des différentes périodes d'estimation. Afin que ces matrices de confusion soient parfaitement homogènes, nous n'avons considéré que les observations communes aux 4 périodes d'estimation. La période d'estimation de 29 ans donnant lieu à trop peu d'observations, nous l'avons exclue de la comparaison. Les résultats sont donnés à la figure 10.5. Afin de tenir compte de la fiabilité des modèles appris avec les trois périodes d'estimation distinctes nous présentons également les résultats obtenus en introduisant un seuil de rejet, fixé comme précédemment à 10% du degré de reconnaissance le plus élevé.

> *Nous avons ajouté une colonne à toutes les matrices de confusion pour regrouper les observations de chaque classe pour lesquelles l'incertitude est trop élevée pour qu'une classe leur soit affectée. Lorsque cela se produit avec un seuil de rejet nul cela signifie que les degrés de reconnaissance des deux classes sont identiques.*

Au vu des résultats de la figure 10.5 la période d'estimation de 7 ans semble préférable aux deux autres. Lorsqu'aucun seuil de rejet n'est fixé, les performances du point de vue de la *F-mesure* sont voisines mais l'écart est plus net une fois introduit un seuil de rejet, surtout avec la période d'estimation de 15 ans.

Si l'on considère notre second critère on constate également que la période d'estimation de 7 ans est celle qui permet d'obtenir les meilleures performances, mais ceci n'est vrai

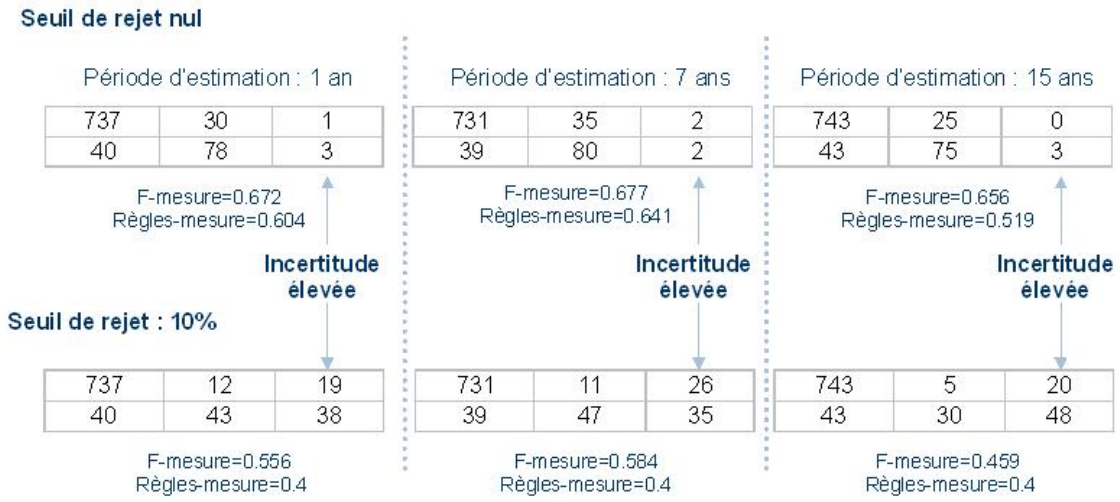

FIG. 10.5 – Comparaison des matrices de confusion obtenues à partir du classement d'observations ne différant que par la période d'estimation utilisée

qu'avec un seuil de rejet nul. Dans le cas contraire, les trois matrices de confusion sont jugées équivalentes. Ceci met en évidence les différences de comportement des deux critères de performance. La méthode à base de règles considère comme non significatives des différences jugées relativement importantes par la *F-mesure* dès lors qu'elles ne concernent que des niveaux moyens de rappel et précision. Au contraire, elle renforce des différences même minimes lorsque rappel et précision sont suffisamment élevés.

Les résultats de la figure 10.5 confortent donc notre hypothèse selon laquelle il existe une durée optimale pour la période d'estimation, du moins ils ne permettent pas d'infirmer cette hypothèse. Ils permettent également de prendre la mesure de l'impact de l'introduction du seuil de rejet. Seules les observations affectées à la classe *crise* sont concernées, ce qui signifie que les décisions prises à propos de la classe *crise* sont les moins fiables. Ceci n'est guère étonnant puisque cette classe est la plus difficile à modéliser.

10.4.1.3 Importance des groupes de pays

L'hypothèse relative à l'importance de la période d'estimation n'avait à notre connaissance jamais été testée ni même formulée dans la littérature. Nous avons également développé une nouvelle hypothèse concernant la spécialisation des modèles à des groupes de pays homogènes, l'intuition sous-jacente étant que les facteurs crisogènes ne sont pas les mêmes dans tous les pays. Idéalement il faudrait presque construire un modèle par pays, voire descendre au niveau des régions d'un même pays, mais nous ne disposerions alors que de très peu d'observations pour apprendre de tels modèles. L'objectif est donc de regrouper un certain nombre de pays dont les liens géographiques, historiques, culturels, économiques... soient suffisamment forts pour que l'on puisse considérer que les contextes structurels marqueurs d'instabilité seront voisines d'un pays à l'autre.

Nous avons procédé à la comparaison, par période d'estimation, des matrices de confusion obtenues à partir d'un modèle global d'une part, et à partir des modèles de groupes de pays d'autre part. À l'instar de ce qui a été fait précédemment pour homogénéiser les matrices comparées, nous n'avons considéré que les observations classées par le modèle global qui ont également été classées par le modèle de l'un des groupes de pays. Les résultats obtenus avec chacune des quatre périodes d'estimation sont donnés à la figure 10.6.

Période d'estimation : 1 an

Seuil de rejet nul

770	33	1
44	83	3

F-mesure=0.665
Règles-mesure=0.579

783	20	1
51	79	0

F-mesure=0.638
Règles-mesure=0.426

Seuil de rejet : 10%

Groupes de pays

770	13	21
44	43	43

F-mesure=0.532
Règles-mesure=0.4

Modèle global

783	0	21
51	0	79

F-mesure=0
Règles-mesure=0.1

Période d'estimation : 7 ans

Seuil de rejet nul

737	35	2
39	78	2

F-mesure=0.671
Règles-mesure=0.622

735	39	0
49	70	0

F-mesure=0.598
Règles-mesure=0.4

Seuil de rejet : 10%

Groupes de pays

737	11	26
39	48	32

F-mesure=0.59
Règles-mesure=0.4

Modèle global

735	0	39
49	0	70

F-mesure=0
Règles-mesure=0.1

Période d'estimation : 15 ans

Seuil de rejet nul

705	27	0
41	71	3

F-mesure=0.65
Règles-mesure=0.513

709	23	0
49	66	0

F-mesure=0.601
Règles-mesure=0.4

Seuil de rejet : 10%

Groupes de pays

705	7	20
41	30	44

F-mesure=0.467
Règles-mesure=0.4

Modèle global

709	0	23
49	0	66

F-mesure=0
Règles-mesure=0.1

Période d'estimation : 29 ans

Seuil de rejet nul

59	7	0
6	16	2

F-mesure=0.721
Règles-mesure=0.824

54	12	0
7	17	0

F-mesure=0.68
Règles-mesure=0.761

Seuil de rejet : 10%

Groupes de pays

59	3	4
6	13	5

F-mesure=0.707
Règles-mesure=0.681

Modèle global

54	0	12
7	0	17

F-mesure=0
Règles-mesure=0.1

FIG. 10.6 – Comparaison des matrices de confusion obtenues par un modèle global et un ensemble de modèles spécifiques à des groupes de pays

Quelle que soit la période d'estimation considérée nous constatons que les résultats confortent tous notre hypothèse selon laquelle il est préférable de différencier les modèles selon les groupes de pays. Nous observons en effet que les performances du modèle global, quel que soit le critère considéré, sont systématiquement inférieures à celles que l'on obtient à partir de modèles spécialisés.

L'introduction du seuil de rejet a des effets encore plus nets que ce que nous avons vu jusqu'à présent. En effet le modèle global ne classe correctement aucun pays de la classe *crise*, ce qui dénote un manque de fiabilité de ce modèle. Obtenir de meilleures performances en prédiction à partir de modèles spécialisés peut paraître intuitif. Ces modèles sont à même d'identifier des liens plus fins entre des contextes structurels régionaux et le déclenchement des crises. La plus grande fiabilité de ces modèles est cependant nettement moins intuitive. Ils sont en effet appris sur des bases de données contenant beaucoup moins d'exemples que celle à partir de laquelle le modèle global est construit. Ce point constitue selon nous la conclusion la plus importante à retenir de l'étude comparative de la figure 10.6.

Cette étude vient également appuyer certaines autres remarques que nous avions énoncées lors de la comparaison des périodes d'estimation. D'une part, les décisions les plus incertaines se rapportent toutes à des observations affectées à la classe *crise* par notre système, ce qui confirme la difficulté de modélisation de cette classe. D'autre part, les performances obtenues à partir d'une période d'estimation de 7 ans sont supérieures à celles qui correspondent aux périodes de 1 et 15 ans. Ceci se vérifie aussi bien avec le critère *F-mesure* qu'avec *Règles-mesure*. Notons qu'en prenant une période d'estimation de 29 ans, les résultats sont encore meilleurs. Ceci suggère que la durée d'estimation de 7 ans n'est peut-être pas optimale et suffisamment longue. Nous tenons cependant à modérer ce constat à cause du faible nombre d'observations à partir desquelles il se fonde. Il faudrait procéder à des tests sur des bases de données plus larges pour le confirmer.

10.4.1.4 Influence de la Guerre froide

La dernière hypothèse que nous avons voulu tester concerne l'influence de la fin de la Guerre froide sur l'émergence des conflits armés intra-étatiques. Nous avons vu à la section 10.2 que cette hypothèse fait l'objet d'une controverse dans le domaine des sciences politiques, différentes études quantitatives ayant abouti à des conclusions contradictoires. Nous présentons dans le tableau 10.10 les matrices de confusion correspondant au classement des exemples avant et après la fin de la Guerre froide.

TAB. 10.10 – Comparaison des performances obtenues par des modèles appris sur des données concernant la période de la Guerre froide, notée *pre* avec des modèles appris sur des données concernant l'après-Guerre froide, notée *post*

		rappel	précision	F-mesure	Règles-mesure	profondeur
pre	moyenne	0.61	0.72	0.62	0.56	1.79
	écart-type	0.22	0.18	0.2	0.27	1.28
	min	0.27	0.41	0.3	0.13	1
	max	1	1	0.93	0.9	6
post	moyenne	0.61	0.75	0.63	0.55	2.24
	écart-type	0.12	0.17	0.11	0.21	1.34
	min	0.42	0.4	0.42	0.4	1
	max	0.87	1	0.87	0.9	5

Les performances moyennes sur les deux époques considérées sont sensiblement identiques. On constate cependant qu'elles sont plus stables après la fin de la Guerre froide. Les écarts-type des quatre mesures sont tous inférieurs durant cette période. La plus grande variabilité des prédictions effectuées pour la période de la Guerre froide se traduit par des valeurs extrêmes plus importantes. La profondeur moyenne des arbres construits est également quelque peu différente entre les deux périodes, les modèles de l'après-Guerre froide étant plus complexes. Le comportement de nos modèles relativement à ces deux périodes n'est donc pas tout à fait similaire.

Pour estimer l'influence de la fin de la Guerre froide d'un point de vue quantitatif, il eût été nécessaire de construire un modèle à partir des observations relatives aux deux époques. Ceci nous aurait permis de mener une étude semblable à celle que nous avons conduite pour évaluer l'impact des groupes de pays. N'ayant pas réalisé ces expériences, il est difficile en l'état de conclure quant à l'intérêt de l'introduction des périodes d'analyse et ce d'autant plus que les différences observées sur le tableau 10.10 sont assez minces. Nous essaierons à la section suivante, dévolue à l'analyse qualitative des résultats expérimentaux, d'apporter de nouvelles informations nous permettant de juger de la validité de cette dernière hypothèse.

10.4.2 Analyse qualitative

L'évaluation quantitative des performances de notre système en prédiction assure un contrôle objectif de son efficacité. Obtenir des performances satisfaisantes est une condition nécessaire pour que le système soit accepté. Cette condition n'est cependant absolument pas suffisante. Nous avons longuement insisté à la partie I sur l'importance de la transparence du système afin que l'expert puisse conserver un regard critique sur chaque inférence effectuée par le système. Aussi allons-nous dans cette section adopter le point de vue de l'expert en charge de l'analyse et de l'interprétation des résultats.

Voyons dans un premier temps quels sont les principaux facteurs de risque mis en

évidence lors de nos expérimentations. Ceci nous permettra de confronter les interprétations qu'autorise notre système avec les théories les plus répandues sur l'émergence des conflits armés intra-étatiques, abordées à la section 10.2. Parmi les différents attributs d'une base de données, seront considérés comme des facteurs de risque ceux qui n'ont été supprimés ni par le filtrage d'attributs ni par la sélection effectuée par *Salammbô*. Ce sont ceux que l'on retrouve dans les arbres de décision qui constituent nos modèles de détection des crises. Ces arbres de décision contiennent non pas l'ensemble des attributs les plus pertinents vis-à-vis de la classe mais seulement un sous-ensemble qui s'avère être de taille suffisante pour assurer de bonnes performances. Il est fort possible que des performances voisines eussent été obtenues à partir d'un sous-ensemble d'attributs légèrement différent.

Afin d'offrir une vue plus complète des facteurs de risque extraits par notre système nous avons décidé de considérer pour chaque base de données l'ensemble des attributs identifiés comme tels, non pas par le seul modèle issu du processus de sélection, mais par les trois meilleurs. Rappelons que l'ensemble des indicateurs structurels que nous avons envisagés sont décrits à l'annexe E. Les sources à partir desquelles ces indicateurs ont été recueillis sont quant à elles précisées à l'annexe F.

10.4.2.1 Facteurs de risque communs à tous les modèles

Nous avons indiqué précédemment que l'introduction des groupes de pays avait permis de construire des modèles spécifiques plus performants dans l'ensemble qu'un modèle global. Affinons désormais notre analyse en précisant les caractéristiques propres aux modèles de chaque groupe de pays ainsi que leurs caractéristiques communes.

Quatre indicateurs sont majoritairement reconnus comme des facteurs de risque et ce quel que soit le groupe de pays considéré :

- *nombre de morts liés aux combats passés*,
- *nombre d'années écoulées depuis la dernière guerre civile*,
- *nombre d'années écoulées depuis l'accès à l'indépendance*,
- *nombre d'années durant lesquelles le pays a été privé de son autonomie depuis son premier accès à l'indépendance.*

Les effets observés sont ceux auxquels nous nous attendions, aggravant pour les 1er et 4e et stabilisant pour les 2e et 3e.

Un nombre élevé de morts liés aux conflits passés accroît le risque de déclenchement d'un nouveau conflit. Ce risque est par ailleurs plus élevé pour les pays qui sont en conflit ou qui viennent juste d'en sortir. On retrouve ainsi la notion de *piège de la guerre civile* évoquée par Collier et Hoefler. Il est cependant difficile en l'état de trancher entre les diverses interprétations possibles de ce phénomène. D'une part, les conflits attisent les antagonismes, renforçant les motivations d'éventuels rebelles. D'autre part, l'émergence de nouveaux conflits est rendue plus facile lorsqu'un conflit est déjà en cours ou qu'il vient juste de se terminer : armes et personnel entraîné sont disponibles, l'attention du gouvernement se porte sur le conflit en cours ou sur la reconstruction du pays, ce qui l'affaiblit.

Le 3e facteur traduit la fragilité des États récemment constitués. Ce point avait été mis en évidence par Laitin et Fearon. Enfin plus un pays a été privé de son autonomie et plus le ressentiment de la population est grand. Il est en effet vraisemblable qu'une frange de la population soit jugée responsable, à tort ou à raison, de la perte d'autonomie subie par le pays. Lorsque le pays recouvre son indépendance, les tensions peuvent alors être très violentes comme ce fut le cas par exemple en France après la Libération en 1944.

Abordons à présent les spécificités de chacun des groupes de pays. Nous préciserons pour chacun des facteurs de risque s'il a un effet aggravant (des valeurs élevées de ce facteur

sont associées à un risque de crise plus important) ou stabilisant (des valeurs élevées de ce facteur sont associées à un risque de crise moindre) sur le déclenchement des conflits. Ceci suppose que le lien entre le facteur de risque et la classe est linéaire. Lorsque tel n'est pas le cas, nous l'indiquerons clairement.

10.4.2.2 Afrique du Nord, Proche-Orient

Parmi les indicateurs ayant un effet stabilisant, nombre d'entre eux sont le reflet de la puissance ou de la légitimité de l'État selon l'interprétation que l'on souhaite apporter. Plus l'État est fort et mieux il sera armé pour contenir une insurrection, ce qui dissuade d'éventuels rebelles de passer à l'acte. Les coûts de la rébellion sont trop élevés. Plus la légitimité de l'État est grande et plus il sera difficile de convaincre les citoyens du bien-fondé de l'insurrection. Dans les deux cas c'est le recrutement des rebelles qui est affecté. Nous avons ainsi la *moyenne du taux d'alphabétisation des 15-24 ans*, la *moyenne du taux de vaccination contre la rougeole* ou encore la *moyenne de l'indice des prix de la nourriture.*

La *part de la superficie sur laquelle la densité est inférieure à 2 hab/km*^{2} a également un effet stabilisant. Les pays dont l'étendue des zones de faible densité est réduite ont un risque de crise plus élevé. Plus l'étendue des zones peu denses est faible et plus la part du territoire occupée par des zones de densité moyenne ou élevée est importante. Or ce sont des zones dans lesquelles la compétition pour l'accès aux ressources est plus forte.

Parmi les facteurs ayant un effet aggravant, il est intéressant de constater qu'intervient la *part de la population vivant dans des zones de forte densité (entre 1000 et 10000 hab/km*^{2}*)*, ce qui semble corroborer notre précédente remarque relative à l'influence de la densité de population.

Notons enfin que le *rapport entre les tailles des deux principaux groupes ethniques* est également tel que le risque de conflit est plus élevé lorsque ce rapport est proche de 1. On retrouve ici l'idée selon laquelle les sociétés polarisées, comportant quelques groupes de taille comparable, ont un risque plus élevé que les sociétés très homogènes ou au contraire très hétérogènes.

La *part de la superficie occupée par les régions d'altitude élevée (entre 1500 et 3000 mètres)* joue également un rôle important. Ainsi quatre pays de cette région ont principalement été en crise avant la fin de la Guerre froide : l'Iran, le Liban, le Maroc et la Turquie qui ont tous une part non négligeable de leur territoire occupée par les montagnes. Avec une mesure nettement moins fine que celle employée par Collier et Hoeffler ou encore Laitin et Fearon, nous parvenons donc également à mettre en évidence l'impact du relief sur les guerres civiles. Comme eux nous constatons qu'un relief accidenté favorise l'émergence d'une insurrection.

Mentionnons enfin la *magnitude des tremblements de terre.* L'occurrence de telles catastrophes naturelles fragilise l'État, ou du moins sa capacité à contenir une rébellion. Les infrastructures sont endommagées et ses efforts se portent sur l'aide aux victimes. De tels événements[18] facilitent donc le passage à l'offensive d'armées rebelles. De plus la légitimité du gouvernement peut être entamée s'il ne réagit pas assez vite.

10.4.2.3 Afrique subsaharienne

La légitimité de l'État semble également être importante pour expliquer le déclenchement des conflits pour les pays d'Afrique subsaharienne. La *moyenne de la formation brute de capital fixe (% du PIB)* a ainsi un effet stabilisant. Ce facteur regroupe l'ensemble des

[18] La généralisation à d'autres types de catastrophes naturelles est assez intuitive, même si par manque de données adéquates nous n'avons pas pu tester cette hypothèse avec d'autres types de catastrophes.

dépenses de l'État engagées pour améliorer la qualité des infrastructures (routes, voies ferrées, hôpitaux, écoles...). Les États dans lesquels le gouvernement intervient peu dans ces domaines sont plus fragiles. Ce manque d'investissements dans des domaines d'utilité publique peut susciter le mécontentement de la population et faciliter le discours d'opposition des rebelles. Il est également possible de considérer que la capacité du gouvernement à contenir une insurrection est moindre lorsque les infrastructures de communication sont moins développées. Il lui est en effet plus difficile de couvrir efficacement l'ensemble du territoire. L'effet stabilisant de la *part des routes goudronnées* conforte cette seconde interprétation.

En évoquant les caractéristiques communes à l'ensemble des groupes de pays nous avons insisté sur l'influence de l'historique des conflits sur le déclenchement de nouvelles crises. Pour les pays d'Afrique subsaharienne cette influence est particulièrement prégnante et se manifeste par certaines conséquences des conflits passés : les flux de réfugiés. Ainsi la *moyenne du nombre de réfugiés originaires du pays* ainsi que la *tendance du nombre de personnes déplacées à l'intérieur du pays* ont un effet aggravant. D'une part, ces deux facteurs contribuent à renforcer les motivations d'insurgés potentiels. D'autre part, ils traduisent un certain désordre dans le pays qui peut être exploité par des rebelles. Pour appuyer cette seconde interprétation précisons que la *moyenne du nombre de réfugiés accueillis par le pays* est également un facteur important ayant un effet aggravant. Or ces flux de réfugiés ne peuvent pas être le fruit de conflits passés dans le pays. Ce sont vraisemblablement les pays voisins qui sont concernés. L'instabilité régionale peut donc également être invoquée pour justifier l'impact de ce facteur.

La pression démographique semble également jouer un rôle en Afrique subsaharienne puisque la *moyenne de la part des moins de 14 ans dans la population* a un effet aggravant. Si la compétition accrue pour l'accès aux ressources est sûrement une réalité, on ne peut également occulter le rôle joué par les enfants soldats dans nombre de pays de cette région. Les moins de 14 ans constituent un vivier de recrues potentielles pour les armées rebelles.

Mentionnons enfin l'influence des échanges commerciaux pour ces pays. Ainsi la *moyenne des échanges (importations et exportations de biens et services en % du PIB)* et plus particulièrement la *moyenne des échanges (importations et exportations) de marchandises (% du PIB)* ont tous deux un effet stabilisant sur le déclenchement des crises. On retrouve ici des résultats mis en évidence par la *State Failure Task Force* à propos de l'ouverture du marché comme facteur de stabilité.

10.4.2.4 Amérique latine, Caraïbes

On retrouve également pour les pays d'Amérique latine l'idée selon laquelle la fragilité du gouvernement accentue les risques de crise. On observe ainsi que la *tendance des revenus du gouvernement (% PIB)* a un effet stabilisant. Plus les revenus du gouvernement ont tendance à décroître et plus le risque croît. L'effet aggravant du *taux de mortalité infantile* que nous observons peut être interprété de manière similaire. Notons que cette variable est l'une des plus importantes dans les modèles développés par la *State Failure Task Force*.

Autre constatation importante, le *nombre de pays voisins en guerre civile* a un effet aggravant, ce qui dénote l'importance de l'instabilité régionale que nous avons évoquée à propos des pays d'Afrique subsaharienne. La démographie influe également. Le *taux de dépendance*, *rapport entre le nombre d'habitants âgés de moins de 15 ans ou de plus de 64 ans et le nombre des 15-64 ans*, a en effet un impact aggravant. Plus le nombre d'habitants âgés de 15 à 64 ans est important et plus le risque de crise est élevé. Ceci suggère que ce n'est pas tant la pression démographique et la compétition pour l'accès aux ressources qui importent, mais plutôt l'existence d'un vivier de combattants potentiels. Les individus

aptes au combat sont en effet ceux qui ont plus de 15 ans et moins de 64 ans, exception faite des enfants soldats. Il s'agit là d'une interprétation directement liée à la théorie de Collier et Hoeffler mettant en avant les facteurs facilitant l'émergence d'une insurrection.

Notons enfin l'effet aggravant de la *moyenne de la part des femmes dans l'industrie*. Selon nous ceci s'explique par le fait que ce facteur reflète l'historique des conflits, dont l'importance a été mentionnée à plusieurs reprises. Conséquence des conflits, la diminution du nombre d'hommmes aptes au travail (morts au combat et invalides) se répercute sur le marché du travail. Ce phénomène a ainsi été observé après la première Guerre mondiale en Europe.

10.4.2.5 Asie du Sud

La *moyenne du nombre de pannes téléphoniques* a un effet aggravant pour les pays d'Asie du Sud. Ces pannes caractérisent une certaine faiblesse de l'État, qui investit insuffisamment ou inefficacement dans les infrastructures de télécommunications pour remédier à ces problèmes, à moins que ces dysfonctionnements ne soient que le reflet des conflits passés qui ont endommagé ces infrastructures. L'importance de la fragilité de l'État est également marquée par la *variation annuelle moyenne des dépenses générales du gouvernement (dollars constants, 1995)*. Le risque est plus élevé lorsque cette variation est faible. Le niveau moyen des dépenses étant peu élevé, de faibles variations dans les dépenses du gouvernement traduisent son incapacité à faire évoluer la situation.

Ceci influe d'une part sur sa capacité de répression puisque peu de dépenses sont engagées dans ce sens. D'autre part, les perspectives réduites d'une amélioration de la situation peuvent contribuer à dégrader le climat social. On retrouve donc les deux interprétations dominantes de l'origine des conflits : les griefs de la population à l'égard du gouvernement et les occasions facilitant le déclenchement d'une insurrection sont les facteurs déterminants.

La *tendance des avoirs nets étrangers* et la *moyenne des prêts publics consentis au secteur privé (% PIB)* ont tous deux un effet aggravant. Plus ils sont élevés et plus le risque de conflits est grand. Les fortes valeurs de ces variables sont associées à des pays en plein développement économique. Il est également possible de considérer qu'un fort et brusque développement économique s'accompagne d'une certaine instabilité qui favorise l'action des rebelles.

10.4.2.6 Asie du Sud-Est, Pacifique

Les pays d'Asie du Sud-Est sont essentiellement marqués par l'impact des conflits passés. Outre les facteurs liés à l'historique des conflits, la *tendance des subventions du gouvernement (% des dépenses totales)* se distingue par son effet stabilisant. Lorsque la part des dépenses dévolues aux subventions a tendance à baisser le risque de crise s'accroît. Cela peut s'expliquer par le fait qu'une diminution des subventions affecte le climat social du pays.

