Extraction d'informations à partir de documents juridiques : application à la contrefaçon de marques

Pierre Renaux

UNIVERSITÉ de CAEN/BASSE-NORMANDIE

U.F.R. : Sciences

ÉCOLE DOCTORALE : SIMEM

THÈSE

présentée par

Pierre RENAUX

et soutenue

le 11 juillet 2006

en vue de l'obtention du

DOCTORAT de l'UNIVERSITÉ de CAEN

spécialité : Informatique

(Arrêté du 25 avril 2002)

Extraction d'informations à partir de documents juridiques : application à la contrefaçon de marques

MEMBRES du JURY

Directeur :	M. Khaldoun Zreik	Professeur	Université de Caen
Rapporteurs :	M. Mohamed Quafafou	Professeur	Université de Marseille
	M. Jean Sallantin	Directeur de Recherche CNRS	Université de Montpellier
Examinateurs :	Mme Anne Nicole	Professeur	Université de Caen
	M. Imad Saleh	Professeur	Université de Paris VIII
	M. Thomas Lebarbé	Maître de Conférence	Université de Grenoble 3

Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen — CNRS UMR 6072

Mis en page avec la classe thloria.

Table des matières

Table des figures

Liste des tableaux

Résumé Le cadre de nos recherches repose sur l'extraction et l'analyse de connaissances à partir d'une source de données documentaire de type juridique caractérisant les contrefaçons de marques nominatives. Cette discipline reflète parfaitement toutes les contraintes appartenant aux différents domaines intervenant dans le cadre de l'extraction de connaissances à partir de documents : document électronique, bases de données, statistiques, intelligence artificielle et interaction homme/machine. Cependant, les performances de ces méthodes sont étroitement liées à la qualité des données utilisées. Dans notre contexte de recherche, chaque décision est supervisée par un rédacteur (le magistrat) et dépend étroitement du contexte rédactionnel, limitant les procédés d'extraction d'information. Nous nous intéressons donc aux décisions susceptibles de biaiser l'apprentissage des documents. Nous observons les fondements de celles-ci, déterminons leur importance stratégique et le cas échéant nous proposons des solutions adaptées afin de réorienter le biais observé vers une meilleure représentation des documents. Nous proposons une approche exploratoire supervisée pour évaluer la qualité des données impliquées, en déterminant les propriétés biaisant la qualité de la connaissance établie ainsi qu'une plate-forme interactive et collaborative de modélisation des processus conduisant à l'extraction de connaissances afin d'intégrer efficacement le savoir-faire de l'expert.

Mots-clés : Apprentissage automatique, Systèmes d'aide à la décision, Base de données

---

Title

Information extraction from forensic documents : a trade-marks infringement decision support system

Abstract

Our research framework focuses on the extraction and analysis of induced knowledge from legal corpus databases describing the nominative trade-mark infringement. This discipline deals with all the constraints arising from the different domains of knowledge discovery from documents : the electronic document, databases, statistics, artificial intelligence and human computer interaction. Meanwhile, the accuracy of these methods are closely linked with the quality of the data used. In our research framework, each decision is supervised by an author (the magistrate) and relies on a contextual writing environment, thus limiting the information extraction process. Here we are interesteding in decisions which direct the document learning process. We observe their surrounding, find their strategic capacity and offer adapted solutions in order to determine a better document representation. We suggest an explorative and supervised approach for calculating the data quality by finding properties which corrupt the knowledge quality. We have developped an interactive and collaborative platform for modelling all the processes concluding to the knowledge extraction in order to efficiently integrate the expert's know-how and practices.

Keywords : Machine Learning, Databases, Decision tools

Discipline : Informatique

Laboratoire : Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen (UMR 6072), Université de Caen Basse-Normandie, France.

Préambule

Les pages qui suivent sont le résultat de travaux de recherche effectués dans le cadre d'un doctorat en informatique à l'Université de Caen - Basse Normandie.

Toutefois, ces travaux ne portent pas uniquement sur la thématique informatique, mais plus précisément sur l'informatique au service du domaine juridique.

Ce court préambule ne présente pas les travaux décrits dans ce mémoire mais les objectifs que nous nous sommes fixés dans l'approche scientifique de la rédaction.

Objectifs scientifiques de ce mémoire

Ce mémoire a pour but de retracer un parcours de recherche tout en mettant en valeur des réflexions théoriques et critiques des avancées scientifiques effectuées.

Par conséquent, nous avons rassemblé nos observations en trois ensembles qui, à notre avis, représentent mieux une démarche scientifique pragmatique :

1. Analyse
2. Réalisation technique
3. Validation / critique

Ces trois étapes pouvant être reprises de manière cyclique. C'est cette démarche que nous retrouvons dans la structure logique de ce mémoire, présentée dans la figure 1 page suivante. Outre les introductions et conclusions, le mémoire est divisé en deux parties : l'une relative au processus d'induction de connaissances à partir d'une base de données (partie modélisation), l'autre relative à la correction de ce processus (partie correction).

Chacune des parties est décomposée en chapitres reposant sur une structure ternaire :

- **Analyse** des étapes de modélisation (première partie), des principes de correction (deuxième partie).
- **Réalisation technique** de chacune de ces étapes, que cela soit pour la modélisation ou la correction du modèle.
- **Validation / critique** des choix retenus et des résultats observés, quelque soit la partie.

Structure du mémoire

Nos travaux sont basés sur la mise en place d'une plate-forme d'évaluation et de correction des différents processus d'apprentissage et plus spécifiquement sur la nécessité de redonner à l'expert son autonomie face aux outils d'aide à la décision pour la modélisation d'un domaine d'étude. Par conséquent, nous avons structuré ce mémoire en deux parties majeures, chacune suivant la structure logique présentée ci-dessus.

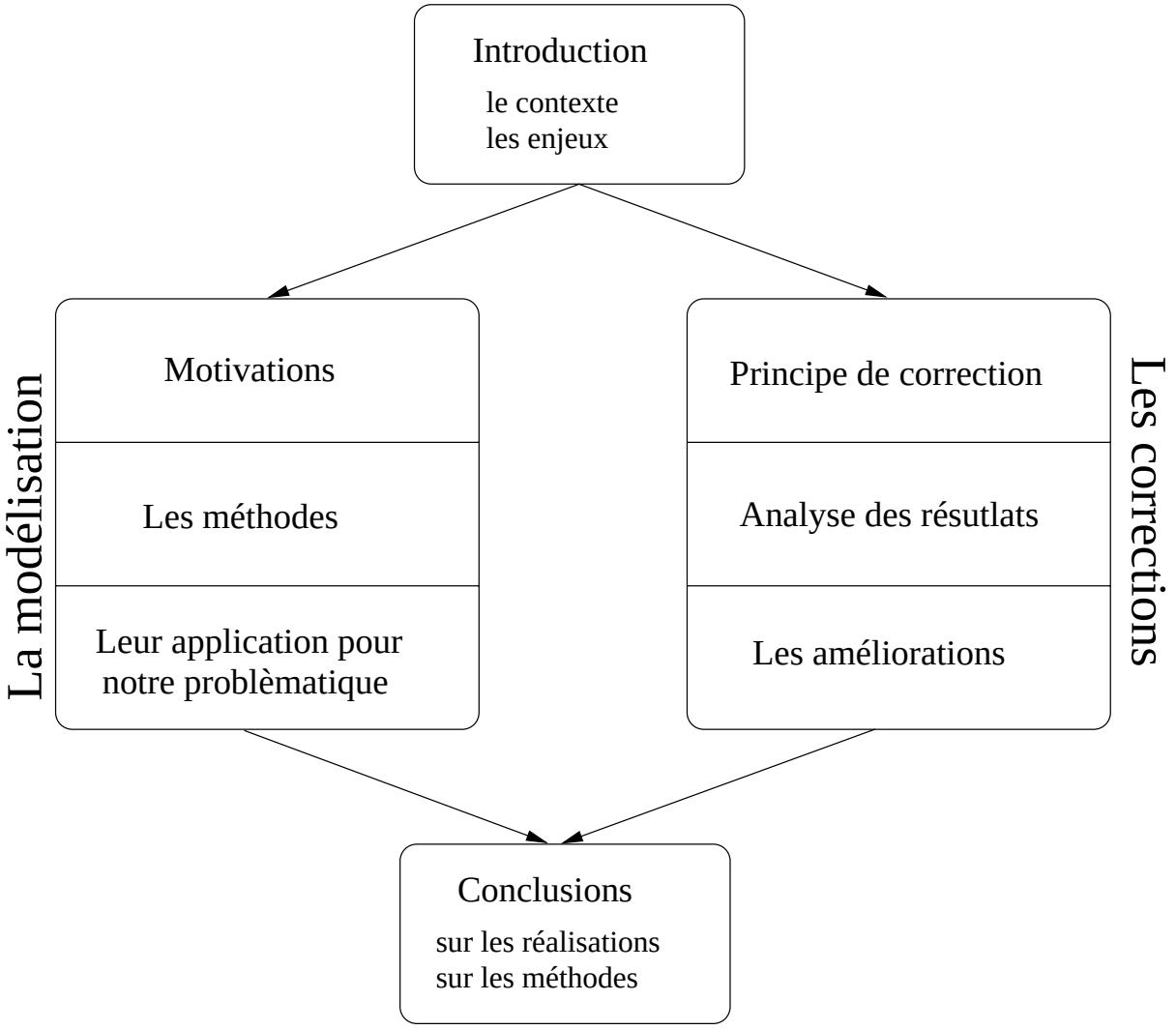


FIG. 1 – Structure logique du mémoire de doctorat

- Une partie présentant les outils permettant la modélisation d'un domaine. De cette modélisation découlent alors des possibilités d'apprentissage permettant une aide à la décision du domaine modélisé.
  - Dans un premier chapitre, nous présentons le principe de sélection d'un sous-ensemble de documents électroniques à partir d'une base documentaire. Ces travaux nous ont permis d'établir un jeu de données ciblé pour la modélisation du domaine.
  - Dans le deuxième chapitre, nous étudions les principes de préparation des documents précédemment sélectionnés afin d'établir une base documentaire rigoureuse de travail.
  - Dans le troisième chapitre, nous présentons les principes de transformation des données favorisant l'intégration du savoir-faire de l'expert aux données documentaire afin d'établir un modèle du domaine. Cette prise en compte du savoir-faire nous a permis de mettre

en avant des attributs améliorant la comparaison documentaire.
  - Dans le quatrième chapitre, nous considérons les différents procédés permettant l'obtention d'un modèle d'une base documentaire, et proposons des solutions afin d'établir des modèles de notre domaine d'étude.
  - Dans le cinquième chapitre, les principes d'évaluation des connaissances dérivées des modèles sont présentés, et adaptés à notre cas d'étude.
- Une partie présentant la réalisation technique de ces outils ainsi que les différentes analyses conduisant à une révision du modèle induit afin d'en améliorer la qualité.
  - Dans le premier chapitre nous étudions la validité du modèle induit et nous le comparons à d'autres principes algorithmiques afin d'en déterminer les limites qualitatives. De ces comparaisons, nous sommes alors en mesure d'orienter les corrections à pratiquer pour atteindre ces objectifs.
  - Dans le deuxième chapitre, les corrections ciblées sont réalisées et une critique sur la validité de la représentation des documents est proposée. De cette étude, nous proposons alors des pistes pour l'amélioration de la représentation du jeu de données documentaire.
  - Dans le troisième chapitre, nous présentons la réalisation technique du domaine (`CATMInE`, son évolution (`DeTTMInE`), ainsi que tous les outils connexes intervenant dans la modélisation. Ces outils permettent alors à l'expert d'agir seul (l'expert redevient autonome face aux outils d'aide à la décision) sur la représentation du domaine sans intervention de l'informaticien.

La figure 2 page suivante montre sous forme arborescente la structure de ce mémoire. Nous privilégions l'approche ascendante permettant d'établir un modèle (première partie) : des données aux connaissances, et l'approche descendante de correction (deuxième partie) : des connaissances aux données.

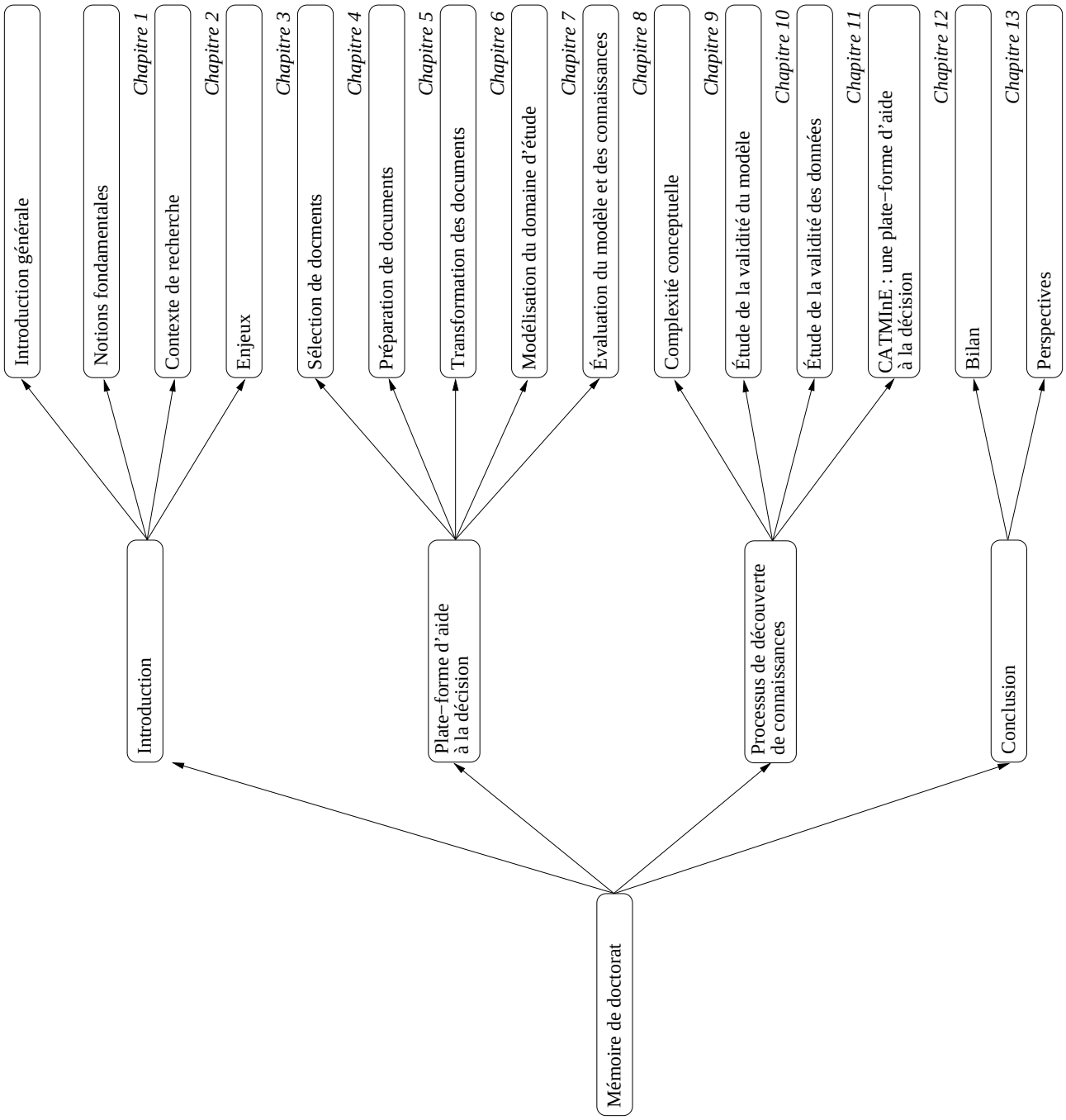


FIG. 2 – Structure arborescente du mémoire de doctorat

Introduction

Introduction générale

Contexte

Nos recherches reposent sur l'extraction et l'analyse de connaissances à partir d'une source de données documentaire de type juridique. Cette jeune discipline reflète parfaitement toutes les contraintes appartenant aux différents domaines intervenant dans le cadre de **l'extraction de connaissances à partir de documents** :

1. le *document électronique*, contenant une information propre à un domaine
2. les *bases de données*, favorisant l'accès, l'organisation et la gestion des documents électroniques
3. les *statistiques*, permettant le traitement de grands volumes de données conservés dans des bases de données
4. l'*intelligence artificielle*, exploitant les résultats de la statistique afin de produire des outils d'aide à la décision
5. l'*interaction homme/machine*, pour la conception des systèmes d'information et l'exploitation des résultats issus des procédés d'intelligence artificielle

Les documents juridiques utilisés dans ces travaux sont relatifs à des décisions de justice en matière de contrefaçon de marques nominatives, pour lesquelles nous devons être en mesure de proposer à l'expert exploitant de tels documents des connaissances caractérisant des propriétés communes entre ces documents.

Ces décisions sont à aborder avec une granularité supérieure à celle du simple document électronique. Les éléments mis en jeux dans une décision de contrefaçon de marques vont au delà des mots. La notion de marque fait intervenir des aspects linguistiques, phonétiques et visuels et la notion de décision, des aspects normatifs employés par les juges. Ces propriétés intrinsèques au document impliquent alors le traitement d'une information présente dans le document mais non explicitée en tant que telle. Cette particularité fait alors glisser d'un problème simple en apparence vers un problème bien plus délicat à traiter.

Avec plus de 1,7 millions de marques enregistrées en France et 8500 décisions de justices archivées sur un siècle, évaluer le risque de contrefaçon d'une nouvelle marque par rapport à

l'ensemble existant représente un travail délicat. Cette difficulté est complétée par :

1. le besoin de prendre en compte les différents pays composant la communauté européenne, voire la communauté internationale pour les multinationales
2. les marques ne doivent pas s'enfreindre
3. le juge prend une décision en fonction des décisions existantes

Cette tâche devient alors complexe et impose l'utilisation d'outils adaptés afin de traiter, efficacement, l'intégralité de l'information disponible.

Ces outils de modélisation, établis par une communauté d'informaticiens ou de statisticiens, se doivent d'être fortement orientés utilisateur afin d'intégrer les connaissances de ces utilisateurs ou experts. Dans ce mémoire, nous nous orientons sur la définition d'une plate-forme d'aide à la décision permettant la prise en compte d'un savoir-faire précis par le biais d'outils interactifs. Ces outils permettent un gain d'objectivité et de compréhension des choix réalisés tout au long du processus d'extraction de connaissances à partir de documents.

Motivations

Les outils d'extraction de connaissances à partir de documents sont nombreux et variés. Tous ont été validés sur de nombreuses bases de données réelles ou artificielles et sont utilisés dans de nombreux domaines industriels, économiques ou scientifiques. Cependant, les performances de ces outils sont étroitement liées à la qualité des données utilisées, et si celle-ci n'est pas optimale, les connaissances extraites seront affectées par ces lacunes, dégradant dans le même temps la précision des procédés d'aide à la décision. Cette dégradation rend alors les outils inexploitables car leur fiabilité est remise en cause et l'expert ne peut avoir une confiance dans les décisions proposées sans avoir à refaire tout le cheminement conduisant à la décision. En revanche, lorsque la fiabilité est accrue, l'expert peut alors se concentrer sur des points précis pour comprendre les fondements de la décision proposée.

Dans notre contexte de recherche, la matière première est le document électronique juridique témoignant des décisions françaises en matière de contrefaçons de marques nominatives. Ces décisions ne sont pas rédigées de façon systématique et ne sont pas issues d'un raisonnement explicite. Bien au contraire, chaque décision est supervisée par un rédacteur (le magistrat) et dépend étroitement du contexte rédactionnel.

Ce problème rédactionnel limite les procédés d'extraction d'informations à partir des documents électroniques juridiques. Chaque document est rédigé par un magistrat caractérisé par un style d'écriture personnel et pouvant avoir déjà rédigé de tels documents auparavant. La rédaction de ces documents n'étant pas constante, le procédé d'extraction d'information et de production de connaissances doit être envisagé comme plus flexible afin de pouvoir capturer et modéliser

les fondements contextuels du texte juridique : l'information permettant de caractériser chacune des décisions.