Autre point intéressant, l'indice de la diversité religieuse exerce un effet aggravant. Sont principalement concernés l'Indonésie, la Thaïlande, les Philippines et le Myanmar. L'influence de cette variable fait l'objet de débats controversés dans la littérature ainsi que nous avons pu le voir à la section 10.2. Pour notre part, nous n'observons pas l'existence d'un lien quadratique (en U inversé) entre cet indice et le risque de conflit. Les règles faisant intervenir la diversité religieuse ne font apparaître qu'un seul seuil situé autour de 0.6 au-delà duquel le risque de conflit est bien plus élevé.

10.4.2.7 Europe de l'Est, Asie centrale

L'importance des facteurs géographiques concernant les anciennes républiques de l' Union soviétique est assez marquée. La *part de la superficie sur laquelle la densité est comprise entre 5 et 10 hab/km*^{2} ainsi que la *différence entre l'altitude la plus élevée et la plus basse* ont un effet aggravant sur le déclenchement des conflits. Contrairement à ce que nous avons observé à propos des pays d'Afrique du Nord et du Proche-Orient, des zones peu denses plus étendues sont associées à un risque plus élevé de crise. Il est difficile de couvrir ces zones pour le gouvernement, et ce, d'autant plus que le terrain y est accidenté, ce que nous avons essayé de traduire par la différence entre altitude la plus élevée et la plus basse.

Corroborant également la théorie de la faisabilité de la guerre civile, nous constatons que la *moyenne de la proportion de la population âgée de 15 à 64 ans* a un effet aggravant. Les combattants potentiels sont en effet issus de cette frange de la population.

La *moyenne du flux des investissements étrangers directs (%PIB)* exerce un effet stabilisant. Les pays pour lesquels ces investissements ne représentent qu'une faible part du PIB ont un risque de crise plus élevé. Cet effet peut s'expliquer par le fait que des entreprises étrangères refusent d'investir dans des pays trop instables, le retour sur investissement risquant d'être nul. Il est donc possible que l'on soit face à un problème d'endogénéité, due à une causalité inverse. Rappelons que nous ne mettons en évidence que des liens entre variables. Il incombe à l'utilisateur d'interpréter correctement les résultats de l'apprentissage. C'est la raison pour laquelle nous avons tant insisté sur la nécessité de fournir des résultats aussi clairs que possible pour que la phase d'interprétation soit facilitée.

Un dernier facteur de crise mérite d'être mentionné à propos de ce groupe de pays : le *nombre de groupes ethniques*. Il exerce un effet aggravant dans les populations majoritairement composées d'individus âgés de 15 à 64 ans. La Russie, la Géorgie, la Macédoine, la Serbie sont des exemples de pays ayant été en crise et ayant de telles caractéristiques.

10.4.2.8 Pays occidentaux

Ce groupe de pays est nettement plus problématique que les précédents. Si l'on considère les performances moyennes sur l'ensemble des pays de ce groupe, une fois qu'un seuil de rejet a été introduit, aucun pays de la classe *crise* n'est reconnu correctement. Les décisions concernant ces pays sont hautement incertaines.

Deux explications peuvent être apportées. Le nombre de pays en crise au sein de ce groupe est très faible. Il s'agit essentiellement du Royaume-Uni à cause du conflit irlandais, de l'Espagne en proie au terrorisme basque, et enfin de l'Afrique du Sud que nous estimions pouvoir classer dans ce groupe du fait de sa relative prospérité économique par rapport aux autres pays d'Afrique subsaharienne. L'introduction de ce pays dont les caractéristiques économiques et démographiques sont assez différentes de celles des autres pays est peut-être à l'origine des mauvaises performances concernant ce groupe. On dénombre huit déclenchements de crise pour ce seul pays entre 1970 et 2002, dont 5 correspondent non pas à un conflit armé intra-étatique mais à un usage unilatéral de la violence. Étant donné le faible nombre de déclenchements de crise dans les autres pays du groupe, l'introduction de l'Afrique du Sud a considérablement modifié la distribution de la variable *classe*.

L'étude des facteurs de risque pour ce groupe vient confirmer ces suppositions. Nous constatons en effet que la *moyenne de la consommation des ménages par habitant (dollars constants, 1995)* ainsi que la *moyenne du nombre d'abonnés (téléphonie fixe et mobile)* sont parmi les principaux facteurs de risque, exception faite de ceux qui reflètent l'histoire des conflits. Or ces facteurs ne permettent que de discriminer l'Afrique du Sud des autres pays du groupe. Cela suggère que ce pays est une exception au sein du groupe.

Se fonder sur notre intuition pour biaiser l'analyse automatique n'a donc pas donné des résultats très heureux. En opérant ainsi nous avons ouvertement transgressé l'un des principes que nous avions décidé de nous imposer, à savoir l'objectivité. Cette erreur nous permet cependant de souligner l'importance de cette contrainte qui fait partie intégrante de notre méthodologie générique d'évaluation des risques. Sans un maximum d'objectivité il est assez aisé d'extraire des régularités complètement artificielles et de biaiser ainsi l'interprétation des résultats.

10.4.2.9 Influence de la fin de la Guerre froide

Après avoir étudié l'influence sur les conflits de la dimension spatiale nous pouvons poursuivre l'analyse qualitative afin de prendre la mesure de l'impact de la dimension temporelle. Nous avons en effet construit nos bases de données de telle sorte qu'un modèle spécifique soit appris non seulement pour chaque groupe de pays mais également pour chaque période d'analyse. Ces périodes ont été constituées de telle manière que puisse être testée l'hypothèse selon laquelle la fin de la Guerre froide a eu des répercussions non négligeables sur les conflits armés intra-étatiques et en particulier sur leur genèse.

Nous avons vu lors de l'analyse quantitative que la profondeur moyenne des arbres correspondant aux modèles des conflits de la Guerre froide était sensiblement inférieure à celle des arbres modélisant les conflits de l'après-Guerre froide. Ne serait-ce que par leur structure les modèles appris ne sont donc pas identiques selon la période considérée. D'un point de vue qualitatif, si les quatre facteurs de risque synthétisant l'historique des conflits sont communs aux modèles des deux époques, de nombreux autres facteurs diffèrent. Ces différences suggèrent que l'hypothèse relative à la fin de la Guerre froide est loin d'être dénuée de fondements.

Facteurs influents durant la Guerre froide Parmi les facteurs de risque identifiés pour la période 1970-1990 et dont l'influence semble avoir disparu ou du moins diminué, citons la *moyenne de la proportion de terres arables* et la *moyenne du montant de l'aide de l'Association Internationale de Développement (AID : institution dépendante de la Banque mondiale)*. On observe un effet aggravant du second facteur. Ceci s'explique par le fait que les institutions telles que la Banque mondiale accordent généralement des prêts à des pays en difficulté. Si nous poursuivons notre analyse nous sommes amené à conclure que pendant la Guerre froide l'AID parvenait à cibler correctement les pays en difficulté. En revanche, ces prêts ne semblent pas à avoir été efficaces dans la mesure où des conflits ont tout de même éclaté dans les pays destinataires de l'aide. Que ce facteur n'intervienne que pour la période de la Guerre froide laisse supposer que l'échec des aides durant cette période peut être imputable à la lutte d'influence à laquelle se sont livrés les deux blocs. Sans pour autant sombrer dans le cynisme, précisons que d'un point de vue purement logique la perte d'influence de ce facteur après la fin de la Guerre froide peut également signifier que l'AID ne parvient plus à identifier correctement les pays en difficulté.

Le rôle joué par la *moyenne de la proportion de terres arables* est plus ambigu. En effet, de faibles valeurs en Afrique subsaharienne sont associées à un risque plus élevé de crise tandis que l'effet est inverse pour les pays d'Afrique du Nord et du Proche-Orient. Pour les pays d'Afrique subsaharienne l'effet observé traduit un problème d'accès aux ressources. Lorsque les terres arables ne représentent qu'une petite partie de la superficie, les ressources sont plus rares ce qui peut générer des tensions.

Pour les pays d'Afrique du Nord et du Proche-Orient, l'interprétation à donner à ce facteur est bien différente. Selon nous une proportion plus importante de terres arables dans cette région, dans les années 70-80, est caractéristique de pays en plein développement

agricole et économique (Liban, Maroc, Turquie par exemple). Or au début de la phase de développement d'un pays, les ressources disponibles sont plus nombreuses et plus facilement exploitables par des rebelles, le gouvernement n'ayant pas encore eu le temps de développer et structurer ses forces coercitives pour accompagner la rapide évolution de la société.

Facteurs influents après la fin de la Guerre froide Parmi les facteurs de risque nouvellement apparus dans la période suivant la fin de la Guerre froide, il nous semble important de mentionner l'*indice de diversité religieuse* ainsi que le *rapport entre la longueur des frontières fluviales et la longueur totale des frontières terrestres*. Tous deux exercent un effet aggravant sur le déclenchement des conflits. L'influence de la diversité religieuse sur les conflits depuis la fin de la Guerre froide est un lieu commun journalistique depuis que Huntington a énoncé sa théorie du choc des civilisations. Les études quantitatives à ce sujet sont contradictoires. Si les résultats de nos expérimentations semblent confirmer les thèses de Huntington, il nous faut cependant préciser que l'influence de la diversité religieuse ne concerne pas l'ensemble des pays mais uniquement ceux d'Asie du Sud-Est et du Pacifique.

L'effet aggravant du *rapport entre la longueur des frontières fluviales et la longueur totale des frontières terrestres* peut s'expliquer par le fait que les frontières fluviales sont souvent sources de tensions entre pays frontaliers, pour des problèmes d'accès à l'eau par exemple. Ces tensions fragilisent le gouvernement dans la mesure où une partie de ses ressources se focalisent sur ces problèmes frontaliers. Cette fragilité peut alors être exploitée par des rebelles. Il est également possible d'envisager que les frontières fluviales constituent un atout économique indéniable pour la région frontalière. D'une part, l'accès à l'eau est garanti. D'autre part, le commerce est favorisé. Enfin les plaines fluviales sont généralement fertiles et propices au développement de l'agriculture. De telles régions ont un avantage concurrentiel sur les autres régions du pays et peuvent être amenées à déclencher un conflit pour obtenir leur indépendance.

La guerre civile qui a touché la république du Congo en 1993 a ainsi été déclenchée par une tentative de sécession du quartier de Bacongo à Brazzaville. Comme le note Piermay (2005)

> la sécession – qui fut aussi politique – du quartier de Bacongo (Brazzaville, Congo) ne fut possible que par le fonctionnement du beach, port sommaire aménagé sur les rives du fleuve, permettant la liaison avec l'autre rive et avec la ville voisine – et capitale de l'autre Congo –, Kinshasa.

Isolant les conflits intra-étatiques séparatistes de ceux dont l'objectif est le contrôle de l'autorité centrale, Buhaug (2006) note une recrudescence des tentatives de sécession dans les années 90 au lendemain de la Guerre froide. La dislocation de l'Union soviétique qui garantissait une certaine stabilité au sein de son bloc d'influence est l'une des explications possibles de ce phénomène. Ce constat sur la prévalence des conflits séparatistes dans les années 90 tend à appuyer notre seconde interprétation sur le rôle des frontières fluviales. Pour tester expérimentalement cette hypothèse il serait utile de savoir si tel ou tel conflit correspond à une tentative de sécession ou non, à l'image de ce qui a été fait par Buhaug.

Notons enfin qu'une dernière hypothèse peut être avancée pour expliquer le lien entre la recrudescence des conflits territoriaux et les frontières fluviales. Les bassins fluviaux correspondent bien souvent à des zones de peuplement ancien marquées par une histoire et une culture forte. Un découpage des frontières arbitraire, effectué sans tenir compte de l'histoire et de la géographie locales, est donc susceptible d'attiser les tensions dans la région. Ces tensions qui se traduisent par des revendications territoriales portant sur le tracé de certaines frontières ont pu éclater une fois que l'équilibre de la terreur rompu.

10.4.2.10 Analyse critique du comportement de notre système

Jusqu'ici nous nous sommes focalisé sur les facteurs de risque mis en évidence par le système. Pour compléter l'analyse qualitative de nos résultats il nous semble important de rendre compte du comportement de notre système vis-à-vis de quelques pays afin d'illustrer les forces et faiblesses de notre approche.

Parmi les observations en crise classées correctement par la quasi-totalité des modèles, nous avons entre autres le Rwanda, le Burundi, la Sierra Leone, l'Angola pour la période 1999-2000 ou encore le Myanmar et les Philippines pour la période 2001-2002. Toutes ces observations correspondent à des pays qui ont été en guerre civile presque chaque année depuis 1970, avec de multiples déclenchements de nouveaux conflits. Pour toutes ces observations, le poids de l'historique des conflits est très important, ce qui a été parfaitement identifié par notre système. Rappelons que les quatre facteurs de risque communs à la grande majorité des modèles se rattachent tous à cette dimension historique.

Si l'identification de l'importance de l'historique des conflits est l'une des réussites de notre système, le poids excessif qui lui est attribué en constitue en revanche une faiblesse. Les pays sortant à peine d'une crise seront fréquemment considérés à risque, parfois à tort. On observe ainsi que la plupart des modèles prédisent que le Burundi est en crise en 1993-1994, de même que la Sierra Leone en 2001-2002 ou encore la Russie et le Liberia en 1997-1998. Or ce sont tous des pays en crise les années précédentes. Ces erreurs ne sont pas excessivement embarassantes dans la mesure où elles ne choqueront vraisemblablement pas un expert en sciences politiques. La situation dans ces pays, durant les périodes considérées, était en effet toujours instable et méritait d'être surveillée.

D'autres erreurs fréquentes sont révélatrices d'une faiblesse de notre méthodologie. Parmi les observations en crise qui ont été presque systématiquement affectées à la classe *non-crise*, citons le Nigéria pour la période 2001-2002, l'Indonésie et le Bangladesh pour la période 1995-1996, le Cameroun pour la période 1993-1994, la Chine pour la période 1989-1990 ou encore la Côte d'Ivoire pour la période 1999-2000. Ces observations partagent toutes une même caractéristique. Elles appartiennent toutes à la classe *crise* non pas parce qu'un conflit armé intra-étatique a eu lieu, mais parce qu'il a été fait usage de la violence de manière unilatérale. Nous avons inclus ces phénomènes dans nos bases de données afin d'être capable d'anticiper différents types de crise. Force est de constater qu'il s'agit là d'une erreur méthodologique. Les deux phénomènes étant de nature très différente, il est illusoire de chercher des associations communes entre les variables explicatives et ces deux phénomènes. Il eût été plus opportun de constituer une troisième classe plutôt que de fusionner les deux types de crise.

10.4.2.11 Étude de cas sur le Rwanda

Pour clore notre analyse nous proposons d'étudier plus en détail les résultats concernant le Rwanda. Nous avons choisi ce pays car il est assez représentatif des réussites et échecs de notre approche. Nous avons représenté sur la figure 10.7 l'évolution de 1970 à 2002 des degrés de reconnaissance des deux classes estimés par nos modèles, pour les périodes d'estimation de 1 et 7 ans. Une ellipse a été placée sur les courbes à chaque fois qu'une erreur de prédiction a été commise.

L'allure générale des courbes sur les deux graphiques de la figure 10.7 est plutôt satisfaisante. Le degré de reconnaissance de la classe *non-crise* très élevé dans les années 70 a tendance à diminuer dans les années 80 puis 90, tandis que le degré de reconnaissance de la classe *crise* suit une progression inverse. Ce dernier est nul ou presque jusqu'en 1985, date à partir de laquelle il commence à augmenter avant de devenir supérieur au degré de

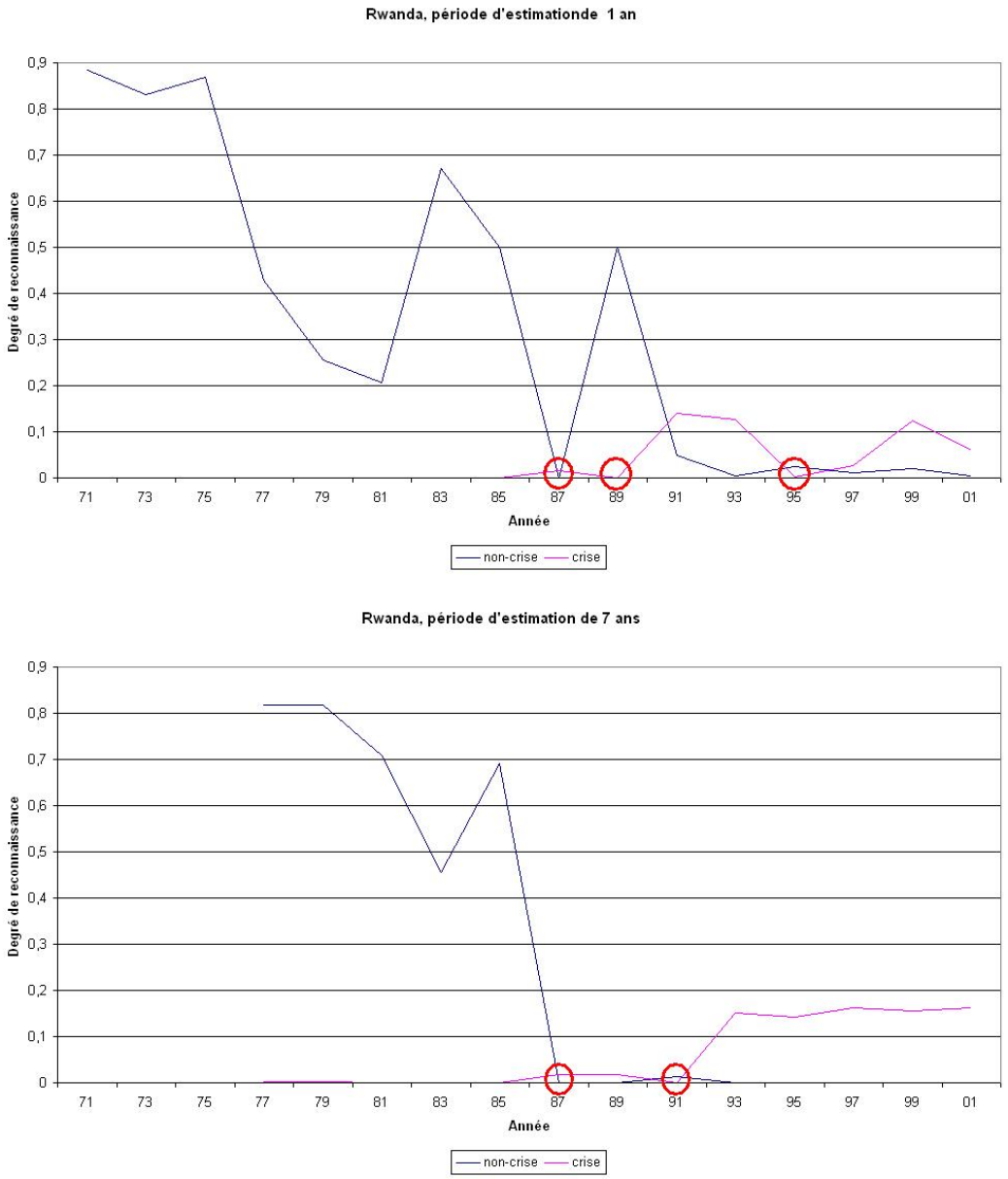

FIG. 10.7 – Évolution du risque de conflit pour le Rwanda, avec des périodes d'estimation de 1 et 7 ans

reconnaissance de la classe *non-crise* dans les années 90. Ces remarques générales sont valables pour les deux périodes d'estimation et traduisent une montée progressive du niveau de risque qui devient alarmant à la fin des années 80.

Précisons que le pays n'est considéré en crise qu'à partir de 1990 dans notre base de données, mais que depuis longtemps des tensions affectent le climat social dans le pays. L'année 1973 est ainsi marquée par une violente campagne anti-tutsis dans le milieu scolaire et par le coup d'État du général Habyarimana. On peut constater, sur les courbes relatives à la période d'estimation de 1 an, une brusque chute du degré de reconnaissance de la classe *non-crise* qui coïncide parfaitement avec ces événements. Fin 1982 l'Ouganda expulse les réfugiés rwandais qui se voient refuser le retour au pays par le gouvernement rwandais, ce qui entraîne une catastrophe humanitaire.

À partir de cette époque l'opposition en exil s'organise en Ouganda et les tensions

avec le gouvernement rwandais vont croître jusqu'au conflit qui sera déclenché en 1990. Là encore, l'évolution générale des degrés de reconnaissance traduit assez fidèlement l'évolution historique. En effet, à partir de 1983 le degré de reconnaissance de la classe *non-crise*, qui était revenu à un niveau assez élevé, chute à nouveau. C'est à partir de 1987-1988 que le degré de reconnaissance de la classe *crise* dépasse celui de la classe *non-crise*, soit deux ans avant le déclenchement effectif du conflit mais exactement à la même époque que les premiers signes forts d'une crise ethnique dans le Burundi voisin. De nombreux réfugiés burundais gagnent à cette époque le Rwanda. Cette erreur ne nous paraît donc pas poser un problème majeur, dans la mesure où l'instabilité de la région était déjà forte en 1988.

Selon la période d'estimation considérée, le tableau diffère ensuite légèrement dans les années 90. Dans l'ensemble, le niveau de risque est élevé dans les deux cas, ce qui correspond à la réalité de l'époque. Cependant, alors que le degré de reconnaissance de la classe *non-crise* reste quasi nul entre 1990 et 2002 avec une période d'estimation de 7 ans, deux pics sont observables lorsque l'on considère la période d'estimation de 1 an. Le risque est considéré moindre en 1995-1996, voire mineur en 1989-1990.

La règle utilisée pour classer le Rwanda en 1987-1988 est la suivante :

Si le nombre de morts liés aux conflits passés est inférieur à 800 et **si** la formation brute de capital fixe est supérieure à 6% du PIB et **si** le montant de l'aide officielle au développement est inférieur à 300 millions de dollars courants et **si** l'inflation est inférieure à 2% **alors** le pays sera en crise.

En 1986 le Rwanda n'avait pas connu récemment de guerre civile, le gouvernement investissait une partie non négligeable de ses revenus pour le développement des infrastructures. Ce sont plutôt là des caractéristiques de pays relativement stables. À cette époque, les pays d'Afrique subsaharienne qui étaient les plus stables sur le plan politique connaissaient cependant des niveaux d'inflation bien plus élevés que le Rwanda. Ainsi l'inflation au Sénégal ou au Kenya se situait entre 6 et 8% contre -7% au Rwanda. La baisse des prix observée au Rwanda a donc été interprétée par nos modèles comme le signe d'un développement fragile insuffisamment soutenu par les aides extérieures. Il est d'ailleurs étonnant de constater que le montant des aides attribuées au Kenya ou au Sénégal était alors près de 2 ou 3 fois supérieur à celui des aides octroyées au Rwanda ou au Burundi, dans lesquels les tensions étaient pourtant perceptibles, avec des troubles ethniques de faible intensité qui ont marqué le pays depuis son indépendance.

Disposer d'indicateurs permettant de mesurer les troubles sociaux afin de pouvoir rendre compte de leur évolution aurait été fort utile pour anticiper les crises rwandaise et burundaise. Rappelons à ce sujet que l'étude quantitative de Lichbach *et al.* (2004) mentionnée à la section 10.2 a permis de mettre en évidence l'intérêt de tels indicateurs.

À partir de 1987-1988, l'inflation est passée au-dessus du seuil des 2% sans que le montant des aides n'évolue significativement, ce qui a conduit notre système à baisser le niveau d'alerte du Rwanda pour la période 1989-1990.

En 1995-1996, le Rwanda est considéré comme faisant partie de la classe *crise* à cause de la poursuite du génocide de 1994. Or nous avons vu précédemment les difficultés que notre système rencontre pour modéliser ce type de phénomène. La règle utilisée pour classer incorrectement le Rwanda ces années-là est la suivante :

Si la dernière guerre civile en date a eu lieu il y a moins de deux ans et **si** le nombre de réfugiés accueillis l'année dernière par le pays est inférieur à 2500 **alors** le pays ne sera pas en crise.

Or le flux de réfugiés enregistré au Rwanda en 1993-1994 est négligeable. Ceci peut s'expliquer par le fait qu'en 1992 et 1993 le Burundi, en conflit depuis 1990, a connu ses premières élections libres avec le retour du multipartisme. S'ils furent de courte durée, certains signes de retour à la stabilité dans la région étaient donc bien perceptibles à cette

époque.

On peut constater que le modèle appris à partir d'une période d'estimation de 7 ans maintient pour sa part un niveau de risque élevé pour 1995-1996. La différence avec le modèle précédent peut s'expliquer par le fait qu'une période d'estimation de 7 ans a tendance à lisser la distribution des variables explicatives. Ce comportement présente l'avantage de tenir compte de l'influence à plus ou moins long terme de l'évolution de certains facteurs, mais il est beaucoup plus difficile de tenir compte des variations sur le court terme de ces mêmes facteurs. Or, ces variations reflètent une modification brusque du contexte structurel et constituent donc bien souvent de bons signaux d'alerte. Pour chaque variable, nous avons introduit cette notion de variabilité, mais elle est diluée sur l'ensemble de la période d'estimation. Cette relative faiblesse explique peut-être pourquoi le gain apporté par la période d'estimation de 7 ans par rapport à la période de 1 an, classiquement employée dans le domaine, est assez faible.

Chapitre 11

Discussion

Ainsi que nous avons pu le constater à la partie II, le prétraitement des données exerce une influence non négligeable sur la qualité des modèles de classification tels que ceux que nous employons en détection de crise. Devant la difficulté de choisir *a priori* une chaîne de prétraitement pour un problème donné nous avons proposé une plate-forme d'évaluation, de comparaison et de sélection de modèles.

La phase de sélection s'appuie sur une analyse de rangs pour éliminer l'ensemble des méthodes dont les performances sont statistiquement inférieures à au moins une méthode, les différentes sous-tâches à traiter étant considérées dans leur ensemble. Lorsque ce tri est insuffisant pour isoler une chaîne d'apprentissage optimale, ce que nous avons constaté lors de nos expérimentations, il est procédé à une analyse de la confiance que l'on peut placer dans les différents modèles construits. Les performances des méthodes issues de la première phase de sélection étant dans l'ensemble équivalentes, l'analyse de la confiance s'effectue sur chaque base de données prise séparément. L'objectif est d'identifier, pour chaque sous-tâche, le modèle le plus fiable. Cette seconde analyse se fonde sur l'introduction d'un seuil de rejet correspondant à la différence minimale acceptable entre les degrés de reconnaissance de chacune des classes. Lorsque cette différence est trop faible la décision prise par le modèle n'est pas fiable, l'incertitude sous-jacente étant trop élevée. Nous avons également employé le score de Brier, qui mesure l'erreur quadratique moyenne de prédiction, pour quantifier le degré d'incertitude globale des décisions prises par le système.