La conséquence directe de cette problématique nous conduit à nous interroger sur la pertinence de l'information retenue pour la représentation des documents et donc pour la recherche de connaissances permettant de caractériser la contrefaçon de marques nominatives. Plus particulièrement, nous voulons déterminer les décisions susceptibles de biaiser l'apprentissage des documents afin d'observer les fondements de celles-ci, de déterminer leur importance stratégique et le cas échéant de proposer des solutions adaptées pour réorienter le biais observé vers une meilleure représentation des documents.

Contributions

Pour répondre à la problématique de dépôt de marques et d'évaluation du risque de contrefaçon, nous proposons d'étudier l'intégralité du processus de modélisation de l'ensemble documentaire caractérisant les contrefaçons de marques nominatives. Puis, au travers d'une approche exploratoire supervisée, nous répondons à la motivation principale de ces travaux de recherche consistant à évaluer la qualité des données impliquées tout au long de ce processus en déterminant les exemples et leurs propriétés biaisant la qualité de la connaissance établie. Cet aspect est motivé par l'analyse des données (les décisions) fournies par l'expert. Enfin nous proposons une plate-forme interactive de modélisation des processus conduisant à l'extraction de connaissances.

Afin d'intégrer efficacement le savoir-faire de l'expert dans l'extraction de connaissances à partir des documents électroniques juridiques, nous avons privilégié une interaction forte entre experts et outils. Cette interaction se justifie par l'aspect contextuel des décisions, ne répondant pas à un raisonnement explicite et ne pouvant être mimée par un ordinateur. Pour extraire l'information du document électronique nous privilégions une extraction semi-supervisée : le système établit de façon automatique l'information nécessaire à la caractérisation d'une décision et le soumet à l'expert pour une validation. Si cette information est incomplète ou erronée, l'expert peut alors déterminer lui même ces propriétés par lecture du document.

La transformation des données caractérisant une décision afin de pouvoir établir des éléments de comparaison sur l'ensemble des décisions se fait par intégration du savoir-faire de l'expert. Celui-ci est assisté par un processus hautement interactif tout au long de la transformation des données. Ce processus permet de définir les traitements à réaliser en fonction de ces attentes et de ces habitudes d'interprétation des documents issus du domaine de définition.

La production de connaissances se fait alors de façon automatique en gardant un aspect interactif pour la validation des résultats obtenus. Des cartographies des décisions ainsi que les fondements sur lesquels repose la connaissance sont établis et proposés à l'expert afin qu'il puisse en juger de leur pertinence.

Enfin, l'intégralité de ce processus de modélisation repose sur une approche cyclique permettant de corriger les actions réalisées afin de garantir au mieux la consistance du modèle. Un partage de l'ensemble des ressources disponibles entre experts du domaine assure à chacun de pouvoir comparer et utiliser la totalité des solutions possibles en intégrant le savoir-faire de l'ensemble des intervenants d'un même domaine.

Deux réalisations techniques ont été menées durant cette activité de recherche. Un transfert de technologie entre l'Université de Caen et le cabinet de conseil en propriété industrielle et intellectuelle Breese Dérambure Majerowicz de Paris a été réalisée, conduisant ces derniers à l'utilisation d'une plate-forme de travail collaborative et interactive d'aide à la décision pour l'évaluation du risque de contrefaçon entre marques nominatives.

Notions fondamentales

Au travers de ce mémoire, nous allons aborder plusieurs notions d'un processus de modélisation. Ces notions permettront au lecteur de mieux comprendre ce type de processus permettant de décrire un phénomène relatif à un domaine.

Ces procédés de modélisation reposent sur une base de données documentaire. Nous appliquons à cette source de données des algorithmes permettant d'induire automatiquement de la connaissance en fonction des croyances d'un expert du domaine. Cette induction repose sur des partis pris à la fois par l'algorithme lors de l'induction du modèle mais aussi par l'expert qui a guidé cette induction.

L'intégration de l'ensemble de ces contraintes : base de données documentaire, algorithme de modélisation, expert et partis pris, a été réalisé dans le cadre de la plate-forme `CATMInE`. Ce chapitre définit ces notions nécessaires à la bonne lecture de ce mémoire et permet de cadrer le contexte de recherche des ces travaux de doctorat.

1 Traitement des données fortement structurées

Dans les systèmes d'aide à la décision, la matière première utilisée pour l'apprentissage est une base de données. Cette base de données est définie par des dimensions :

- hauteur : le nombre d'enregistrements
- largeur : le nombre d'attributs

Afin d'obtenir un modèle de cette base de données, les algorithmes vont déterminer des relations entre attributs à travers tous les exemples. Évidemment si pour certains exemples, des attributs ne sont pas renseignés, il devient délicat de déterminer à partir de l'ensemble des données ces relations intrinsèques. C'est pourquoi, dans tout processus d'apprentissage, il est nécessaire de s'efforcer à avoir la base de données la plus homogène possible, où pour chaque enregistrement l'ensemble des descripteurs est renseigné. Nous parlons alors de base de données rigoureuse. Cette appellation sous-entend que chaque attribut est bien renseigné et que l'information correspond bien au type d'information attendue (un descripteur numérique ne peut par exemple pas contenir de modalités composées de caractères alphabétiques).

2 Induction et apprentissage automatique

À la différence de la déduction qui est une inférence logique légale, l'induction conduit du particulier au général et n'a de validité que psychologique (car elle peut être démentie par les faits) et constitue une règle raisonnable en l'absence de contre-exemple, et une hypothèse de travail à explorer (accréditer par d'autres exemples, ou infirmer par un contre-exemple).

L'induction repose sur une supposition. Le syllogisme inductif est alors définit comme hypothétique. À partir d'observations (qui sont toujours des propositions particulières), l'induction produit des propositions générales (hypothétiques) qui seraient réfutables.

Nous nous situons dans un contexte d'apprentissage inductif. Celui-ci fait référence au développement, à l'analyse et l'implémentation de méthodes qui permettent à une machine (au sens large) d'évoluer et de remplir des tâches associées à une intelligence artificielle grâce à un processus d'apprentissage. Cet apprentissage permet d'avoir un système qui s'optimise en fonction de l'environnement, les expériences et les résultats observés à partir de faits observés. Cet ensemble de contraintes impose alors d'avoir un processus interactif afin de les intégrer les plus efficacement possible.

3 Apprentissage automatique et découverte de connaissances à partir de bases de données

En intelligence Artificielle, nous parlons d'apprentissage automatique (*Machine Learning*) ou encore de découverte de connaissance à partir de base de données (*Knowledge Database Discovery*). Il s'agit de deux spécialités très proches (la première étant une partie de la seconde), cherchant à établir des régularités, des motifs ou des concepts à partir d'un ensemble de données. Cependant, si nous les regardons indépendamment, il existe quelques différences [Mannila, 1996].

1. L'apprentissage automatique et la découverte de connaissances à partir de bases de données ont tout deux pour objectif d'identifier des régularités, des motifs ou concepts à partir d'un ensemble de données. Toutefois, dans la littérature, l'apprentissage automatique sous-entend qu'il existe une structure, des relations au sein d'un jeu de données et que celui-ci est issu d'un mécanisme (notion de régularité) de production (notion d'automate). Contrairement à cet aspect, la découverte de connaissances implique plus d'appréhender le jeu de données en tant que tel, sans pour autant avoir des aprioris sur les moyens d'obtention de celui-ci.

   La conséquence directe est que l'apprentissage automatique va produire de la connaissance à partir de l'ensemble des données, alors que la découverte de connaissance pourra déterminer des connaissances en utilisant qu'une partie des données. L'utilisation de l'intégralité ou non des données est une des différences entre ces procédés.

2. La seconde différence est liée aux objectifs. Les systèmes d'apprentissage automatiques ont plus vocation à déterminer des relations complexes et non identifiables humainement alors que les systèmes de découverte de connaissances ont des objectifs plus modestes sur la connaissance obtenue, connaissance qui doit pouvoir être identifiée par un expert du domaine si le temps n'est plus une contrainte.

L'apprentissage automatique est au cœur de la découverte de connaissance, celle-ci regroupant en plus les notions de compréhension du domaine, de préparation des données et d'interprétation des résultats. C'est pourquoi dans notre exposé, nous ne ferons pas de distinction entre les deux philosophies : chacune conduisant à la production de connaissances (interprétables ou non) et toutes deux sont complémentaires l'une de l'autre.

4 Le rôle de l'expert

Tout au long du processus d'apprentissage dans le cadre de nos recherches, nous faisons intervenir l'expert. L'expert est généralement spécialisé dans un domaine particulier. Il peut exceller dans la résolution de problèmes, la recherche ou les solutions de rechange.

L'expert est considéré comme maître d'un savoir (contrairement au savant qui lui maîtrise la connaissance d'un domaine pouvant généralement être formalisée et considérée comme objective), qui intègre naturellement des éléments de connaissance, mais qui prend en compte une expérience et des savoirs transmis non formalisés le rendant porteur de son savoir.

L'expert est donc caractérisé par un savoir-faire implicite non verbalisé relatif à son expérience du domaine d'étude. À ce savoir-faire est ajouté en plus une notion de connaissance qui peut-être formalisée et considérée comme objective. Son rôle est plus que nécessaire, c'est à partir de ce savoir-faire et de ses habitudes qu'il lui est possible d'orienter la construction d'un modèle prenant en compte ces éléments.

5 Les partis pris : une orientation dans l'établissement d'un modèle

Dans tout processus d'apprentissage automatique, l'établissement de connaissances dépend de choix stratégiques et techniques intervenant dans un processus décisionnel. Ces choix sont strictement dépendant de la modélisation ou sont exprimés par l'expert en fonction de son savoir-faire et de ses connaissances.

Un parti pris peut par exemple conduire un expert à accepter ou refuser une preuve, non pas sur la force des arguments eux-mêmes, justifiant la preuve, mais sur les relations que peut avoir l'expert entre la preuve et ses convictions.

Les algorithmes ont des partis pris (façon de segmenter les données, regrouper les enregistrements, etc...) lorsqu'ils établissent des modèles, tout comme l'expert prenant des décisions en fonction de son savoir-faire et donc de ses convictions pour une situation donnée.

6 CATMInE : une plate-forme d'aide à la décision

Enfin, ce mémoire de thèse présente nos travaux dont les retombées applicatives sont une plate-forme d'aide à la décision nommée `CATMInE`[1]. Ce projet applicatif a fait l'objet d'un contrat de recherche et développement en 2001 entre l'Université de Caen par le biais de Thomas Lebarbé et le cabinet juridique Breese Derambure Majerowicz de Paris.

Cette plate-forme a été développée dans une approche évolutive. Ces corrections, `DeTTMInE`[2], en sont des évolutions possibles. Par commodité, nous avons conservé le terme de `CATMInE` pour citer le cadre applicatif de nos travaux. Lorsqu'il est fait référence aux éléments intervenant dans `DeTTMInE` le lecteur en sera averti.

[1] **C**omputer **A**ssisted **T**rade **M**arks **In**fringement **E**valuation
[2] **De**cision **T**ree for **T**rade **M**arks **In**fringement **E**valuation

Chapitre 1

Contexte de recherche : le document électronique juridique

Sommaire

Le cadre de nos recherches repose sur l'analyse et l'extraction de connaissances à partir d'une source de données documentaire juridique, réalisant la liaison entre les disciplines d'interaction homme-machine et l'apprentissage automatique. Cette place au sein de ce carrefour interdisciplinaire est présentée dans la figure 1.1 page suivante.

Dans un premier temps, le rédacteur (expert dans son domaine de travail) construit et rédige le document, le rendant alors supervisé dans son contenu : celui-ci est orienté pour caractériser un phénomène précis. Cette rédaction influence alors tout ce qui peut être réalisé en exploitant le document : consultation, archivage et aide à la décision. De même, ces exploitations du document permettent à leur tour d'influencer le rédacteur en proposant des corrections (par rapport à des critères de qualité et de pertinence qui sont subjectifs) en vue d'améliorer la structure du document. Ces améliorations étant principalement axées pour optimiser l'exploitation du document.

Nous nous plaçons ici dans l'optique d'une source de données documentaires supervisées ca-

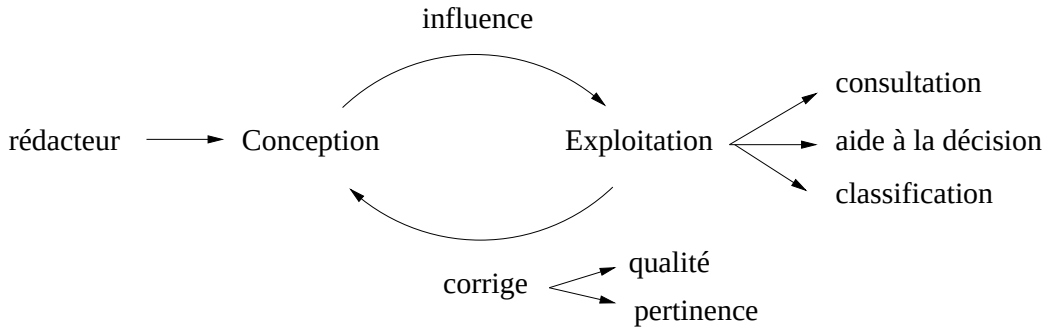


FIG. 1.1 – De la conception à l'exploitation du document : une problématique interdisciplinaire.

ractérisant des décisions rendues par des tribunaux français pour des litiges en droit des marques nominatives. Ces litiges représentent un sous-ensemble des documents électroniques juridiques (des décisions) numérisés par une société privé (Jouve) pour une base de données privée (celle de l'INPI) à usage de juristes conseil (le cabinet Breese Derambure Majerowicz)

1.1 Propriétés des documents électroniques

1.1.1 Des documents structurés

Le document électronique présente différents niveaux structurels. Le premier est lié à la structure même de la rédaction : il s'agit de la relation entre les titres, paragraphes et mots. Cette structure caractérise la sémantique structurelle, ou structure logique du document. Nous pouvons représenter cette structure par le biais d'un graphe orienté sans cycle, où un nœud contient un élément structurel et un arc une relation d'appartenance entre deux nœuds. Cette représentation étant non circulaire, nous obtenons alors une structure arborescente du document comme le présente la figure 1.2 page ci-contre.

Le second niveau de structuration est une mise en forme à partir d'un langage à balises comme le `HTML` permettant de changer la granularité de la représentation du document. Nous parlons alors de structure physique du document. L'information est alors encadrée par des éléments structurels apportant une information supplémentaire au contenu qu'ils encadrent comme par exemple des balises de mise en forme textuelle, structurelle ou d'hyperliens. Ce dernier point lorsqu'il est interne au document introduit une nouvelle relation structurelle dans le document et lorsqu'il est externe au document complète les relations au sein d'un volume documentaire. Le document est alors enrichi d'une nouvelle structure de niveau supérieur à celle caractérisant la sémantique structurelle.

Le dernier niveau, structuration pervasive, est lié à l'utilisation d'une formalisation de type `XML` d'un document électronique. Ce type de formalisation est proche du `HTML` dans le sens où il

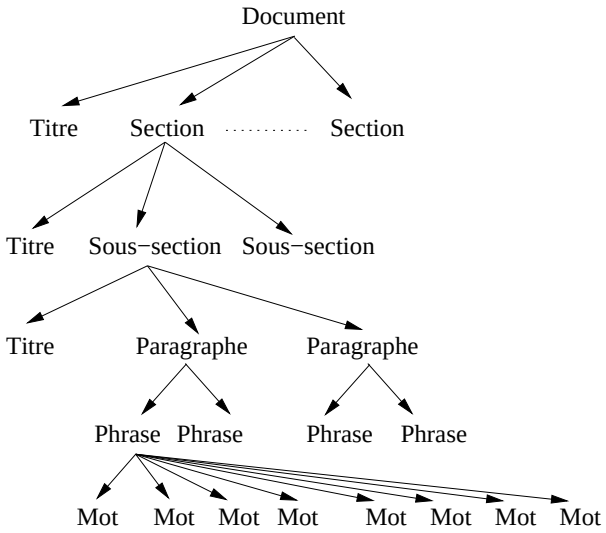


FIG. 1.2 – Exemple d'une structure logique d'un document dit "structuré".

repose sur un langage à balises. L'apport majeur ici est la possibilité de définir les balises afin de créer des éléments propres au document et permettant de renforcer la sémantique de celui-ci. Ce type de formalisation renforce la description d'un document en représentant conjointement l'information textuelle et l'information structurelle d'un document. De plus en utilisant des langages dérivés tels que `X-Link`, la définition d'un lien se trouve complètement changée en permettant l'intégration de sources, de destinations et d'arcs multiples pour un même lien.

1.1.2 Des documents supervisés

Une autre particularité des documents juridiques est leur aspect supervisé. Nous entendons par supervisé le fait que ces documents sont rédigés en vue de caractériser une classe de décision. Cette rédaction est donc orientée afin d'exprimer et de décrire l'intégralité ou une partie du domaine pour lequel ils ont été déployés. En linguistique, nous parlons alors de documents performatifs : ils réalisent une action.

1.1.3 Le document électronique juridique supervisé

Le document électronique juridique supervisé répond aux deux aspects énoncés précédemment : une structure forte et une supervision de la rédaction.

Pour la particularité de structuration, celle-ci apparaît sur deux niveaux : un niveau portant sur la sémantique guidée par la représentation et un relatif à la sémantique structurelle. Le premier niveau prend forme par le biais d'une représentation `XML`. Le document est en effet décomposé en deux parties liées. Le second niveau correspond à un cartouche résumant la décision

et la seconde correspond au texte complet de la décision. Cette structure est vraie uniquement pour les documents que nous utilisons et ne représente en aucun cas les publications officielles.

Le cartouche résumant la décision n'introduit pas de sémantique structurelle. Les éléments sont étiquetés via des balises caractérisant une information ciblée : parties, marques, niveau juridictionnel, date de la décision et un résumé en style télégraphique de la décision. La table 1.1 page suivante présente un extrait de ce cartouche résumant une décision.

C'est dans le texte complet de la décision que nous retrouvons le second aspect de structuration du document : la sémantique structurelle. Quelle que soit la décision, le cheminement rédactionnel reste le même d'une décision à une autre. En revanche, le vocabulaire utilisé variera en fonction du rédacteur. Nous retrouvons alors généralement une décision en six points :

1. rappel des faits et de la procédure
2. validité de la marque plaignante
3. validité de la marque du défendant
4. atteinte à la dénomination sociale et au nom commercial de la société plaignante
5. les mesures réparatrices
6. l'action en contrefaçon

Le tableau C page 207 présente le texte complet de la décision présentée dans le tableau 1.1 page suivante[3].

1.1.4 Pertinence des documents

Les documents utilisés pour la modélisation du phénomène doivent être considérés comme adaptés au problème car sélectionnés par l'expert. Par adaptés, nous entendons que les documents sont caractéristiques du phénomène étudié. L'expert les a retenus pour leurs caractéristiques, l'information qu'ils véhiculent et leur pertinence[4].

Cependant, il est aussi possible que ces documents soient détournés de leur utilisation primaire : le document est pensé pour de la consultation, puis utilisé pour de l'aide à la décision. L'information contenue dans le document peut alors couvrir plus que le domaine étudié, ou inversement, il est possible que le phénomène ne soit pas intégralement décrit et qu'une information manque. Pour être cohérent, la conception du document aurait dû se faire en prenant en compte l'utilisation de ce dernier comme source pour un processus décisionnel.

Ces points précis peuvent trouver leur réponse dans les documents que nous qualifions de connaissance stratégique ou de bruit (en fonction des interprétations de l'expert). Ceux-ci étant situés à la frontière entre deux ou plusieurs concepts. La figure 1.3 page 22 présente cette problématique. Les régions AB, AC, BC et ABC sont des zones critiques où les documents présents

[3] Les coquilles présentent dans l'extrait ne sont pas une erreur.