L'ensemble du processus de sélection dépend fortement du critère de performance choisi pour évaluer la qualité d'un modèle. Afin que l'utilisateur de notre système puisse guider le processus de sélection, nous avons construit un système d'inférences simple permettant de définir une mesure de performance à partir de règles floues. Ce principe permet à l'utilisateur d'exprimer relativement facilement ses préférences.

L'application de cette méthodologie à la détection des conflits armés intra-étatiques a été riche d'enseignements. D'un point de vue purement quantitatif, notre approche a permis d'obtenir des performances plus que satisfaisantes. Elles ne sont en effet que légèrement inférieures à celles que la *State Failure Task Force* a pu obtenir, et ce, malgré quelques erreurs méthodologiques dans la construction des bases de données. D'un point de vue qualitatif nous avons pu observer que nombre de facteurs de risque jugés importants dans la littérature ont été identifiés. De nouveaux facteurs corroborant certaines explications théoriques sur l'émergence des conflits ont également été mis en évidence. Nous avons ainsi trouvé des éléments confirmant aussi bien les arguments insistant sur l'importance des griefs de la population à l'égard du gouvernement que ceux qui mettent l'accent sur l'importance des occasions facilitant le déclenchement d'une rébellion. Nous avons de plus proposé de nouvelles hypothèses de travail pour l'étude des conflits. Si certaines nous semblent devoir être abandonnées, comme par exemple la fusion en une même catégorie

des conflits armés intra-étatiques et des actes de violence commis de manière unilatérale, d'autres méritent selon nous l'attention de la communauté. Plutôt que de développer un modèle générique adapté à l'ensemble des pays, il semble ainsi préférable de construire des modèles spécifiques pour chaque groupe de pays et pour une période donnée[1]. Le nombre d'années considérées pour estimer la valeur des variables explicatives est communément fixé à 1 an dans la littérature économétrique. Nous avons vu qu'il peut être judicieux de considérer des périodes dites d'estimation plus longues afin de mieux tenir compte de l'évolution historique du contexte structurel.

Afin de juger plus finement la qualité prédictive de nos modèles, il serait bon de mener à bien des tests de robustesse en envisageant différentes définitions de la classe *crise*. D'une part, nous pourrions utiliser des bases de données sur les conflits autres que celle qui a été développée par l'université d'Uppsala afin d'envisager des définitions de la notion de conflits différente. La base *Correlates of War* ou celles qui ont été construites par Laitin et Fearon, Collier et Hoeffler ou encore la *State Failure Task Force* en sont quelques exemples. D'autre part, nous pourrions envisager des horizons d'alerte de plus de deux ans.

La notion de robustesse est fondamentale en apprentissage et plus généralement dans les sciences expérimentales. Elle assure une meilleure compréhension des phénomènes étudiés et permet de délimiter le domaine de validité des hypothèses théoriques testées empiriquement. Aussi pourrait-il s'avérer fort bénéfique d'inclure dans la phase de sélection des modèles une étape d'analyse de leur robustesse. Une telle analyse aurait pour objectif d'identifier les variations limites de chaque attribut au-delà desquelles les décisions prises par le modèle sont modifiées. Alvarez *et al.* (2007) ont proposé une méthodologie pour mettre en place une telle analyse dans le cadre de l'apprentissage d'arbres de décision. En adaptant ces travaux aux arbres de décision flous, nous serions à même d'évaluer la robustesse des modèles construits par *Salammbô*.

[1] Nous faisons ici référence à la période que nous avons nommée période d'analyse.

Conclusion et perspectives

11.1 Conclusion

Partie intégrante de l'évaluation des risques, la prévision du déclenchement des crises est l'objet principal de cette thèse. Dans cette optique nous avons proposé un système générique d'aide à la décision destiné à des experts en veille stratégique. Notre système repose sur l'apprentissage automatique d'arbres de décision flous à partir de données historiques décrites par un ensemble de facteurs structurels potentiellement crisogènes. Ces arbres constituent des modèles de prédiction permettant de décider si une situation donnée est à risque ou non. L'incertitude liée à la décision qui est prise est employée pour estimer le degré de risque associé à cette situation.

Outre l'automatisation de l'analyse des données, l'intérêt de notre approche réside dans la facilité avec laquelle un utilisateur sera à même d'interpréter les résultats fournis par notre système. À chaque décision prise correspond en effet une règle exprimable en langage naturel et aisément compréhensible.

La qualité des modèles étant fortement dépendante de la qualité des données à partir desquelles ils sont appris, nous avons étudié l'impact de la chaîne de prétraitement sur le comportement d'un classifieur. Nous nous sommes plus précisément intéressé au traitement des valeurs manquantes et à la sélection d'attributs.

Après avoir analysé les forces et faiblesses des principales méthodes dans chacun des deux domaines, nous avons retenu celles que nous jugions susceptibles d'être employées pour traiter des données déséquilibrées en grande dimension contenant un nombre important de valeurs manquantes réparties sur l'ensemble des attributs. Ce sont là, en effet, les principales caractéristiques des données d'apprentissage pour les problèmes de détection de crise.

Pour le traitement des valeurs manquantes, nous avons opté pour des méthodes de substitution qui permettent de reconstruire une base de données complète, utilisable par n'importe quel algorithme d'apprentissage.

Quant à la sélection d'attributs, après avoir envisagé l'emploi d'un algorithme génétique dans la première partie, nous avons décidé dans la deuxième partie de recourir à des filtres, essentiellement pour des raisons de complexité.

Après avoir étudié empiriquement le comportement des méthodes de substitution et de filtrage de façon indépendante, nous avons procédé à l'analyse de la chaîne d'apprentissage dans son ensemble. Nous avons ainsi pu mettre en évidence l'importance des interactions entre les différents maillons de la chaîne et la difficulté d'identifier une chaîne optimale, ne serait-ce que sur un nombre restreint d'applications.

Afin que notre méthodologie soit aussi générique que possible, nous avons fait en sorte que notre système ne corresponde pas simplement à l'instanciation d'une chaîne d'apprentissage particulière, mais plutôt à une plate-forme d'évaluation, de comparaison et de sélection de modèles. Ce n'est qu'en procédant ainsi que nous sommes à même de choisir la chaîne d'apprentissage la mieux adaptée à un problème donné.

L'application de notre système à la détection des conflits armés intra-étatiques nous a permis d'illustrer l'intérêt de notre approche. Les résultats obtenus sont plus que prometteurs, sur le plan tant quantitatif que qualitatif. D'une part, bien que la modélisation des pays en crise ne soit pas encore totalement satisfaisante, les performances en prédiction de notre système sont du niveau de celles que parvient à obtenir la *State Failure Task Force*, qui est une référence en la matière. D'autre part, malgré les divergences entre notre approche et celles, issues de l'économétrie, qui sont classiquement employées dans le domaine, les facteurs de risque mis en évidence par notre système sont parfaitement compatibles avec les théories dominantes sur l'origine des conflits.

11.2 Originalité de nos travaux

Au gré de nos recherches nous avons été amené à explorer différents domaines.

- l'évaluation des risques de crise
- le traitement des valeurs manquantes
- la sélection d'attributs
- la prévision des conflits armés intra-étatiques
- l'apprentissage automatique

La description synthétique que nous venons de donner de notre travail a permis d'illustrer l'intérêt que nous avons accordé à chacun de ces domaines durant notre thèse. Il nous faut à présent approfondir notre réflexion afin de mettre en valeur les répercussions que nos travaux peuvent avoir sur ces différents domaines.

Évaluation des risques de crise

La méthodologie générique de prévision des crises que nous avons proposée est, selon nous, notre apport principal au domaine de l'évaluation des risques de crise. Formaliser le problème de quantification de l'incertitude liée à l'occurrence d'une crise par un problème de classification supervisée est assez peu fréquent. Nous avons montré expérimentalement que pour certains problèmes, tels que la détection des conflits armés intra-étatiques, cette approche pouvait être efficace.

Mais l'efficacité, mesurée par les performances en prédiction, est loin de suffire pour convaincre des experts qu'une approche automatisée peut leur être utile. Aussi avons-nous mis l'accent sur la transparence de notre système en ayant recours à des arbres de décision flous. Nous sommes ainsi en mesure d'apporter à l'utilisateur, non pas simplement des indices de risques, mais également des règles en langage naturel explicitant la façon dont ces indices ont été estimés. Ce point est essentiel car l'utilisateur doit pouvoir porter un regard critique sur le système. Celui-ci n'a pas vocation à être autonome et est conçu comme un outil d'aide à la décision.

Dans cette optique nous avons développé une interface d'aide au raisonnement et à la compréhension de la situation. L'utilisateur dispose pour chaque observation, du niveau de risque estimé par le système, de la règle ayant permis d'obtenir ce niveau de risque ainsi que des valeurs des attributs qui sont considérés comme étant les principaux facteurs de risque.

Pour faciliter le raisonnement contrefactuel et plus généralement conditionnel, nous avons fait en sorte qu'il ait la possibilité de modifier ces valeurs afin de se rendre compte de l'influence de ces modifications sur le niveau de risque.

Nous avons également insisté sur la nécessité d'adapter notre système aux besoins de l'utilisateur. Pour y parvenir, nous avons proposé de guider la sélection de modèles par un critère de performance à base de règles floues. Ces règles, définies par l'utilisateur, permettent de prendre en compte ses préférences.

Traitement des données manquantes

De nos recherches sur le traitement des valeurs manquantes, deux points importants sont à retenir. Nous avons introduit une nouvelle taxinomie des méthodes de substitution, bien plus détaillée et globale que celles, embryonnaires[2], qui ont pu être présentées dans la littérature. En dégageant de nouveaux critères de discrimination, nous avons mis en évidence certains traits caractéristiques des principales méthodes. Ces caractéristiques constituent autant de degrés de liberté sur lesquels peut agir un chercheur pour construire de nouvelles techniques de substitution. L'intérêt de notre taxinomie ne réside donc pas uniquement dans sa capacité à décrire synthétiquement les méthodes du domaine.

Nous avons également insisté sur l'objectif sous-jacent de l'emploi d'une méthode de substitution. La plupart des techniques existantes ont pour objet de trouver des valeurs de remplacement qui soient aussi proches que possible des valeurs d'origine, valeurs inconnues en pratique.

Avec Thanh Ha Dang nous avons cherché à ce que les valeurs de remplacement permettent d'optimiser les performances d'un classifieur qui serait appris à partir des données complétées. Ce changement de point de vue nous a conduit au développement d'une nouvelle technique basée sur la minimisation de l'entropie conditionnelle, qui vise à maximiser le pouvoir discriminant de chaque attribut. Nous avons pu observer empiriquement que cette nouvelle méthode était performante sur un certain nombre de problèmes.

Sélection d'attributs

Dans le domaine de la sélection d'attributs, notre apport est également double. Après avoir réduit notre champ d'étude aux filtres pour des raisons de complexité, nous avons proposé une extension du filtre rapide basé sur la corrélation, développé par Liu et Yu. En utilisant le test de Kolmogorov-Smirnov pour analyser tant la pertinence des attributs, que la redondance entre les couples d'attributs, nous sommes parvenu à maintenir une complexité réduite tout en proposant une méthode capable de traiter directement les attributs continus. Une comparaison empirique de différents filtres a permis de montrer que les performances de notre filtre étaient équivalentes à celles de filtres répandus dans la littérature.

Outre l'introduction de cette nouvelle méthode de filtrage, notre travail s'est distingué de la littérature en proposant une étude globale de la chaîne d'apprentissage. À notre connaissance, algorithmes de sélection d'attributs et de substitution des valeurs manquantes n'ont été, jusqu'à présent, étudiés qu'indépendamment les uns des autres. Or un système d'apprentissage global, tel que celui que nous avons mis en place, intègre les deux étapes de prétraitement avant que l'apprentissage à proprement parler soit effectué.

Nos expérimentations ont mis en évidence l'importance d'une telle analyse globale. Pour un problème donné, la chaîne d'apprentissage optimale n'est en effet pas forcément composée du filtre et de la méthode de substitution qui, indépendamment l'un de l'autre, sont les mieux adaptés au problème considéré. La stratégie de combinaison, qui indique si le filtrage doit avoir lieu avant la substitution, et si oui, s'il doit ignorer ou non les valeurs

[2] L'emploi de cet adjectif ne se justifie que par comparaison avec ce qui a été fait dans d'autres domaines comme la sélection d'attributs par exemple.

manquantes, influe également sur les performances de l'ensemble de la chaîne. Seule une analyse globale de la chaîne permet d'identifier la stratégie optimale de combinaison en fonction des méthodes de prétraitement qui sont considérées.

Prévision des conflits armés intra-étatiques

De notre approche de la prévision des conflits armés intra-étatiques nous retiendrons quatre innovations méthodologiques importantes. Les bonnes performances de notre système en prédiction et la facilité d'analyse des résultats ont contribué, du moins nous l'espérons, à montrer l'intérêt de l'utilisation des méthodes d'apprentissage dans ce domaine, lorsque l'objectif fixé est l'anticipation des conflits. Rappelons qu'à de rares exceptions près, seules les techniques de régression sont employées. Ceci s'explique par le fait que les chercheurs du domaine ont pour principal objectif d'expliquer l'émergence des conflits et non de les prédire.

Les trois autres points sur lesquels nous souhaitons revenir concernent les choix de modélisation qui s'expriment lors de la création des bases de données. Celle-ci guide en effet l'apprentissage des modèles prédictifs. Au travers de nos expérimentations nous avons vu qu'il pouvait s'avérer fort judicieux de construire des modèles spécifiques à un groupe de pays donné et à une période historique donnée. L'idée sous-jacente est que les conditions structurelles propices au déclenchement d'un conflit fluctuent au cours du temps et ne sont pas identiques pour tous les pays.

Nous avons également pu constater que la durée de la période d'estimation influait sur les résultats. Ainsi, pour prédire l'émergence d'un conflit, il peut être utile de considérer les données relatives à la situation dans le pays plusieurs années avant la date à laquelle les prédictions sont effectuées, et pas uniquement l'année précédant cette date, comme cela est le cas dans la grande majorité des travaux.

Apprentissage automatique

En évoquant nos travaux relatifs à la substitution des valeurs manquantes et à la sélection d'attributs, nous avons implicitement mentionné certaines de nos contributions dans le domaine de l'apprentissage automatique. Nous avons en effet étudié le problème du prétraitement des données exclusivement dans le cadre de ce domaine. Outre le développement de nouvelles techniques de prétraitement, nous estimons que la mise en exergue de l'importance d'une étude globale de la chaîne d'apprentissage constitue un apport plus conséquent au domaine. Selon nous notre principale contribution est d'ordre méthodologique. Rien de ce que nous avons proposé d'un point de vue méthodologique n'est réellement nouveau. L'importance que nous y avons attachée l'est nettement plus.

Nous nous sommes efforcé tout au long de cette thèse de souligner l'importance de la méthodologie dans le processus de recherche en apprentissage automatique. L'expérimentation y joue un rôle centrale ainsi que l'affirme Langley (1988) dès 1988. Cette remarque a deux conséquences importantes.

Premièrement, il convient d'appliquer dans notre discipline la méthode expérimentale dont les principes ont été clairement énoncés par Bernard (1984).

> Le savant complet est celui qui embrasse à la fois la théorie et la pratique expérimentale. 1° Il constate un fait ; 2° à propos de ce fait, une idée naît dans son esprit ; 3° en vue de cette idée, il raisonne, institue une expérience, en imagine et en réalise les conditions matérielles. 4° De cette expérience résulte de nouveaux phénomènes qu'il faut observer et ainsi de suite.

On retrouve dans cette description la notion d'allers-retours successifs entre expérimentation et formalisation théorique que nous avons mise en avant au chapitre 4 et que nous nous sommes efforcé de suivre durant cette thèse.

D'une première tentative de modélisation, nous avons avancé pas à pas vers la réalisation d'une plate-forme globale de sélection de modèles. Chacune des étapes de notre cheminement a été guidée par le besoin de remédier à certaines faiblesses mises en évidence durant les étapes précédentes. L'expérience est cruciale[3] en ce qu'elle permet de tester des hypothèses. La force de ces tests réside uniquement dans la réfutation potentielle d'une théorie et non dans sa confirmation[4]. C'est pour cette raison que nous avons essayé de formuler avec précaution certaines de nos conclusions expérimentales, en précisant bien à chaque fois que celles-ci n'étaient valables que dans un cadre expérimental particulier et en aucun cas général.

Le fait que nous ayons opté pour une approche basée sur la classification supervisée répond également à cette nécessité de mettre l'expérience au centre de la méthodologie. Pour construire des expériences en vue du test de certaines hypothèses, il faut que les faits déduits d'une théorie particulière puissent être confirmés ou infirmés par des faits. Dans le domaine de la prévision des risques de crise, il est impossible de comparer directement un niveau de risque prédit avec la réalité. Seules les prédictions de déclenchement de crises sont vérifiables. Aussi construire des modèles de régression pour estimer le niveau de risque suppose que l'on introduise un seuil d'alerte pour se ramener *in fine* à une tâche de classification. Ce n'est donc qu'indirectement que les modèles de régression pourront être soumis au contrôle de l'expérience.

Deuxièmement, nous avons insisté sur la méthodologie à employer pour construire des expériences sans biais, pour observer et interpréter les faits expérimentaux le plus objectivement possible. Claude Bernard parle à ce propos de « l'art d'obtenir des faits exacts au moyen d'une investigation rigoureuse ». C'est dans cette optique que nous avons mis l'accent sur les protocoles expérimentaux. Ils doivent être clairement énoncés, pour que les expériences puissent être reproduites par d'autres chercheurs. Une expérience non reproductible n'a que peu de valeurs. Il est en effet impossible de vérifier qu'aucune erreur méthodologique n'est venue entacher le recueil des faits.

La procédure d'analyse de ces faits doit également être transparente pour qu'il soit possible de limiter et d'isoler clairement la part de subjectivité nécessairement introduite par l'expérimentateur. À cet effet, nous avons employé un certain nombre de tests statistiques. Notre objectif était alors d'identifier la part de hasard intervenant dans les différences observées entre les méthodes comparées, afin d'éviter d'accorder une importance indue à certaines d'entre elles. Si nous nous sommes beaucoup appuyé sur ces tests, nous avons également été soucieux de limiter l'impact des conclusions qu'ils permettent de tirer. Il est en effet extrêmement facile de mésinterpréter ces conclusions.

Drummond (2006), plus de quinze ans après l'article de Langley, note qu'une importance démesurée a été accordée aux tests statistiques dans le domaine de l'apprentissage automatique[5]. Le problème, selon lui, ne vient pas tant des tests statistiques que de l'interprétation qui en est faite. Il conseille de ne pas se restreindre à l'utilisation de ces tests, mais d'indiquer chaque fois que cela est possible les intervalles de confiance associés à

[3] Claude Bernard parle d'*experimentum crucis*

[4] Depuis les travaux de Popper, cette idée est bien ancrée dans la démarche réflexive de la science. Précisons cependant que cinquante plus tôt Claude Bernard ne disait pas autre chose : « un expérimentateur qui voit son idée confirmée par une expérience, doit douter encore et demander une contre-épreuve ».

[5] Certains chercheurs font le même constat dans des disciplines telles que la psychologie expérimentale (Cohen, 1994).

telle ou telle mesure. Ce point est essentiel et constitue, selon nous, l'un des axes majeurs d'améliorations de notre méthodologie. Détaillons à présent les autres améliorations importantes que nous pourrions apporter à notre travail. Ceci nous permettra de souligner les perspectives ouvertes par notre thèse.

11.3 Perspectives

Nous avons considéré qu'une chaîne d'apprentissage était composé d'une méthode de substitution des valeurs manquantes, d'une méthode de filtrage des attributs non pertinents ou redondants et d'un algorithme d'apprentissage supervisé. Si nous avons déjà mentionné un certain nombre d'améliorations potentielles lors des conclusions intermédiaires, plusieurs points importants ont cependant été occultés.

La chaîne d'apprentissage

Premièrement, la phase de substitution n'est obligatoire que dans la mesure où nous refusons d'envisager des modifications de *Salammbô* qui lui permettraient de traiter des bases de données incomplètes. Travailler à l'adaptation de *Salammbô* aux données incomplètes est une piste qu'il pourrait être intéressant d'approfondir.

Deuxièmement, le filtrage d'attributs ne sélectionne qu'un sous-ensemble des attributs initialement recueillis. Il est fort possible que certaines combinaisons d'attributs soient plus pertinentes que tel ou tel attribut pris indépendamment. Aussi souhaiterions-nous étudier les techniques d'extraction d'attributs. Nous avons soulevé précédemment le problème de l'interprétabilité des résultats, qui est souvent affectée par la création de nouveaux attributs par des méthodes telles que l'ACP. Nous pensons cependant que cette difficulté peut être surmontée par programmation génétique contrainte, l'idée étant de fixer un certain nombre de contraintes sur les opérations de combinaison d'attributs.

Troisièmement, seul l'algorithme de construction d'arbres de décision flous, *Salammbô*, a été envisagé pour réaliser l'apprentissage des modèles de classification. Les progrès réalisés dans l'apprentissage de structure des réseaux bayésiens mériteraient d'être considérés un peu plus attentivement. Il serait également bon de proposer dans notre système l'apprentissage de modèles par régression logistique. Dans notre contexte applicatif, ceci nous permettrait de confronter notre approche avec celle qui est fréquemment employée dans la littérature économétrique sur l'émergence des conflits.

Un quatrième axe d'amélioration de notre système concerne la prise en compte du déséquilibre de la répartition des exemples dans les différentes classes. Ce déséquilibre est fréquent dans les problèmes d'évaluation de risque, quel que soit le domaine d'application. Les techniques de rééchantillonnage, surtout le sous-échantillonnage, offrent des perspectives intéressantes. La *State Failure Task Force* l'a mis en place avec un certain succès.

Domaine d'application d'une méthode

Les expérimentations que nous avons menées durant cette thèse nous ont permis de tester un certain nombre d'hypothèses et d'avancer ainsi vers une solution admissible au problème que nous nous étions posé. Lorsque nous avons étudié certains points précis de la chaîne d'apprentissage nous avons été amené à comparé expérimentalement un certain nombre de méthodes, y compris les nouvelles que nous avons introduites. Si nous avons insisté sur l'importance de l'expérimentation en tant qu'outil permettant de mieux comprendre ces méthodes, force est de constater que nous ne sommes parvenu que très médiocrement à délimiter le domaine d'application adéquat pour chacune d'entre elles.

La plupart des méthodes étant équivalentes si l'on considère l'ensemble des problèmes possibles, il est fondamental de pouvoir identifier pour chacune d'elles les types de problème pour lesquels elle est la mieux adaptée. Y parvenir permettrait de faciliter la comparaison et la sélection des méthodes. Il faut pour cela caractériser finement les différents types de problèmes auxquels on peut être confronté. C'est ce que nous avons entrepris pour le traitement des valeurs manquantes avec notre taxinomie présentée du point de vue de l'utilisateur (voir figure 6.6). Un gros travail mériterait d'être mené pour essayer de compléter une telle taxinomie en indiquant, pour chaque type de problèmes, quelles sont les méthodes les plus efficaces. Langley ne disait rien d'autre quand il écrivait : « In any science, the goal of experimentation is to better understand a class of behaviors and the conditions under which they occur »[6].

Analyse de la robustesse

D'un point de vue méthodologique, il nous semble également essentiel d'inclure dans notre plate-forme de sélection de modèles une étape d'analyse de la robustesse. La sélection doit se faire à partir d'un ensemble de critères. Nous avons, dans cette thèse, considéré la performance des modèles, en insistant sur la définition de la mesure de performance, ainsi que la confiance que l'on peut leur accorder. Analyser plus finement la variance de ces modèles ainsi que leur sensibilité aux conditions initiales serait un atout non négligable qui offrirait à l'utilisateur une information de meilleure qualité pour juger de la pertinence du système.

Dans la lignée de cet effort qu'il conviendrait de fournir sur le plan de l'analyse de la variance, il serait bon de tirer profit des recherches menées sur les forêts d'arbres (Marsala, 1998). Elles correspondent à la combinaison d'un ensemble d'arbres de décision et permettent, sans que l'interprétabilité soit grandement affectée, de réduire la variance des arbres de décision simples, variance qui est leur principale faiblesse (Geurts *et al.*, 2006).

Interface d'aide au raisonnement

Nous avons mis en évidence l'importance de la notion de période d'analyse pour l'évaluation des risques. Il est ainsi important d'adapter les modèles prédictifs à une période historique donnée. Cette remarque soulève une importante question que nous n'avons pas évoquée jusqu'à présent : quelle est la durée de validité d'un modèle appris à un instant t ? Cette question a une portée bien plus générale que ce ne laisse supposer notre discours.

Pour construire un système générique d'aide à l'anticipation des crises, qui soit en partie autonome et efficace, il faudrait développer des outils permettant d'identifier un changement de paradigme tel qu'il est nécessaire de revoir complètement les modèles existants. Il s'agirait de ne pas simplement procéder à un réapprentissage pour affiner les modèles, mais bien d'en construire de nouveaux *ex nihilo*. L'apprentissage automatique nécessite cependant des données historiques. Aussi conviendrait-il de recourir à l'expertise pour spécifier dans un premier temps les nouvelles règles de décision.

Dans cette optique nous pensons qu'il serait bon d'améliorer l'interface de notre outil, surtout du point de vue de l'interactivité. Un utilisateur devrait pouvoir modifier les règles apprises automatiquement, en supprimer certaines ou en ajouter de nouvelles en fonction de sa propre expertise. Mêler l'approche automatique à une approche experte nous paraît indispensable, ne serait-ce que pour que le système soit accepté par les utilisateurs.

[6] Dans toute science, le but de l'expérimentation est de parvenir à une meilleure compréhension d'une classe de comportements ainsi que des conditions dans lesquelles ils sont susceptibles de se produire.

Généricité de notre approche

De la tâche globale d'évaluation des risques, nous ne nous sommes préoccupé que de la sous-tâche relative à la quantification de l'incertitude liée à l'occurrence d'une crise. Nous avons négligé sciemment l'estimation de la gravité d'une crise potentielle. Il est évident que s'attaquer à cette seconde dimension des crises est un axe de recherche futur. Si nous disposons de mesures précises et chiffrées de l'ampleur des conséquences des crises passées, il est envisageable de recourir à des techniques de régression.

Si, au contraire, nous ne disposons que d'informations qualitatives sur les dommages engendrés par les crises du passé, il nous paraît plus opportun d'étendre nos modèles de classification en considérant plus de deux classes. Chacune des classes correspondrait alors, non plus à la présence ou à l'absence d'une crise, mais plutôt au niveau d'intensité d'une crise. Cette seconde approche ne demande que peu de modifications à notre système. L'essentiel du travail réside dans l'étiquetage des données et le choix de nouvelles mesures de performance qui soient adaptées à cette nouvelle formalisation du problème.

Couplage alerte rapide - veille stratégique

Nous avons défendu l'idée que notre approche de l'évaluation des risques de crise était générique, mais nous ne l'avons appliquée qu'à la tâche de détection des conflits armés intra-étatiques. Aussi, pour appuyer notre argumentation, faudrait-il tester notre système sur des problèmes distincts, comme par exemple la détection de crises financières, énergétiques, ou encore diplomatiques. La plupart des indicateurs que nous avons recueillis pourraient être réutilisés. Le principal travail consisterait, ici aussi, à revoir l'étiquetage des données pour l'adapter à la tâche souhaitée.

La construction d'un outil d'aide à la veille stratégique peut être considérée comme une finalité en soi. Cependant, dès le début de cette thèse nous avons indiqué qu'elle avait été motivée par de précédents travaux sur l'alerte rapide ayant pour objectif d'identifier les signaux annonciateurs de crise dans un flux de documents textuels. Aussi ne considérerons-nous ce travail comme accompli qu'une fois qu'auront été couplées les deux approches et que notre système de veille à long-terme sera employé pour contextualiser les événements pris en compte par un système d'alerte à court-terme. Ce couplage sera d'autant plus efficace que l'analyse à long-terme sera capable de discriminer différents types de crise et d'identifier les facteurs de risque afférents. Chacun des deux systèmes, pris indépendamment l'un de l'autre, peut avoir des répercussions pratiques importantes en intelligence économique, veille stratégique, veille sanitaire... Mais ce n'est qu'au travers de leur couplage que leur intégration dans une méthodologie globale d'analyse des risques contribuera de manière significative à faire évoluer cette discipline.

Bibliographie

E. Acuna et C. Rodriguez : The treatment of missing values and its effect in the classifier accuracy. *In Classification, Clustering and Data Mining Applications*, pages 639–648. Springer-Verlag, 2004.

D.W. Aha et R.L. Bankert : A comparative evaluation of sequential feature selection algorithm. *In* D. Fisher et J.H. Lenz, éditeurs : *Artificial Intelligence and Statistics*. Springer Verlag, 1996.

A. Al-Shahib, R. Breitling et D. Gilbert : Feature selection and the class imbalance problem in predicting protein function from sequence. *Applied Bioinformatics*, 4(3):195–203, 2005.

H. Almuallim et T.G. Dietterich : Learning with many irrelevant features. *Artificial Intelligence*, 69(1-2):279–305, 1994.

I. Alvarez, S. Bernard et G. Deffuant : Keep the decision tree and estimate the class probabilities using its decision boundary. *In International Joint Conference on Artificial Intelligence (IJCAI)*, pages 654–659. Morgan Kaufmann, 2007.

S. Ampleford, D. Carment, G. Conway et A. Ospina : Risk assessment template. Rapport technique, CIFP, 2001. http://www.carleton.ca/cifp/docs/studra1101.pdf.

T. Back : *Evolutionary Algorithms in Theory and Practice : Evolution Strategies, Evolutionary Programming, Genetic Algorithms*. Oxford University Press, 2004.

T. Bass et R. Robichaux : Defense-in-depth revisited : Qualitative risk analysis methodology for complex network-centric operations, 2001. IEEE MILCOM 2001, Policy, Systems & Security Track.

G. Batista et M.C. Monard : An analysis of four missing data treatment methods fo supervised learning. *Applied Artificial Intelligence*, 6(3):309–327, 2003.

P. Baybutt : Assessing risks from threats to process plant : Threat and vulnerability analysis. *Process Safety Progress*, 21(4):269–275, 2002.

N. Beck, G. King et L. Zeng : Improving quantitative studies of international conflict : A conjecture. *American Political Science Review*, 94(1):21–36, 2000.

D.A. Bell et H. Wang : A formalism for relevance and its application in feature subset selection. *Machine Learning*, 41(2):175–195, 2000.

C. Bernard : *Introduction à l'étude de la médecine expérimentale*. Flammarion, 1984. (1865).

H. Beyer et H. Schewefel : Evolution strategies - a comprehensive introduction. *Natural Computing*, 1(1):3–52, 2002.

J. Biesiada et W. Duch : Feature selection for high-dimensional data : A Kolmogorov-Smirnov correlation-based filter solution. *In Advances in Soft Computing*, pages 95–104. Springer, 2005.

J. Bins et B.A. Draper : Feature selection from huge features sets. *In International Conference on Computer Vision (ICCV)*, volume 2, pages 159–165, 2001.

A. Blum et P. Langley : Selection of relevant features and examples in machine learning. *In* R. Greiner et D. Subramanian, éditeurs : *Artificial Intelligence, special issue on Relevance*, pages 245–271. Elsevier Science Publishers Ltd., 1997.

A.L. Blum et R.L. Rivest : Training a 3-node neural networks is NP-complete. *Neural Networks*, 5:117–127, 1992.

T. Bouchon-Meunier : *La logique floue.* Que sais-je ? Presses universitaires de France, 2007. Quatrième édition.

P.B. Brazdil et C. Soares : A comparison of ranking methods for classification algorithm selection. *In European Conference on Machine Learning (ECML)*, volume 1810 de *Lecture Notes in Computer Science*, pages 63–74. Springer-Verlag, 2000.

H. Buhaug : Relative capability and rebel objective in civil war. *Journal of Peace Research*, 43(6):691–708, 2006.

S. Butler et P. Fischbeck : Multi-attribute risk assessment. *In Symposium on Requirements Engineering for Information Security (SREIS 2002)*, 2002.

E. Cantu-Paz : Feature subset selection, class separability, and genetic algorithms. *In Genetic and Evolutionary Computation (GECCO)*, volume 3102 de *Lecture Notes in Computer Science*, pages 959–970. Springer Verlag, 2004.

P. Capet : *Logique du mensonge.* Thèse de doctorat, Université Paris III, Sorbonne, 2006.

D. Carment : Assessing country risk : Creating an index of severity. Background discussion paper prepared for CIFP risk assessment template, CIFP, 2001. http://www.carleton.ca/cifp/docs/IndexOfSeverity.pdf.

R.A. Caruana et D. Freitag : Greedy attribute selection. *In International Conference on Machine Learning (ICML)*, pages 28–36, 1994.

F. Caselli et W.J. Coleman : On the theory of ethnic conflict. Nber working paper, National Bureau of Economic Research, Inc., 2006. http://faculty.fuqua.duke.edu/~coleman/web/ethnic.pdf.

Y.Y. Chen : Fuzzy anaylsis of statistical evidence. *IEEE Transactions on Fuzzy Systems*, 8(6), 2000.

J. Cohen : The earth is round ($p < 0.05$). *American Psychologist*, 49(12):997–1003, 1994.

S. Cohen, G. Dror et E. Ruppin : Playing the game of feature selection. *In International Joint Conference on Artificial International (IJCAI)*, 2005.

P. Collier et A. Hoeffler : Economic causes of civil war. *Oxford Economic Papers*, 50 (4):563–573, 1998.

P. Collier et A. Hoeffler : Greed and grievance in civil war. *Oxford Economic Papers*, 56(4):563–595, 2004.

P. Collier, A. Hoeffler et D. Rohner : Beyond greed and grievance : Feasibility and civil war. Working paper, Centre for the Study of African Economies, 2006. http://www.csae.ox.ac.uk/workingpapers/pdfs/2006-10text.pdf.

C. Conversano et R. Siciliano : Incremental tree-based imputation with lexicographic ordering. *In Interface*, 2003.

A. Cullen et M. Small : Uncertain risk : The role and limits of quantitative assessment. *In* T. McDaniels, éditeur : *Risk Analysis and Society : an Interdisciplicary Characterization of the Field*, pages 163–212. Cambridge University Press, 2004.

K. Dahal, Z. Hussain et A. Hossain : Loan risk analyzer based on fuzzy logic. *In IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'05)*, pages 363–366, 2005.

T.H. Dang : *Mesures de discrimination et leurs applications en apprentissage inductif.* Thèse de doctorat, Université Paris VI Pierre et Marie Curie, 2007.

T.H. Dang et T. Delavallade : Utilisation de l'entropie pour substituer des valeurs manquantes symboliques dans un problème de classification supervisée. *In Systèmes Intelligents : Théorie et Applications*, pages 45–54, 2006.

T.H. Dang, C. Marsala, B. Bouchon-Meunier et A. Boucher : Discrimination-based criteria for the evaluation of classifiers. *In International Conference on Flexible Query Answering Systems (FQAS)*, pages 552–563, 2006.

S. Das : Filters, wrappers and a boosting-based hybrid for feature selection. *In International Conference on Machine Learning (ICML)*, pages 74–81, 2001.

M. Dash et H. Liu : Feature selection for classification. *Intelligent Data Analysis : an International Journal*, 1(3):131–156, 1997.

M. Dash, H. Liu et H. Motoda : Consistency based feature selection. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 98–109, 2000.

S. Daskalaki, I. Kopanas et N. Avouris : Evaluation of classifiers for an uneven class distribution problem. *Applied Artificial Intelligence*, 20:381–417, 2006.

I. de Soysa : Globalization, social insurance, and civil conflict, 1975-2002, 2004. Meeting of the Polarization and Conflict research group, Barcelona, 10-12 décembre 2004 http://www.polarizationandconflict.org/bcn04/6de%20Soysa.pdf.

T. Delavallade, B. Bouchon-Meunier, P. Capet et C. Marsala : Country risk ratings : A new methodology to assess internal conflict risk. *In European Conference on Risk Management*, 2005.

T. Delavallade et T.H. Dang : Using entropy to impute missing data in a classification task. *In FUZZIEEE*, 2007.

T. Delavallade, L. Mouillet, B. Bouchon-Meunier et E. Collain : Monitoring event flows and modelling scenarios for crisis prediction, application to ethnic crisis forecasting. *International Journal of Uncertainty and Fuzziness Knowledge-Based Systems (IJUFKS)*, 15:83–110, 2007.

H. Delgrande et T.G. Pelletier : A formal analysis of relevance. *Erkenntnis*, 49 (2):137–173, 1998.

A. Dempster, N. Laird et D. Rubin : Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

J. Demsar : Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research (JMLR)*, (7):1–30, 2006.

M. Detyniecki : *Opérateurs mathématiques d'agrégation et leur application à la recherche d'information dans la vidéo*. Thèse de doctorat, Université Paris VI Pierre et Marie Curie, 2000.

T.G. Dietterich : Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, (10):1895–1924, 1998.

C. Ding et H. Peng : Minimum redundancy feature selection from microarray gene expression data. *In Computational Systems Bioinformatics*, pages 523–528, 2003.

J. Doak : An evaluation of feature selection methods and their application to computer security. Technical Report, Davis CA : University of California, Department of Computer Science, 1992. http://www.bis.org/publ/bcbs118.htm.

P. Domingos : The role of occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3(4):409–425, 1999.

J. Dougherty, R. Kohavi et M. Sahami : Supervised and unsupervised discretization of continuous features. *In International Conference on Machine Learning (ICML)*, pages 194–202, 1995.

C. Drummond : Machine learning as an experimental science (revisited). *In Evaluation Methods for Machine Learning Workshop of the Twenty-First National Conference on Artificial Intelligence*, 2006.

W. Duch : Filter methods. *In Feature Extraction : Foundations and Applications*, Studies in Fuzziness and Soft Computing, chapitre 3. Physica-Verlag, Springer, 2006.

B. Dufour, A.M. Hattenberger et A. Martin : Appréciation qualitative du risque et expertise collégiale. *Épidémiologie et santé animale*, 41:45–52, 2002.

H.R. Ekbia et A.G. Maguitman : Context and relevance : a pragmatic approach. *In International and Interdisciplinary Conference on Modeling and Using Context*, volume 2116 de *Lecture Notes in Computer Science*, pages 156–169. Springer-Verlag, 2001.

J. Epstein : Modeling civil violence : An agent-based computational approach. *In National Academy of Science of the USA*, volume 99, 2002. Suppl 3.

A. Farhangfar, L. Kurgan et W. Pedrycz : Experimental analysis of methods for imputation of missing values in databases. *In SPIE, Intelligent Computing : Theory and Applications II*, volume 5421, pages 172–182, 2004.

U.M. Fayyad et K.B. Irani : Multi-interval discretization of continuous-valued attributes for classification learning. *In International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1022–1027, 1993.

J.D. Fearon : Ethnic structure and cultural diversity around the world : A cross-national data set on ethnic groups, 2002. Annual Meeting of the American Political Science Association, Princeton University, Boston, www.stanford.edu/group/ethnic/workingpapers/egroups.pdf.

J.D. Fearon : Primary commodity exports and civil war. *Journal of Conflict Resolution*, 49(4):483–507, 2005.

J.D. Fearon et D.D. Laitin : Ethnicity, insurgency, and civil war. *American Political Science Review*, 97:75–90, 2003.

A.J. Feelders : Handling missing data in trees : Surrogate splits or statistical imputation. *In European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'99)*, 1999.

F. Ferri, P. Pudil, M. Hatef et J. Kittler : Comparative study of techniques for large-scale feature selection. *In* E.S. Gelsema et L.S. Kanal, éditeurs : *Pattern Recognition in Practice IV, Multiple Paradigms, Comparative Studies and Hybrid Systems*, pages 403–413. Elsevier, 1994.

G. Forman : An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research (JMLR)*, 3:1289–1305, 2003. special issue on special feature.

O. François et P. Leray : Étude comparative d'algorithmes d'apprentissage de structure dans les réseaux bayésiens. *Journal électronique d'intelligence artificielle*, 5(39):1–19, 2004.

J. Galindo et P. Tamayo : Credit risk assessment using statistical and machine learning : Basic modeling applications. *Computational Economics*, 15(1-2):107–143, 2000.

J. Gandhi et J. Vreeland : Political institutions and civil war : Unpacking anocracy, 2004. http://www.yale.edu/macmillan/ocvprogram/Gandhi_and_Vreeland1.pdf.

J. Garcia-Montalvo et M. Reynal-Querol : Why ethnic fractionalization ? polarization, ethnic conflict and growth. Economics Working Papers 660, Department of Economics and Business, Universitat Pompeu Fabra, 2002. http://www.econ.upf.edu/docs/papers/downloads/660.pdf.

P. Geurts : *Contributions to Decision Tree Induction : Bias/Variance Tradeoff and Time Series Classification*. Thèse de doctorat, Université de Liège, 2002.

P. Geurts, D. Ernst et L. Wehenkel : Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

Z. Ghahramani et M.I. Jordan : Supervised learning from incomplete data via an EM approach. *In* J.D. Cowan, G. Tesauro et J. Alspector, éditeurs : *Advances in Neural Information Processing Systems 6*, pages 120–127. Morgan Kaufman, 1994.

E.E. Ghiselli : *Theory of Psychological Measurement*. McGraw-Hill, 1964.

N.P. Gleditsch, H. Strand et H. Hegre : Democracy and civil violence, 2006. Polarization and Conflict Workshop, avril 2006.

N.P. Gleditsch, P. Wallensteen, M. Ericksson, M. Sollenberg et H. Strand : Armed conflict 1946-2001 : A new dataset. *Journal of Peace Research*, 39(5):615–637, 2002.

J. Goldstone, T.R. Gurr, B. Harff, M.A. Levy, M.G. Marshall, R.H. Bates, D. Epstein, C.H. Kahl, P.T. Surko, J.C. Ulfelder et A.N. Unger : State Failure Task Force report : Phase III findings. Rapport technique, State Failure Task Force, 2000. http://globalpolicy.gmu.edu/pitf/EFTF%20Phase%20III%20Report%20Final.pdf.

M. Grabisch et P. Perny : Agrégation multicritère. *In* B. Bouchon-Meunier et C. Marsala, éditeurs : *Utilisations de la logique floue*. Hermès, 1999.

P. Gärdenfors : On the logic of relevance. *Synthese*, 37:351–367, 1978.

W.H. Greene : *Econometric Analysis*. Prentice Hall, 2003. Cinquième édition.

J.W. Grzymala-Busse et M. Hu : A comparison of several approaches to missing attribute values in data mining. *In RSCTC '00 : Revised Papers from the Second International Conference on Rough Sets and Current Trends in Computing*, pages 378–385. Springer-Verlag, 2001.

N. Guler, O.T. Yidiz, F. Gurgen, F. Varol et E. Alpaydin : Discriminant functions and decision tree induction techniques for antenatal fetal risk assessment. *In International Joint Conference on Neural Networks (IJCNN)*, volume 4, pages 2712–2717, 2001.

H. Guo, A.K. Nandi et L.B. Jack : Multi-class nonlinear feature extraction by genetic programming. *In World Congress of International Fuzzy Systems Association (IFSA)*, pages 1347–1352, 2005.

P. Gupta, D. Doermann et D. DeMenthon : Beam search for feature selection in automatic SVM defect classification. *In International Conference on Pattern Recognition*, pages 212–215, 2002.

T.R. Gurr : *Why Men Rebel*. Princeton University Press, 1971.

T.R. Gurr et B. Harff : Systematic early warning of humanitarian emergencies. *Journal of Peace Research*, 35(5):551–579, 1998.

I. Guyon et A. Elisseeff : Introduction to variable and feature selection. *Journal of Machine Learning Research (JMLR)*, (3):1157–1182, 2003.

M. Hall : *Correlation-Based Feature Subset Selection for Machine Learning*. Thèse de doctorat, Department of Computer Science, University of Waikato, 1999.

M. Hall : Correlation-based feature selection for discrete and numeric class machine learning. *In International Conference on Machine Learning (ICML)*, pages 359–366, 2000.

M. Hall et L.A. Smith : Feature subset selection : a correlation based filter approach. *In International Conference on Neural Information Processing and Intelligent Information Systems*, pages 855–858, 1997.

T. Hastie, R. Tibshirani et J. Friedman : *Elements of Statistical Learning*. Springer-Verlag, 2001.

H. Hegre, T. Ellingsen, S. Gates et N.P. Gleditsch : Toward a democratic peace ? democracy, political change, and civil war, 1816-1992. *American Political Science Review*, 95(1):33–48, 2001.

H. Hegre et N. Sambanis : Sensitivity analysis of empirical results on civil war onset. *Journal of Conflict Resolution*, 50(4):508–535, 2006.

C.S. Hendrix et S.M. Glaser : Trends and triggers : Climate change and civil conflict in sub-saharan Africa, 2005. Human and Security Climate Change international workshop, Oslo, 21-23 juin 2005, http://www.cicero.uio.no/humsec/papers/Hendrix\&Glaser.pdf.

A. Hämmerli, R. Gattiker et R. Weyermann : Conflict cooperation in an actors' network of chechenya based on event data. *Journal of Conflict Resolution*, 50(2):159–175, 2006.

S.J. Hong : Use of contextual information for feature ranking and discretization. *IEEE Transactions on Knowledge and Data Engineering*, 9:718–730, 1997.

D.L. Horowitz : *Ethnic Groups in Conflict*. University of California Press, 1985.

M. Hu, S.M. Salvucci et M.P. Cohen : Evaluation of some popular imputation algorithms. *In Section on Survey Research Methods*, pages 309–313, 2000. American Statistical Association.

D. Hull : Using statistical testing in the evaluation of retrieval experiments. *In ACM-SIGIR*, pages 329–338, 1993.

S.P. Huntington : The clash of civilizations. *Foreign Affairs*, 72(3), 1993.

L. Hyafil et R.L. Rivest : Constructing optimal binary decision trees is NP-complete. *Information Processing Letters*, 5(1):15–17, 1976.

I. Inza, P. Larranaga, R. Etxeberria et B. Sierra : Feature subset selection by bayesian network-based optimization. *Artificial Intelligence*, 123(1-2):157–184, 2000.

A.K. Jain et D. Zongker : Feature selection : Evaluation, application and small sample performance. *Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, 1997.

A. Jakulin et I. Bratko : Analyzing attribute dependencies. *In Practice of Knowledge Discovery in Databases (PKDD)*, pages 229–240, 2003.

A. Jakulin et I. Bratko : Testing the significance of attribute interactions. *In International Conference on Machine Learning (ICML)*, 2004.

G.H. John, R. Kohavi et K. Pfleger : Irrelevant features and the subset selection problem. *In International Conference on Machine Learning (ICML)*, pages 121–129, 1994.

M.I. Jordan, éditeur. *Learning in Graphical Models*. Kluwer Academic Publishers, 1998.

R.D. Kaplan : The coming anarchy. *The Atlantic Monthly*, 273(2):44–76, 1994.

H. Kim, G.H. Golub et H. Park : Missing value estimation for DNA microarray gene expression data : local least square. *Bioinformatics*, 21(2):187–198, 2005.

K. Kira et L. Rendell : A practical approach to feature selection. *In International Conference on Machine Learning (ICML)*, pages 249–256, 1992.

S. Kmenta, P. Fitch et K. Ishii : Advanced failure modes and effects analysis of complex processes. *In ASME Design Engineering Technical Conferences*, 1999.

R. Kohavi et G.H. John : Wrappers for feature selection. *In Artificial Intelligence, special issue on Relevance*, pages 273–324. Elsevier, 1997.

R. Kohavi, P. Langley et Yun Y. : The utility of feature weighting in nearest-neighbor algorithms. *In European Conference on Machine Learning (ECML)*, 1997.

D. Koller et M. Sahami : Toward optimal feature selection. *In International Conference on Machine Learning (ICML)*, pages 284–292, 1996.

I. Kononenko : Estimating attributes : Analysis and extensions of RELIEF. *In European Conference on Machine Learning (ECML)*, volume 784 de *Lecture Notes in Computer Science*, pages 171–182. Springer-Verlag, 1994.

I. Kononenko : On biases in estimating the multivalued attributes. *In International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1034–1040. Morgan Kaufmann, 1995.

I. Kononenko et S.J. Hong : Attribute selection for modelling. *Future Generation Computer Systems*, 13:181–195, 1997.

S.B. Kotsiantis et P.E. Pintelas : Hybrid feature selection instead of ensemble of classifiers in medical decision support. *In Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, pages 269–276, Perrugia, Italy, 2004.

T. Kuhn : *The Structure of Scientific Revolutions.* University of Chicago Press, 1970. Deuxième édition.

B. Lacina et N.P. Gleditsch : Monitoring trends in global combat : A new dataset of battle deaths. *European Journal of Population*, 21(2-3):145–166, 2005.

D.D. Laitin : Language policy and civil war. *In* P. van Parijs, éditeur : *Cultural Diversity versus Economic Solidarity.* Brussels : Deboeck Université, Francqui Scientific Library, 2004.

P. Langley : Machine learning as an experimental science. *Machine Learning*, 3(1):5–8, 1988.

P. Langley et W. Iba : Average-case analysis of a nearest neighbor algorithm. *In International Joint Conference on Artificial Intelligence (IJCAI)*, pages 113–117, 1993.

J. Lemoine, H. Benhadda et J. Ah-Pine : Classification non supervisée de documents hétérogènes : application au corpus 20 newsgroup. *In Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, Paris, France, 2006.

P. Leray et O. François : Réseaux bayésiens pour la classification - méthodologie et illustration dans le cadre du diagnostic médical. *Revue d'Intelligence Artificielle*, 18:169–193, 2004.

D.D. Lewis et W. Gale : A sequential algorithm for training text classifiers. *In ACM-SIGIR*, pages 3–12, 1994.

M.I. Lichbach, C. Davenport et D.A. Armstrong : Contingency, inherency, and the onset of civil war, 2004. Annual Meeting of the Midwest Political Association, avril 2004, http://www.bsos.umd.edu/gvpt/davenport/dcawcp/paper/ContingencyInherencyandOnset.pdf.

A. Linder et C. Santiso : Assessing the predictive power of country risk ratings and governance indicators. Working paper, Paul H Nitze School of Advanced International Studies (SAIS), John Hopkins University, USA, 2002. http://www.sais-jhu.edu/workingpapers/WP-02-02b.pdf.

R.J. Little et D.B. Rubin : *Statistical Analysis with Missing Data*. John Wiley and Sons, 2002. Deuxième édition.

H. Liu, H. Motoda et M. Dash : A monotonic measure for optimal feature selection. *In European Conference on Machine Learning (ECML)*, pages 319–327, 1998.

H. Liu, H. Motoda et L. Yu : Feature selection with selective sampling. *In International Conference on Machine Learning (ICML)*, pages 395–402, 2002.

H. Liu et R. Setiono : Dimensionality reduction via discretization. *Knowledge-Based Systems*, 9:67–72, 1996a.

H. Liu et R. Setiono : Feature selection and classification : A probabilistic approach. *Knowledge-Based Systems*, 9:67–72, 1996b.

H. Liu et R. Setiono : A probabilistic approach for feature selection : A filter solution. *In International Conference on Machine Learning (ICML)*, pages 319–327, 1996c.

H. Liu et L. Yu : Feature selection for data mining. Survey draft, Department of computer Science and Engineering, Arizona State University, 2002. http://www.public.asu.edu/~huanliu/sur-fs02.ps.

H. Liu et L. Yu : Feature selection for high dimensional data : A fast correlation-based filter solution. *In International Conference on Machine Learning (ICML)*, pages 856–863, 2003.

H. Liu et L. Yu : Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491–502, 2005.

M. Magnani : Techniques for dealing with missing data in knowledge discovery tasks. Research report, University of Bologna, Computer Science Department, 2003. http://magnanim.web.cs.unibo.it/data/pdf/missingdata.pdf.

K.F. Man, K.S. Tang et S. Kwong : *Genetic Algorithms*. Springer, 1999.

R. Marchal et C. Messiant : De l'avidité des rebelles, l'analyse économique de la guerre civile selon Paul Collier. *Critique internationale (Paris)*, (16):58–69, 2002.

J.L. Marichal : Behavioural analysis of aggreagation in multicriteria decision. *In* J. Fodor, B. De Baets et P. Perny, éditeurs : *Preferences and Decisions under incomlete Knowledge*, volume 51 de *Studies in Fuziness and Soft Computing*. Physica Verlag, Heidelberg, 2000.