[4] La pertinence se rapporte au fait que toutes les informations des documents portent sur un seul sujet.

jurinpi

Référence M20010655

Domaine MARQUE

Nature de la décision DECISION FRANCAISE

Juridiction TRIBUNAL DE GRANDE INSTANCE DE PARIS (CH.03), 2001-05-18

Date de la décision 2001-05-18

Noms des parties EXACOD (SA) / LA POSTE

Marque EXACOD ;HEXACODES

Analyse DENOMINATION SOCIALE ET NOM COMMERCIAL (EXACOD) - MARQUE DE FABRIQUE ET DE SERVICES - MARQUE COMPLEXE - PARTIE VERBALE (EXACOD) - APPAREILS POUR LE TRAITEMENT DE L'INFORMATION, LA GESTION DES AFFAIRES COMMERCIALES, LES CONSEILS, L'INFORMATION ET LE RENSEIGNEMENT D'AFFAIRES, LA GESTION DE DOSSIER INFORMATIQUE, LA PROGRAMMATION POUR ORDINATEUR ET LA LOCATION DE TEMPS D'ACCES A UN SERVEUR DE BASES DE DONNEES - **CL09, CL16, CL35, CL36, CL41** - NUMERO D'ENREGISTREMENT 99 796 368 - MARQUE VERBALE (HEXACODES) - LOGICIELS DE GESTION DE FICHIERS D'ADRESSES DESTINES AUX ROUTEURS ET AU GRANDS EMETTEURS DE COURRIERS, LES FICHES, LA CONSTITUTION, LA CENTRALISATION, LA TENUE ET LA MISE A JOUR DE FICHIERS D'ADRESSES POUR VALIDATION DU CONTENU DES BASES DE DONNEES (VOIES, BOITES POSTALES, LOCALITES ET CEDEX) - **CL09, CL16, CL35** - NUMERO D'ENREGISTREMENT 3 009 533

ACTION EN CONTREFACON

MARQUE (EXACOD) - VALIDITE (OUI) - PORTEE - ETENDUE DE LA PROTECTION - ARTICLE L 712-2 CODE DE LA PROPRIETE INTELLECTUELLE - LIBELLE - PRECISION (OUI) - DESIGNATION DES CATEGORIES AUXQUELLES APPARTIENNENT LES PRODUITS - IDENTIFICATION DU CONTENU DES SERVICES (NON) - CARACTERE ABUSIF DE LA DESIGNATION DES SERVICES (NON) - FRAUDE (NON) - DOMAINE EXCEDANT L'ACTIVITE SOCIALE (NON)

MARQUE (HEXACODES) - VALIDITE (NON) - DISPONIBILITE (NON) - ARTICLE L 711-4 CODE DE LA PROPRIETE INTELLECTUELLE - DROIT ANTERIEUR (OUI) - MARQUE ANTERIEURE ENREGISTREE (EXACOD) - RISQUE DE CONFUSION (OUI) - SIMILITUDE VISUELLE ET PHONETIQUE - RISQUE DE CONFUSION SUR ORIGINE DES SERVICES

ATTEINTE A LA DENOMINATION SOCIALE ET AU NOM COMMERCIAL (NON) - ACTIVITES DIFFERENTES - RISQUE DE CONFUSION (NON)

CONTREFACON - PREJUDICE - EVALUATION - ELEMENT PRIS EN CONSIDERATION - DEVALORISATION - DIFFUSION DE BROCHURES AUPRES DE PROFESSIONNELS

Numéro(s) 99796368 ;3009533

Classification des produits et services CL09 ;CL16 ;CL35 ;CL36 ;CL41

Produits et services APPAREILS POUR LE TRAITEMENT DE L'INFORMATION, LA GESTION DES AFFAIRES COMMERCIALES, LES CONSEILS, L'INFORMATION ET LE RENSEIGNEMENT D'AFFAIRES, LA GESTION DE DOSSIER INFORMATIQUE, LA PROGRAMMATION POUR ORDINATEUR ET LA LOCATION DE TEMPS D'ACCES A UN SERVEUR DE BASES DE DONNEES - LOGICIELS DE GESTION DE FICHIERS D'ADRESSES DESTINES AUX ROUTEURS ET AU GRANDS EMETTEURS DE COURRIERS, LES FICHES, LA CONSTITUTION, LA CENTRALISATION, LA TENUE ET LA MISE A JOUR DE FICHIERS D'ADRESSES POUR VALIDATION DU CONTENU DES BASES DE DONNEES (VOIES, BOITES POSTALES, LOCALITES ET CEDEX) - CL09, CL16, CL35 - ...

TAB. 1.1 – Exemple de jurisprudence issu de la base documentaire `JURINPI`

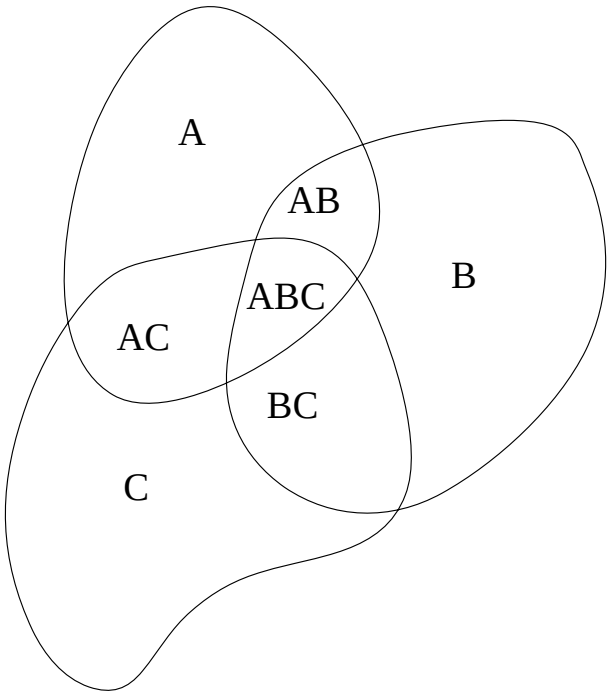


FIG. 1.3 – Exemple de documents situés à la frontière entre trois concepts.

nécessitent une étude approfondie. Cela laisse supposer qu'ils ont des propriétés communes à ces ensembles. Si tel est le cas, alors il peut y avoir eu erreur durant le processus extraction d'informations. L'erreur n'est bien sûr pas à imputer au processus d'extraction en tant que tel, mais aux documents ne présentant pas une information constante, en terme structurel et terminologique, à travers l'échantillon retenu. Bien évidement, le système d'extraction peut être plus robuste aux variations, mais il faut alors établir le coût d'une telle démarche relativement à des documents établis rigoureusement et donc sans erreur. Cette déduction ne peut se faire qu'après avoir reconsidéré la cohérence de la modélisation et plus particulièrement celle des descripteurs. Si c'est le cas, alors la représentation des documents utilisés n'est pas suffisamment fidèle aux documents d'origine.

1.1.5 Similarité des documents

Les documents servant à établir une base de données sont généralement issus d'une base documentaire. Les structures logiques et physiques de ces documents y sont généralement identiques d'un document à l'autre, et il est possible que celle-ci puisse servir à représenter plusieurs familles de documents. Le risque à ce niveau est d'avoir des documents fortement similaires : une même structure physique et logique pour une information différente. Cette similarité peut poser des problèmes pour distinguer les documents. Nous pouvons imaginer des documents identiques

à 95%, et les éléments faisant la différence entre les concepts deviennent alors plus délicats à mettre en valeur. Dans le cadre de documents générés de façon automatique à partir de capteurs (processus industriel par exemple), identifier les variations sera une tâche bien plus simple que de repérer les différences dans des documents textuels présentant une terminologie et une structure identique, mais dont la rédaction varie d'un auteur à l'autre.

Dans un autre cas de figure, la base de données peut être issue d'une compilation de plusieurs types de documents, rendant l'exploration plus délicate, mais néanmoins possible. La difficulté va résider, pour l'expert, dans la sélection de la bonne information, permettant de comparer ce qui est comparable. Pour cela, il lui faudra déterminer pour chaque famille les éléments saillants qui la caractérisent. Ces éléments doivent bien entendu être identiques d'une famille à l'autre. Cela implique un partage des caractéristiques communes. Si certaines modalités sont modifiées, transformées, adaptées pour permettre la comparaison, il n'est plus évident de garantir que la sémantique a été conservée tout au long de la transformation. Seul l'expert peut garantir une telle démarche, mais cela nécessitera de sa part une vigilance accrue pour éviter tout oubli, toute erreur.

1.2 Bilan du contexte de recherche

Nous avons présenté dans la première section la notion de document structuré et supervisé. Nous avons aussi présenté certains problèmes que peuvent poser les sources documentaires mettant en jeu des documents fortement similaires, où il peut être délicat de mettre en valeur les faibles différences entre documents (différences pourtant capitales, car à l'origine de la connaissance à établir).

Cette définition de document structuré supervisé, à la fois par une structure physique et une structure pervasive, caractérise la source documentaire de ces travaux de recherche sur le document électronique juridique. Avec de telles propriétés, les défis en terme d'extraction de connaissances peuvent s'en trouver moins complexes. L'information est fortement enrichie par une structure venant la compléter. Cette structure est établie par un savoir-faire et une habitude rédactionnelle afin de caractériser et d'identifier l'information pertinente portée par les documents.

1.3 Problématiques

Les bases de données ont pris leur essor voilà deux décennies (depuis 1980), et ce en parallèle de l'explosion du trafic internet. Toute entreprise soucieuse de connaître sa clientèle, ses fournisseurs, son marché, cherche à conserver le maximum d'informations pour s'assurer une traçabilité, un moyen de comparaison et d'étude statistique des différents intervenants. Toutefois, ce volume

d'enregistrements est de plus en plus conséquent et devient de moins en moins humainement exploitable.

Les modèles relationnels de base de données ont permis d'archiver de façon structurée l'information en autorisant des consultations et en restreignant le nombre de réponses à l'aide de contraintes et de filtres. Mais ces modèles sont plus que limités et ne sont optimisés que pour l'archivage et la consultation. Pourtant, l'information ainsi conservée peut révéler par exemple des regroupements en fonction de propriétés communes. Maîtriser les données devient un atout commercial fort par rapport à une concurrence qui n'anticipe pas ce genre de problématique.

Cet aspect des bases de données est ici résumé à une exploitation commerciale, mais peut très bien conserver de l'information sur un processus automatique de fabrication (les machines outils d'une chaîne de production par exemple), ou encore sur les propriétés d'un produit, pour déterminer si celui-ci est conforme aux normes de qualité fixées. Le cas échéant, il faut pouvoir détecter les pannes ou faiblesses du processus de fabrication pour améliorer la qualité du produit fabriqué.

Bref, les champs d'application du stockage d'information paraissent sans limite, mais l'exploitation de ces données en terme de recherche de connaissances implicites n'est pas encore une discipline très répandue. L'utilisation de tels procédés dépend du degré de maîtrise des systèmes d'informations par l'expert. Pourtant, avec des bases de données bien ciblées, il est possible d'étudier les effets d'un médicament (dans le cadre d'une enquête médicale), de prédire un comportement (dans un processus de fabrication), d'automatiser la classification de données dans un modèle relationnel (si le document électronique est source d'activité), de mieux connaître ses clients.

Mais l'introduction de tels systèmes dans une entreprise implique la présence d'un informaticien formé à ces techniques d'étude et de manipulation de données. Or l'informaticien n'est pas nécessairement un fin commercial, un médecin qualifié, un avocat conseil ou encore un archiviste de documents numériques. Trois solutions sont alors possibles :

1. avoir un informaticien possédant une double compétence en informatique et dans le domaine de l'étude.
2. faire collaborer l'informaticien avec un expert du domaine, avec tout ce que cela implique d'organisation.
3. avoir des outils adaptés à proposer à l'expert pour qu'il mène les études de données lui même.

La première solution tend à se répandre. La deuxième solution est une expérience enrichissante, mais la mobilisation de tous les intervenants n'est pas aisée. Chacun des protagonistes a ses propres contraintes et objectifs. Il est difficile de dégager des horaires de travaux communs pour partager les savoirs mis en jeu et les planifications de rendez-vous peuvent se faire sur de

longues périodes.

La dernière possibilité, consistant à désengager l'informaticien du domaine d'étude, semble être la plus prometteuse. Il semble plus facile de définir des outils adaptés à l'exploration de données, au classement et permettant à l'expert du domaine d'opérer quand et comme bon lui semble. L'objectif d'une telle démarche permet un retour d'informations à l'expert en fonction de ses choix de représentation, d'étude de qualité et d'hypothèses de travail, formulés à partir de son savoir-faire tout en minimisant les interactions.

Les travaux de recherche présentés au travers de ce document ont été réalisés au sein du laboratoire GREYC, de l'Université de Caen, et en collaboration avec le cabinet en propriété industrielle et intellectuelle, Breese Derambure Majerowicz de Paris. Ces travaux s'orientent sur la troisième démarche consistant à désengager l'informaticien du processus de modélisation d'un domaine, avec de bons outils, ergonomiques, rapides et donc fonctionnels.

Avec des outils dédiés, l'expert peut aisément mener une étude sur les comportements de ses clients, la qualité de son processus de fabrication, ou son cœur de métier. L'objectif fixé est d'obtenir des outils, en gardant à l'esprit un aspect générique, d'interdisciplinarité, et de retour d'information le plus abouti possible.

Quelque soit l'étude réalisée, la matière première nécessaire de la fouille de données est l'information. Celle-ci peut être issue de capteurs, d'observations ou encore de documents. Les deux premiers aspects peuvent être assimilés en gardant bien à l'esprit que pour le premier cas de figure, les capteurs sont purement mécaniques et que dans le second cas de figure, les observations, c'est l'homme qui joue un rôle de capteur. Pour ces deux types d'intégration de données, le système peut être défectueux.

Vient alors se poser le problème de ce qui est à formaliser. Il existe des différences entre les problèmes formalisables (une chaîne de production automatisée par exemple) et les problèmes afférents à une décision humaine (parfois subjective et pouvant reposer sur des documents ambigus). Dans le premier cas de figure, il est possible d'étudier les données : elles caractérisent ce que nous voulons observer, mesurer. Pour les problèmes humains, la difficulté est toute autre. Il faut pouvoir être en mesure de traiter des problèmes de sémiotiques, de pluridisciplinarités et de motivation du *capteur* humain.

Dans ce mémoire, nous ne nous intéressons qu'aux problèmes humains, impliquant un raisonnement particulier, subjectif et étroitement lié aux données sur lesquelles ils reposent. Ce genre de tâche correspond à une grande majorité des problèmes à traiter. En effet, les systèmes mettant en jeu des procédés automatiques sont plus faciles à étudier : les données sont dans un domaine défini précisément et ne sortent pas des limites fixées par leur nature même. Si les performances des capteurs des "données à formalisées" sont en question, nous pourrons toujours trouver les possibilités pour les optimiser. Dans notre cas, les problèmes issus de décisions humaines imposent de comprendre ou d'être en mesure de traiter des capteurs humains basés essentiellement

sur des facultés de perception et de raisonnements purement contextuels ainsi que de mettre en valeur la sémantique et la sémiotique du raisonnement de l'expert.

A cela s'ajoutent les problèmes d'interactions entre informaticiens et experts présents pour la mise en place d'outils d'aide à la décision. Ces interactions impliquent une pleine et entière collaboration entre les intervenants, d'être capable de dégager du temps de travail pour la réalisation de cette tâche. La réalité est tout autre et l'implication des experts reste une action difficile.

Pour répondre à ces attentes, nous nous proposons de mettre à la disposition des experts des outils et des méthodes de travail afin de leur permettre d'étudier et de maîtriser de bout en bout un processus décisionnel adapté à leur domaine et à leurs contraintes.

Chapitre 2

L'extraction de connaissances à partir d'une base de données documentaire

Sommaire

Le déploiement d'une application reposant sur un ensemble volumineux de données à des fins de consultation ou d'archivage nécessite le passage des documents utilisés à un formalisme de base de données. Ce passage s'inspire d'un protocole bien précis de découvertes de connaissances à partir d'un ensemble documentaire (Knowledge Database Discovery). Ce protocole a pour motivation l'étude, l'adaptation et la mise en place d'outils permettant le traitement des données étudiées.

Cependant, dans le domaine juridique, en général, et dans notre projet de recherche en particulier, l'unité d'information de base n'est pas une donnée structurée, supervisée, mais un document peu structuré et supervisé comme nous l'avons présenté précédemment. Cette différence relative à la structure du document entraîne le besoin de disposer d'outils adaptés pour extraire de ces documents, l'information utile et pertinente afin de répondre au problème de la mise en place d'un processus d'apprentissage.

Nous allons aborder au cours de ce chapitre les enjeux de la recherche et de la découverte de connaissances au travers des bases de données documentaire (2.1 page suivante). Nous sommes

conscient que la fiabilité de transformer le document juridique en base de données risque d'être problématique tant les documents électroniques imposent des contraintes dans leur traitement. Toutefois, cette étape est indispensable pour toute démarche d'apprentissage interactive.

Ces enjeux sont de plusieurs natures telles que le gain de temps pour l'archivage des jugements, l'accessibilité à ces jugements, la synthèse d'informations à partir des décisions, et assurent à qui les maîtrise une avance commerciale ou technologique non négligeable, voir indispensable.

Dans un second temps, nous continuerons notre présentation sur le rôle de l'extraction de connaissances dans une approche globale et volontairement vulgarisée de la chaîne de traitement de l'information intervenant dans le protocole d'étude de bases de données (section 2.4 page 31). Après cette vue d'ensemble, nous aborderons les connaissances obtenues par l'étude d'une base de données documentaire : quels en sont les objectifs, comment doivent elles être structurées, qu'en attend l'expert.

2.1 Enjeux

Les enjeux de l'extraction d'informations à partir de bases de données documentaires sont étroitement liés à la dématérialisation de l'information. Celle-ci est de moins en moins exprimée sous forme de documents papiers pour prendre un aspect immatériel : le document électronique ou *e-document*.

Nous sommes alors passés d'une information peu structurée (chacun mettant en place des mécanismes propriétaires de représentation d'information), à une standardisation de la représentation, influencée par l'émergence de modèles relationnels tels que les bases de données `SQL` et maintenant `XML`. Ce type de standard permet un "langage universel" et est adaptable pour exprimer les connaissances et archiver l'information. Cela permet aussi d'obtenir des documents hétérogènes et multimodaux. Les documents sont ainsi enrichis, mais il incombera alors au concepteur du document de tenir compte de certains traitements d'extraction d'informations aux différents supports modaux inclus (l'extraction d'informations à partir d'un tableau ne repose pas sur les mêmes principes qu'à partir d'un texte). Avec le temps et un peu de pratique Internet, il est aisé de s'apercevoir que tous les domaines sont concernés : commerce, e-gouvernement ou encore santé et dans le cadre de ces recherches le domaine juridique. Quel que soit le domaine, les besoins sont généralement les mêmes : découvrir de nouvelles connaissances sur un ensemble volumineux de données (tant en nombre d'enregistrements qu'en nombre de propriétés) par des méthodes statistiques.

Le domaine juridique s'inscrit également dans cette problématique de mutation vers le document électronique, apportant de nouvelles sources documentaires, témoignant du savoir-faire et de la connaissance des différents experts impliqués dans le processus de création du document électronique tant sur le fond que sur la forme. Disposant d'une telle masse de données, les ma-

gistrats et avocats sont désormais désireux de l'exploiter à des fins de modélisation pour étude, d'aide à la décision ou encore de compréhension de raisonnement.

C'est dans ce domaine que la découverte de connaissances à partir de bases de données s'est fixée. Selon [Feldman et al., 1998], l'objectif de ce domaine est d'extraire de façon non triviale des connaissances implicites, inconnues auparavant, et potentiellement utiles pour l'étude d'un phénomène ciblé. Les grandes applications de l'extraction de connaissances pour le domaine juridique sont typiquement de cibler des jugements répondant à des spécifications bien précises que l'on ne peut observer sans un traitement automatique. La découverte de comportements décisionnels inattendus, l'aide à la prise de décision pour répondre à des situations reproductibles mais variant dans le temps, telle une décision juridique, en sont des exemples. L'extraction de connaissances est aussi reconnue pour permettre de retrouver de l'objectivité dans une prise de décision conditionnée au fil du temps par un savoir-faire de plus en plus précis.

2.2 Objectifs de connaissances recherchées

Définir les objectifs de connaissance recherchée par un processus d'extraction de connaissances dépend étroitement de son utilisation finale et très souvent nécessite l'intervention de l'expert. Cependant, la connaissance doit être exprimée de manière à ce que l'expert puisse la comprendre et la manipuler aisément. Cela peut par exemple impliquer la mise en place d'un dictionnaire (ou d'une ontologie) afin d'établir des relations entre valeurs numériques et connaissances sur ces valeurs. Dans le cas de `CATMInE`[5], la découverte de connaissances est orientée sur la caractérisation des traits qualifiant ou non une contrefaçon de marques nominatives. Le problème est par exemple soulevé lorsque l'expert cherche à établir une relation de similarité entre marques nominatives. S'il établit comme relation le nombre commun de caractères entre deux marques, à partir de quelle proportion peut-il fonder une relation de contrefaçon entre celles-ci ? Ces tâches ne peuvent donc qu'être réalisées par un expert.

Dans un autre cadre d'utilisation de la connaissance induite, les objectifs peuvent être l'optimisation de l'accès aux données par l'intermédiaire d'outils d'exploration de données reposant sur ces connaissances. Les propriétés de comparaison et de sélection doivent caractériser la connaissance du domaine. Ici aussi, `CATMInE`, par le biais d'outils d'exploration et de visualisation de données, assiste l'expert à mieux identifier les données (par le biais d'indice de qualité) et à accéder de façon intuitive aux documents électroniques juridiques composant la base de données.

[5] `Computer Assisted Trade-Mark Infringement Evaluation`

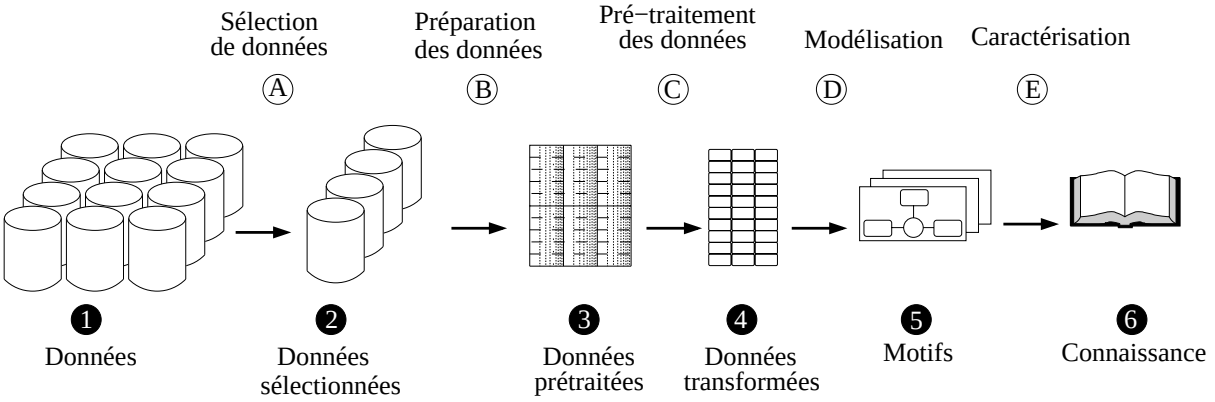


FIG. 2.1 – Vue d'ensemble des différentes étapes constituant la processus de fouille de données et d'extraction de connaissances.

2.3 Extraction de connaissances : présentation des principes généraux

L'obtention de connaissances par la fouille de données (acquisition ou extraction) est un processus, comprenant une série de traitements réalisés sur un ensemble de données. Les données sont usuellement obtenues à partir d'observations. Dans le cadre de `CATMInE`, ces observations sont d'ordre juridique, et sont caractérisées par le biais de jugements décrivant une décision de justice. La série de traitements réalisée permet de préparer et de raffiner les données en vue d'établir une collection, la plus rigoureuse possible et donc homogène, de l'ensemble de départ. Dans notre cadre applicatif, les traitements appliqués aux données permettent l'intégration d'un ensemble de décisions de justice (qui sont des documents électroniques) en éléments exploitables par un processus décisionnel. Lorsque les données sont préparées, il est alors possible d'appliquer un algorithme de classification ou de catégorisation afin d'en extraire de la connaissance relative aux relations intrinsèques des données. Le schéma 2.1 synthétise ces différentes étapes est repose sur les travaux de Usama Fayyad ([Fayyad et al., 1996b], [Fayad et al., 1996]), et Sushmita Mitra ([Mitra et al., 2002]).

D'un ensemble de données (1) plus ou moins volumineux, l'expert sélectionne (A) un échantillon considéré par la suite comme représentatif du phénomène à étudier (2). Cet ensemble peut constituer l'intégralité de l'information disponible tout comme une infime partie. Cette première étape (A) est dite de sélection des données. Nous reviendrons plus en détails sur cette étape dans le chapitre 3 page 42 afin de présenter les différentes méthodes applicables.

Dans un deuxième temps, l'ensemble des données sélectionnées (2) est préparé (B). Cette étape de préparation des données, présentée dans le chapitre 4 page 54, permet de se séparer des enregistrements incomplets et d'obtenir une base considérée comme rigoureuse (3). Cette base a

la particularité de ne contenir aucun enregistrement incomplet et dont l'ensemble des données a été validé.

La troisième étape, pré-traitement (C) des données (chapitre 5 page 70), est appliquée s'il est nécessaire de changer l'échelle de certains descripteurs, de segmenter des valeurs continues, de combiner des descripteurs ou encore d'éclater des descripteurs (4). Cette étape est étroitement liée au savoir-faire acquis par le professionnel sur ce domaine d'étude.

La quatrième étape (D) correspond à l'utilisation d'un outil de catégorisation (identifiant un phénomène) ou de classification (explicitant un phénomène identifié). Cette étape recherche des motifs récurrents et fréquents dans le jeu de données afin de regrouper ceux-ci selon des caractéristiques communes (5). Le chapitre 6 page 82 y est consacré.

Enfin, la cinquième et dernière étape (E) est l'obtention de connaissances formalisées (6) pour une compréhension par l'expert et une utilisation future dans un algorithme de prise de décision (chapitre 7 page 102). Ce processus est orienté selon deux axes : soit il est réalisé de manière supervisée (approche explicative), soit de manière non supervisée (approche exploratoire). Cette distinction intervient dans de nombreuses étapes du processus et représente un parti pris dans les choix réalisés. Ce mémoire se propose donc de suivre cette logique.

2.4 Pré-requis nécessaires à l'établissement d'un processus décisionnel

L'objet de cette section est de présenter la façon dont l'expert va devoir formaliser son problème pour le traiter selon le protocole présenté dans la section précédente. Ce protocole va devoir définir deux éléments : la finalité de l'apprentissage et les descripteurs pour caractériser la connaissance. En connaissant la finalité, et en ayant raffiné au mieux son expression, l'expert peut alors déterminer par son savoir-faire les éléments (ou descripteurs) permettant de caractériser cette valeur (appelée valeur de classe, ou classe). La valeur de classe est un descripteur comme un autre, mais son utilisation dans le processus de modélisation diffère.

Pour `CATMInE`, la finalité de l'apprentissage est d'identifier une contrefaçon de marques nominatives. Les marques nominatives n'étant que des chaînes de caractères, un traitement morphologique à défaut de syntaxique sur celles-ci semble naturel : la contrefaçon de marque concerne le signifiant et non le signifié. Afin d'économiser les frais de la mise en place d'une ontologie pour résoudre le problème sémantique, celui-ci n'est pas pris en compte dans la réalisation de l'application.

Pour suivre notre démarche, nous allons aborder le choix des descripteurs pour formaliser les données, puis nous reviendrons plus en détails sur cette notion de valeur de classe et de son utilisation.

2.4.1 Choix des descripteurs

Dans un processus de modélisation d'un domaine, l'expert cherche à mettre en valeur un phénomène afin de le caractériser. Cette caractérisation permettra par la suite une meilleure compréhension du phénomène étudié.

Descripteurs, modalités et définitions

Pour réaliser une telle démarche, il est nécessaire d'extraire les éléments caractérisant le phénomène. En fonction de ses croyances et connaissances sur le domaine et le phénomène, il devient possible pour l'expert d'identifier des éléments dans le flux de données participant à l'expression du phénomène. Ces éléments caractéristiques sont appelés descripteurs ou encore attributs.

La caractérisation d'une contrefaçon de marques nominatives dans `CATMInE` est exprimée par la recherche de traits communs entre les mots impliqués dans les marques. Nous pouvons alors dégager des propriétés propres aux caractères utilisés dans chaque marque, telles que le nombre de caractères en commun. Cette hypothèse corrobore le fait que s'il y a contrefaçon c'est qu'il y a similitude. En droit, nous parlons alors de confusion dans l'esprit du consommateur. Or entre deux mots, la ressemblance est en partie exprimée par l'unité caractère. Dans les mots `co-voiturage` et `voiturette`, l'ensemble des caractères {v ;o ;i ;t ;u ;r ;e} est partagé. Autre exemple, le descripteur prenant en compte la position de la plus longue sous-chaîne commune entre les deux marques permet d'indiquer si celle-ci se trouve en début, en milieu ou en fin de marque. L'importance est alors portée sur le fait que si la sous-chaîne la plus longue commune est en attaque de marque (en début de marque) alors l'impact sur le consommateur sera plus forte que si elle est en fin de marque.

Un descripteur est une relation entre un nom (ou identifiant) définissant une propriété et des valeurs ou modalités propres à cette dernière. Le nom désigne et qualifie une caractéristique du phénomène étudié et les modalités, des valuations de cette caractéristique. Les modalités d'un descripteur, peuvent être de deux natures : symboliques ou numériques tel que l'expliquent [Fayyad & Irani, 1993].

Les modalités symboliques sont généralement des mots, décrivant une propriété de l'attribut. L'ensemble des modalités constituant un descripteur permettent de caractériser une propriété en fonction d'un contexte. Dans `CATMInE`, le descripteur prenant en compte la position de la plus longue sous-chaîne commune est une utilisation de valeurs symboliques pour le définir (début, milieu, fin). Les modalités symboliques ont la particularité d'être distinctes les unes des autres.

Opposées aux modalités symboliques, les modalités numériques caractérisent généralement un ensemble continu et borné ou plus généralement un attribut ordonné linéairement. Pour faire le parallèle avec notre application `CATMInE`, l'exemple du descripteur représentant le nombre

de caractères en commun entre deux marques est un rapport entre le nombre de caractères en commun et la taille de la marque. Nous pouvons ainsi observer si une marque est reprise pour contrefaçon à plus de 60% par exemple. Les descripteurs continus sont riches en informations car ils correspondent à une donnée brute, non raffinée par l'identification de valeurs charnières. Ces valeurs charnières représentent une connaissance supplémentaire sur le descripteur. L'expert pourrait caractériser une contrefaçon par le nombre de caractères en commun entre les marques incriminées, une simple mesure numérique (un rapport, donc un descripteur continu) et ensuite affiner ce descripteur en ajoutant qu'il a la connaissance de ce qui distingue une contrefaçon d'une non-contrefaçon : le pourcentage de caractères en commun doit être supérieur à 60% [6], selon les attentes de l'expert pour qu'il y ai contrefaçon et pour une non-contrefaçon, une proportion inférieure à 30%[7]

Le passage d'attributs continus à attributs symboliques s'appelle la segmentation. [Fayyad & Irani, 1993] rappellent le principe de segmentation et proposent une généralisation de la discrétisation binaire. Cependant, segmenter des valeurs continues en fonction de méthodes statistiques n'est peut-être pas optimal. L'intervention de l'expert pour établir une segmentation d'attributs peut introduire une dimension sémantique à cette étape en proposant des seuils pour caractériser l'intervalle de définition des attributs.

Identification des descripteurs

Les descripteurs mis en place permettent de caractériser au mieux un jeu de données. Cependant, il faut pouvoir être en mesure de bien définir les descripteurs. L'exemple suivant (extrait de [Gordonn & des Jardins, 1995]), d'aspect simpliste et pédagogique, présente cet intérêt de bien définir et choisir les descripteurs.

Si l'on considère un monde fait d'objets, et que l'on s'intéresse à ceux capables de tenir sur une table. L'expert dispose de plusieurs descripteurs dont ceux décrivant la couleur et la taille des objets. Il dispose aussi d'un autre, définissant la forme des objets. Deux exemples sont présentés au système d'apprentissage : un petit cube bleu (stable), et une petite boule rouge (instable). Le système peut alors induire deux règles, toutes les deux consistantes au vu des exemples :

1. *règle 1 :* tout ce qui est bleu sera stable, tout ce qui est rouge sera instable
2. *règle 2 :* Tout ce qui est cubique est stable, tout ce qui est sphérique ne l'est pas

Supposons maintenant que l'on soumette une petite boule bleue à ce système composé de ces deux règles. Avec la *règle1*, le système déclarera la boule stable (alors qu'elle ne l'est pas) et la seconde règle révélera l'instabilité de l'objet. Évidement, la *règle 2* est meilleur oracle que la première règle.

[6] Cette valeur est juste là, à titre pédagogique, il n'en est probablement rien dans la réalité.
[7] idem.

Ce petit exemple démontre bien qu'une sélection judicieuse des descripteurs en fonction du savoir-faire est plus que nécessaire pour garantir la pertinence du modèle et réduire le risque d'erreur. Dans l'exemple, la forme suffisait à résoudre le problème avec un descripteur porteur de sens pour la compréhension du phénomène de stabilité. L'établissement de ces premiers descripteurs résulte d'une construction établie par l'expert en fonction des besoins liés à l'étude. Il est donc important de définir :

- ce sur quoi l'apprentissage porte : quel phénomène cherche t-on à d'écrire ?
- quels sont les motifs factuels décrivant le phénomène ?

Ainsi, en fonction des croyances et connaissances de l'expert sur le domaine, celui-ci peut identifier des éléments participant à la caractérisation du phénomène que l'expert cherche à modéliser. Le rôle de l'expert, dans nos travaux de modélisation de la contrefaçon de marques nominatives, a consisté alors à déterminer aux travers des documents électroniques disponibles les informations utiles pour caractériser le phénomène. Outre les noms des marques, les références aux décisions antérieures pour les jugements de Cour d'Appel ou de Cours de cassation permettent de ne tenir compte que de la dernière décision appliquée. Cet aspect du phénomène soulève alors le problème de tenir compte ou pas de l'intégralité des décisions et donc de la base documentaire. Si tel était le cas, le système réaliserait une modélisation complète du domaine mais il devrait composer avec des décisions paradoxales dans le cas d'appel ou de pourvoit où la décision seraient "cassée". En utilisant que la dernière décision appliquée dans une décision pour la modélisation du phénomène, le système se conformera alors à la définition du phénomène, mais l'intégralité du domaine ne sera pas mis en valeur, et donc certains raisonnements juridiques seront omis (même si un juge est revenu sur la décision).

Ainsi, dans `CATMInE`, l'expert a identifié les informations suivantes comme descripteurs permettant de caractériser la contrefaçon de marques nominatives :

1. le nombre de caractères en commun entre la marque du plaignant et la marque du défendant
2. le nombre de caractères en commun entre la marque du défendant et la marque du plaignant
3. la plus longue sous-chaîne graphémique commune pondérée par la position (début, milieu, fin) entre la marque du plaignant et la marque du défendant
4. la plus longue sous-chaîne graphémique commune pondérée par la position (début, milieu, fin) entre la marque du défendant et la marque du plaignant
5. le nombre de phonèmes en commun entre la marque du plaignant et la marque du défendant
6. le nombre de phonèmes en commun entre la marque du défendant et la marque du plaignant
7. la plus longue sous-chaîne phonémique commune pondérée par la position (début, milieu, fin) entre la marque du plaignant et la marque du défendant
8. la plus longue sous-chaîne phonémique commune pondérée par la position (début, milieu, fin) entre la marque du défendant et la marque du plaignant

Nous obtenons alors pour la décision *Golf plus vs. Golf'us (Contrefaçon - TGI Paris 01/09/1999)*, les descripteurs suivants :

1. *Golf plus* contient tous les caractères de *Golf'us*
2. *Golf'us* contient six caractères sur les neuf de *Golf plus*
3. *Golf* est la plus longue sous-chaîne commune en mot d'ouverture de *Golf plus*
4. *Golf* est la plus longue sous-chaîne commune en mot d'ouverture de *Golf'us*
5. *Golf plus* contient tous les phonèmes de *Golf'us*
6. *Golf'us* contient huit phonèmes sur les neuf de *Golf plus*
7. *golf* est la plus longue sous-chaîne commune en mot d'ouverture de *Golf plus*
8. *golf* est la plus longue sous-chaîne commune en mot d'ouverture de *Golf'us*

En plus des marques intervenant dans la décision, des informations connexes comme le lieu de la décision, la date, le niveau décisionnel, la référence et les parties ont été retenues afin de constituer le jeu de données.

A cette difficulté d'établir les bons descripteurs pour caractériser un domaine, s'ajoute celle de l'obtention des modalités pour constituer le jeu de données. Les modalités sont souvent obtenues à partir de la base documentaire, par le biais d'extractions d'informations textuelles. Ces informations sont isolées par la recherche de motifs précis dans le document ([Nédellec, 2000], [Azé & Roche, 2003]) et caractérisées par des règles d'extraction.

Dans le cadre des bases documentaires juridiques, la qualité de l'extraction des données ne peut être garantie (sauf cas exceptionnel bien évidement). Il faudrait pour cela que toutes les décisions soient rédigées de la même manière, avec le même vocabulaire. Évidement cet aspect est loin d'être possible, chaque juge utilise son propre style pour rédiger une décision. Cependant, certains magistrats commencent à se tourner vers des outils de rédaction à base de clauses, y voyant un gain de temps dans la rédaction. Mais pour d'autres, utiliser de tels outils revient à les priver de leur liberté rédactionnelle. Proposer alors des règles d'extraction d'informations devient délicat, il faut pouvoir adapter les règles à chaque rédacteur. Dans le chapitre 1 page 17 de ces recherches nous avons présenté ce phénomène en explicitant que les documents sont très proches les uns des autres car ils caractérisent la contrefaçon de marques, mais que leur différence porte sur la rédaction des décisions, et que la réalisation technique de cette extraction est particulièrement complexe.

Enfin, la base de données documentaire initiale avant tout traitement oriente implicitement le choix des descripteurs. L'expert ne peut proposer des descripteurs que si l'information qui les constitue est présente dans le jeu de données. Il faut alors pouvoir extraire cette information et la formaliser le mieux possible.

Parmi tous les descripteurs mis en place pour caractériser et étudier un jeu de données, il y en a un particulier, le descripteur caractérisant ce que l'expert cherche à mettre en valeur : la

valeur de classe. Cette valeur de classe représente la finalité de l'apprentissage et donc de la modélisation du phénomène permettant au système de répondre à une question bien identifiée.

2.4.2 Finalité de l'apprentissage

La finalité de l'apprentissage correspond au phénomène que l'expert cherche à mettre en valeur afin d'en identifier les fondements, d'en approfondir la connaissance, voire d'en comprendre la démarche intellectuelle de raisonnement : sur quels principes les décisions sont-elles retenues ? Cette finalité est implicitement définie par les documents structurés et supervisés utilisés comme support de la modélisation. En effet, les documents étant supervisés, ils répondent alors à une ou plusieurs contraintes : la première étant de caractériser ce pour quoi ils ont été établis. Lorsque l'expert sait ce qu'il cherche à mettre en valeur, il va alors orienter la structure du document (aspect supervisé) pour permettre un accès facilité à l'information portée par le document. Les décisions de justice utilisées dans nos travaux présentent une structure tant sur la mise en forme du document que sur l'information contenue afin de permettre une lecture du document à deux vitesses. Dans un premier temps un résumé du document (lecture rapide) et dans un second temps la décision complète du tribunal. Cette décision est là aussi structurée afin de répondre aux éléments procéduraux d'une décision de justice. En l'occurrence le document doit répondre à des questions précises sur :

1. quels sont les faits ?
2. quelle a été la procédure ?
3. le jugement a-t-il raison d'être ?
4. y-a-t-il contrefaçon ?
5. la concurrence déloyale, s'il y a.
6. la décision rendue et ses motivations.