C. Marsala : *Apprentissage inductif en présence de données imprécises : construction et utilisation d'arbres de décision flous*. Thèse de doctorat, Université Paris VI Pierre et Marie Curie, 1998.

C. Marsala et B. Bouchon-Meunier : Fuzzy partitioning using mathematical morphology in a learning scheme. *In IEEE Conference on Fuzzy Systems*, volume 2, pages 1512–1517, 1996.

Z. Michalewicz : *Genetic Algorithms + Data Structure = Evolution Programs*. Springer, 1996.

E. Miguel, S. Satyanath et E. Sergenti : Economic schocks and civil conflict : An instrumental variables approach. *Journal of Political Economy*, 112(4):725–753, 2004.

A.J. Miller : *Subset Selection in Regression*. Chapman & Hall/CRC, 2001. Deuxième édition.

T. Mitchell : *Machine Learning*. McGraw Hill, 1997.

L.C. Molina, L. Belanche et A. Nebot : Feature selection algorithms : A survey and experimental evaluation. *In IEEE International Conference on Data Mining*, pages 306–313, 2002.

W.H. Moore et T.R. Gurr : Assessing the risks of ethnorebellion in the year 2000 : Three empirical approaches. *In* S. Schmeidl et H. Adelman, éditeurs : *Early Warning and Early Response*. Columbia University Press, 1998.

M. Morita, R. Sabourin, F. Bortolozzi et C.Y. Yuen : Unsupervised feature selection using multi-objective genetic algorithms for handwritten word recognition. *In International Conference on Document Analysis and Recognition*, pages 666–670, 2003.

L. Mouillet : *Modélisation, apprentissage et reconnaissance de scénarios de conflits ethno-politiques*. Thèse de doctorat, Université Paris VI Pierre et Marie Curie, 2005.

F. Moutou, B. Dufour et Y. Ivanov : A qualitative assessment of the risk of introducing foot and mouth disease into russia and europe from georgia, armenia and azerbaijan. *Revue scientifique et technique de l'office international des épizooties*, 20(3):723–730, 2001.

P. Narendra et K. Fukunaga : A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computer*, 26(9):917–922, 1977.

A.Y. Ng : On feature selection : Learning with exponentially many irrelevant features as training examples. *In International Conference on Machine Learning (ICML)*, pages 404–412. Morgan Kaufmann, 1998.

S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara et S. Ishii : A bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19 (16):2088–2096, 2003.

S.P. O'Brien : Anticipating the good, the bad, and the ugly : An early warning approach to conflict and instability analysis, 1975-2015. *Journal of Conflict Resolution*, 46(6), 2001.

I. Oh, J. Lee et B. Moon : Hybrid genetic algorithms for feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11):1424–1437, 2004.

C. Olaru et L. Wehenkel : Bias-variance tradeoff of soft decision trees. *In Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, pages 385–392, Perrugia, Italy, 2004.

Y. Papadopoulos, D. Parker et C. Grante : Automating the failure modes and effects analysis of safety critical systems. *In IEEE International Symposium on High Assurance Systems Engineering (HASE'04)*, 2004.

T.R. Payne et P. Edwards : Implicit feature selection with the value difference metric. *In European Conference on Artificial Intelligence*, pages 450–454, 1998.

J. Pearl : *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference.* Morgan Kaufmann, 1988.

M. Pechenizkiy, S. Puuronen et A. Tsymbal : Feature extraction for classification in knowledge discovery systems. *In International Conference on Knowledge-Based Intelligent Information and Engineering Systems*, pages 526–532, 2003.

H. Peng, F. Long et C. Ding : Feature selection based on mutual information : Criteria of max-dependency, max-relevance and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(27):1226–1238, 2005.

P. Perner : Improving the accuracy of decision tree induction by feature pre-selection. *Applied Artificial Intelligence*, 15(8):747–760, 2001.

J. Petrark, R. Trappl et J. Fürnkranz : The possible contribution of AI to the avoidance of crises and wars : Using CBR methods with the KOSIMO database of conflicts. Technical report tr 94-32, Austrian Research Insitute for Artificial Intelligence, 1994. http://citeseer.ifi.unizh.ch/petrak94possible.html.

C. Picard : *Graphes et questionnaires.* Gauthier-Villars, 1972.

J.L. Piermay : Nouvelles frontières ? *Outre Terre*, (11):57–71, 2005.

W.D. Potter, X. Deng, J. Li, M. Xu, Y. Wei, I. Lappas, M.J. Twery et D.J. Bennett : A web-based expert system for gypsy moth risk assessment. *Computers and Electronics in Agriculture*, 27(1):95–105, 2000.

W.H. Press, S.A. Teukolsky, W.T. Vetterling et B.P. Flannery : *Numerical Recipes in C++, The Art of Computing.* Cambridge University Press, 2002. Deuxième édition.

F.J. Provost et P. Domingos : Tree induction for probability-based ranking. *Machine Learning*, 52(3):199–215, 2003.

P. Pudil, F. Ferri, J. Novovicova et J. Kittler : Floating search methods for feature selection with nonmontonic criterion functions. *In International Conference on Computer Visison & Image Processing*, volume 2 de *Pattern Recognition*, pages 279–283, 1994.

J.R. Quinlan : Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

J.R. Quinlan : *C4.5 Programs for Machine Learning.* Morgan Kaufmann, 1993.

P. Radivojac, Z. Obradovic, A.K. Dunker et S. Vucetic : Feature selection filters based on the permutation test. *In European Conference on Machine Learning (ECML)*, volume 3201 de *Lecture Notes in Computer Science*, pages 334–346. Springer-Verlag, 2004.

A. Ragel et B. Crémilleux : Treatment of missing values for association rules. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 258–270, 1998.

A. Ragel et B. Crémilleux : Mvc - a preprocessing method to deal with missing values. *Knowledge-Based Systems*, 12(5):285–291, 1999.

B. Raman et T.R. Ioerger : Instance based filter for feature selection. *Journal of Machine Learning Research (JMLR)*, 1:1–23, 2002.

M.L. Raymer, W.F. Punch, E.D. Goodman et L.A. Kuhn : Genetic programming for improved data mining - application to the biochemistry of protein interactions. *In Conference on Genetic Programming*, pages 375–380, 1996.

M.L. Raymer, W.F. Punch, E.D. Goodman, L.A. Kuhn et A.K. Jain : Dimensionality reduction using genetic algorithms. *IEEE Transactions on Evolutionnary Computation*, 4:164–171, 2000.

P. Refaeilzadeh, L. Tang et H. Liu : On comparison of feature selection algorithms. *In Association for the Advancement of Artificial Intelligence (AAAI), Workshop on Evaluation Methods for Machine Learning II*, 2007.

A. Renteria et R. Tanscheit : Comparison of feature selection algorithms based on information theory. *In IFSA World Congress*, pages 1414–1418, 2005.

J. Reunanen : Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Reseach*, 3:1371–1382, 2003.

S.J. Rhee et K. Ishii : Life cost-based FMEA using empirical data. *In Design Engineering Technical Conferences (DETC2003)*, 2003.

S. Salzberg : On comparing classifiers : Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1(3):317–328, 1997.

N. Sambanis : What is civil war ? : Conceptual and empirical complexities of an operational definition. *Journal of Conflict Resolution*, 48(6):814–858, 2004.

G. Saporta : *Probabilités, analyse des données et statistique*. Technip, 2006. Deuxième édition.

J.L. Schafer et J.W. Graham : Missing data : Our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002.

G. Schneider et N. Wiesehomeier : Ethnic polarization, potential conflict, and civil wars : Comment, 2006. Polarization and Conflict Workshop, avril 2006.

K. Schrader-Frechette : *Risk and Rationality*. University of California Press, 1991.

K. Schrader-Frechette : How some risk frameworks disenfranchise the public. *Risk, Health, Safety and Environment*, (8), 1997.

P. Schrodt : Pattern recognition of international crises using hidden markov models. *In* Diana Richards, éditeur : *Political Complexity : Nonlinear Models of Politics*, pages 296–328. Ann Arbor : University of Michigan Press, 2000.

M. Seban et R. Nock : Impact of learning set quality and size on decision tree performances. *International Journal of Computers, Systems and Signals*, pages 85–105, 2001.

J. Sepulveda-Sanchis, G. Camps-Valls, E. Soria-Olivas, S. Salcedo-Sanz, C. Bousono-Calzon, G. Sanz-Romero et J. Marrugat de la Iglesia : Support vector machines and genetic algorithms for detecting unstable angina. *Computers in Cardiology*, pages 413–416, 2002.

C.E Shannon : A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, 1948.

J. Sherrah, R.E. Bogner et A. Bouzerdoum : The evolutionary pre-processor : Automatic feature extraction for supervised classification using genetic programming. *In Conference on Genetic Programming*, pages 304–312, 1997.

W. Siedlecki et J. Sklansky : A note on genetic algorithms for large-scale feature selection. *IEEE Transactions on Computers*, 10:157–346, 1993.

S.K. Singhi et H. Liu : Feature subset selection bias for classification learning. *In International Conference on Machine Learning (ICML)*, pages 849–856, 2006.

H. Situngkir : On massive conflict : Macro-micro link. Working paper, Bandung Fe Institute, Computational Sociology Department, 2004. http://www.ekonofisika.com/bfi/2004d.pdf.

D. Skalak : Prototype and feature selection by sampling and random mutation hill climbing algorithms. *In International Conference on Machine Learning (ICML)*, pages 293–301, 1994.

P. Slovic : Perception of risk. *Science*, (236):280–285, 1987.

M.G. Smith et L. Bull : Genetic programming with a genetic algorithm for feature construction and selection. *Genetic Programming and Evolvable Machines*, 6(3):265–281, 2005.

Q. Song et M. Shepperd : A new imputation method for small software project data sets. *Journal of Systems and Software*, 80(1):51–62, 2007.

M. Spanos, G. Dounias, N. Matsatsinis et C. Zopounidis : A fuzzy knowledge-based decision aiding method for the assessment of financial risks. *In European Symposium on Intelligent Techniques (ESIT'99)*, pages 1–7, 1999.

S.D. Stearns : On selecting features for pattern recognition. *In International Conference on Pattern Recognition*, pages 71–75, 1976.

D. Straub : Natural hazard risk assessment using bayesian networks. *In International Conference on Structural Safety and Reliability (ICOSSAR'2005)*, pages 2509–2516, 2005.

S. Supattathum, S. Olejnik et J. Li : Statistical power of modified Bonferroni methods. Annual Meeting of the American Educational Research Association, avril 1994.

P.B. Thompson et W.R. Dean : Competing conceptions of risk. *Risk, Health, Safety and Environment*, (7):361–384, 1996.

H. Timm, C. Döring et R. Kruse : Differentiated treatment of missing values in fuzzy clustering. *In International Fuzzy Systems Association World Congress*, volume 2715, pages 354–361, Istanbul, Turkey, 2003. Springer-Verlag.

R. Trappl, J. Fürnkranz et J. Petrark : Digging for peace : Using machine learning methods for assessing international conflict databases. *In European Conference on Artificial Intelligence*, pages 453–457, 1996.

G.V. Trunk : A problem of dimensionality : A simple example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(3):306–307, 1977.

P. Utgoff et J. Clouse : A Kolmogorov-Smirnov metric for decision-tree induction. Technical report, Department of Computer Science, University of Massachusetts, 1996. http://www.citeseer.comp.nus.edu.sg/84811.html.

H. Vafaie et F. Imam : Feature selection methods : Genetic algorithms vs. greedy-like search. *In International Conference on Fuzzy and Intellignet Control Systems*, 1994.

V.N. Vapnik : *The Nature of Statistical Learning.* Springer, 1995.

H. Wang et D.A. Bell : Axiomatic approach to feature subset selection based on relevance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(3):271–277, 1999.

M.D. Ward et K. Bakke : Predicting civil conflicts : On the utility of empirical research, 2005. Conference on Disaggregating the Study of Civil War and Transnational Violence, mai 2005.

L. Wehenkel et M. Pavella : Decision trees and transient stability of electric power systems. *Automatica*, 27(1):115–134, 1991.

I.H. Witten et E. Frank : *Data Mining : Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2005. Deuxième édition.

D.H. Wolpert et W.G. Macready : No free-lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–87, 1997.

Y. Wu et A. Zhang : Feature selection for classifying high-dimensional numeric data. *In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, volume 2, pages 251–258, 2004.

E. Xing, M. Jordan et R. Karp : Feature selection for high-dimensional genomic microarray data. *In International Conference on Machine Learning (ICML)*, pages 601–608, 2001.

J. Xuan, Y. Dong, J. Kahn, E. Hoffman, R. Clarke et Y. Wang : Robust feature selection by weighted Fisher criterion for multiclass prediction in gene expression profiling. *In International Conference on Pattern Recognition (ICPR)*, pages 291–294, 2004.

J. Yang et V. Hononvar : Feature subset selection using a genetic algorithm. *In* H. Liu et H. Motoda, éditeurs : *Feature Extraction, Construction, Subset Selection : A Data Mining Perspective*, pages 117–136. Kluwer Academic Publishers, 1998.

L. Yu et H. Liu : Efficient feature selection via analysis of relevance and redundancy. *Oxford Economic Papers*, 5:1205–1224, 2004.

B. Zadrozny et C. Elkan : Obtaining calibrated probability estimates from decision tress and naive bayesian classifiers. *In International Conference on Machine Learning (ICML)*, pages 609–616, 2001.

J. Zar : *Biostatistical Analysis.* Prentice Hall, 1999. Quatrième édition.

C. Zepeda Sein : Méthode d'évaluation des risques zoosanitaires lors d'échanges internationaux. *In* Office International des épizooties, éditeur : *Séminaire sur la sécurité zoosanitaire des échanges dans les Caraïbes*, pages 2–17. 1998.

Z. Zhao et H. Liu : Searching for interacting features. *In International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

R. Zimmerman et V.M. Bier : Risk assessment of extreme events, 2002. Columbia-Wharton/Penn Roundtable on Risk Management Strategies in an Uncertain World, IBM Palisades Executive Conference Center, Palisades, New-York, 12-13 avril 2002, http://www.ldeo.columbia.edu/chrr/documents/meetings/roundtable/white_papers/zimmerman_wp.pdf.

Y. Zou, A. An et X. Huang : Evaluation and automatic selection of methods for handling missing data. *In IEEE International Conference on Granular Computing*, 2005.

Annexe A

Notations

Description d'une base de données Une base de données est constituée d'un ensemble d'exemples décrits par un ensemble d'attributs. L'attribut que l'on cherche à modéliser par apprentissage supervisé est appelé classe. Ayant été influencé par des travaux appartenant au domaine de la fouille de données, mais aussi au domaine de la statistique, nous utilisons indifféremment dans notre manuscrit la terminologie propre à chacun de ces domaines. Ainsi nous évoquerons les trois concepts clés que nous venons d'introduire par les termes suivants :

- exemple : observation, instance
- attribut : variable explicative, variable indépendante
- classe : variable cible, variable dépendante, variable à expliquer

Après ce point terminologique, passons aux notations que nous avons employées dans ce manuscrit.

- e_i : i-ième exemple
- n : nombre d'observations.
- $\mathcal{E} = \{e_1, ..., e_n\}$: ensemble des n exemples constituant la base de données
- y : variable cible
- c_i : i-ième classe. On utilise également le terme classe pour désigner l'une des modalités de la variable *classe* y.
- K : nombre de classes
- $\mathcal{C} = \{c_1, ..., c_K\}$: ensemble des K modalités que peut prendre y
- y_i : classe de l'exemple e_i
- D_y : domaine de définition de la variable y
- $Classe$: fonction qui associe à chaque exemple de $\mathcal{E}$ sa classe.
  On a $Classe\,(e_i) = y_i$
- v_i : i-ième attribut
- p : nombre de variables.
- $\mathcal{V} = \{v_1, ..., v_p\}$: ensemble des p attributs utilisés pour décrire les exemples de la base de données.
- v_{ij} : valeur de v_j pour l'exemple e_i
- V : matrice représentant la base de données : l'élément (i, j) correspond à v_{ij}
- m_{ij} : j-ième modalité de l'attribut symbolique v_i
- n_i : nombre de modalités de l'attribut symbolique v_i
- $\mathcal{M}_i = \{m_{i1}, ..., m_{in_i}\}$: ensemble des n_i modalités que peut prendre l'attribut v_i
- n_{ij} : nombre d'exemples prenant la modalité m_{ij}
- $n_{ij}^{c_r}$: nombre d'exemples de la classe c_r qui prennent la modalité m_{ij}

- d_i : domaine de définition de l'attribut v_i
- g : fonction inconnue qui lie les variables explicatives à la variable cible. $y = g(v_1, ..., v_p)$
- f : fonction apprise automatiquement pour approximer g

Apprentissage automatique, statistique et logique floue

- Acc : taux de bonnes classifications
- $BalAcc$: Moyenne des taux de reconnaissance de chaque classe
- AUC : aire sous la courbe ROC. Nous utilisons la version développée dans Weka 3.4.7
- L : nombre minimum d'exemples que doit contenir un nœud pour pouvoir être partitionné lors de l'induction d'arbres de décision
- rappel (i) : taux de rappel de la classe c_i
- précision (i) : taux de précision de la classe c_i
- F-mesure (i) : F-mesure associée à la classe c_i. Lorsqu'aucune classe n'est spécifiée, la F-mesure est appliquée à la classe minoritaire.
- Agg_m : opérateur d'agrégation comportant m paramètres
- $\mathcal{N}(m, s)$: loi normale de moyenne m et d'écart-type s
- $\mathcal{U}(a, b)$: loi uniforme sur l'intervalle $[a; b]$
- P : probabilité
- E : espérance mathématique
- p_v : densité de probabilité associée à la variable aléatoire v
- F_v : fonction de répartition associée à la variable aléatoire v.
- $\mu(v)$: moyenne de la variable v pour l'ensemble de la population (inconnue en général)
- $\overline{v}$: moyenne empirique de la variable v
- $\sigma(v)$: écart-type de la variable v pour l'ensemble de la population (inconnue en général)
- $s(v)$: écart-type empirique de la variable v
- α : taux d'erreurs de type I dans un test statistique. Lorsque plusieurs tests sont effectués simultanément (comparaison multiple), α correspond au taux d'erreurs global (erreur de type I pour l'ensemble des tests).
- α^* : taux d'erreurs de comparaison par opposition au taux d'erreurs global. C'est l'erreur de type I pour l'un des tests d'une comparaison multiple.
- β : taux d'erreur de type II dans un test statistique. $1 - \beta$ est appelée *puissance* du test.
- I : entropie de Shannon
- δ_{KL} : divergence de Küllback-Leibler
- SU : incertitude symétrique
- δ_{KS} : distance de Kolmogorov-Smirnov
- $Dist(e_i, e_j)$: distance entre les exemples e_i et e_j dans l'espace de dimension p défini par $v_1, ..., v_p$
- $dist_l(e_i, e_j)$: distance élémentaire entre e_i et e_j, dans l'espace de dimension 1 défini par v_l
- P^* : probabilité d'événements flous
- $\top$: t-norme
- μ_A : fonction d'appartenance du sous-ensemble flou A

Comparaison de classifieurs

- k : nombre de classifieurs à comparer
- C_i : i-ième classifieur
- X_i : performance du i-ième classifieur
- X_{ij} : performance du i-ième classifieur, évaluée sur la j-ième base de données
- d : différence entre les performances de deux classifieurs.
- n : nombre de bases de données utilisées pour évaluer les performances de chaque classifieur. C'est le nombre d'observations des variables X_i.
- k_i : nombre d'exemples composant la i-ième base de données
- m : nombre de paires de bases (apprentissage,test) générées par base de données
- r_i^j : rang de la performance du classifieur C_j sur la i-ième base de données
- R_j : rang moyen des performances du classifieur C_j, estimé sur l'ensemble des n bases de données

Données manquantes

- ? : symbole indiquant que la valeur correspondante est manquante
- v^o : partie observée de l'attribut v
- v^m : partie manquante de l'attribut v
- $\hat{v}_{ij}$: valeur de substitution correspondant à la valeur manquante v_{ij}
- $\mathcal{E}_i^o$: ensemble des exemples pour lesquels la valeur de v_i est observée.
  On a $\mathcal{E}_i^o = \{e_j \in \mathcal{E}, v_{ji} \neq ?\}$
- n_i^o : nombre de valeurs observées de l'attribut v_i.
  On a $n_i^o = |\mathcal{E}_i^o|$
- n^o : nombre total de valeurs observées
- $\mathcal{E}_i^m$: ensemble des exemples pour lesquels la valeur de v_i est manquante.
  On a $\mathcal{E}_i^m = \{e_j \in \mathcal{E}, v_{ji} = ?\}$
- n_i^m : nombre de valeurs manquantes de l'attribut v_i.
  On a $n_i^m = |\mathcal{E}_i^m|$
- n^m : nombre total de valeurs manquantes
- S_i : ensemble des substitutions possibles pour l'attribut v_i
- Q : matrice indicatrice des valeurs manquantes. L'élément (i, j) de cette matrice vaut 1 si $v_{ij} =?$ et 0 sinon.

Sélection d'attributs

- d : nombre d'attributs que l'on souhaite sélectionner
- r : relation de pertinence
- $\overline{r}$: négation de la relation de pertinence
- $\mathcal{W} = \{w_1, ..., w_k\}$: ensemble de k attributs. Les majuscules calligraphiques sont utilisées pour désigner des ensembles.
- $W = (w_1, ..., w_k)$: variable aléatoire associée à la loi jointe des k variables aléatoires w_i. Une majuscule classique est utilisée pour distinguer la variable aléatoire jointe de l'ensemble d'attributs correspondant
- J : critère de performance d'un ensemble d'attributs qu'il convient de maximiser.
- $\mathcal{S}_i$: ensemble des variables autres que v_i.
  On a $\mathcal{S}_i = \mathcal{V} - \{v_i\}$
- m_{cor} : mesure générique de corrélation. Ce peut être une mesure de corrélation statistique, une mesure de divergence ou une mesure d'information.

- *pert* : mesure de la pertinence d'un ensemble d'attributs vis-à-vis de la classe y
- *red* : mesure de la redondance d'un ensemble d'attributs

Annexe B

Bases de données UCI

Substitution des valeurs manquantes Les tableaux[1] B.1 et B.2 décrivent respectivement les bases de données numériques et symboliques utilisées à la section 6.6.4

TAB. B.1 – Description des bases de données symboliques

Nom de la base	Nb attributs	Nb observations	Nb classes
Car Evaluation (*car*)	6	1728	4
Congressional Voting Records (*hv*)	16	435	2
Tic Tac Toe (*tic_ tac_ toe*)	9	958	2
Zoo (*zoo*)	16	100[1]	7
Promoter Gene Sequence (*promoters*)	57	106	2

[1] Il y a 101 observations, mais nous avons supprimé l'un des deux doublons *frog*.

TAB. B.2 – Description des bases de données numériques

Nom de la base	Nb attributs	Nb observations	Nb classes
Iris	4	150	3
Wine	13	178	3
Ionosphere	32[1]	351	2
Bupa	6	345	2
Pima Indians Diabetes	8	768	2
Breast Cancer	9	683[2]	2
Glass	9	214	2
Yeast	8	1484	10

[1] Il y a 34 attributs, mais nous avons supprimé les deux premiers attributs, suivant la suggestion faite par Acuna et Rodriguez (2004).
[2] Il y a 699 observations mais nous avons supprimé les 16 observations contenant des valeurs manquantes.

[1] Toutes les bases de données décrites dans cette annexe sont issues de l'*UCI Repository* http://www.ics.uci.edu/~mlearn/MLRepository.html

Sélection d'attributs Le tableau B.3 décrit les bases de données numériques utilisées à la section 7.6.2

TAB. B.3 – Description des bases de données sur lesquelles ont été comparés les filtres

Nom de la base	Nb attributs	Nb observations	Nb classes
Ionosphere	32[1]	351	2
WDBC	30	569	2
Waveform+noise	40	1000	3
Spam	57	4601	2
Yeast	166	6598	2

[1] Il y a 34 attributs, mais nous avons supprimé les deux premiers attributs, suivant la suggestion faite par Acuna et Rodriguez (2004).

Annexe C

Caractéristiques générales des bases de données étudiées

Les caractéristiques des 53 bases de données que nous avons constituées sont données dans les tableaux de cette annexe. Chacun de ces tableaux synthétise les informations relatives à une période d'estimation donnée. Les différentes bases de données sont décrites à l'aide des attributs suivants :

- groupe : nom du groupe pays auquel se réfère la base de données
- per. id : identifiant de la période d'analyse. Le nom du groupe, l'identifiant de la période d'analyse ainsi que la durée de la période d'estimation forment le triplet nécessaire à l'identification d'une base de données. Des exemples de tels triplets sont donnés à l'annexe G.
- attr. : nombre d'attributs
- obs. : nombre d'observations
- % manq. : taux de valeurs manquantes
- non-crise : nombre d'observations appartenant à la classe *non-crise*
- crise : nombre d'observations appartenant à la classe *crise*
- en-cours : nombre d'observations correspondant à un pays pour lequel une crise est en cours. Ces observations ont été exclues de notre base de données mais pourraient être ajoutées si l'on voulait étudier l'occurrence et non le déclenchement d'une crise.

La durée de la période d'estimation influe sur le choix des périodes d'analyse. Ainsi avec une période de 29 ans, les données étant disponibles de 1970 à 2002, pour chaque pays, nous ne pouvons construire que 2 observations, concernant les années 1999-2000 d'une part et 2001-2002 d'autre part. Avec cette période d'estimation, nous sommes donc contraints à n'avoir qu'une seule période d'analyse correspondant aux années 1999-2002.

Concernant la période d'estimation de 15 ans, nous avons défini deux périodes d'analyse.

- 0 : ensemble des observations antérieures à la fin de la Guerre froide. Pour chaque pays, nous avons 3 observations potentielles relatives aux années 1985-1986, 1987-1988 et 1989-1990.
- 1 : ensemble des observations postérieures à la fin de la Guerre froide. Pour chaque pays, nous avons 6 observations potentielles relatives aux années 1991-1992, 1993-1994 jusqu'à 2001-2002.

Disposant d'un plus grand nombre d'observations sur une période plus longue, nous avons défini 3 périodes d'analyse pour les périodes d'estimation de 1 et 7 ans en prenant

soin de construire des périodes d'analyse de telle sorte que puisse être étudiée l'influence de la fin de la Guerre froide sur le déclenchement des conflits.