Il est donc important de :

- définir ce sur quoi l'apprentissage porte : quel phénomène cherche-t-on à décrire ?
- définir quels sont les faits décrivant le phénomène
- définir la façon d'extraire ces éléments des données factuelles
- s'assurer de l'efficacité de l'extraction (perte d'exemples si le procédé est trop sélectif)

Le premier point de cette réflexion consiste à déterminer le phénomène dans le flux de données textuelles. Le phénomène s'exprime toujours sur deux à plusieurs modalités opposées : soit A l'expression d'un phénomène, alors $\neg A$ (non A) en représente le contraire. Dans beaucoup de domaines, l'intervalle entre A et $\neg A$ autorise des flexions du phénomène (et donc de son expression). Ces flexions représentent alors une sous-partie du phénomène, celui-ci n'ayant pas nécessairement deux issues possibles, et peut être détaillé en plusieurs issues distinctes (plus ou moins). Dans

CATMInE, la contrefaçon de marques nominatives admet comme flexions la contrefaçon par reproduction (reproduction d'articles) ou encore la contrefaçon par imitation. Les oppositions ainsi établies correspondent à la définition d'un attribut de classe. Chaque enregistrement du jeu de données doit ensuite pouvoir être associé à une des flexions (ou modalités) du descripteur de classe. S'il n'est pas possible de le faire, alors la valeur de classe est mal identifiée (elle ne décrit pas tous les exemples), et nécessite d'être redéfinie afin que chaque exemple puisse être convenablement identifié et caractérisé.

2.5 Conclusion

Ce chapitre est axé sur la nécessité de bien formaliser ce que l'expert cherche à mettre en valeur au travers d'un jeu de données qui est généralement produit avant d'avoir envisagé de mettre en place un processus de découverte de connaissances, ou d'apprentissage.

Nous avons aussi présenté les procédés que l'expert doit suivre pour mettre en place une base de données d'étude à partir de la base documentaire initiale. Ce passage implique de définir des descripteurs, et de rechercher l'intégralité des modalités possibles disponibles à travers la base documentaire. Cette opération doit être rigoureuse afin de garantir une sémantique pertinente et utile à la compréhension du phénomène.

Toutefois, la fiabilité de l'expert dépend strictement de la qualité et de la quantité des données à traiter. Le choix de la fenêtre d'exploration (l'échantillonnage) est primordial pour garantir une pleine réussite de la modélisation. Un système interactif semble alors nécessaire pour assister l'expert dans sa tâche. Cette interactivité doit se faire au travers d'interfaces et de connaissances ciblées.

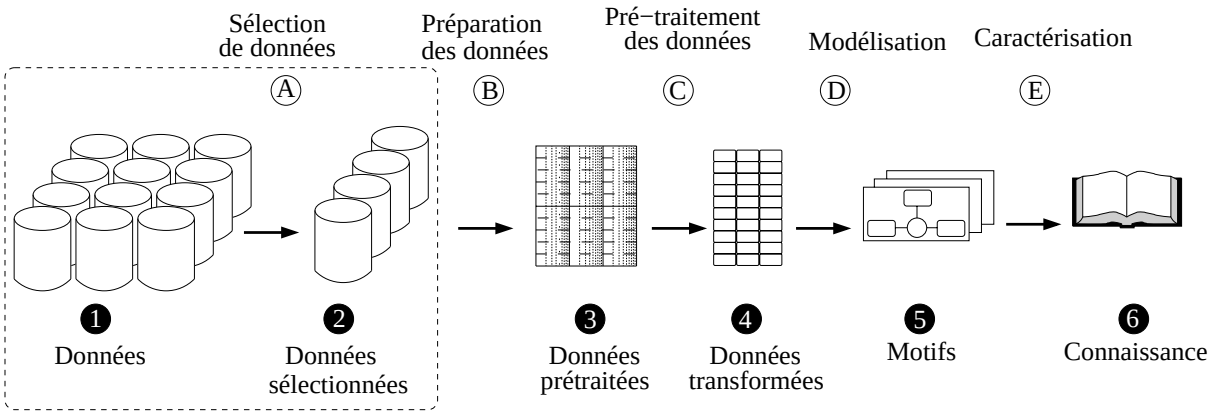
Les éléments caractéristiques obtenus par cette plate-forme interactive sont ensuite utilisés pour produire un modèle du domaine, principe présenté dans les parties suivantes.

Première partie

Processus de modélisation d'une base de données documentaire

Chapitre 3

La sélection de données à partir de bases documentaire (data archeology)



La sélection de données : échantillonnage d'un ensemble documentaire en un sous-ensemble ciblé.

Sommaire

Le processus de fouille de données que nous avons présenté dans le chapitre précédent propose, avant tout travail de modélisation, de déterminer le jeu de données utilisé pour la suite de l'étude d'un phénomène : nous parlons alors d'échantillonnage du jeu de données. En effet, lors de cette étude, l'ensemble des données disponibles (ou encore enregistrements) peut ne pas être adapté. Le domaine d'étude peut être une partie d'un domaine beaucoup plus vaste dans lequel l'information utile est à isoler : les bases de données contenant des millions d'enregistrements sont désormais courantes et peuvent exprimer plusieurs phénomènes, qui sont guidés par la structure des documents utilisés.

Dans notre cadre applicatif, la sélection des jugements utiles à la modélisation de la contrefaçon de marques nominatives permet d'éliminer de l'ensemble des jugements tous ceux ne correspondant pas à notre cadre de recherche. Ainsi, les jugements caractérisant les contrefaçons de marques figuratives (logotypes et enseignes) peuvent être écartés de la base. D'autres jugements présentant des défauts ou absences de données comme le texte de la décision ne sont pas retenus.

Il est donc nécessaire de sélectionner parmi tous les enregistrements ceux qui correspondent à la cible de l'étude. Pour mieux en saisir l'intérêt, nous proposons dans un premier temps de présenter les motivations de la sélection de données et l'apport de cette étape sur la suite du processus. Enfin, différentes techniques de sélection seront présentées, pour conclure sur la méthode retenue dans ce projet.

3.1 Les motivations

La sélection de données, dénommée *data archeology* (par analogie à l'archéologie), est résumée dans [Brachman et al., 1993] par le fait qu'elle a pour vocation de dériver de la connaissance à partir de l'étude d'autres données (pour l'archéologue, d'artefacts), replacées dans leur contexte (en archéologie, le contexte est historique, culturel) et à partir desquelles des hypothèses sont établies, des croyances adaptées. La sélection de données a deux objectifs : faciliter l'utilisation des algorithmes d'apprentissage, et faciliter la compréhension des résultats produits. Nous allons détailler ces deux points pour ensuite présenter les deux méthodologies de sélection de données. La première, dite non supervisée qui est à opposer bien évidement à la seconde, supervisée.

3.1.1 Faciliter l'apprentissage

La première des motivations de la sélection de données repose sur la complexité croissante des bases de données. Avec le temps et les besoins, les bases sont de plus en plus volumineuses en terme de dimensions (nombre d'enregistrements et de descripteurs). Or les algorithmes d'apprentissage ont pour objectif la recherche de relations intrinsèques entre enregistrements d'une base, voire de relations entre descripteurs. La complexité de ces algorithmes est alors exponentielle en fonction des dimensions du jeu de données, d'où la nécessité de les réduire au possible.

Philip Chan et Salvatore Stolfo dans [Chan & Stolfo, 1996] expliquent brièvement que plus la base est conséquente, plus les algorithmes perdent en performance. Les algorithmes de modélisation ont souvent besoin de charger intégralement le jeu de données en mémoire. Quand l'ensemble de données est plus important que la mémoire disponible (physique et virtuelle), les calculs deviennent irréalisables. Ces même auteurs [Chan & Stolfo, 1995] proposent comme solution de segmenter la base en sous-ensembles afin d'alléger les traitements en n'utilisant qu'un échantillon de la base de données. Cette segmentation garantit alors que les sous-ensembles produits peuvent être chargés en mémoire. Le principe consiste à produire des ensembles disjoints de données respectant les propriétés du jeu initial.

Hannu Toivonen [Toivonen, 1996] propose, quand à lui, une sélection aléatoire du jeu de données. Enfin, Paul S. Bradley et al [Bradley et al., 1998] innovent en identifiant dans une base de données des zones similaires pouvant être compressées, de celles indispensables ou encore inutiles pour les futurs traitements. En sélectionnant un sous-ensemble de données, considéré comme représentatif, il peut devenir plus aisé pour les algorithmes de déterminer des propriétés (le processus ne sera pas biaisé par des données caractérisant un autre phénomène) interférant sans pour autant avoir à parcourir l'intégralité d'une large base. Les algorithmes gagnent alors en efficacité de traitement et en temps de calcul.

L'ensemble des données disponibles pour `CATMInE` ne dispose pas d'une structure suffisamment enrichie pour permettre la sélection des contrefaçons nominatives parmi l'ensemble des jugements

disponibles. Il est donc nécessaire d'identifier les décisions impliquant deux marques déposées et se contrefaisant parmi l'ensemble des décisions disponibles (comme la contrefaçon d'une marque par l'enseigne d'un magasin).

3.1.2 Faciliter la compréhension des résultats

La seconde motivation de la sélection de données est liée à la production des résultats. Lorsqu'une base de données est conséquente en terme de dimensions, par exemple plusieurs centaines de milliers d'enregistrements, pour plus d'un millier de descripteurs, la qualité des résultats sera étroitement liée à ces dimensions : quantité, pertinence, etc... En effet, sur de gros volumes, la combinatoire entre les descripteurs explose de manière exponentielle. Le volume de connaissances induit est associé à ce volume : la connaissance est liée aux relations entre enregistrements et descripteurs. La compréhension des résultats devient donc une tâche ardue, nécessitant plusieurs heures de traitements afin d'en isoler les informations les plus pertinentes (et qu'il reste à qualifier comme telles). L'objectif est donc de réduire l'ensemble des résultats produits et, dans un même temps, de cibler plus finement l'étude de ceux-ci. La suite de ce chapitre s'attache donc à présenter les différentes méthodes de sélection de données (en terme d'enregistrements), les deux grands axes admis dans ce domaine : la sélection non supervisée et celle dite supervisée.

3.2 Les méthodes non supervisées d'échantillonnage de données

La sélection de données, réalisée de manière non supervisée, consiste en une sélection d'exemples sans *a priori* ni parti pris sur l'issue ou la catégorie des données. Ce principe de sélection ne garantira pas la représentativité des données retenues relativement à l'ensemble. Il est judicieux de s'interroger sur de telles méthodes. En ne réalisant pas un apprentissage fidèle aux données proposées, le système modélise alors des connaissances qui ne seront peut être pas utiles (tout dépendra du jugement de l'expert). Les principes de sélection peuvent être présentés sur deux axes : le tirage aléatoire de celles-ci, et la sélection d'un jeu par l'expert. Dans le premier cas de figure, une sélection à l'aveugle, dans le second cas, l'expert biaisera l'apprentissage en orientant la modélisation des connaissances, orientation induite par son savoir-faire.

3.2.1 Par sélection aléatoire

La sélection aléatoire est la plus naturelle et la plus naïve des méthodes de sélection de données, en fixant le volume de données souhaité (tout ou partie des données disponibles). Cependant, aucune connaissance sur les données n'est utilisée pour la sélection, entraînant une représentativité des données peu fiable. La connaissance induite ne sera pas nécessairement réaliste par rapport au phénomène étudié, et encore moins par rapport aux données de départ.

Si les enregistrements se distinguent en deux catégories, et si l'algorithme de sélection ne retient que majoritairement des données d'une des deux issues possibles, le jeu constitué ne sera pas représentatif du domaine car la répartition des exemples observés sur la base initiale n'est pas respectée. Cependant, si le jeu est volumineux et hétérogène, une sélection aléatoire non représentative est peu probable.

Cette méthode ne peut pas s'appliquer à notre cadre de recherche. Ici, nous ne sommes pas confronté au problème de volume de données disponibles et répondant toutes à un domaine bien identifié. Bien au contraire, nous avons affaire à une source documentaire répondant à plusieurs phénomènes proches. Les documents doivent donc être sélectionnés de manière à n'identifier que la contrefaçon de marques nominatives enregistrées.

3.2.2 Par l'intervention de l'expert

Avec de gros volumes de données, les systèmes d'apprentissage tendent à produire trop de connaissances qui ne sont peut-être pas en relation avec le phénomène étudié. Cette surproduction de connaissances est étroitement liée au potentiel de corrélations intrinsèques des données. La recherche de motifs caractérisant des sous-ensembles de données peut conduire à établir de nombreux motifs, avec peu d'enregistrements (le support étant faible). Avi Silberschatz et Alexander Tuzhilin [Silberschatz & Tuzhilin, 1996] exposent la nécessité d'avoir des outils de mesure de qualité et d'intérêt pour la connaissance induite. Selon eux, cette mesure doit à la fois être objective (la structure des enregistrements, les relations entre descripteurs) et subjective (qui dépend de l'expert). Or ce qui est vrai au niveau des connaissances l'est aussi au niveau de la sélection de données. L'intervention de l'utilisateur dans la sélection de données de manière non supervisée (l'expert ne sait pas ce que caractérisent les enregistrements mais possède de l'expertise sur le document tant au niveau structurel qu'informationnel) permet d'orienter celle-ci sans pour autant prendre en compte la finalité des éléments. Ainsi l'expert peut retenir un ensemble de données pour des caractéristiques remarquables sans pour autant être capable d'en définir la représentativité au sein de la collection. De plus, l'expert ne sera pas capable de définir leur appartenance exacte à un concept ciblé, cette tâche résulte généralement d'un travail d'analyse, de statistique et de comparaison entre catégories possibles. Enfin, une telle stratégie implique aussi l'omniscience de l'expert sur son domaine, ce qui n'est pas nécessairement acquis.

Sur de gros volumes de données, cette tâche devient très délicate pour l'utilisateur sans l'utilisation d'outils capables de l'assister. A cet effet, Ronald Brachman [Brachman et al., 1993] présente les limites des outils actuels de sélection de données tels que le `SQL` (langage de requêtes). Pour répondre à ces contraintes, il propose une approche reposant sur une représentation formelle de la connaissance. Cette représentation sert alors de support pour en produire une nouvelle, déclarative, du domaine d'application. Celle-ci est ensuite mise en valeur par différents processus

d'inférence réalisant une modélisation hiérarchique de classe afin de produire un langage d'interrogation compréhensif des données.

Ici encore, cette méthode n'est pas nécessairement adaptée à la réalisation de notre étude de la contrefaçon de marques nominatives. Pour la plate-forme `CATMInE`, l'expert a une connaissance de la finalité des documents (contrefaçon ou non). Il est alors en mesure d'orienter l'échantillonnage des données vers un sous-ensemble représentatif du phénomène étudié en ciblant des décisions qu'il jugera comme représentatives.

3.3 Les méthodes supervisées d'échantillonnage de données

La sélection de données de manière supervisée s'oppose à la méthode précédente par la connaissance d'une information supplémentaire sur les enregistrements : la façon dont ceux-ci sont regroupés. En effet, comme nous l'avons vu précédemment (chapitre 2.1 page 28), les données utilisées pour ces travaux de recherches sont regroupées en catégories (définies implicitement par les documents supervisés) : contrefaçon de marques enregistrées, contrefaçon par le biais d'une enseigne et contrefaçon par reproduction. En connaissant une telle information, il devient possible de prendre en compte des informations relatives à la représentativité des données en plus d'une sélection reposant sur la finalité du document. Une telle information permet de construire par la suite un modèle de connaissances plus fidèle aux données que par l'intermédiaire de la méthode précédente, sans avoir pour autant à utiliser l'ensemble des données. Comme pour les sondages d'opinion, un sous-ensemble représentatif peut suffire.

3.3.1 Par validation croisée

La validation croisée (*cross validation*) est une des méthodes de sélection supervisée. Par validation croisée, nous présentons le principe de sélection de plusieurs sous-ensembles de données de façon aléatoire en respectant la représentativité des classes au sein des sous-ensembles créés. Sur chaque sous-ensemble produit, nous pouvons alors appliquer des algorithmes de modélisation de façon raisonnable en terme de ressources : c'est le principe présenté dans [Chan & Stolfo, 1995].

Cette méthode est plus destinée à tester des algorithmes de classifications en créant plusieurs paires de sous-ensembles apprentissage/test à partir d'un jeu de données. Il est donc possible d'adapter ce type d'algorithme pour sélectionner de manière aléatoire les données tout en respectant leur représentativité. Avec une telle méthode, un gros volume de données peut être segmenté en plusieurs éléments de taille quasi identique (dépendant de l'arité des données) sur lesquels il devient possible d'appliquer des algorithmes d'apprentissage de façon raisonnable en terme de temps de calculs. La validation croisée permet de réduire le temps d'apprentissage : la combinatoire entre descripteurs et exemples étant réduite, l'apprentissage d'une base de volume N est plus consommateur de ressources que l'apprentissage de la même base segmentée

en p ensembles de taille identiques et plus petite que l'ensemble de départ. Ainsi, si nous ap-

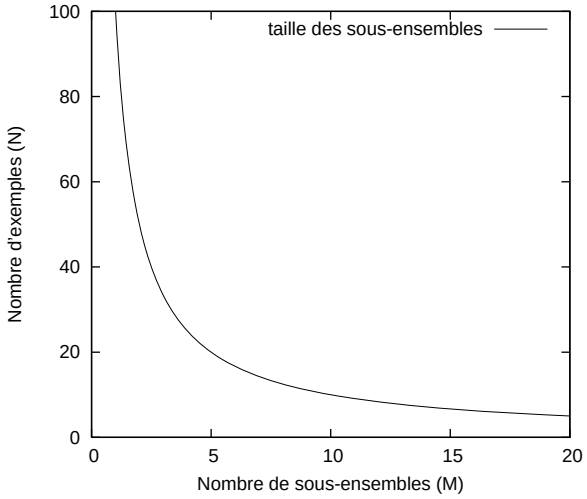


FIG. 3.1 – Évolution de la taille des ensembles d'apprentissage en fonction du nombre de paquets produits.

pliquons cet algorithme sur une base contenant N enregistrements, avec comme paramètre M sous-ensembles souhaités, alors les sous-ensembles seront de taille N/M. Si M tend vers N, les ensembles d'apprentissages seront de plus en plus larges et les ensembles de test de plus en plus petits. Le graphique 3.1 présente cette propriété. Nous réduisons ainsi de l'espace de recherche une des composantes définissant la combinatoire entre descripteurs et exemples.

La figure 3.2 présente le principe de la validation croisée. Pour une base documentaire juridique, nous souhaitons produire six sous-ensembles de données : trois pour l'apprentissage et trois pour les tests de validation de l'apprentissage. Les contraintes appliquées sont :

1. chaque jeu d'apprentissage ne partage aucun exemple avec les autres jeux d'apprentissage,

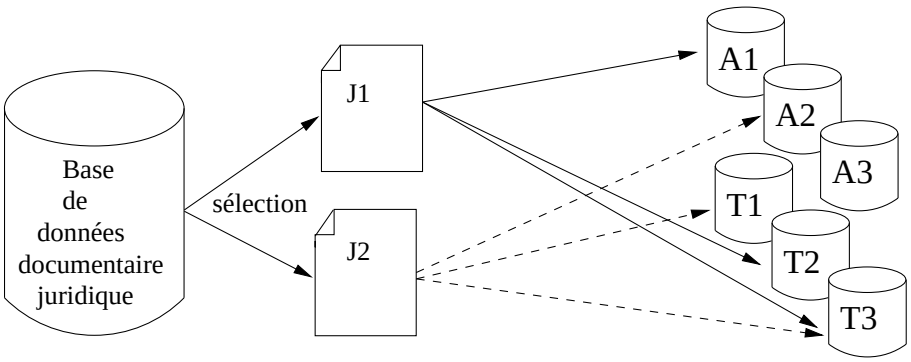


FIG. 3.2 – Principe de la validation croisée.

2. les jeux d'apprentissage et de validation appairés non aucun exemple en commun
3. les répartitions entre classes doivent être respectées.

Ainsi dans notre exemple, le jugement $J1$ est inséré dans la base d'apprentissage $A1$ et dans les deux jeux de validation $T2$ et $T3$. Ensuite, le jugement $J2$ est sélectionné et inséré dans la base d'apprentissage $A2$ et les bases de validation $T1$ et $T3$. L'algorithme continu ainsi jusqu'à ce qu'il n'y ait plus d'exemples dans la base documentaire. Nous disposons alors de trois jeux d'apprentissage/validation $A1/T1$, $A2/T2$ et $A3/T3$ pour réaliser des expériences ou établir des méta-modèles de décision (ces algorithmes sont présentés dans le chapitre 6 page 82).

Cette méthode n'a pas été utilisée dans l'échantillonnage des décisions disponibles à partir de la source documentaire. La finalité des décisions étant établie par la structure des documents, elles sont supervisées mais peu identifiables sans l'intervention de l'expert. En revanche, une fois que les décisions sont connues, nous pouvons les enrichir de ce complément d'informations et appliquer la méthode de validation croisée afin de réaliser des tests de qualité du modèle.

3.3.2 Par Regroupement

Le regroupement (ou *clustering*) a pour principe la division de données en groupes d'objets similaires (*clusters*) selon deux grands principes : par partitionnement ou par hiérarchie. Il existe d'autres méthodes comme le clustering à base de contraintes, à descente de gradient ou encore par projection. Ici, nous nous limitons aux deux plus importantes classes d'algorithmes.

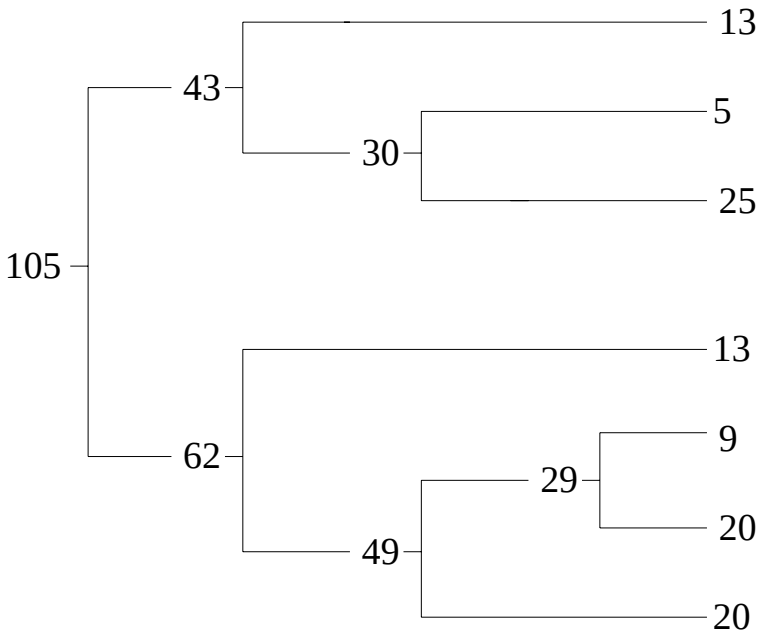


FIG. 3.3 – Dendogramme des partitions associées aux regroupements obtenus sur la base de données de `CATMInE`.

La première, hiérarchique, consiste à établir une hiérarchie de clusters (sous forme d'arbres) appelée dendogramme. Cette méthode permet d'explorer le jeu de données avec une notion de granularité de celui-ci. La figure 3.3 page ci-contre présente ce type de représentation permettant d'assister dans la sélection de données. Il existe deux grandes familles de méthodes hiérarchiques :

- ascendante ou agglomérative (`bottom-up`)
- descendante ou soustractive (`top-down`)

Le principe agglomératif est initialisé avec un cluster qui est caractérisé par un singleton et ajoute de nouveaux enregistrements, sans en dégrader la qualité. La méthode soustractive est opposée à la précédente : l'algorithme démarre avec un cluster contenant l'ensemble des données et obtient une solution en le segmentant récursivement. Le processus continu jusqu'à avoir atteint le nombre de clusters désiré.

La seconde méthode, dite de partitionnement, repose sur une division du jeu de données en plusieurs sous-ensembles. Cette division est établie à partir d'heuristiques précises, réallouant itérativement les enregistrements en fonction du nombre de clusters recherché.

Une bonne vue d'ensemble de ces différentes méthodes est résumée dans [Berkin, 2002]. Les deux grandes classes d'algorithmes y sont présentées ainsi que leurs variantes et innovations.

Bien que ces deux méthodes soient issues de la modélisation non-supervisée (présentée dans la suite de ce mémoire au chapitre 6 page 82), à partir des résultats, il est possible pour l'expert d'étudier les regroupements proposés et de sélectionner parmi eux un ensemble représentatif. Cette intervention de l'utilisateur en fait une méthode supervisée. Celui-ci sélectionne les données par une étude des caractéristiques observées pour chaque regroupement. L'expert peut ainsi "nettoyer" les données afin d'améliorer la qualité des connaissances recherchées.

3.4 La sélection de données appliquée à la contrefaçon de marques nominatives

La première source documentaire qui nous a été fournie par l'expert contenait des décisions erronées (erreur dans les noms de marques). Nous n'avons pas remis en cause l'expert (fatigue, temps, outils informatiques inadaptés), mais nous avons cherché à utiliser les outils disponibles afin de produire une nouvelle source documentaire, que nous pourrions utiliser sans avoir à la remettre en cause. Nous allons donc voir dans un premier temps que ces méthodes ne peuvent pas être appliquées sur des sources de données documentaire hétérogènes, et qu'il a fallu mettre en place de nouveaux outils adaptés à de telles sources de données.

3.4.1 Des méthodes inadaptées

Dans notre cas d'étude, la sélection de données est un problème délicat. Parmi l'ensemble des données disponibles, seul un sous-ensemble précis nous intéresse. La base documentaire regroupant plusieurs types de jugements relatifs à la contrefaçon de marques, la sélection de données doit être abordée d'une manière différente.

Les méthodes présentées dans ce chapitre sous-entendent que :

1. la base documentaire est composée de données homogènes et ne caractérisant qu'un seul phénomène (pour les méthodes non supervisées)
2. ou bien que les documents, s'ils appartiennent à des domaines différents, soient caractérisés de façon explicite pour en permettre un regroupement aisé (pour les méthodes supervisées).

Nous aurions donc pu utiliser les méthodes non supervisées si l'ensemble des données avait décrit la contrefaçon de marques nominatives dont les marques sont toutes enregistrées. Il en est de même pour les méthodes supervisées si une information clairement identifiée avait permis de distinguer les jugements les uns des autres.

3.4.2 Les solutions apportées

Comme nous l'avons souligné précédemment, la finalité des documents est implicitement liée au fait que ces derniers sont supervisés et structurés. C'est au travers de cette structure et de l'information portée qu'est exprimée la contrefaçon de marques nominatives. Tous les documents partagent une structure commune pour décrire des décisions n'appartenant pas toutes au même domaine. Cela est possible car ces domaines sont proches sur la finalité (il s'agit toujours de contrefaçon, mais dans des contextes différents, comme par exemple la contrefaçon par reproduction).

Trier les documents électroniques de façon automatique pour en extraire le domaine d'étude n'est pas chose aisée. Il est très délicat de mettre en valeur ce qui distingue les décisions. Cette information est d'ordre sémantique et implique un raisonnement complexe pour l'extraire. Ainsi, pour pouvoir identifier un sous-ensemble utile pour mener nos expériences, il a fallu l'intervention de l'avocat conseil.

3.4.3 La sélection de données pour `JURINPI`

La base de données `JURINPI`, utilisée pour ces travaux de recherche, est caractérisée comme étant de type documentaire et hétérogène. Différentes jurisprudences y sont présentées, allant de la contrefaçon de marques nominatives et figuratives, à la contrefaçon de brevets ou encore de dessins et modèles. Cet ensemble de jugements représente un total dépassant les 18.000 enregistrements. La base est alimentée hebdomadairement par les décisions rendues par les tribunaux

français.

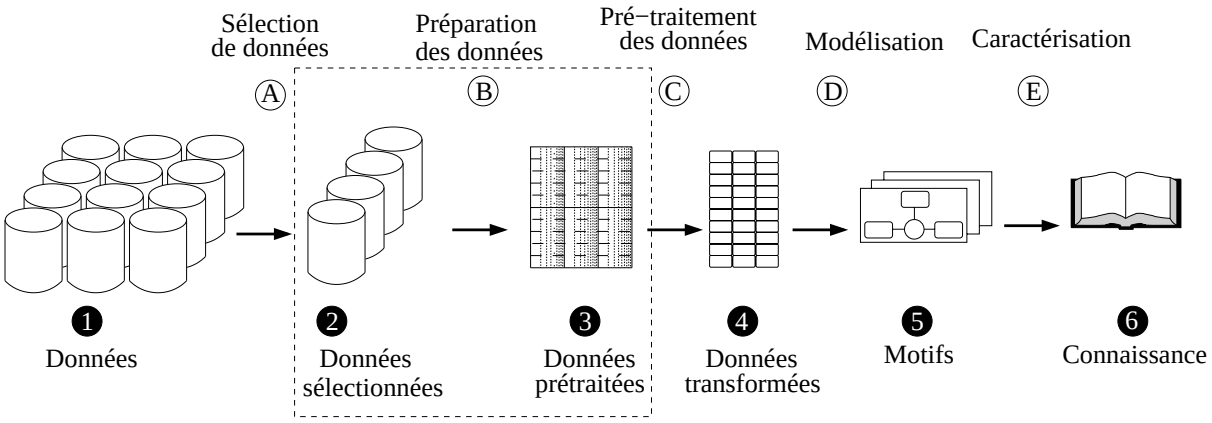
Dans le cadre de nos travaux de recherche, nous nous limitons à l'étude de la contrefaçon de marques nominatives. Cette étude représente alors un sous-ensemble de 8450 jugements (en 2001), couvrant la période 1904 à nos jours. À partir de 1997, les décisions sont en plus renseignées du texte intégral du jugement. Cependant, les jugements de 1904 ne sont absolument pas pertinents par rapport à nos problèmes : ils ne décrivent pas la société actuelle. De plus, en 1977, le droit des marques a été fondamentalement modifié. Nous nous limiterons donc à une étude des cas les plus récents. Pour ce faire, nous proposons l'année 1997 comme année charnière. Ce choix n'est pas anodin, c'est à cette année que les jugements présentent en plus des informations usuelles, le texte intégral du jugement. Le volume de données se restreint alors à un ensemble de 2576 décisions. Ainsi, sur cet ensemble de 2576 jugements, l'expert a isolé un ensemble de 800 décisions estimées pertinentes pour la description du phénomène.

A partir de ce sous-ensemble de 800 décisions, il devient possible d'appliquer les méthodes de sélection de données présentées. Toutefois, le volume de données est peu conséquent, les algorithmes de modélisation sont capables d'opérer sur un tel volume. Même avec l'intégralité de la source documentaire, l'échantillonnage n'aurait pas été nécessaire. Cependant, nous pensons que l'étape de sélection de données, telle qu'elle est abordée dans la littérature omet de prendre en compte la situation à laquelle nous avons été confronté. Cette étape implique d'avoir des documents caractérisant tous le même domaine. Évidement, il peut nous être reproché de vouloir résoudre la situation rencontrée dans `CATMInE` dans cette étape, au motif que la base n'était pas conforme au pré-requis implicite : les documents décrivent le même phénomène. Toutefois, nous pensons que nombre d'applications sont établies à partir de bases de données réelles et que si l'on se conforme à la démarche présentée dans le chapitre 2.1 page 28, la source documentaire initiale ne doit contenir que des données homogènes. Or ces ensembles documentaires réels ne sont pas nécessairement homogènes (et c'est le cas de `CATMInE`).

Nous avons donc exposé la nécessité d'appliquer une sélection supervisée par l'expert pour déterminer une fenêtre d'échantillonnage valide. Cette sélection se fait par le biais d'un système interactif permettant de confronter les résultats d'une sélection automatique à ceux issus de l'expertise d'un avocat conseil. Cette comparaison permet de corriger les erreurs issues du traitement automatique et de s'assurer de la pertinence des documents sélectionnés.

Chapitre 4

Préparation des données documentaires sélectionnées



La préparation de données : passage d'une base sélectionnée à une base pré-traitée.

Sommaire

À l'aide des stratégies présentées dans le chapitre précédent, le phénomène à caractériser est maintenant isolé par un sous-ensemble d'enregistrements issus de la masse de données disponible initialement. Ce sous-ensemble peut ne pas correspondre à l'intégralité des enregistrements possibles pour le domaine ciblé et peut n'en être qu'une vue partielle. Comme pour tout jeu de données, l'imperfection existe et la base peut ne pas être optimale. Pour pallier ce problème, ce chapitre aborde la seconde étape du processus de modélisation : la préparation des données. Cette opération va permettre de rendre la base la plus homogène possible. La finalité de cette démarche implique que tous les enregistrements soient caractérisés par des descripteurs partagés, aux modalités connues. Ainsi, entre l'étape précédente d'échantillonnage et celle-ci, nous avons réalisé une extraction de l'information jugée nécessaire pour la modélisation du domaine. Cette extraction implique, pour chaque document, d'avoir identifié l'information, de l'avoir conservé dans une base de données tabulaire : en ligne les enregistrements, en colonnes les descripteurs représentant l'information isolée.

4.1 Motivations

Les motivations de cette étape de pré-traitement(s) des données sont d'éliminer toute source d'erreur, de confusion, ou de manque d'information d'un jeu de données destiné à l'analyse. Cette étape va permettre de ne plus remettre en question les données dans le processus décisionnel en rendant la base rigoureuse (robuste, fiable, et complète). En effet, si ces données caractérisent bien le domaine d'étude - comprendre qu'elles sont d'une qualité irréprochable - les problèmes futurs ne pourront alors être liés au jeu lui-même.

Dans la sélection de données (chapitre précédent), un échantillon de la base est déterminé afin de caractériser un domaine. Chaque enregistrement est alors transformé en un ensemble de tuples d'attributs/valeurs. Ce glissement de représentation a été présenté dans la section 2.4

page 31 afin de pouvoir appliquer par la suite des algorithmes de modélisation. Bien évidement, la qualité du jeu de données qui va être préparé dépend étroitement des processus d'extraction d'informations et de leur capacité à retrouver ou non l'intégralité de l'information ciblée.

A partir d'une représentation des données sous la forme d'un ensemble de descripteurs fixés, et d'une extraction d'informations afin de compléter les descripteurs, la préparation des données (*data cleaning*) apporte des méthodes et algorithmes permettant de garantir la qualité du jeu de données.

4.2 La préparation de données documentaire

La préparation des données (*data cleaning*) est une étape importante dans le processus de découverte de connaissances. La finalité de cette phase de préparation est l'obtention d'une base de travail répondant à des exigences fonctionnelles précises. Ces contraintes sont de plusieurs natures telles que le traitement du bruit présent dans une base ou encore la gestion des données incomplètes. Les travaux sur ce champ d'investigation ont dégagé trois grandes pistes de recherche :

1. La gestion du bruit est généralement caractérisée par des enregistrements indésirables relativement aux attentes de la base. Ces enregistrements doivent alors être isolés et une décision concernant leur contenu doit être appliquée. En effet, l'étude d'un phénomène ciblé peut impliquer, par exemple, de filtrer toutes les données sur une modalité (2.4.1 page 32) de descripteurs. Cette opération détermine les enregistrements dont la modalité observée est supérieure à celle retenue (ou inférieure, tout dépend de la contrainte imposée par l'expert), et les considère comme bruit.

2. La seconde piste concerne la gestion des données présentant des informations incomplètes. Ce cas de figure peut être aussi éventuellement assimilé à du bruit (ou silence). Les données étant incomplètes, celles-ci ne satisfont pas les attentes de la base. Cependant, un enregistrement assimilé à du bruit (contenant des valeurs hors champs par exemple) ne peut être équivalent à un enregistrement dont les valeurs présentes correspondent aux attentes mais dont une ou deux des informations sont manquantes. Pour ce genre de données des traitements statistiques permettent d'atténuer les problèmes consécutifs à leur absence.

3. Enfin, le troisième cas de figure est relatif à la création de bases de données à partir de sources multiples. Il n'est pas rare, en effet, dans l'industrie, d'avoir des bases documentaires réparties sur plusieurs sites et archivant des documents électroniques partageant plus ou moins de descripteurs. Lors de la création de la base de travail, il faut alors être capable de prendre des décisions dans le cas de données communes ou non.

Nous allons donc aborder au cours de ce chapitre l'analyse du bruit, comment l'identifier et le traiter (4.2.1). Puis nous analyserons plus en détails le traitement des valeurs manquantes (4.2.2 page 58) pour terminer par la gestion des sources multiples de données (4.2.3 page 60).

4.2.1 La notion de bruit dans une base de données

La gestion du bruit dans une base de données documentaire nécessite tout d'abord de définir dans un premier temps ce que l'expert entend par bruit. Cette définition va ensuite conduire à un nouvel échantillonnage de la base documentaire issu de la sélection de données. Dans cette étape les données concernent toutes le domaine ciblé par l'étude. Cependant, pour certains enregistrements, l'information portée ne peut satisfaire le domaine. Nous allons donc voir dans un premier temps ce que nous entendons par bruit dans une base documentaire puis dans un second temps comment identifier ce bruit.

Définition du bruit

Les données issues du monde réel (à opposer aux données générées de façon automatique) sont parfois partiellement erronées voire incomplètes. Dans notre cadre applicatif, l'échantillon retenu pour `CATMInE` est constitué de 800 jugements (résultat de la sélection de documents présentée dans le chapitre 3 page 42). Dans cet échantillon certaines décisions ne sont pas de qualité suffisante pour permettre de modéliser le domaine. Ce phénomène peut être expliqué par une défaillance des systèmes d'alimentation de la source documentaire (humains ou automatiques) et impose un tri des données afin d'assurer la qualité de la découverte de connaissances. Le bruit peut alors être défini comme une décision normale au comportement étrange, ou comme un jugement n'appartenant pas aux probabilités des distributions. L'expert doit identifier le bruit, les éléments "hors normes", les valeurs manquantes et les erreurs de typage de modalité d'une information. Il lui est donc nécessaire de mettre en place une stratégie d'étude des documents électroniques, en vue d'identifier toute forme de bruit et d'adapter la réponse du système en fonction des problèmes rencontrés.

[Famili, 1995] avance que la présence de bruit au sein d'une base a deux conséquences pour la sélection d'un jeu de données test : la première implique de vérifier les enregistrements sélectionnés afin de s'assurer qu'ils disposent bien de l'ensemble des attributs retenus, et que ceux-ci correspondent bien aux attentes fixées ; la seconde conséquence est qu'un tel jeu de données ne peut être considéré comme pertinent pour des tests s'il est composé pour l'essentiel de données indésirables. Afin d'atténuer les effets d'une base de données bruitée, nous allons voir dans un premier temps les différentes possibilités d'identification du bruit pour ensuite aborder sa gestion en vue d'améliorer la qualité des données.

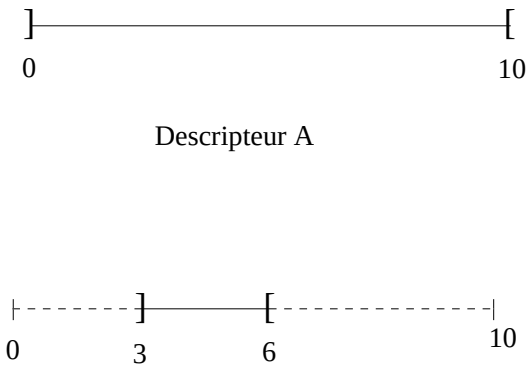


FIG. 4.1 – Redéfinition d'un descripteur

Identification du bruit

L'identification du bruit peut se faire en deux temps. Dans un premier temps, des filtres automatiques peuvent être mis en place afin de dégrossir le travail. L'expert va donc chercher à éliminer tout ce qui en est saillant. Les valeurs hors champs sont, par exemple, faciles à identifier ainsi que les modalités erronées. Le schéma 4.1 mets en application cette propriété en redéfinissant à titre d'exemple un descripteur continu, initialement borné sur l'intervalle $]0;10[$, sur l'intervalle $]3;6[$. Cette opération définit alors comme bruit tous les enregistrements pour lesquels le descripteur A prend une valeur dans la réunion des intervalles $]0;3] \cup [6;10[$. Les enregistrements prenant une valeur sur l'intervalle $]3;6[$ sont sélectionnés pour constituer le jeu de données final.