Pour la période d'estimation de 7 ans, nous avons les périodes d'analyse suivantes :

- 0 : pour chaque pays, nous avons 2 observations potentielles relatives à la fin des années 70 : 1977-1978, 1979-1980.
- 1 : pour chaque pays, nous avons 5 observations potentielles relatives aux années 80 : 1981-1982, 1983-1984 jusqu'à 1989-1990. Les périodes d'analyse 0 et 1 regroupent l'ensemble des observations antérieures à la fin de la Guerre froide.
- 2 : ensemble des observations postérieures à la fin de la Guerre froide. Pour chaque pays, nous avons 6 observations potentielles concernant les années 1991-1992, 1993-1994 jusqu'à 2001-2002.

Pour la période d'estimation de 1 an, nous avons les périodes d'analyse suivantes :

- 0 : pour chaque pays, nous avons 4 observations potentielles relatives aux années 70 : 1971-1972, 1973-1974 jusqu'à 1977-1978.
- 1 : pour chaque pays, nous avons 6 observations potentielles relatives aux années 80 : 1979-1980, 1981-1982, jusqu'à 1989-1990. Les périodes d'analyse 0 et 1 regroupent l'ensemble des observations antérieures à la fin de la Guerre froide.
- 2 : ensemble des observations postérieures à la fin de la Guerre froide. Pour chaque pays, nous avons 6 observations potentielles concernant les années 1991-1992, 1993-1994 jusqu'à 2001-2002.

TAB. C.1 – Description des groupes de pays lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **1** année

groupe	per. id.	attr.	obs.	% manq.	non-crise	crise	en cours
Asie du Sud-Est, Pacifique	0	129	41	15	34	7	18
	1	253	75	15	67	8	20
	2	243	100	11	85	15	10
Europe de l'Est, Asie centrale	2	328	130	14	114	16	2
Amérique latine, Caraïbes	0	217	88	12	82	6	9
	1	301	164	13	147	17	19
	2	340	185	15	173	12	8
Proche-Orient, Afrique du Nord	0	160	60	12	54	6	4
	1	270	84	14	70	14	15
	2	321	93	18	83	10	15
Asie du Sud	0	176	20	12	13	7	4
	1	283	32	12	24	8	12
	2	330	32	15	24	8	15
Afrique subsaharienne	0	181	145	12	133	12	9
	1	266	227	12	199	28	28
	2	277	246	11	183	63	24
pays occidentaux	1	273	134	7	129	5	16
	2	342	148	9	142	6	8
global	0	98	487	7	447	40	54
	1	216	748	10	665	83	110
	2	307	939	15	808	131	82

TAB. C.2 – Description des groupes de pays lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **29** années

groupe	per. id.	attr.	obs.	% manq.	non-crise	crise	en cours
Europe de l'Est, Asie centrale	0	413	54	11	49	5	0
Afrique subsaharienne	0	810	77	11	55	22	11
global	0	801	272	13	240	32	28

TAB. C.3 – Description des groupes de pays lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **7** années

groupe	per. id.	attr.	obs.	% manq.	non-crise	crise	en cours
Asie du Sud-Est, Pacifique	0	278	27	17	22	5	5
	1	707	58	15	52	6	18
	2	688	98	11	83	15	10
Europe de l'Est, Asie centrale	2	708	110	13	99	11	2
Amérique latine, Caraïbes	1	815	133	12	119	14	17
	2	955	184	15	172	12	8
Proche-Orient, Afrique du Nord	0	545	31	16	24	7	3
	1	766	64	11	56	8	14
	2	902	91	17	82	9	15
Asie du Sud	1	771	27	12	19	8	10
	2	885	32	14	24	8	15
Afrique subsaharienne	0	611	66	13	60	6	8
	1	770	183	13	159	24	24
	2	820	243	12	179	64	24
pays occidentaux	2	932	146	8	140	6	8
global	0	313	243	11	213	30	29
	2	881	893	16	774	119	82

TAB. C.4 – Description des groupes de pays lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **15** années

groupe	per. id.	attr.	obs.	% manq.	non-crise	crise	en cours
Asie du Sud-Est, Pacifique	0	567	36	15	32	4	10
	1	682	97	11	82	15	10
Europe de l'Est, Asie centrale	1	365	132	14	116	16	2
Amérique latine, Caraïbes	0	820	75	11	65	10	11
	1	923	184	12	172	12	8
Proche-Orient, Afrique du Nord	1	850	89	16	81	8	15
Asie du Sud	1	876	32	14	24	8	15
Afrique subsaharienne	0	791	108	13	93	15	15
	1	800	240	10	177	63	24
pays occidentaux	1	861	144	6	138	6	8
global	0	624	353	11	312	41	59
	1	832	847	14	732	115	82

Annexe D

Liste des pays étudiés

Les tableaux de cette annexe décrivent l'ensemble des observations de chacune des 53 bases de données que nous avons constituées. 198 États dont le contexte structurel a été estimé entre 1970 et 2002 ont été considérés comme des observations potentielles. Chacun des tableaux ci-après se rapporte à l'un des huit groupes de pays que nous avons construits ainsi qu'à l'une des quatre périodes d'estimation que nous avons envisagées. Chaque État est décrit par les attributs suivants :

- pays : nom de l'État
- existence : intervalle temporel durant lequel l'État a fait partie du système international (source : *Correlates of War*)
- obs. id : nombre d'observations correspondant à cet État dans la base de données relative à la période d'analyse d'identifiant *id*.
- crise id : nombre d'observations correspondant à cet État et appartenant à la classe *crise* dans la base de données relative à la période d'analyse d'identifiant *id*.

Les périodes d'analyse diffèrent en fonction de la période d'estimation considérée. Leur description exacte est donnée à l'annexe C.

TAB. D.1 – Description des pays du groupe « Asie du Sud-Est, Pacifique » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **1** année

pays	existence	obs. 0	obs. 1	obs. 2	crise 0	crise 1	crise 2
Brunei	1984-2002			1			
Cambodia	1970-2002	2		2	2		
China	1970-2002	4	6	6		1	
East Timor	2002-2002						
Federated States of Micronesia	1991-2002						
Fiji	1970-2002	4	6	6			
Indonesia	1970-2002	3	2	5	1	2	4
Japan	1970-2002	4	6	6			
Kiribati	1999-2002						
Laos	1970-2002	2	3	6		1	
Malaysia	1970-2002	3	6	6	1	1	
Marshall Islands	1991-2002						
Mongolia	1970-2002	4	4	6			
Myanmar	1970-2002	1		5	1		5
Nauru	1999-2002						
North Korea	1970-2002	1					
Palau	1994-2002						
Papua New Guinea	1975-2002	1	6	4		1	1
Philippines	1970-2002	2	2	4	2	2	4
Republic of Vietnam	1970-1975						
Samoa	1976-2002	1	6	6			
Singapore	1970-2002	4	6	6			
Solomon Islands	1978-2002		6	6			
South Korea	1970-2002	4	6	6			
Taiwan	1970-2002						
Thailand	1970-2002		4	6			1
Tonga	1999-2002			1			
Tuvalu	2000-2002						
Vanuatu	1981-2002		4	6			
Vietnam	1970-2002	1	2	6			

TAB. D.2 – Description des pays du groupe « Europe de l'Est, Asie centrale » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **1** année

pays	existence	obs. 2	crise 2
Albania	1970-2004	6	
Armenia	1991-2004	5	
Azerbaijan	1991-2004	5	2
Belarus	1991-2004	5	
Bosnia and Herzegovina	1992-2004	2	
Bulgaria	1970-2004	6	
Croatia	1992-2004	5	2
Czech Republic	1993-2004	4	
Czechoslovakia	1970-1992		
Estonia	1991-2004	5	
Georgia	1991-2004	5	2
German Democratic Republic	1970-1990		
Hungary	1970-2004	6	
Kazakhstan	1991-2004	5	
Kyrgyzstan	1991-2004	5	
Latvia	1991-2004	5	
Lithuania	1991-2004	5	
Macedonia	1993-2004	4	1
Moldova	1991-2004	5	
Poland	1970-2004	6	
Romania	1970-2004	6	
Russia	1970-2004	5	4
Slovakia	1993-2004	4	
Slovenia	1992-2004	5	
Tajikistan	1991-2004	5	3
Turkmenistan	1991-2004	4	
Ukraine	1991-2004	5	
Uzbekistan	1991-2004	5	1
Yugoslavia	1970-2004	2	1

TAB. D.3 – Description des pays du groupe « Amérique latine, Caraïbes » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **1** année

pays	existence	obs. 0	obs. 1	obs. 2	crise 0	crise 1	crise 2
Antigua & Barbuda	1981-2002		4	6			
Argentina	1970-2002	3	6	6	2		
Bahamas	1973-2002	2	6	6			
Barbados	1970-2002	4	6	6			
Belize	1981-2002		4	6			
Bolivia	1970-2002	4	6	6			
Brazil	1970-2002	4	6	6			
Chile	1970-2002	4	6	6	1		
Colombia	1970-2002		2	3		2	3
Costa Rica	1970-2002	4	6	6			
Cuba	1970-2002			1			
Dominica	1978-2002		6	6			
Dominican Republic	1970-2002	4	6	6			
Ecuador	1970-2002	4	6	6			
El Salvador	1970-2002	4	2	6	1	2	1
Grenada	1974-2002	2	6	6			
Guatemala	1970-2002		1	5		1	2
Guyana	1970-2002	4	6	6			
Haiti	1970-2002	4	6	6			2
Honduras	1970-2002	4	6	6			
Jamaica	1970-2002	4	6	6			
Mexico	1970-2002	4	6	6			2
Nicaragua	1970-2002	4	3	6	1	3	
Panama	1970-2002	4	6	6		1	
Paraguay	1970-2002	4	6	6		1	
Peru	1970-2002	4	4	2		4	1
St. Kitts and Nevis	1983-2002		3	6			
St. Lucia	1979-2002		5	6			
St. Vincent and the Grenadines	1979-2002		5	6			
Suriname	1975-2002	1	5	6		1	
Trinidad and Tobago	1970-2002	4	6	6		1	
Uruguay	1970-2002	4	6	6	1		
Venezuela	1970-2002	4	6	6		1	1

TAB. D.4 – Description des pays du groupe « Proche-Orient, Afrique du Nord » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **1** année

pays	existence	obs. 0	obs. 1	obs. 2	crise 0	crise 1	crise 2
Algeria	1970-2002	4	6	3			3
Bahrain	1971-2002	3	6	6			
Djibouti	1977-2002		3	5			1
Egypt	1970-2002	4	6	3			1
Iran	1970-2002	4	5	4		5	2
Iraq	1970-2002	3			2		
Jordan	1970-2002	4	6	6			
Kuwait	1970-2002	4	6	6			
Lebanon	1970-2002	4	1	6	2	1	
Libya	1970-2002	4	6	5			
Morocco	1970-2002	3	2	6	2	2	
Oman	1971-2002	1	6	6			
Qatar	1971-2002	3	3	5			
Saudi Arabia	1970-2002	4	6	6		1	
Syria	1970-2002	4	6	6		2	
Tunisia	1970-2002	4	6	6		1	
Turkey	1970-2002	4	4	2		2	2
United Arab Emirates	1971-2002	3	6	6			
Yemen	1990-2002			6			1
Yemen Arab Republic	1970-1990						
Yemen People's Republic	1970-1990						

TAB. D.5 – Description des pays du groupe « Asie du Sud » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **1** année

pays	existence	obs. 0	obs. 1	obs. 2	crise 0	crise 1	crise 2
Afghanistan	1970-2002	4			1		
Bangladesh	1972-2002	1	1	5	1	1	1
Bhutan	1971-2002		4	6			
India	1970-2002	3	4	2	1	4	2
Maldives	1970-2002	1	6	6			
Nepal	1970-2002	4	6	5			2
Pakistan	1970-2002	3	6	6	3		1
Sri Lanka	1970-2002	4	5	2	1	3	2

TAB. D.6 – Description des pays du groupe « Afrique subsaharienne » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **1** année

pays	existence	obs. 0	obs. 1	obs. 2	crise 0	crise 1	crise 2
Angola	1975-2002		1	5		1	5
Benin	1970-2002	4	6	6			
Botswana	1970-2002	4	6	6			
Burkina Faso	1970-2002	4	6	6		1	
Burundi	1970-2002	4	6	4		1	3
Cameroon	1970-2002	4	6	6		1	1
Cape Verde	1975-2002	1	6	6			
Central African Republic	1970-2002	4	6	6			1
Chad	1970-2002		1	3		1	2
Comoros	1975-2002		5	6		1	1
Congo	1970-2002	4	6	6			4
Democratic Republic of the Congo	1970-2002	4	6	6	1	1	6
Equatorial Guinea	1970-2002	4	2	6			
Eritrea	1993-2002			4			
Ethiopia	1970-2002	3	2	4	3	2	3
Gabon	1970-2002	4	6	6			
Gambia	1970-2002	4	6	6		1	
Ghana	1970-2002	4	6	6		2	
Guinea	1970-2002	4	2	5			1
Guinea-Bissau	1974-2002	2	6	6			2
Ivory Coast	1970-2002	4	6	6			2
Kenya	1970-2002	4	6	6		1	
Lesotho	1970-2002	4	6	6			1
Liberia	1970-2002	4	6	5		2	4
Madagascar	1970-2002	4	6	6	1		
Malawi	1970-2002	4	6	6			
Mali	1970-2002	4	6	6		1	2
Mauritania	1970-2002	4	6	6			
Mauritius	1970-2002	4	6	6			
Mozambique	1975-2002		1	5		1	
Namibia	1990-2002			6			
Niger	1970-2002	4	6	5		1	3
Nigeria	1970-2002	4	6	6			1
Rwanda	1970-2002	4	6	6		1	6
Sao Tome and Principe	1975-2002		6	6			
Senegal	1970-2002	4	6	4		1	4
Seychelles	1976-2002	1	6	6			
Sierra Leone	1970-2002	4	6	5			4
Somalia	1970-2002	4	4		1	3	
Sudan	1970-2002	4	3	2	2	1	2
Swaziland	1970-2002	4	6	6			

TAB. D.7 – Description des pays du groupe « Afrique subsaharienne » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **1** année

pays	existence	obs. 0	obs. 1	obs. 2	crise 0	crise 1	crise 2
Tanzania	1970-2002	4	6	6			1
Togo	1970-2002	4	6	6		1	1
Uganda	1970-2002	4	3	3	2	3	3
Zambia	1970-2002	4	6	6			
Zimbabwe	1970-2002	2	5	6	2		

TAB. D.8 – Description des pays du groupe « pays occidentaux » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **1** année

pays	existence	obs. 1	obs. 2	crise 1	crise 2
Andorra	1993-2004				
Australia	1970-2004	6	6		
Austria	1970-2004	6	6		
Belgium	1970-2004	6	6		
Canada	1970-2004	6	6		
Cyprus	1970-2004	6	6		
Denmark	1970-2004	6	6		
Finland	1970-2004	6	6		
France	1970-2004	6	6		
German Federal Republic	1970-1990				
Germany	1990-2004		6		
Greece	1970-2004	6	6		
Iceland	1970-2004	6	6		
Ireland	1970-2004	6	6		
Israel	1970-2004				
Italy	1970-2004	6	6		
Liechtenstein	1990-2004				
Luxembourg	1970-2004	6	6		
Malta	1970-2004	6	6		
Monaco	1993-2004				
Netherlands	1970-2004	6	6		
New Zealand	1970-2004	6	6		
Norway	1970-2004	6	6		
Portugal	1970-2004	6	6		
San Marino	1992-2004				
South Africa	1970-2004	3	6	3	3
Spain	1970-2004	5	6	2	1
Sweden	1970-2004	6	6		
Switzerland	1970-2004	6	6		
United Kingdom	1970-2004		4		1
United States of America	1970-2004	6	6		1

TAB. D.9 – Description des pays du groupe « Asie du Sud-Est, Pacifique » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **7** années

pays	existence	obs. 0	obs. 1	obs. 2	crise 0	crise 1	crise 2
Brunei	1984-2002						
Cambodia	1970-2002	2		2	2		
China	1970-2002	2	5	6		1	
East Timor	2002-2002						
Federated States of Micronesia	1991-2002						
Fiji	1970-2002	2	5	6			
Indonesia	1970-2002	1	1	5	1	1	4
Japan	1970-2002	2	5	6			
Kiribati	1999-2002						
Laos	1970-2002	2	2	6		1	
Malaysia	1970-2002	2	5	6		1	
Marshall Islands	1991-2002						
Mongolia	1970-2002	2	3	6			
Myanmar	1970-2002			5			5
Nauru	1999-2002						
North Korea	1970-2002	2					
Palau	1994-2002						
Papua New Guinea	1975-2002	2	5	4		1	1
Philippines	1970-2002	2	1	4	2	1	4
Republic of Vietnam	1970-1975						
Samoa	1976-2002		5	6			
Singapore	1970-2002	2	5	6			
Solomon Islands	1978-2002		4	6			
South Korea	1970-2002	2	5	6			
Taiwan	1970-2002						
Thailand	1970-2002		4	6			1
Tonga	1999-2002						
Tuvalu	2000-2002						
Vanuatu	1981-2002		3	6			
Vietnam	1970-2002	2		6			

Tab. D.10 – Description des pays du groupe « Europe de l'Est, Asie centrale » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **7** années

pays	existence	obs. 2	crise 2
Albania	1970-2004	6	
Armenia	1991-2004	4	
Azerbaijan	1991-2004	4	1
Belarus	1991-2004	4	
Bosnia and Herzegovina	1992-2004	2	
Bulgaria	1970-2004	6	
Croatia	1992-2004	3	
Czech Republic	1993-2004	3	
Czechoslovakia	1970-1992		
Estonia	1991-2004	4	
Georgia	1991-2004	4	1
German Democratic Republic	1970-1990		
Hungary	1970-2004	6	
Kazakhstan	1991-2004	4	
Kyrgyzstan	1991-2004	4	
Latvia	1991-2004	4	
Lithuania	1991-2004	4	
Macedonia	1993-2004	3	1
Moldova	1991-2004	4	
Poland	1970-2004	6	
Romania	1970-2004	6	
Russia	1970-2004	4	3
Slovakia	1993-2004	3	
Slovenia	1992-2004	3	
Tajikistan	1991-2004	4	2
Turkmenistan	1991-2004	4	
Ukraine	1991-2004	4	
Uzbekistan	1991-2004	4	1
Yugoslavia	1970-2004	3	2

TAB. D.11 – Description des pays du groupe « Amérique latine, Caraïbes » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **7** années

pays	existence	obs. 1	obs. 2	crise 1	crise 2
Antigua & Barbuda	1981-2002	3	6		
Argentina	1970-2004	5	6		
Bahamas	1973-2004	5	6		
Barbados	1970-2004	5	6		
Belize	1981-2004	3	6		
Bolivia	1970-2004	5	6		
Brazil	1970-2004	5	6		
Chile	1970-2004	5	6		
Colombia	1970-2004	1	3	1	3
Costa Rica	1970-2004	5	6		
Cuba	1970-2004				
Dominica	1978-2004	4	6		
Dominican Republic	1970-2004	5	6		
Ecuador	1970-2004	5	6		
El Salvador	1970-2004	1	6	1	1
Grenada	1974-2004	5	6		
Guatemala	1970-2004	1	5	1	2
Guyana	1970-2004	5	6		
Haiti	1970-2004	5	6		2
Honduras	1970-2004	5	6		
Jamaica	1970-2004	5	6		
Mexico	1970-2004	5	6		2
Nicaragua	1970-2004	3	6	3	
Panama	1970-2004	5	6	1	
Paraguay	1970-2004	5	6	1	
Peru	1970-2004	3	2	3	1
St. Kitts and Nevis	1983-2004	2	6		
St. Lucia	1979-2004	4	6		
St. Vincent and the Grenadines	1979-2004	4	6		
Suriname	1975-2004	4	6	1	
Trinidad and Tobago	1970-2004	5	6	1	
Uruguay	1970-2004	5	6		
Venezuela	1970-2004	5	6	1	1

Tab. D.12 – Description des pays du groupe « Proche-Orient, Afrique du Nord » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **7** années

pays	existence	obs. 0	obs. 1	obs. 2	crise 0	crise 1	crise 2
Algeria	1970-2002	2	5	3			3
Bahrain	1971-2002	2	5	6			
Djibouti	1977-2002		1	5			1
Egypt	1970-2002	2	5	3			1
Iran	1970-2002	2	4	4	1	4	2
Iraq	1970-2002						
Jordan	1970-2002	2	5	6			
Kuwait	1970-2002	2	5	6			
Lebanon	1970-2002	2		5	2		
Libya	1970-2002	2	5	6			
Morocco	1970-2002	1	1	6	1	1	
Oman	1971-2002	2	5	6			
Qatar	1971-2002	2		5			
Saudi Arabia	1970-2002	2	5	6	1		
Syria	1970-2002	2	5	6	1	1	
Tunisia	1970-2002	2	5	6	1		
Turkey	1970-2002	2	3	2		2	2
United Arab Emirates	1971-2002	2	5	6			
Yemen	1990-2002			4			
Yemen Arab Republic	1970-1990						
Yemen People's Republic	1970-1990						

Tab. D.13 – Description des pays du groupe « Asie du Sud » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **7** années

pays	existence	obs. 1	obs. 2	crise 1	crise 2
Afghanistan	1970-2004				
Bangladesh	1972-2004	1	5	1	1
Bhutan	1971-2004	3	6		
India	1970-2004	4	2	4	2
Maldives	1970-2004	5	6		
Nepal	1970-2004	5	5		2
Pakistan	1970-2004	5	6		1
Sri Lanka	1970-2004	4	2	3	2

TAB. D.14 – Description des pays du groupe « Afrique subsaharienne » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **7** années

pays	existence	obs. 0	obs. 1	obs. 2	crise 0	crise 1	crise 2
Angola	1975-2002		1	5		1	5
Benin	1970-2002	2	5	6			
Botswana	1970-2002	2	5	6			
Burkina Faso	1970-2002	2	5	6		1	
Burundi	1970-2002	2	5	4		1	3
Cameroon	1970-2002	2	5	6		1	1
Cape Verde	1975-2002		5	6			
Central African Republic	1970-2002	2	5	6			1
Chad	1970-2002		1	3		1	2
Comoros	1975-2002		3	6		1	1
Congo	1970-2002	2	5	6			4
Democratic Republic of the Congo	1970-2002	2	5	6	1	1	6
Equatorial Guinea	1970-2002		1	6			
Eritrea	1993-2002			3			
Ethiopia	1970-2002	1	1	4	1	1	3
Gabon	1970-2002	2	5	6			
Gambia	1970-2002	2	5	6		1	
Ghana	1970-2002	2	5	6		2	
Guinea	1970-2002			5			1
Guinea-Bissau	1974-2002	1	5	6			2
Ivory Coast	1970-2002	2	5	6			2
Kenya	1970-2002	2	5	6		1	
Lesotho	1970-2002	2	5	6			1
Liberia	1970-2002	2	5	4	1	1	4
Madagascar	1970-2002	2	5	6			
Malawi	1970-2002	2	5	6			
Mali	1970-2002	2	5	6		1	2
Mauritania	1970-2002	2	5	6			
Mauritius	1970-2002	2	5	6			
Mozambique	1975-2002			5			
Namibia	1990-2002			4			
Niger	1970-2002	2	5	5		1	3
Nigeria	1970-2002	2	5	6			1
Rwanda	1970-2002	2	5	6		1	6
Sao Tome and Principe	1975-2002		4	6			
Senegal	1970-2002	2	5	4		1	4
Seychelles	1976-2002		5	6			
Sierra Leone	1970-2002	2	5	5			4
Somalia	1970-2002	2	3	1	1	3	1
Sudan	1970-2002	2	2	2		1	2
Swaziland	1970-2002	2	5	6			

TAB. D.15 – Description des pays du groupe « Afrique subsaharienne » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **7** années

pays	existence	obs. 0	obs. 1	obs. 2	crise 0	crise 1	crise 2
Tanzania	1970-2002	2	5	6			1
Togo	1970-2002	2	5	6		1	1
Uganda	1970-2002	2	2	3	2	2	3
Zambia	1970-2002	2	5	6			
Zimbabwe	1970-2002		5	6			

TAB. D.16 – Description des pays du groupe « pays occidentaux » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **7** années

pays	existence	obs. 2	crise 2
Andorra	1993-2004		
Australia	1970-2004	6	
Austria	1970-2004	6	
Belgium	1970-2004	6	
Canada	1970-2004	6	
Cyprus	1970-2004	6	
Denmark	1970-2004	6	
Finland	1970-2004	6	
France	1970-2004	6	
German Federal Republic	1970-1990		
Germany	1990-2004	4	
Greece	1970-2004	6	
Iceland	1970-2004	6	
Ireland	1970-2004	6	
Israel	1970-2004		
Italy	1970-2004	6	
Liechtenstein	1990-2004		
Luxembourg	1970-2004	6	
Malta	1970-2004	6	
Monaco	1993-2004		
Netherlands	1970-2004	6	
New Zealand	1970-2004	6	
Norway	1970-2004	6	
Portugal	1970-2004	6	
San Marino	1992-2004		
South Africa	1970-2004	6	3
Spain	1970-2004	6	1
Sweden	1970-2004	6	
Switzerland	1970-2004	6	
United Kingdom	1970-2004	4	1
United States of America	1970-2004	6	1

TAB. D.17 – Description des pays du groupe « Asie du Sud-Est, Pacifique » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **15** années

pays	existence	obs. 0	obs. 1	crise 0	crise 1
Brunei	1984-2002				
Cambodia	1970-2002		2		
China	1970-2002	3	6	1	
East Timor	2002-2002				
Federated States of Micronesia	1991-2002				
Fiji	1970-2002	3	6		
Indonesia	1970-2002	1	5	1	4
Japan	1970-2002	3	6		
Kiribati	1999-2002				
Laos	1970-2002	3	6	1	
Malaysia	1970-2002	3	6		
Marshall Islands	1991-2002				
Mongolia	1970-2002	2	6		
Myanmar	1970-2002		5		5
Nauru	1999-2002				
North Korea	1970-2002				
Palau	1994-2002				
Papua New Guinea	1975-2002	3	4	1	1
Philippines	1970-2002		4		4
Republic of Vietnam	1970-1975				
Samoa	1976-2002	3	6		
Singapore	1970-2002	3	6		
Solomon Islands	1978-2002	2	6		
South Korea	1970-2002	3	6		
Taiwan	1970-2002				
Thailand	1970-2002	3	6		1
Tonga	1999-2002				
Tuvalu	2000-2002				
Vanuatu	1981-2002	1	6		
Vietnam	1970-2002		5		