Si l'expert recherche des relations entre descripteurs en tant que contraintes particulières, alors l'élimination des descripteurs ne satisfaisant pas ces contraintes peut être réalisée en vérifiant la présence ou non de ces relations pour chaque exemple.

Dans un second temps, une étude statistique du jeu restant va permettre de s'assurer de l'homogénéité des données. Une représentation des données sous forme d'histogrammes (en abscisse les modalités du descripteurs, en ordonnée la population) suffit à avoir une vue d'ensemble des données par descripteurs. Nous pouvons alors identifier la répartition des modalités, et construire un modèle pouvant respecter ou non les proportions observées. Le schéma 4.2 page suivante présente un exemple pédagogique d'un descripteur définit sur l'ensemble $\{A;B;C;D;E;F;G;H;I\}$. Nous pouvons alors observer que les sous-ensembles $\{D;E\}$ et $\{F;G;H;I\}$ les données appartiennent à une classe unique (`Classe 1`) dans le premier sous-ensemble, et `Classe 2 pour le second`). Le dernier sous-ensemble, $\{A;B;C\}$, est constitué d'exemples provenant des deux classes d'enregistrements possibles.

Quelque soit le type de recherche effectuée sur les données, les méthodes semi-supervisées par-

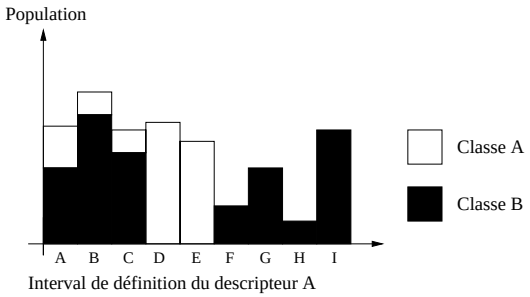


FIG. 4.2 – Étude de la répartition d'un descripteur A prenant ses valeurs dans le domaine de définition $\{A; B; C; D; F; G; H; I\}$.

courent chaque enregistrement du jeu de données, vérifient pour chacun les contraintes imposées par l'expert et le cas échéant écarteront les données ne respectant pas ces contraintes.

4.2.2 Les valeurs manquantes

Les valeurs manquantes caractérisent des enregistrements dont certaines valeurs ne sont pas renseignées. Ces données se trouvent alors inexploitables en tant que telles et ont été longtemps considérées comme du bruit, ce que rappellent [Schafer & Graham, 2002] dans leur étude. Désormais, un autre regard est porté sur ce type de données, et ces deux auteurs expliquent que même si elles sont incomplètes, les informations restantes participent néanmoins à la caractérisation du phénomène étudié. Il faut donc prendre une décision. Celle-ci dépend des valeurs manquantes rencontrées. Sont-elles nécessaires (indispensables) à la bonne mise en œuvre de l'étude ? Peut-on établir un modèle sans les inclure ? Si les valeurs manquantes concernent un descripteur indispensable au processus décisionnel, il faut alors trouver une alternative capable de compléter l'information présente.

Nous allons aborder dans un premier temps la décision à prendre si des valeurs sont manquantes, puis nous présenterons les procédés mis en place afin de pallier ce manque de données pour de tels enregistrements. Il faut aussi garder à l'esprit que, comparativement à une étude menée uniquement sur les données sans valeur manquante, une telle démarche a pour effet de rendre plus complexe le jeu de données et donc peut substantiellement dégrader la qualité de l'apprentissage. En augmentant le nombre d'enregistrements, on augmente alors les relations intrinsèques entre ceux-ci et donc la combinatoire entre nombre d'enregistrements et nombre de descripteurs. De plus, tel que présentée ici, nous considérons que la prédiction des valeurs manquantes est parfaite. Généralement, cela n'est pas respecté, et la qualité des données soumises à l'apprentissage est dégradée. Nous allons donc nous intéresser à la suppression des enregistrements contenant des valeurs manquantes.

Suppression des enregistrements contenant des valeurs manquantes

Usuellement, les enregistrements contenant des valeurs manquantes ont été ignorés dans les processus décisionnels, ce que rappelle Joseph Schafer et John Graham dans [Schafer & Graham, 2002]. Ces données étaient assimilées à du bruit et donc non prises en compte dans l'apprentissage. Cependant, ce traitement peut se révéler un peu barbare. Avant de supprimer un enregistrement contenant des valeurs manquantes, il faut d'abord s'assurer que ces valeurs ne sont pas indispensables pour l'expression du phénomène étudié. Si c'est le cas, la suppression ne devient plus une étape à appliquer avant la sélection des descripteurs comme il est souvent admis mais après. Il est en effet étrange de s'affranchir d'un descripteur sémantiquement fort[8] pour la compréhension d'un domaine d'étude, celui-ci contenant des valeurs manquantes. De plus, avoir des valeurs manquantes n'induit pas pour l'ensemble des enregistrements que le descripteur soit manquant (sinon, il est inutile). Dans le cas contraire (les valeurs ne sont pas indispensables), la suppression semble alors toute indiquée : le descripteur n'apporte rien (même s'il est partiellement présent) à la modélisation et compréhension de l'étude menée sur les données.

Cette méthode est trop radicale, dans certaines bases documentaire, l'information peut être redondante et donc déterminée par d'autres procédés d'extraction.

Dans d'autres cas de figure, l'information manquante peut aussi être déterminée car corrélée à une autre valeur. Nous présentons ces deux aspects dans la section suivante.

Estimation des enregistrements présentant des valeurs manquantes

L'estimation des valeurs manquantes n'est pas une idée nouvelle, et Ioannis Kopanas ([Kopanas et al., 2002]) le rappelle. Une des méthodes envisageables est de réaliser un apprentissage de la valeur manquante (sur les exemples pour lesquels elle est présente) en la considérant comme valeur de classe. Une fois le modèle établi, il peut être appliqué sur un jeu de données contenant cette valeur manquante afin de la déterminer. Cette méthode atteint ses limites en fonction du modèle établi pour prédire cette valeur manquante.

Une autre méthode est l'application de la théorie des motifs fréquents pour répondre au problème. Ce traitement, mis en valeur par François Rioult ([Rioult, 2002]), permet par des règles d'association de déterminer ces valeurs manquantes.

Il existe aussi des méthodes de pondération ([Little & Rubin, 1986]) permettant d'approximer les valeurs manquantes. Le principe consiste à pondérer les exemples en fonction de leur distribution. Cela permet de corriger les exemples de façon non paramétrique pour les valeurs manquantes monotones.

Il existe d'autres méthodes de gestion des valeurs manquantes, et nous renvoyons le lecteur

[8] Cette définition est purement arbitraire et dépend étroitement du savoir-faire de l'expert, de ses croyances et des intentions qu'il prête au modèle.

à l'étude menée par Joseph Schafer et John Graham ([Schafer & Graham, 2002]) pour une présentation des techniques usuelles possibles.

Enfin, le dernier cas de figure, autre que l'estimation, est de s'assurer que la valeur manquante n'est pas une information redondante. La base documentaire utilisée dans `CATMInE` en est l'exemple type. L'information sélectionnée est issue d'une partie du document résumant la décision qu'il contient. Si dans cette partie l'information n'est pas présente, il reste possible de l'isoler dans le texte complet de la décision. Comme nous l'avons présenté précédemment (chapitre 3 page 42) ce travail peut vite se révéler fastidieux. Cependant, un traitement supervisé par l'expert peut permettre de retrouver cette information sur le sous-ensemble de documents présentant une telle anomalie. Il est ainsi inutile de chercher à supprimer l'enregistrement ou à déterminer l'information par recherche de corrélations entre informations.

4.2.3 Les sources documentaire multiples

Enfin, un troisième objectif de la préparation de données est relatif à l'utilisation de données provenant de sources multiples. Il n'est pas rare en effet dans le domaine industriel que les données soient issues d'un ensemble de sites de production (et donc de bases). Ces différentes bases peuvent avoir des descripteurs et des transactions en commun, l'ensemble en commun ou encore aucun. Dès lors, cette étape de pré-traitement a pour but d'isoler et d'éliminer d'éventuels doublons du jeu de données soumis à analyse. Un autre cas de figure est spécifique aux bases de données d'un volume ne permettant pas sa prise en charge au sein de la mémoire de l'ordinateur qui traite la modélisation. Découper en parts égales ce volume de données permet alors de s'affranchir de cette difficulté. La solution la plus naturelle consiste à établir un classifieur par base de données. L'expert obtient alors un modèle par série de transactions dont il ne reste plus qu'à calculer la moyenne. Cependant, [Chan & Stolfo, 1996] démontrent que cette stratégie est moins efficace en terme de reconnaissance qu'un apprentissage intégrant la totalité des données. Le seul moyen de réduire cette erreur consiste à intégrer des doublons dans l'ensemble des jeux de données. Pour Paul Bradley, Usama Fayyad et Cory Reina ([Bradley et al., 1998]) la solution proposée pour lever la difficulté consiste à parcourir l'intégralité du jeu de données disponible afin de déterminer les régions à défausser, celles à compresser de celles à conserver telles quelles pour apprentissage. Chaque région compressée est représentée par une transaction résumant la zone qu'elle qualifie. Les zones défaussées ne seront pas utilisées pour la modélisation du domaine. Le modèle est obtenu à partir des régions compressées et des régions conservées, sans application de mécanismes particuliers.

Dans notre cadre applicatif, nous ne sommes pas confronté à ce problème. Toutefois, nous pensons qu'il est judicieux de le présenter. Ce document de thèse ayant pour motivation de présenter une vue globale des différentes méthodes.

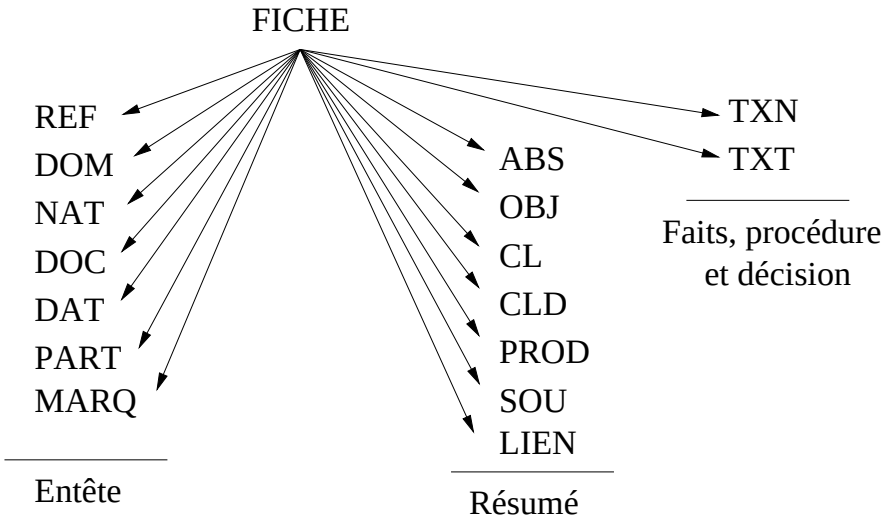


FIG. 4.3 – DTD de JURINPI.

4.3 La préparation des données juridique

Dans nos travaux de recherche, la préparation des données juridiques a impliqué de mettre en place des mécanismes d'extraction d'informations adaptés. Cette extraction d'informations a ensuite guidée l'application des méthodes présentées au tout au long de ce chapitre.

4.3.1 JURINPI : une base de données de jugements

Cette collection de documents électroniques juridiques est disponible (à titre payant) au format `XML`, décrite par une structure de 21 balises. Cet ensemble de balises permet de couvrir l'ensemble de l'information à caractériser dans une telle base. Cependant, la base `JURINPI` ne contient pas que des informations relatives à la contrefaçon de marques nominatives. D'autres informations telles que les brevets et dessins et modèles y sont décrites. Ainsi sur un ensemble de 21 balises, 16 seront réellement utiles pour la suite de ce travail. Les 5 autres étant relatives aux autres domaines présents dans `JURINPI`. La figure 4.3 présente les balises utilisées dans la représentation des documents de décisions relatives aux marques nominatives. Dans ce schéma, les balises ont été regroupées en fonction de leur rôle : appartenance à l'en-tête, au résumé, ou encore au texte complet de la décision. L'ensemble des balises est rattaché au même niveau dans l'arborescence de la `DTD`. Il n'existe donc pas de relation arborescente entre ces balises. La table 4.1 page suivante présente un extrait de jurisprudence disponible dans la base JURINPI (extrait, car le texte complet n'a pas été inséré afin de présenter clairement le document) et utilisant la structure `XML` présentée précédemment. Le texte intégral du jugement permet de confronter les informations extraites afin de s'assurer de la qualité de l'extraction d'informations. Ce travail de recoupement se justifie par le manque de rigueur dans l'utilisation de la structure

jurinpi

Référence M20010655

Domaine MARQUE

Nature de la décision DECISION FRANCAISE

Juridiction TRIBUNAL DE GRANDE INSTANCE DE PARIS (CH.03), 2001-05-18

Date de la décision 2001-05-18

Noms des parties EXACOD (SA) / LA POSTE

Marque EXACOD ;HEXACODES

Analyse DENOMINATION SOCIALE ET NOM COMMERCIAL (EXACOD) - MARQUE DE FABRIQUE ET DE SERVICES - MARQUE COMPLEXE - PARTIE VERBALE (EXACOD) - APPAREILS POUR LE TRAITEMENT DE L'INFORMATION, LA GESTION DES AFFAIRES COMMERCIALES, LES CONSEILS, L'INFORMATION ET LE RENSEIGNEMENT D'AFFAIRES, LA GESTION DE DOSSIER INFORMATIQUE, LA PROGRAMMATION POUR ORDINATEUR ET LA LOCATION DE TEMPS D'ACCES A UN SERVEUR DE BASES DE DONNEES - **CL09, CL16, CL35, CL36, CL41** - NUMERO D'ENREGISTREMENT 99 796 368 - MARQUE VERBALE (HEXACODES) - LOGICIELS DE GESTION DE FICHIERS D'ADRESSES DESTINES AUX ROUTEURS ET AU GRANDS EMETTEURS DE COURRIERS, LES FICHES, LA CONSTITUTION, LA CENTRALISATION, LA TENUE ET LA MISE A JOUR DE FICHIERS D'ADRESSES POUR VALIDATION DU CONTENU DES BASES DE DONNEES (VOIES, BOITES POSTALES, LOCALITES ET CEDEX) - **CL09, CL16, CL35** - NUMERO D'ENREGISTREMENT 3 009 533

ACTION EN CONTREFACON

MARQUE (EXACOD) - VALIDITE (OUI) - PORTEE - ETENDUE DE LA PROTECTION - ARTICLE L 712-2 CODE DE LA PROPRIETE INTELLECTUELLE - LIBELLE - PRECISION (OUI) - DESIGNATION DES CATEGORIES AUXQUELLES APPARTIENNENT LES PRODUITS - IDENTIFICATION DU CONTENU DES SERVICES (NON) - CARACTERE ABUSIF DE LA DESIGNATION DES SERVICES (NON) - FRAUDE (NON) - DOMAINE EXCEDANT L'ACTIVITE SOCIALE (NON)

MARQUE (HEXACODES) - VALIDITE (NON) - DISPONIBILITE (NON) - ARTICLE L 711-4 CODE DE LA PROPRIETE INTELLECTUELLE - DROIT ANTERIEUR (OUI) - MARQUE ANTERIEURE ENREGISTREE (EXACOD) - RISQUE DE CONFUSION (OUI) - SIMILITUDE VISUELLE ET PHONETIQUE - RISQUE DE CONFUSION SUR ORIGINE DES SERVICES

ATTEINTE A LA DENOMINATION SOCIALE ET AU NOM COMMERCIAL (NON) - ACTIVITES DIFFERENTES - RISQUE DE CONFUSION (NON)

CONTREFACON - PREJUDICE - EVALUATION - ELEMENT PRIS EN CONSIDERATION - DEVALORISATION - DIFFUSION DE BROCHURES AUPRES DE PROFESSIONNELS

Numéro(s) 99796368 ;3009533

Classification des produits et services CL09 ;CL16 ;CL35 ;CL36 ;CL41

Produits et services APPAREILS POUR LE TRAITEMENT DE L'INFORMATION, LA GESTION DES AFFAIRES COMMERCIALES, LES CONSEILS, L'INFORMATION ET LE RENSEIGNEMENT D'AFFAIRES, LA GESTION DE DOSSIER INFORMATIQUE, LA PROGRAMMATION POUR ORDINATEUR ET LA LOCATION DE TEMPS D'ACCES A UN SERVEUR DE BASES DE DONNEES - LOGICIELS DE GESTION DE FICHIERS D'ADRESSES DESTINES AUX ROUTEURS ET AU GRANDS EMETTEURS DE COURRIERS, LES FICHES, LA CONSTITUTION, LA CENTRALISATION, LA TENUE ET LA MISE A JOUR DE FICHIERS D'ADRESSES POUR VALIDATION DU CONTENU DES BASES DE DONNEES (VOIES, BOITES POSTALES, LOCALITES ET CEDEX) - CL09, CL16, CL35 - ...

TAB. 4.1 – Exemple de jurisprudence issue de `JURINPI`

XML proposée. L'étude du texte intégral, utile à la validation des informations retenues, nécessite une approche raisonnée du langage et de la structure des textes juridiques afin de rendre efficace la recherche d'informations. Ces documents sont des décisions sur support papier, résultant d'un processus d'acquisition et de structuration automatique, ce qui explique les erreurs observées à travers les documents.

Enfin, relativement à la cohérence des jugements, une étude de la base a mis en évidence que certains d'entre eux décrivent des abandons et des incidents de procédure conduisant à un "non jugement" du cas étudié. Ces jugements sont pertinents pour le juriste, mais pas pour le domaine d'étude. Nous retrouvons aussi dans la base (et malgré une interrogation rigoureuse de celle-ci) des jugements relatifs aux dessins et modèles, aux brevets mais aussi aux marques figuratives. Ce manque de pertinence est lié aux informations contenues dans la structure XML des documents. De plus, dans de nombreux cas, l'information est incomplète : omission d'une des parties, de la date de jugement, et autres ... Ainsi, sur les 800 décisions retenues lors de la sélection des documents 3.4.3 page 50, 695 jugements ont été rejetés car inexploitables de façon rigoureuse. Cette sélection de cas a été réalisée par un non expert du domaine, et ce dernier n'a retenu que les jugements où l'information recherchée était facilement (par observation et courte lecture) identifiable.

Enfin, pour l'aspect statistique de l'étude, nous pouvons retenir les informations présentées dans le tableau 4.2 page suivante, et décrivant la base de données JURINPI. Cette description s'attache à mettre en valeur la différence entre les jugements aboutissant à une contrefaçon et ceux caractérisant une non-contrefaçon, en fonction du niveau décisionnel.

Dans notre cas d'étude, un jugement peut être remis en cause si l'une des deux parties n'est pas satisfaite du jugement. Ainsi, lorsqu'un procès a lieu pour la première fois, le *Tribunal de Grande Instance* (TGI) observe les faits et la loi et rend un jugement. La partie lésée, mécontente ou insatisfaite par le verdict a alors la possibilité de porter le jugement devant la *Cour d'appel* (CA). Si cette dernière invalide le jugement rendu par le TGI, alors celui-ci n'est plus valable, et seul le jugement de la Cour d'Appel prévaut. De nouveau, si l'une des parties n'est toujours pas satisfaite par le jugement rendu en Cour d'Appel, il lui reste la possibilité de porter l'affaire en *Cour de Cassation* (CCASS) qui elle juge les points de droit et non les faits. Pour les mêmes raisons que la Cour d'Appel, si la Cour de Cassation infirme un point de droit relatif au jugement rendu en appel, l'affaire est alors renvoyé en Cour d'Appel.

La base ainsi construite est à peu près homogène. Il n'y a pas de jugement en Cour de Cassation, cette dernière ne statue que sur les points de droit. Or dans notre étude, nous n'observons que les aspects factuels.

Cours	Contrefaçons	Non-Contrefaçons
TGI	42 (40%)	38(35%)
CA	13 (12%)	12 (11%)
CCASS	0 (0%)	0 (0%)

TAB. 4.2 – Répartition des jugements par rapport au verdict et au niveau juridictionnel.