Tab. D.18 – Description des pays du groupe « Europe de l'Est, Asie centrale » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **15** années

pays	existence	obs. 1	crise 1
Albania	1970-2004	6	
Armenia	1991-2004	5	
Azerbaijan	1991-2004	5	2
Belarus	1991-2004	5	
Bosnia and Herzegovina	1992-2004	3	
Bulgaria	1970-2004	6	
Croatia	1992-2004	5	2
Czech Republic	1993-2004	4	
Czechoslovakia	1970-1992		
Estonia	1991-2004	5	
Georgia	1991-2004	4	1
German Democratic Republic	1970-1990		
Hungary	1970-2004	6	
Kazakhstan	1991-2004	5	
Kyrgyzstan	1991-2004	5	
Latvia	1991-2004	5	
Lithuania	1991-2004	5	
Macedonia	1993-2004	4	1
Moldova	1991-2004	5	
Poland	1970-2004	6	
Romania	1970-2004	6	
Russia	1970-2004	5	4
Slovakia	1993-2004	4	
Slovenia	1992-2004	5	
Tajikistan	1991-2004	5	3
Turkmenistan	1991-2004	4	
Ukraine	1991-2004	5	
Uzbekistan	1991-2004	5	1
Yugoslavia	1970-2004	4	2

TAB. D.19 – Description des pays du groupe « Amérique latine, Caraïbes » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **15** années

pays	existence	obs. 0	obs. 1	crise 0	crise 1
Antigua & Barbuda	1981-2002	1	6		
Argentina	1970-2002	3	6		
Bahamas	1973-2002	3	6		
Barbados	1970-2002	3	6		
Belize	1981-2002	1	6		
Bolivia	1970-2002	3	6		
Brazil	1970-2002	3	6		
Chile	1970-2002	3	6		
Colombia	1970-2002	1	3	1	3
Costa Rica	1970-2002	3	6		
Cuba	1970-2002				
Dominica	1978-2002	2	6		
Dominican Republic	1970-2002	3	6		
Ecuador	1970-2002	3	6		
El Salvador	1970-2002		6		1
Grenada	1974-2002	3	6		
Guatemala	1970-2002	1	5	1	2
Guyana	1970-2002	3	6		
Haiti	1970-2002	3	6		2
Honduras	1970-2002	3	6		
Jamaica	1970-2002	3	6		
Mexico	1970-2002	3	6		2
Nicaragua	1970-2002	1	6	1	
Panama	1970-2002	3	6	1	
Paraguay	1970-2002	3	6	1	
Peru	1970-2002	2	2	2	1
St. Kitts and Nevis	1983-2002		6		
St. Lucia	1979-2002	2	6		
St. Vincent and the Grenadines	1979-2002	2	6		
Suriname	1975-2002	2	6	1	
Trinidad and Tobago	1970-2002	3	6	1	
Uruguay	1970-2002	3	6		
Venezuela	1970-2002	3	6	1	1

TAB. D.20 – Description des pays du groupe « Proche-Orient, Afrique du Nord » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **15** années

pays	existence	obs. 1	crise 1
Algeria	1970-2004	3	3
Bahrain	1971-2004	6	
Djibouti	1977-2004	4	
Egypt	1970-2004	3	1
Iran	1970-2004	4	2
Iraq	1970-2004		
Jordan	1970-2004	6	
Kuwait	1970-2004	6	
Lebanon	1970-2004	6	
Libya	1970-2004	6	
Morocco	1970-2004	6	
Oman	1971-2004	6	
Qatar	1971-2004	5	
Saudi Arabia	1970-2004	6	
Syria	1970-2004	6	
Tunisia	1970-2004	6	
Turkey	1970-2004	2	2
United Arab Emirates	1971-2004	6	
Yemen	1990-2004	2	
Yemen Arab Republic	1970-1990		
Yemen People's Republic	1970-1990		

TAB. D.21 – Description des pays du groupe « Asie du Sud » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **15** années

pays	existence	obs. 1	crise 1
Afghanistan	1970-2004		
Bangladesh	1972-2004	5	1
Bhutan	1971-2004	6	
India	1970-2004	2	2
Maldives	1970-2004	6	
Nepal	1970-2004	5	2
Pakistan	1970-2004	6	1
Sri Lanka	1970-2004	2	2

TAB. D.22 – Description des pays du groupe « Afrique subsaharienne » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **15** années

pays	existence	obs. 0	obs. 1	crise 0	crise 1
Angola	1975-2002		4		4
Benin	1970-2002	3	6		
Botswana	1970-2002	3	6		
Burkina Faso	1970-2002	3	6	1	
Burundi	1970-2002	3	4	1	3
Cameroon	1970-2002	3	6		1
Cape Verde	1975-2002	3	6		
Central African Republic	1970-2002	3	6		1
Chad	1970-2002	1	3	1	2
Comoros	1975-2002	2	6	1	1
Congo	1970-2002	3	6		4
Democratic Republic of the Congo	1970-2002	3	6	1	6
Equatorial Guinea	1970-2002		6		
Eritrea	1993-2002		1		
Ethiopia	1970-2002	1	4	1	3
Gabon	1970-2002	3	6		
Gambia	1970-2002	3	6		
Ghana	1970-2002	3	6		
Guinea	1970-2002		5		1
Guinea-Bissau	1974-2002	3	6		2
Ivory Coast	1970-2002	3	6		2
Kenya	1970-2002	3	6		
Lesotho	1970-2002	3	6		1
Liberia	1970-2002	3	5	1	4
Madagascar	1970-2002	3	6		
Malawi	1970-2002	3	6		
Mali	1970-2002	3	6	1	2
Mauritania	1970-2002	3	6		
Mauritius	1970-2002	3	6		
Mozambique	1975-2002		5		
Namibia	1990-2002		2		
Niger	1970-2002	3	5	1	3
Nigeria	1970-2002	3	6		1
Rwanda	1970-2002	3	6	1	6
Sao Tome and Principe	1975-2002	2	6		
Senegal	1970-2002	3	4	1	4
Seychelles	1976-2002	3	6		
Sierra Leone	1970-2002	3	5		4
Somalia	1970-2002	2	2	2	1
Sudan	1970-2002		2		2
Swaziland	1970-2002	3	6		

TAB. D.23 – Description des pays du groupe « Afrique subsaharienne » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **15** années

pays	existence	obs. 0	obs. 1	crise 0	crise 1
Tanzania	1970-2002	3	6		1
Togo	1970-2002	3	6	1	1
Uganda	1970-2002	1	3	1	3
Zambia	1970-2002	3	6		
Zimbabwe	1970-2002	3	6		

TAB. D.24 – Description des pays du groupe « pays occidentaux » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **15** années

pays	existence	obs. 1	crise 1
Andorra	1993-2004		
Australia	1970-2004	6	
Austria	1970-2004	6	
Belgium	1970-2004	6	
Canada	1970-2004	6	
Cyprus	1970-2004	6	
Denmark	1970-2004	6	
Finland	1970-2004	6	
France	1970-2004	6	
German Federal Republic	1970-1990		
Germany	1990-2004	2	
Greece	1970-2004	6	
Iceland	1970-2004	6	
Ireland	1970-2004	6	
Israel	1970-2004		
Italy	1970-2004	6	
Liechtenstein	1990-2004		
Luxembourg	1970-2004	6	
Malta	1970-2004	6	
Monaco	1993-2004		
Netherlands	1970-2004	6	
New Zealand	1970-2004	6	
Norway	1970-2004	6	
Portugal	1970-2004	6	
San Marino	1992-2004		
South Africa	1970-2004	6	3
Spain	1970-2004	6	1
Sweden	1970-2004	6	
Switzerland	1970-2004	6	
United Kingdom	1970-2004	4	1
United States of America	1970-2004	6	1

TAB. D.25 – Description des pays du groupe « Europe de l'Est, Asie centrale » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **29** années

pays	existence	obs. 0	crise 0
Albania	1970-2002	2	
Armenia	1991-2002	2	
Azerbaijan	1991-2002	2	
Belarus	1991-2002	2	
Bosnia and Herzegovina	1992-2002	2	
Bulgaria	1970-2002	2	
Croatia	1992-2002	2	
Czech Republic	1993-2002	2	
Czechoslovakia	1970-1992		
Estonia	1991-2002	2	
Georgia	1991-2002	2	
German Democratic Republic	1970-1990		
Hungary	1970-2002	2	
Kazakhstan	1991-2002	2	
Kyrgyzstan	1991-2002	2	
Latvia	1991-2002	2	
Lithuania	1991-2002	2	
Macedonia	1993-2002	2	1
Moldova	1991-2002	2	
Poland	1970-2002	2	
Romania	1970-2002	2	
Russia	1970-2002	2	2
Slovakia	1993-2002	2	
Slovenia	1992-2002	2	
Tajikistan	1991-2002	2	
Turkmenistan	1991-2002	2	
Ukraine	1991-2002	2	
Uzbekistan	1991-2002	2	1
Yugoslavia	1970-2002	2	1

TAB. D.26 – Description des pays du groupe « Afrique subsaharienne » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **29** années

pays	existence	obs. 0	crise 0
Angola	1975-2002	2	2
Benin	1970-2002	2	
Botswana	1970-2002	2	
Burkina Faso	1970-2002	2	
Burundi	1970-2002	1	1
Cameroon	1970-2002	2	
Cape Verde	1975-2002	2	
Central African Republic	1970-2002	2	1
Chad	1970-2002		
Comoros	1975-2002	2	
Congo	1970-2002	2	2
Democratic Republic of the Congo	1970-2002	2	2
Equatorial Guinea	1970-2002	2	
Eritrea	1993-2002		
Ethiopia	1970-2002	1	1
Gabon	1970-2002	2	
Gambia	1970-2002	2	
Ghana	1970-2002	2	
Guinea	1970-2002	1	1
Guinea-Bissau	1974-2002	2	1
Ivory Coast	1970-2002	2	2
Kenya	1970-2002	2	
Lesotho	1970-2002	2	
Liberia	1970-2002	1	1
Madagascar	1970-2002	2	
Malawi	1970-2002	2	
Mali	1970-2002	2	
Mauritania	1970-2002	2	
Mauritius	1970-2002	2	
Mozambique	1975-2002	2	
Namibia	1990-2002		
Niger	1970-2002	2	
Nigeria	1970-2002	2	1
Rwanda	1970-2002	2	2
Sao Tome and Principe	1975-2002	2	
Senegal	1970-2002	1	1
Seychelles	1976-2002	2	
Sierra Leone	1970-2002	2	1
Somalia	1970-2002	2	2
Sudan	1970-2002		
Swaziland	1970-2002	2	

TAB. D.27 – Description des pays du groupe « Afrique subsaharienne » lorsque les moyenne, tendance et variabilité des attributs sont estimées sur **29** années

pays	existence	obs. 0	crise 0
Tanzania	1970-2002	2	1
Togo	1970-2002	2	
Uganda	1970-2002		
Zambia	1970-2002	2	
Zimbabwe	1970-2002	2	

Annexe E

Liste des variables utilisées

Cette annexe liste l'ensemble des 476 indicateurs de base que nous avons utilisés pour décrire le contexte structurel d'un État. Pour la période d'estimation de 1 an, ces indicateurs ont été pris tels quels, tandis que pour les périodes d'estimation plus longues (7, 15 ou 29 ans) leurs moyenne, tendance et variabilité ont été construites. Dans chacune de nos bases de données nous avons ainsi potentiellement $476 \times 3 = 1428$ attributs. Les tableaux suivants rappellent le nom de ces 476 indicateurs ainsi que leur source. L'ensemble des sources à partir desquelles nous avons réuni les différents indicateurs sont décrites à l'annexe F.

Tab. E.1 – Liste des variables collectées, et de la source associée

Number of battle related deaths	UCDP/PRIO
Involved in interstate conflict	UCDP/PRIO
Number of years in civil war	UCDP/PRIO
Number of years since last civil war	UCDP/PRIO
Surrounding countries in conflict	COW2
Number of years since autonomy	COW2
Number of years autonomy was lost	COW2
Percent area where population density < 2	CIESIN
Percent area where population density [2,5[	CIESIN
Percent area where population density [5,10[	CIESIN
Percent area where population density [10,15[	CIESIN
Percent area where population density [1000,10000[	CIESIN
Percent area where population density [10000,50000[	CIESIN
Percent area where population density > 50000	CIESIN
Percent area where elevation [800,1500[	CIESIN
Percent area where elevation [1500,3000[	CIESIN
Percent area where elevation [3000,5000[	CIESIN
Percent area where elevation > 5000	CIESIN
Percent population, in areas where population density < 2	CIESIN
Percent population, in areas where population density [2,5[	CIESIN
Percent population, in areas where population density [5,10[	CIESIN
Percent population, in areas where population density [10,15[	CIESIN
Percent population, in areas where population density [1000,10000[	CIESIN
Percent population, in areas where population density [10000,50000[	CIESIN
Percent population, in areas where population density > 50000	CIESIN
Percent population, in areas where elevation [800,1500[	CIESIN
Percent population, in areas where elevation [1500,3000[	CIESIN
Percent population, in areas where elevation [3000,5000[	CIESIN
Percent population, in areas where elevation > 5000	CIESIN
Total river boundaries to total land boundaries ratio	Toset
Number of shared rivers	Toset
Earthquakes Magnitude	USGS
Number of refugees originating the country (x1000)	USCRI
Number of internally displaced persons in the country (x1000)	USCRI
Number of refugees hosted by the country (x1000)	USCRI
Crisis Class	UCDP/PRIO
Herfindahl Ethnic Diversity Score	Fearon
Number of Ethnic Groups	Fearon
Dominant vs second ethnic group size ratio	Fearon
Coastline to Total Land Boundaries ratio	CIA World Factbook
Delta Altitude	CIA World Factbook
Herfindahl Religious Diversity Score	CIA World Factbook
Number of Religious Groups	CIA World Factbook
Dominant vs second religious group size ratio	CIA World Factbook

Tab. E.2 – Liste des variables collectées, et de la source associée

Age dependency ratio (dependents to working-age population)	Banque Mondiale
Agricultural machinery, tractors per 100 hectares of arable land	Banque Mondiale
Agricultural raw materials exports (% of merchandise exports)	Banque Mondiale
Agricultural raw materials imports (% of merchandise imports)	Banque Mondiale
Agriculture value added per worker (constant 1995 US$)	Banque Mondiale
Agriculture, value added (% of GDP)	Banque Mondiale
Agriculture, value added (annual % growth)	Banque Mondiale
Aid (% of central government expenditures)	Banque Mondiale
Aid (% of GNI)	Banque Mondiale
Aid (% of gross capital formation)	Banque Mondiale
Aid (% of imports of goods and services)	Banque Mondiale
Aid per capita (current US$)	Banque Mondiale
Air transport, freight (million tons per km)	Banque Mondiale
Air transport, passengers carried	Banque Mondiale
Aircraft departures	Banque Mondiale
Apparent intake rate in grade 1, female (% of relevant age group)	Banque Mondiale
Apparent intake rate in grade 1, male (% of relevant age group)	Banque Mondiale
Apparent intake rate in grade 1, total (% of relevant age group)	Banque Mondiale
Arms exports (constant 1990 US$)	Banque Mondiale
Arms imports (constant 1990 US$)	Banque Mondiale
Bank and trade-related lending (PPG + PNG) (NFL, current US$)	Banque Mondiale
Bank liquid reserves to bank assets ratio	Banque Mondiale
Birth rate, crude (per 1,000 people)	Banque Mondiale
Births attended by skilled health staff (% of total)	Banque Mondiale
Cable television subscribers (per 1,000 people)	Banque Mondiale
Capital expenditure (% of total expenditure)	Banque Mondiale
Central government debt, total (% of GDP)	Banque Mondiale
Cereal yield (kg per hectare)	Banque Mondiale
Changes in inventories (current US$)	Banque Mondiale
Changes in net reserves (BoP, current US$)	Banque Mondiale
Chemicals (% of value added in manufacturing)	Banque Mondiale
Claims on governments and other public entities (current LCU)	Banque Mondiale
Claims on governments, etc. (annual growth as % of M2)	Banque Mondiale
Claims on private sector (annual growth as % of M2)	Banque Mondiale
CO2 emissions (kg per 1995 PPP $ of GDP)	Banque Mondiale
CO2 emissions (metric tons per capita)	Banque Mondiale
Combustible renewables and waste (% of total energy)	Banque Mondiale
Commercial service exports (current US$)	Banque Mondiale
Commercial service imports (current US$)	Banque Mondiale
Communications, computer, etc. (% of service exports, BoP)	Banque Mondiale
Communications, computer, etc. (% of service imports, BoP)	Banque Mondiale
Computer, communications and other services (% of commercial service exports)	Banque Mondiale
Computer, communications and other services (% of commercial service imports)	Banque Mondiale
Consumer price index (1995 = 100)	Banque Mondiale

TAB. E.3 – Liste des variables collectées, et de la source associée

Container port traffic (TEU : 20 foot equivalent units)	Banque Mondiale
Contraceptive prevalence (% of women ages 15-49)	Banque Mondiale
Crop production index (1989-91 = 100)	Banque Mondiale
Current account balance (% of GDP)	Banque Mondiale
Current expenditure (current LCU)	Banque Mondiale
Current revenue, excluding grants (% of GDP)	Banque Mondiale
Current transfers, receipts (BoP, current US$)	Banque Mondiale
Daily newspapers (per 1,000 people)	Banque Mondiale
Death rate, crude (per 1,000 people)	Banque Mondiale
DEC alternative conversion factor (LCU per US$)	Banque Mondiale
Deposit interest rate (%)	Banque Mondiale
Discrepancy in expenditure estimate of GDP (constant LCU)	Banque Mondiale
Domestic credit provided by banking sector (% of GDP)	Banque Mondiale
Domestic credit to private sector (% of GDP)	Banque Mondiale
Domestic financing, total (% of GDP)	Banque Mondiale
Electric power consumption (kwh per capita)	Banque Mondiale
Electric power transmission and distribution losses (% of output)	Banque Mondiale
Electricity production (kwh)	Banque Mondiale
Electricity production from coal sources (% of total)	Banque Mondiale
Electricity production from hydroelectric sources (% of total)	Banque Mondiale
Electricity production from natural gas sources (% of total)	Banque Mondiale
Electricity production from nuclear sources (% of total)	Banque Mondiale
Electricity production from oil sources (% of total)	Banque Mondiale
Employees, agriculture, female (% of female employment)	Banque Mondiale
Employees, agriculture, male (% of male employment)	Banque Mondiale
Employees, industry, female (% of female employment)	Banque Mondiale
Employees, industry, male (% of male employment)	Banque Mondiale
Employees, services, female (% of female employment)	Banque Mondiale
Employees, services, male (% of male employment)	Banque Mondiale
Employment in agriculture (% of total employment)	Banque Mondiale
Employment in industry (% of total employment)	Banque Mondiale
Employment in services (% of total employment)	Banque Mondiale
Energy imports, net (% of commercial energy use)	Banque Mondiale
Energy production (kt of oil equivalent)	Banque Mondiale
Energy use (kg of oil equivalent per capita)	Banque Mondiale
Energy use (kt of oil equivalent)	Banque Mondiale
Energy use per PPP GDP (kg of oil equivalent per constant 1995 PPP $)	Banque Mondiale
Expenditure per student, primary (% of GDP per capita)	Banque Mondiale
Expenditure per student, secondary (% of GDP per capita)	Banque Mondiale
Expenditure per student, tertiary (% of GDP per capita)	Banque Mondiale
Expenditure, total (% of GDP)	Banque Mondiale
Export duties (% of tax revenue)	Banque Mondiale
Exports as a capacity to import (constant LCU)	Banque Mondiale
Exports of goods and services (% of GDP)	Banque Mondiale
Exports of goods and services (annual % growth)	Banque Mondiale

Tab. E.4 – Liste des variables collectées, et de la source associée

External balance on goods and services (% of GDP)	Banque Mondiale
External debt, total (DOD, current US$)	Banque Mondiale
Fertility rate, total (births per woman)	Banque Mondiale
Fertilizer consumption (100 grams per hectare of arable land)	Banque Mondiale
Final consumption expenditure, etc. (% of GDP)	Banque Mondiale
Final consumption expenditure, etc. (annual % growth)	Banque Mondiale
Financing from abroad (% of GDP)	Banque Mondiale
Fixed line and mobile phone subscribers (per 1,000 people)	Banque Mondiale
Food exports (% of merchandise exports)	Banque Mondiale
Food imports (% of merchandise imports)	Banque Mondiale
Food price index (1995 = 100)	Banque Mondiale
Food production index (1989-91 = 100)	Banque Mondiale
Food, beverages and tobacco (% of value added in manufacturing)	Banque Mondiale
Foreign direct investment, net inflows (% of GDP)	Banque Mondiale
Foreign labor force (% of total labor force)	Banque Mondiale
Foreign population (% of total population)	Banque Mondiale
Forest area (% of land area)	Banque Mondiale
Fuel exports (% of merchandise exports)	Banque Mondiale
Fuel imports (% of merchandise imports)	Banque Mondiale
GDP (constant 1995 US$)	Banque Mondiale
GDP deflator (base year varies by country)	Banque Mondiale
GDP growth (annual %)	Banque Mondiale
GDP per capita, PPP (constant 1995 international $)	Banque Mondiale
GDP per unit of energy use (constant 1995 PPP $ per kg of oil equivalent)	Banque Mondiale
General government final consumption expenditure (% of GDP)	Banque Mondiale
General government final consumption expenditure (annual % growth)	Banque Mondiale
GINI index	Banque Mondiale
GNI per capita, PPP (current international $)	Banque Mondiale
Goods and services expenditure (% of total expenditure)	Banque Mondiale
Goods exports (BoP, current US$)	Banque Mondiale
Goods imports (BoP, current US$)	Banque Mondiale
Gross capital formation (% of GDP)	Banque Mondiale
Gross capital formation (annual % growth)	Banque Mondiale
Gross domestic income (constant LCU)	Banque Mondiale
Gross domestic savings (% of GDP)	Banque Mondiale
Gross fixed capital formation (% of GDP)	Banque Mondiale
Gross fixed capital formation (annual % growth)	Banque Mondiale
Gross foreign direct investment (% of GDP)	Banque Mondiale
Gross national expenditure (% of GDP)	Banque Mondiale
Gross national savings, including NCTR (% of GDP)	Banque Mondiale
Gross national savings, including NCTR (% of GNI)	Banque Mondiale
Gross private capital flows (% of GDP)	Banque Mondiale
Gross value added at factor cost (constant 1995 US$)	Banque Mondiale
Health expenditure per capita (current US$)	Banque Mondiale
Health expenditure, private (% of GDP)	Banque Mondiale
Health expenditure, public (% of GDP)	Banque Mondiale

Tab. E.5 – Liste des variables collectées, et de la source associée

Health expenditure, total (% of GDP)	Banque Mondiale
Highest marginal tax rate, corporate rate (%)	Banque Mondiale
Highest marginal tax rate, individual (on income exceeding, US$)	Banque Mondiale
Highest marginal tax rate, individual rate (%)	Banque Mondiale
High-technology exports (% of manufactured exports)	Banque Mondiale
Hospital beds (per 1,000 people)	Banque Mondiale
Household final consumption expenditure per capita (constant 1995 US$)	Banque Mondiale
Household final consumption expenditure per capita growth (annual %)	Banque Mondiale
Household final consumption expenditure, etc. (% of GDP)	Banque Mondiale
Household final consumption expenditure, etc. (annual % growth)	Banque Mondiale
IBRD loans and IDA credits (PPG DOD, current US$)	Banque Mondiale
Immunization, DPT (% of children ages 12-23 months)	Banque Mondiale
Immunization, measles (% of children ages 12-23 months)	Banque Mondiale
Import duties (% of tax revenue)	Banque Mondiale
Imports of goods and services (% of GDP)	Banque Mondiale
Imports of goods and services (annual % growth)	Banque Mondiale
Improved sanitation facilities (% of population with access)	Banque Mondiale
Improved sanitation facilities, rural (% of rural population with access)	Banque Mondiale
Improved sanitation facilities, urban (% of urban population with access)	Banque Mondiale
Improved water source (% of population with access)	Banque Mondiale
Improved water source, rural (% of rural population with access)	Banque Mondiale
Improved water source, urban (% of urban population with access)	Banque Mondiale
Income payments (BoP, current US$)	Banque Mondiale
Income receipts (BoP, current US$)	Banque Mondiale
Income share held by fourth 20%	Banque Mondiale
Income share held by highest 10%	Banque Mondiale
Income share held by highest 20%	Banque Mondiale
Income share held by lowest 10%	Banque Mondiale
Income share held by lowest 20%	Banque Mondiale
Income share held by second 20%	Banque Mondiale
Income share held by third 20%	Banque Mondiale
Industry, value added (% of GDP)	Banque Mondiale
Industry, value added (annual % growth)	Banque Mondiale
Inflation, consumer prices (annual %)	Banque Mondiale
Inflation, food prices (annual %)	Banque Mondiale
Inflation, GDP deflator (annual %)	Banque Mondiale
Inflows of asylum seekers	Banque Mondiale
Inflows of foreign population	Banque Mondiale
Inflows of foreign workers	Banque Mondiale
Information and communication technology expenditure (% of GDP)	Banque Mondiale
Information and communication technology expenditure per capita (US$)	Banque Mondiale
Insurance and financial services (% of commercial service exports)	Banque Mondiale
Insurance and financial services (% of commercial service imports)	Banque Mondiale
Insurance and financial services (% of service exports, BoP)	Banque Mondiale
Insurance and financial services (% of service imports, BoP)	Banque Mondiale
Interest payments (% of current revenue)	Banque Mondiale

Tab. E.6 – Liste des variables collectées, et de la source associée

Interest payments (% of total expenditure)	Banque Mondiale
Interest rate spread (lending rate minus deposit rate)	Banque Mondiale
International telecom, outgoing traffic (minutes per subscriber)	Banque Mondiale
International tourism, expenditures (% of total imports)	Banque Mondiale
International tourism, number of arrivals	Banque Mondiale
International tourism, number of departures	Banque Mondiale
International tourism, receipts (% of total exports)	Banque Mondiale
Internet users (per 1,000 people)	Banque Mondiale
Labor force with primary education (% of total)	Banque Mondiale
Labor force with primary education, female (% of female labor force)	Banque Mondiale
Labor force with primary education, male (% of male labor force)	Banque Mondiale
Labor force with secondary education (% of total)	Banque Mondiale
Labor force with secondary education, female (% of female labor force)	Banque Mondiale
Labor force with secondary education, male (% of male labor force)	Banque Mondiale
Labor force with tertiary education (% of total)	Banque Mondiale
Labor force with tertiary education, female (% of female labor force)	Banque Mondiale
Labor force with tertiary education, male (% of male labor force)	Banque Mondiale
Labor force, children 10-14 (% of age group)	Banque Mondiale
Labor force, female (% of total labor force)	Banque Mondiale
Land area (sq km)	Banque Mondiale
Land use, arable land (% of land area)	Banque Mondiale
Land use, arable land (hectares per person)	Banque Mondiale
Land use, area under cereal production (hectares)	Banque Mondiale
Land use, irrigated land (% of cropland)	Banque Mondiale
Land use, permanent cropland (% of land area)	Banque Mondiale
Lending interest rate (%)	Banque Mondiale
Life expectancy at birth, female (years)	Banque Mondiale
Life expectancy at birth, male (years)	Banque Mondiale
Life expectancy at birth, total (years)	Banque Mondiale
Liquid liabilities (M3) as % of GDP	Banque Mondiale
Listed domestic companies, total	Banque Mondiale
Literacy rate, adult female (% of females ages 15 and above)	Banque Mondiale
Literacy rate, adult male (% of males ages 15 and above)	Banque Mondiale
Literacy rate, adult total (% of people ages 15 and above)	Banque Mondiale
Literacy rate, youth female (% of females ages 15-24)	Banque Mondiale
Literacy rate, youth male (% of males ages 15-24)	Banque Mondiale
Literacy rate, youth total (% of people ages 15-24)	Banque Mondiale
Livestock production index (1989-91 = 100)	Banque Mondiale
Living on less than $1 a day (PPP) (% of people)	Banque Mondiale
Living on less than $2 a day (PPP) (% of people)	Banque Mondiale
Long-term debt (DOD, current US$)	Banque Mondiale
Machinery and transport equipment (% of value added in manufacturing)	Banque Mondiale
Malnutrition prevalence, height for age (% of children under 5)	Banque Mondiale
Malnutrition prevalence, weight for age (% of children under 5)	Banque Mondiale
Manufactures exports (% of merchandise exports)	Banque Mondiale
Manufactures imports (% of merchandise imports)	Banque Mondiale

Tab. E.7 – Liste des variables collectées, et de la source associée

Manufacturing, value added (% of GDP)	Banque Mondiale
Manufacturing, value added (annual % growth)	Banque Mondiale
Market capitalization of listed companies (% of GDP)	Banque Mondiale
Merchandise exports (current US$)	Banque Mondiale
Merchandise imports (current US$)	Banque Mondiale
Military expenditure (% of central government expenditure)	Banque Mondiale
Military expenditure (% of GDP)	Banque Mondiale
Military personnel (% of total labor force)	Banque Mondiale
Military personnel, total	Banque Mondiale
Mobile phones (per 1,000 people)	Banque Mondiale
Money (current LCU)	Banque Mondiale
Money and quasi money (M2) as % of GDP	Banque Mondiale
Money and quasi money (M2) to gross international reserves ratio	Banque Mondiale
Money and quasi money growth (annual %)	Banque Mondiale
Mortality rate, adult, female (per 1,000 female adults)	Banque Mondiale
Mortality rate, adult, male (per 1,000 male adults)	Banque Mondiale
Mortality rate, infant (per 1,000 live births)	Banque Mondiale
Mortality rate, under-5 (per 1,000)	Banque Mondiale
Multilateral debt service (% of public and publicly guaranteed debt service)	Banque Mondiale
Net barter terms of trade (1995 = 100)	Banque Mondiale
Net capital account (BoP, current US$)	Banque Mondiale
Net current transfers (BoP, current US$)	Banque Mondiale
Net current transfers from abroad (current US$)	Banque Mondiale
Net domestic credit (current LCU)	Banque Mondiale
Net financial flows, IBRD (current US$)	Banque Mondiale
Net financial flows, IDA (current US$)	Banque Mondiale
Net financial flows, IMF concessional (current US$)	Banque Mondiale
Net financial flows, IMF nonconcessional (current US$)	Banque Mondiale
Net financial flows, RDB concessional (current US$)	Banque Mondiale
Net financial flows, RDB nonconcessional (current US$)	Banque Mondiale
Net foreign assets (current LCU)	Banque Mondiale
Net income (BoP, current US$)	Banque Mondiale
Net income from abroad (current US$)	Banque Mondiale
Net intake rate in grade 1 (% of official school-age population)	Banque Mondiale
Net intake rate in grade 1, female (% of official school-age population)	Banque Mondiale
Net intake rate in grade 1, male (% of official school-age population)	Banque Mondiale
Net taxes on products (current US$)	Banque Mondiale
Net trade in goods (BoP, current US$)	Banque Mondiale
Net trade in goods and services (BoP, current US$)	Banque Mondiale
Nontax revenue (% of current revenue)	Banque Mondiale
Official development assistance and official aid (current US$)	Banque Mondiale
Official exchange rate (LCU per US$, period average)	Banque Mondiale
Ores and metals exports (% of merchandise exports)	Banque Mondiale
Ores and metals imports (% of merchandise imports)	Banque Mondiale
Organic water pollutant (BOD) emissions (kg per day per worker)	Banque Mondiale

Tab. E.8 – Liste des variables collectées, et de la source associée

Other taxes (% of current revenue)	Banque Mondiale
Overall budget balance, including grants (% of GDP)	Banque Mondiale
Passenger cars (per 1,000 people)	Banque Mondiale
Patent applications, nonresidents	Banque Mondiale
Patent applications, residents	Banque Mondiale
Persistence to grade 5, female (% of cohort)	Banque Mondiale
Persistence to grade 5, male (% of cohort)	Banque Mondiale
Persistence to grade 5, total (% of cohort)	Banque Mondiale
Personal computers (per 1,000 people)	Banque Mondiale
Personal computers installed in education	Banque Mondiale
Physicians (per 1,000 people)	Banque Mondiale
Population ages 0-14 (% of total)	Banque Mondiale
Population ages 15-64 (% of total)	Banque Mondiale
Population ages 65 and above (% of total)	Banque Mondiale
Population density (people per sq km)	Banque Mondiale
Population density, rural (people per sq km)	Banque Mondiale
Population growth (annual %)	Banque Mondiale
Population, female (% of total)	Banque Mondiale
Population, total	Banque Mondiale
Portfolio investment, bonds (PPG + PNG) (NFL, current US$)	Banque Mondiale
Portfolio investment, equity (DRS, current US$)	Banque Mondiale
Portfolio investment, excluding LCFAR (BoP, current US$)	Banque Mondiale
Poverty gap at $1 a day (%)	Banque Mondiale
Poverty gap at $2 a day (%)	Banque Mondiale
Poverty headcount, national (% of population)	Banque Mondiale
Poverty headcount, rural (% of population)	Banque Mondiale
Poverty headcount, urban (% of population)	Banque Mondiale
PPG debt service (% of central government current revenue)	Banque Mondiale
PPP conversion factor to official exchange rate ratio	Banque Mondiale
Present value of debt (% of exports of goods and services)	Banque Mondiale
Present value of debt (% of GNI)	Banque Mondiale
Primary completion rate, female (% of relevant age group)	Banque Mondiale
Primary completion rate, male (% of relevant age group)	Banque Mondiale
Primary completion rate, total (% of relevant age group)	Banque Mondiale
Primary education, pupils (% female)	Banque Mondiale
Primary education, teachers (% female)	Banque Mondiale
Private capital flows, net total (DRS, current US$)	Banque Mondiale
Private investment in energy (current US$)	Banque Mondiale
Private investment in telecoms (current US$)	Banque Mondiale
Private investment in transport (current US$)	Banque Mondiale
Private investment in water and sanitation (current US$)	Banque Mondiale
Private nonguaranteed debt (% of external debt)	Banque Mondiale
Public and publicly guaranteed debt service (% of exports)	Banque Mondiale
Public and publicly guaranteed debt service (% of GNI)	Banque Mondiale
Public spending on education, total (% of GDP)	Banque Mondiale
Pump price for diesel fuel (US$ per liter)	Banque Mondiale

TAB. E.9 – Liste des variables collectées, et de la source associée

Pump price for super gasoline (US$ per liter)	Banque Mondiale
Pupil-teacher ratio, primary	Banque Mondiale
Purchasing power parity conversion factor (LCU per international $)	Banque Mondiale
Quasi money (current LCU)	Banque Mondiale
Quasi-liquid liabilities (% of GDP)	Banque Mondiale
Radios (per 1,000 people)	Banque Mondiale
Rail lines, electric (km)	Banque Mondiale
Rail lines, total (km)	Banque Mondiale
Rail traffic density (passengers and freight/km)	Banque Mondiale
Railway employee productivity (traffic units per employee)	Banque Mondiale
Ratio of commercial service exports to merchandise exports (%)	Banque Mondiale
Ratio of girls to boys in primary and secondary education (%)	Banque Mondiale
Ratio of rail passenger tariffs to freight tariffs	Banque Mondiale
Ratio of young literate females to males (% ages 15-24)	Banque Mondiale
Real effective exchange rate index (1995 = 100)	Banque Mondiale
Real interest rate (%)	Banque Mondiale
Repetition rate, primary (% of total enrollment)	Banque Mondiale
Repetition rate, primary, female (% of total enrollment)	Banque Mondiale
Repetition rate, primary, male (% of total enrollment)	Banque Mondiale
Research and development expenditure (% of GDP)	Banque Mondiale
Researchers in R&D (per million people)	Banque Mondiale
Risk premium on lending (%)	Banque Mondiale
Roads, goods transported (million ton-km)	Banque Mondiale
Roads, paved (% of total roads)	Banque Mondiale
Roads, total network (km)	Banque Mondiale
Royalty and license fees, payments (BoP, current US$)	Banque Mondiale
Royalty and license fees, receipts (BoP, current US$)	Banque Mondiale
Rural population (% of total population)	Banque Mondiale
Rural population growth (annual %)	Banque Mondiale
School enrollment, preprimary (% gross)	Banque Mondiale
School enrollment, primary (% gross)	Banque Mondiale
School enrollment, primary (% net)	Banque Mondiale
School enrollment, primary, female (% gross)	Banque Mondiale
School enrollment, primary, female (% net)	Banque Mondiale
School enrollment, primary, male (% gross)	Banque Mondiale
School enrollment, primary, male (% net)	Banque Mondiale
School enrollment, secondary (% gross)	Banque Mondiale
School enrollment, secondary (% net)	Banque Mondiale
School enrollment, secondary, female (% gross)	Banque Mondiale
School enrollment, secondary, female (% net)	Banque Mondiale
School enrollment, secondary, male (% gross)	Banque Mondiale
School enrollment, secondary, male (% net)	Banque Mondiale
School enrollment, tertiary (% gross)	Banque Mondiale
School enrollment, tertiary, female (% gross)	Banque Mondiale
School enrollment, tertiary, male (% gross)	Banque Mondiale
Scientific and technical journal articles	Banque Mondiale

Tab. E.10 – Liste des variables collectées, et de la source associée

Secondary education, pupils (% female)	Banque Mondiale
Service exports (BoP, current US$)	Banque Mondiale
Service imports (BoP, current US$)	Banque Mondiale
Services, etc., value added (% of GDP)	Banque Mondiale
Services, etc., value added (annual % growth)	Banque Mondiale
Short-term debt (% of total external debt)	Banque Mondiale
Social security taxes (% of current revenue)	Banque Mondiale
Stocks traded, total value (% of GDP)	Banque Mondiale
Stocks traded, turnover ratio (%)	Banque Mondiale
Subsidies and other current transfers (% of total expenditure)	Banque Mondiale
Surface area (sq km)	Banque Mondiale
Tax revenue (% of GDP)	Banque Mondiale
Taxes on goods and services (% of current revenue)	Banque Mondiale
Taxes on goods and services (% value added of industry and services)	Banque Mondiale
Taxes on income, profits and capital gains (% of current revenue)	Banque Mondiale
Taxes on income, profits and capital gains (% of total taxes)	Banque Mondiale
Taxes on international trade (% of current revenue)	Banque Mondiale
Technicians in R&D (per million people)	Banque Mondiale
Telephone faults (per 100 mainlines)	Banque Mondiale
Telephone mainlines (per 1,000 people)	Banque Mondiale
Telephone mainlines in largest city (per 1,000 people)	Banque Mondiale
Telephone mainlines per employee	Banque Mondiale
Telephone mainlines, waiting list	Banque Mondiale
Telephone revenue per mainline (current US$)	Banque Mondiale
Television sets (per 1,000 people)	Banque Mondiale
Terms of trade adjustment (constant LCU)	Banque Mondiale
Textiles and clothing (% of value added in manufacturing)	Banque Mondiale
Total debt service (% of exports of goods and services)	Banque Mondiale
Total debt service (% of GNI)	Banque Mondiale
Total reserves (includes gold, current US$)	Banque Mondiale
Total reserves in months of imports	Banque Mondiale
Total reserves minus gold (current US$)	Banque Mondiale
Trade (% of GDP)	Banque Mondiale
Trade in goods (% of GDP)	Banque Mondiale
Trade in goods (% of goods GDP)	Banque Mondiale
Trademarks, applications filed	Banque Mondiale
Trademarks, nonresidents	Banque Mondiale
Trademarks, residents	Banque Mondiale
Transport services (% of commercial service exports)	Banque Mondiale
Transport services (% of commercial service imports)	Banque Mondiale
Transport services (% of service exports, BoP)	Banque Mondiale
Transport services (% of service imports, BoP)	Banque Mondiale
Travel services (% of commercial service exports)	Banque Mondiale
Travel services (% of commercial service imports)	Banque Mondiale

TAB. E.11 – Liste des variables collectées, et de la source associée

Travel services (% of service exports, BoP)	Banque Mondiale
Travel services (% of service imports, BoP)	Banque Mondiale
Two-wheelers (per 1,000 people)	Banque Mondiale
Unemployment, female (% of female labor force)	Banque Mondiale
Unemployment, male (% of male labor force)	Banque Mondiale
Unemployment, total (% of total labor force)	Banque Mondiale
Unemployment, youth female (% of female labor force ages 15-24)	Banque Mondiale
Unemployment, youth male (% of male labor force ages 15-24)	Banque Mondiale
Unemployment, youth total (% of total labor force ages 15-24)	Banque Mondiale
Urban population (% of total)	Banque Mondiale
Urban population growth (annual %)	Banque Mondiale
Use of IMF credit (DOD, current US$)	Banque Mondiale
Vehicles (per 1,000 people)	Banque Mondiale
Vehicles (per km of road)	Banque Mondiale
Wages and salaries (% of total expenditure)	Banque Mondiale
Water pollution, chemical industry (% of total BOD emissions)	Banque Mondiale
Water pollution, clay and glass industry (% of total BOD emissions)	Banque Mondiale
Water pollution, food industry (% of total BOD emissions)	Banque Mondiale
Water pollution, metal industry (% of total BOD emissions)	Banque Mondiale
Water pollution, other industry (% of total BOD emissions)	Banque Mondiale
Water pollution, paper and pulp industry (% of total BOD emissions)	Banque Mondiale
Water pollution, textile industry (% of total BOD emissions)	Banque Mondiale
Water pollution, wood industry (% of total BOD emissions)	Banque Mondiale
Workers' remittances, receipts (BoP, current US$)	Banque Mondiale

Annexe F

Liste des sources utilisées

Nous détaillons dans cette annexe les sources qui nous ont permis de recueillir les indicateurs structurels listés à l'annexe E.

UCDP/PRIO Collaboration entre l'institut international de recherche sur la paix d'Oslo (*International Peace Research Institute, Oslo*) et le département de recherche sur la paix et les conflits de l'université d'Uppsala dans le cadre du programme sur les données relatives aux conflits de l'université d'Uppsala (*Uppsala Conflict Data Program*). Cette collaboration a pour but de recueillir des données relatives aux conflits armés. Nous avons employé la base de données nommée *Battle Deaths Dataset* version 1.0 (Lacina et Gleditsch, 2005), disponible à l'adresse suivante : http://new.prio.no/CSCW-Datasets/Data-on-Armed-Conflict/Battle-Deaths-Data2

COW2 Projet intitulé *Correlates of War* regroupant de nombreuses bases de données relatives aux conflits armés. Nous nous sommes servi des bases de données suivantes : *State System Membership* version 2004.1 et *Direct Contiguity* version 3.0. La première indique pour chaque État les dates d'entrée et sortie du système international tandis que la seconde indique pour chaque État ceux qui lui sont frontaliers. Nous avons utilisé ces informations pour créer la variable donnant le nombre de pays voisins en crise : *Surrounding countries in conflict*. Ces bases de données sont disponibles à l'adresse suivante : http://cow2.la.psu.edu

CIESIN Réseau international créé par la NASA (*National Aeronautics and Space Agency* pour collecter des informations géospatiales (CIESIN est le sigle correspondant à *Center for International Earth Science Information Network*). Nous avons travaillé à partir de la base *Population, Landscape, and Climate Estimates* (PLACE) pour construire des variables reflétant la répartition de la population sur le territoire, dans les zones d'altitude élevée, de très faible et très forte densité. Ces données sont disponibles à l'adresse suivante : http://sedac.ciesin.columbia.edu/plue/nagd/place

Toset H.P.W. Toset est à l'origine d'une base de données recensant les frontières fluviales. Nous avons employé la version de 1998 de cette base qui est disponible à l'adresse suivante : http://new.prio.no/CSCW-Datasets/Geographical-and-Resource-Datasets-/Shared-rivers

USGS Organisation scientifique américaine chargée d'étudier les questions et problèmes liés aux ressources naturelles et à l'environnement (USGS est le sigle correspondant à

United States Geological Survey). Nous avons récupéré les données collectées dans le cadre du programme sur les tremblements de terre (*Earthquake Hazards Program*) à propos des principaux tremblements de terre. Ces données sont disponibles à l'adresse suivante http://earthquake.usgs.gov/eqcenter/historic_eqs.php

USCRI *United States Committee for Refugees and Immigrants.* Les données fournies par cet organisme ont permis aux chercheurs du *Center for Systemic Peace*, sous la direction de M.G. Marshall, de construire une base recensant pour chaque pays, depuis 1964, le nombre de réfugiés accueillis, le nombre de personnes déplacées à l'intérieur du territoire ainsi que le nombre de réfugiés originaire de ce pays. Cette base se nomme *Forcibly Displaced Populations (1964-2002)* est disponible à l'adresse *www.systemicpeace.org/inscr.htm.*

Fearon Fearon (2005) a recensé les principaux groupes ethno-linguistiques composant un grand nombre de pays. Nous avons utilisé ces informations afin de construire l'indice de diversité ethnique. Ces données sont disponibles dans la base *egroupsrepdata*, disponible à l'adresse suivante : http://www.stanford.edu/group/ethnic/publicdata/publicdata.html

CIA World Factbook Synthèse du contexte structurel des pays du monde, remise à jour chaque année par la CIA. Nous avons utilisé les données de 2002 afin de récupérer la proportion de frontières maritimes, la différence entre l'altitude la plus élevée et la plus basse et le nombre de groupes religieux ainsi que la proportion de la population que chacun d'eux représente afin de construire l'indice de diversité religieuse. Le World Factbook de la CIA est disponible à l'adresse suivante : https://www.cia.gov/library/publications/the-world-factbook/

Banque Mondiale La majorité des indicateurs structurels dont nous disposons proviennent du CD-ROM de la Banque mondiale intitulé *World Development Indicators 2004*. Ce CD-ROM contient 581 indicateurs de développement estimés pour 208 pays entre 1960 et 2002.

Annexe G

Résultats de la sélection de modèles

Tab. G.1 – Méthodes sélectionnées pour chacune des bases de données relatives aux conflits armés intra-étatiques

Base de données	Méthode sélectionnée
1 Europe de l'Est, Asie centrale 2	(A) 1ppv SansFiltre
1 Asie du Sud-Est, Pacifique 0	(A) 5ppv KSF
1 Asie du Sud-Est, Pacifique 1	(B) EF-Entropie KSCBF
1 Asie du Sud-Est, Pacifique 2	(A) Moyenne KSF
1 Global 0	(C) 1LLSI KSF
1 Global 1	(A) Moyenne KSCBF
1 Global 2	(A) Médiane CFS
1 Amérique latine, Caraïbes 0	(B) Médiane KSF
1 Amérique latine, Caraïbes 1	(A) 1ppv FCBF
1 Amérique latine, Caraïbes 2	(B) EF-Entropie FCBF
1 Afrique du Nord, Proche-Orient 0	(B) EF-Entropie FCBF
1 Afrique du Nord, Proche-Orient 1	(A) 1ppv SansFiltre
1 Afrique du Nord, Proche-Orient 2	(B) EF-Entropie KSCBF
1 Asie du Sud 0	(A) 1LLSI SansFiltre
1 Asie du Sud 1	(B) EF-Entropie KSCBF
1 Asie du Sud 2	(C) CMoyenne KSCBF
1 Afrique subsaharienne 0	(A) 5ppv KSF
1 Afrique subsaharienne 1	(A) 5ppv SansFiltre
1 Afrique subsaharienne 2	(B) 5ppv CFS
1 pays occidentaux 1	(A) 5ppv KSF
1 pays occidentaux 2	(B) CMoyenneA CFS
15 Europe de l'Est, Asie centrale 1	(B) Médiane KSF
15 Asie du Sud-Est, Pacifique 1	(B) Moyenne KSF
15 Global 0	(A) 1ppv CFS
15 Global 1	(B) Moyenne CFS
15 Amérique latine, Caraïbes 0	(A) 1ppv FCBF
15 Amérique latine, Caraïbes 1	(C) Médiane KSF
15 Afrique du Nord, Proche-Orient 1	(B) 5ppv KSF

TAB. G.2 – Méthodes sélectionnées pour chacune des bases de données relatives aux conflits armés intra-étatiques

Base de données	Méthode sélectionnée
15 Asie du Sud 0	(C) Médiane KSF
15 Asie du Sud 1	(A) 5ppv CFS
15 Afrique subsaharienne 0	(B) CMoyenne CFS
15 Afrique subsaharienne 1	(C) 1ppv FCBF
15 pays occidentaux 1	(B) Médiane FCBF
29 Europe de l'Est, Asie centrale 0	(B) Médiane FCBF
29 Global 0	(A) Médiane KSF
29 Afrique subsaharienne 0	(A) AléatoireMM SansFiltre
7 Europe de l'Est, Asie centrale 2	(C) Médiane KSCBF
7 Asie du Sud-Est, Pacifique 0	(B) 5ppv KSF
7 Asie du Sud-Est, Pacifique 1	(A) CMoyenneA FCBF
7 Asie du Sud-Est, Pacifique 2	(A) 1LLSI KSF
7 Global 0	(C) 1ppv FCBF
7 Global 2	(B) Médiane KSCBF
7 Amérique latine, Caraïbes 1	(B) 5LLSI CFS
7 Amérique latine, Caraïbes 2	(B) AléatoireMM FCBF
7 Afrique du Nord, Proche-Orient 0	(C) 1ppv KSCBF
7 Afrique du Nord, Proche-Orient 1	(C) Médiane KSF
7 Afrique du Nord, Proche-Orient 2	(A) 1ppv FCBF
7 Asie du Sud 1	(C) Médiane KSF
7 Asie du Sud 2	(B) 5ppv KSF
7 Afrique subsaharienne 0	(B) Médiane KSF
7 Afrique subsaharienne 1	(C) 5ppv KSCBF
7 Afrique subsaharienne 2	(C) 5LLSI KSCBF
7 pays occidentaux 2	(B) EF-Entropie CFS
15 Asie du Sud 1	(A) 5ppv CFS
15 Afrique subsaharienne 0	(B) CMoyenne CFS
15 Afrique subsaharienne 1	(C) 1ppv FCBF
15 pays occidentaux 1	(B) Médiane FCBF
29 Europe de l'Est, Asie centrale 0	(B) Médiane FCBF
29 Global 0	(A) Médiane KSF
29 Afrique subsaharienne 0	(A) AléatoireMM SansFiltre

Index des sigles et noms de méthodes

AB&B, 125
ACP, 123
AD, 85
AG, 132
AléatoireHD, 80
AléatoireMM, 80
ALLSI, 99
ANOVA, 62
APV, 80

B&B, 125
BBS, 131
BFBS, 130
BFFS, 130
BFS, 131
B_g, 140

C4.5, 34
CAléatoireHD, 80
CAléatoireMM, 80
CAPV, 81
CART, 34
CD, 75
CFS, 130
Ckppv, 82
CkppvI, 82
CLIP4, 85
clustering, 82
CM, 149
CMédiane, 79
CMode, 79
CMoyenne, 79
CMoyenneA, 80
CorrFS, 175

DDP, 99
Dunn-Bonferroni, 64

EF, 99
EM, 86
Entropie, 89
EW, 99

FASE, 26
FCBF, 157
FOCUS, 125
FWER, 62

G, 34
GR, 34

Holland-Copenhaver, 65
Hot Deck, 80

IB1, 99
ID3, 34, 99
IM, 119

J48, 99

kppv, 39, 81
KSCBF, 159
KSF, 159

LLS, 84
LLSI, 99
LQR, 84
LR, 84
LV, 132

MAR, 70
MCAR, 70
Médiane, 79
MI, 87
MINI, 82
MMRE, 96
Mode, 79
Moyenne, 79
MoyenneA, 80

NB, 85
NI, 71
NMAR, 71
NRMSE, 96

Pareto Rappel, 46
Pareto Reco, 46
PMC, 25
Probit, 24
PTA, 129

RELIEF, 147
ReliefF, 149
RN, 85
Rocchio, 40

Salammbô, 32
SansFiltre, 175
SBS, 127
SFBS, 129
SFFS, 129
SFS, 127
SHC, 132
SU, 140
SVDR, 84
SVM, 85

TDIDT, 32
Test binomial, 59
Test de Dunnet, 64
Test de Fisher, 63
Test de Friedman, 66
Test de Kolmogorov-Smirnov, 166
Test de McNemar, 57
Test de Nemenyi, 67
Test de Student, 55
Test de Tukey, 64
Test de Wilcoxon, 60
Test du Chi2, 57