4.3.2 L'information ciblée

Comme nous l'avons expliqué précédemment, la base `JURINPI` est de type documentaire et hétérogène. De plus, cette base est au format `XML`. Ainsi, avant d'énoncer l'ensemble des traitements à effectuer, il est utile de réfléchir aux besoins, et donc aux informations nécessaires pour établir les descripteurs précédemment présentés (2.4.1 page 32).

Tout d'abord, notre étude porte sur la contrefaçon de marques nominatives. Il est donc important de récupérer certaines informations telles que les marques, les parties et le verdict. C'est ce que présente le tableau 4.3 page ci-contre, faisant le parallèle entre l'information à récupérer, la balise `XML` la contenant, et la forme prise par cette information dans cette balise[9]. D'autres éléments comme le niveau décisionnel, la date de jugement ou encore la référence du jugement permettent de recouper certaines informations mais aussi d'apporter une information supplémentaire pour l'étude des résultats.

Enfin, il est important de garder à l'esprit que l'ensemble documentaire caractérisé par `JURINPI` a été numérisé. Cela implique du bruit dans la base, les procédés ne sont pas efficaces à 100%. Ainsi, il n'est pas rare de trouver des données où par exemple le mot CONTREFAÇON est présent mais ne peut être trouvé car il en manque des lettres. Il en va de même pour tous les autres éléments recherchés. Ces problèmes peuvent être lié au procédé de numérisation du document et plus particulièrement au principe de reconnaissance de caractère.

La date du jugement

La date d'un jugement, marquée par la balise `<DAT>` n'est pas présente dans une large partie des enregistrements. La balise est tout simplement omise. Pour résoudre ce problème, il est alors nécessaire d'extraire la date dans une autre balise, celle du lieu (`<DOC>`).

Lorsque la date est bien présente dans la balise, l'information est décrite ainsi :

– $< DAT > 2001 - 05 - 18 < /DAT >$

Lorsque la date n'est pas présente dans la balise date, nous la retrouvons alors dans la balise `DOC` comme suit :

– *<DOC>TRIBUNAL DE GRANDE INSTANCE DE PARIS (CH.03) ; 2001-05-18</DOC>*

[9] Attention, ceci n'est pas une liste complète, mais juste des exemples de cas possibles

Information	balise	expressions
Référence	<REF>	*M19950787* *M* = Marque *1995* = année de jugement *0787* = numéro d'ordre
Date de la décision	<DAT>	*Année-Mois-Jour*
Les parties	<PART>	*Plaignant1* ; *Plaignant2* / *Défendant1* ;...
Les marques	<MARQ>	*Marque1* ; *Marque2*
Le verdict	<ABS>	*Contrefaçon (OUI)* *Contrefaçon OUI* *Contrefaçon NON* *Confirmation* *Dommages et intérêts*
Le tribunal	<DOC>	niveau *lieu (Chambre)*, date
Le niveau	<DOC>	*niveau* lieu (Chambre), date
Le jugement de référence s'il y a.	<LIEN>	*niveau lieu (Chambre), date* ou une référence : *M19950787*
Le numéro d'enregistrement des deux marques	<OBJ>	*14417607;1532107*

TAB. 4.3 – Présentation des types, balises, et expressions des informations nécessaires.

Les parties

En ce qui concerne les parties, plusieurs cas de figure s'offrent à nous :

- soit les deux parties sont présentes selon le formalisme précisé dans le tableau 4.3 : *<PART> EXACOD (SA)/LA POSTE </PART>*
- soit, une seule partie est mentionnée : *<PART>LA POSTE</PART>*
- soit aucune partie n'est citée : *< PART >< /PART >*

Si une seule partie est mentionnée, une recherche de la seconde est possible au sein de l'analyse du jugement (balise <ABS> : *[...] DENOMINATION SOCIALE ET NOM COMMERCIAL EXACOD - MARQUE DE FABRIQUE [...]*), ou encore dans les faits et procédure du jugement (balise <TXN> : *[...] La société EXACOD, est inscrite au registre [...]*). Dans ces deux cas de figure, les efforts d'extraction d'informations sont alors plus lourds et plus coûteux.

De plus, lorsqu'une seule partie est citée, il est important de vérifier si le niveau décisionnel du jugement n'est pas différent du Tribunal de Grande Instance. Si c'est le cas, la balise <LIEN> est alors présente dans le document. Cette balise contient une référence au jugement de cours in-

férieure[10]. Il est alors possible de compléter les informations manquantes en faisant une recherche des jugements précédents et en étudiant la balise adéquate (balise `<ABS>`).

Les marques

Les marques nominatives présentent exactement les mêmes problèmes que les parties. Ainsi, nous pouvons trouver qu'une seule marque (voir aucune), au quel cas une étude approfondie du texte de la décision, ou encore une recherche de jugement antérieure (si le jugement en cours d'étude est d'un niveau supérieur au TGI) permet d'y répondre.

Le cas échéant, les marques sont étiquetées par la balise `<MARQ>` comme suit : *<MARQ> EXACOD;HEXACODES </MARQ>*.

Les jugements de référence

L'intérêt de cette balise est surtout dans un objectif de consultation des jugements pour l'étude des résultats. Il se décompose en trois partie :

1. une lettre : `M`, `D`, `B`. Cette lettre décrit le domaine : `M` pour marque, `D` pour dessin et modèle et `B` pour brevet
2. l'année de jugement, sur quatre chiffres.
3. un numéro d'ordre du jugement dans l'année de jugement. Ce numéro est établit sur 4 chiffres aussi.

Cette recherche de jugement de référence peut toutefois poser des difficultés : le lien proposé doit normalement correspondre à une référence de jugement (contenu dans une balise `<REF>` : *<REF>M20010655</REF>*). Or ce n'est pas obligatoirement le cas, et nous pouvons y retrouver le contenu de la balise `<DOC>` du jugement de niveau inférieur (*<DOC>TRIBUNAL DE GRANDE INSTANCE DE PARIS (CH.03); 2001-05-18</DOC>*). Si tel est le cas, il n'est pas possible de reconstruire l'équivalent du contenu de la balise `<REF>`. En effet, il n'est pas possible de déduire le numéro d'ordre du jugement dans l'année (d'après le principe même de construction de la référence). Cependant, il est possible de rechercher l'ensemble des jugements rendus le même jour que la référence dont nous disposons (pour notre exemple, le 18 mai 2001), et de trouver des points de comparaison afin d'isoler le bon jugement pour en extraire la référence définitive.

Isoler le verdict

La recherche du verdict est à elle seule très délicate. Comme l'a présenté le tableau précédent[11], il n'y a pas de balise dédiée, au même titre que la date par exemple. Le verdict est donc

[10] Tribunal de Grande Instance pour un verdit de Cour d'Appel, et Cour d'Appel pour une Cour de Cassation

[11] cf. 4.3 page précédente

exprimé dans l'analyse du jugement (balise `<ABS>`), mais aussi dans le texte de la décision (balise `<TXT>`). La partie d'analyse présente la succession des événements, décisions, et constats pris par la cours. Ainsi, il faut rechercher dans cet amas d'informations, celle qui met en valeur le verdict. Quoi qu'il en soit, la seule possibilité est de rechercher une séquence textuelle susceptible de caractériser le verdict. Cette séquence peut prendre une forme différente de celle espérée (pour mémoire, CONTREFAÇON), tel que : "Confirmation" ou encore "Dommages et intérêts". La première écriture fait référence à un jugement antérieur (`TGI` ou `CA`, en fonction du niveau actuel). La seconde sous-entend, quant à elle, le déboutement d'une des deux parties, exprimé par une sanction pécuniaire : nous avons donc à faire à une contrefaçon où des dommages et intérêts peuvent être attribués au défendant que le plaignant a accusé à tort de contrefaçon. Pour remarque, à aucun moment nous ne cherchons à déterminer quelle partie a gagné. Nous cherchons seulement l'issue du jugement : contrefaçon on non.

Le numéro d'enregistrement des marques

L'utilisation du numéro d'enregistrement de la marque est un élément supplémentaire et utile afin de déterminer l'antériorité d'une marque par rapport à une autre (principe du premier arrivé, premier servi). Cependant, cette information encadrée par la balise `<OBJ>` (*<OBJ>99796368 ; 3009533</OBJ>*) n'apparaît pas dans la majorité des enregistrements. De plus, lorsque cette balise est renseignée, la présence des deux numéros (un par marque) n'est pas non plus un aspect rigoureux de la base. Il est courant de ne trouver que celui de la marque contrefaite, laissant alors supposer que l'autre marque n'est tout simplement pas déposée. Dans les cas où la balise `<OBJ>` n'est pas présente, l'information peut être retrouvée pour un certain nombre de jugements dans l'analyse de l'affaire (balise `<ABS>` : *(...)<ABS>NUMERO D'ENREGISTREMENT 99 796 368(...)</ABS>*) ou encore dans le texte des faits et procédures (balise `<TXN>` : *la marque EXACOD déposée le 3 juin 1999 et enregistrée sous le n° 99.796.368*). Mais cela n'est pas le cas pour la totalité des jugements.

De plus, relativement aux numéros de produits, un autre problème se pose : les deux premiers chiffres d'un numéro d'enregistrement sont censés représenter l'année de dépôt. Si l'on étudie pour exemple le cas des marques "p@riscope, une semaine de Paris" (numéro 95576855) déposée le 21 juin 1995, et "pariscope" (numéro 1554231) déposée le 6 octobre 1989, nous nous apercevons que pour la première marque la règle dit vrai, en revanche, pour la seconde cela n'est pas le cas. La mécanique de création des références a changé, et est moins évidente à interpréter pour identifier un jugement par sa référence. Ce numéro permet ensuite de faire la jonction entre la base de données relationnelle et les documents électroniques juridiques pour l'accès à l'information dans `CATMInE`.

4.4 Conclusions

A travers ce chapitre nous avons exposé les différents aspects de la préparation des données afin de disposer d'une base complète (sans valeur manquante) et cohérente. Plus cette démarche est rigoureuse, meilleur sera le processus décisionnel. En effet, si l'étude du jeu de données est réalisée à l'aide des méthodes présentées, alors l'expert aura une meilleure appréhension des enregistrements qu'il manipulera. La connaissance des descripteurs et de leurs modalités est alors plus qu'importante.

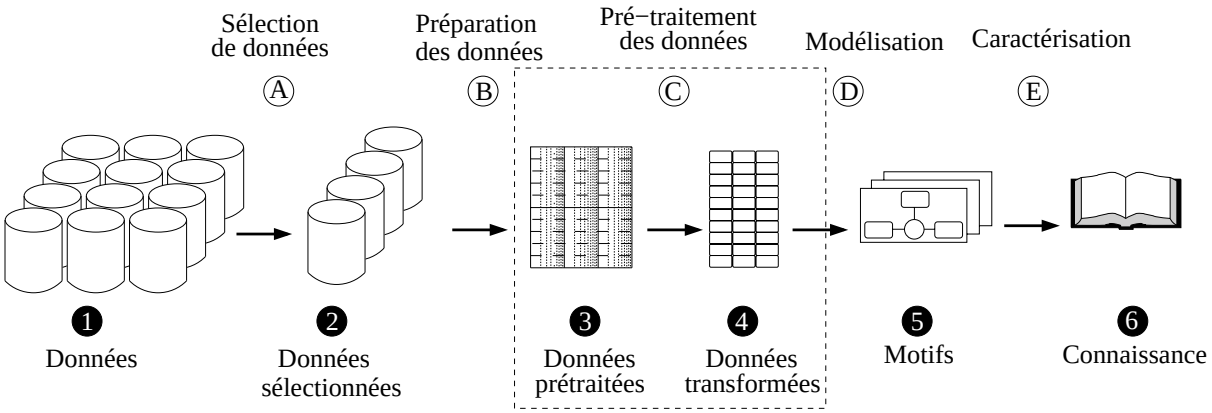
Toutefois, il ne faut pas non plus chercher à simplifier le jeu de données. C'est pour cela que le rôle de l'expert est capital. Il est le seul à être en mesure de comprendre des valeurs pouvant sembler aberrantes et à décider de leur utilisation ou non. Certes, dans certains cas un processus automatique peut répondre aux besoins, comme repérer les descripteurs invariants ou encore les doublons provenant de la réunion de plusieurs bases de données. Mais un tel traitement ne doit pas être complètement automatisé. Pour garantir le succès de cette étape l'expert doit avoir un droit de regard et une prise de contrôle à tout instant, impliquant alors un processus semi-supervisé afin que chaque choix soit validé.

Pour `CATMInE`, nous avons appliqué le principe que toutes données dont l'information recherchée n'était pas présente (malgré les différents mécanismes d'extraction d'informations) ne seront pas conservées pour la modélisation du domaine. Ainsi, dès qu'il y a valeur manquante, ou doute sur la qualité de l'information, le document électronique est écarté afin de garantir une base de données fiable pour les traitements futurs. Contrairement à notre définition du bruit, qui s'inscrit dans un cadre générale, ici, si nous observons une décision normale au comportement étrange, il est nécessaire de la conserver, celle-ci caractérise une information importante : l'expert a interprété différemment un phénomène.

Une fois la base mise en place, un travail d'adaptation des données (chapitre 5 page 70) est nécessaire pour exprimer au mieux le phénomène étudié. De là, la mise en place d'un algorithme de modélisation (chapitre 6 page 82) permettra de formaliser la connaissance issue des données, et une étude de celle-ci par l'expert (chapitre 7 page 102) validera ou non le modèle établi.

Chapitre 5

Transformation des données



La transformation de données : passage d'une base pré-traitée à une base transformée.

Sommaire

5.1 Motivations

Les enregistrements issus d'une base de données sont rarement exploitables en l'état. Nous avons présenté dans les chapitres précédents les intérêts de commencer par sélectionner et préparer un jeu de données répondant à un ensemble de contraintes déterminées par l'expert. Dans ce jeu de données, l'expert va pouvoir renommer ses descripteurs afin que les informations exprimées par la base soient porteuses de sens pour lui. Il s'agit ici, d'apporter une valeur sémantique aux objets (descripteurs et modalités) manipulés.

L'expert peut aussi décider de regrouper des modalités si son savoir-faire le suggère. La segmentation des valeurs continues est une autre tâche dépendant là aussi du savoir-faire de l'expert, si ce dernier en connaît les valeurs charnières, porteuses de sens. Ces deux transformations (regroupement de modalités et segmentation de valeurs continues) font glisser la représentation du domaine, des données brutes vers les objectifs fixés par l'expert. Le savoir-faire de l'expert est censé garantir la qualité de ces opérations, bien que nous ne puissions évaluer finement le degré de connaissance de cet expert relativement aux données et à leurs relations intrinsèques. Enfin, une fois les données transformées, l'ensemble des descripteurs disponibles n'est pas nécessaire pour répondre au problème. Un sous-ensemble peut suffire pour qualifier le problème. Cela évite ainsi de compliquer la recherche d'associations entre descripteurs (susceptibles de ne pas en avoir) et réduit l'expression de la base de données en un minimum suffisant. En procédant ainsi, nous agissons directement sur la combinatoire entre le nombre d'exemples et le nombre d'attributs.

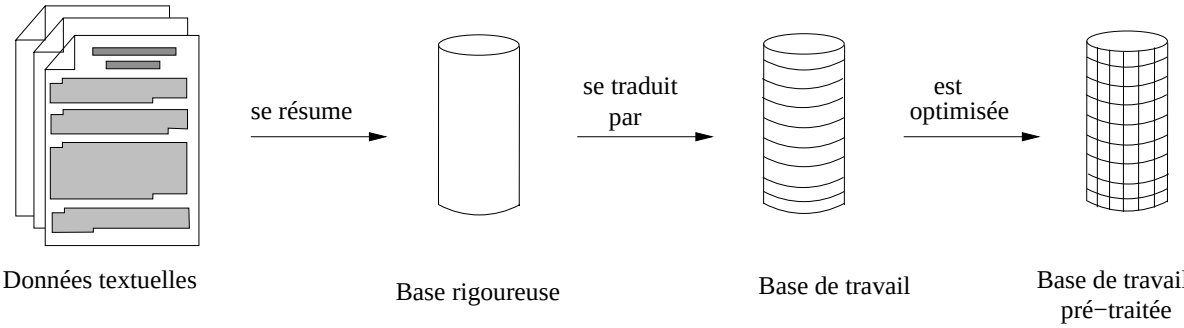


FIG. 5.1 – Du document électronique à une base de travail.

Cette fois et contrairement aux chapitres précédents nous agissons sur la composante nombre d'attributs. Le nombre d'exemples ayant été réduit par la sélection et la préparation des données.

La finalité de la transformation des données est une meilleure compréhension des résultats (connaissance) par l'expert dans la mise en place d'un processus décisionnel. La figure 5.1 est adaptée de la description présentée dans la section 2.3 page 30 et résume l'étape de transformation des données. Le passage d'un ensemble de données à une base rigoureuse a été présenté dans le chapitre précédent. Le présent chapitre explique alors la méthodologie pour exploiter cette base. Nous allons aborder dans la section 5.2 les différentes transformations opérables sur les descripteurs initiaux afin d'obtenir une base de travail. Puis dans la suite de ce chapitre (section 5.3 page 74), nous présenterons les différents principes de sélection de descripteurs conduisant à la base de travail pré-traitée et considérée comme opérationnelle pour réaliser une modélisation du domaine.

5.2 Transformation des descripteurs

La transformation des descripteurs est une étape importante dans le sens où elle apporte de la sémantique à quelque chose qui ne semble pas en avoir car mal définie. Cette étape permet aussi une meilleure compréhension des résultats exprimés par le processus décisionnel. Ces adaptations vont influer sur les ressources nécessaires pour la modélisation en réduisant l'espace de recherche décrivant le domaine. Les axes de transformation sont au nombre de trois : repérer les invariants, éliminer ou regrouper les modalités, et segmenter les modalités continues.

5.2.1 Gestion des invariants

Lorsqu'un descripteur présente des caractéristiques communes au travers d'exemples associés à la même valeur de classe, nous pouvons alors supposer que la modélisation du phénomène étudié semble être cohérente. Il existe des relations entre les enregistrements, cependant, si un

descripteur reste invariant quelque soit la finalité de l'enregistrement, le problème se pose alors d'établir l'impact du descripteur sur la base. Celui-ci étant identique à tous les enregistrements, s'en passer n'aura pas d'incidence dans les relations auxquelles il participe. Ces relations seront caractérisées par les autres descripteurs qui les composent.

Cependant, avant de supprimer ce descripteur des éléments caractérisant les enregistrements, l'expert doit s'assurer que ce descripteur est bien invariant et ce quelque soit le degré de précision (pour une valeur décimale par exemple) utilisé pour représenter le descripteur. Si un changement d'échelle du descripteur fait apparaître des différences notables pour les modalités observées ce descripteur ne peut plus être considéré comme invariant et doit être traité en conséquence : comme un descripteur portant une information caractérisant le phénomène.

5.2.2 Réduire les modalités

Dans les systèmes d'apprentissage, les limites de calcul sont essentiellement posées par les jeux de données. Les algorithmes cherchent à déterminer des relations intrinsèques aux données par combinaisons des descripteurs. Ainsi, si la base de données possède plusieurs milliers d'enregistrements, avec pour chacun plusieurs centaines de descripteurs, composés d'un nombre conséquent de modalités, la recherche des relations intrinsèques aura un coût en ressources informatiques lié à cette taille. Il se peut que la base de données ne soit pas adaptée pour bénéficier de la totalité de l'information disponible. Cependant, l'expert du domaine d'étude est capable de cibler les modalités utiles en fonction des besoins. Si, pour un phénomène étudié, toutes les modalités d'un descripteur inférieures à une borne sont inutiles, il est judicieux de les supprimer en vue d'améliorer, en temps cette fois, l'algorithme de découverte de connaissances.

5.2.3 Segmentation des valeurs continues

Les valeurs continues représentent une contrainte encore plus gênante qu'un nombre conséquent de modalités. En effet, par valeurs continues, il faut comprendre une valeur définie dans un intervalle et offrant une infinité de modalités possibles (comme un nombre réel par exemple).

Dans `CATMInE`, ce problème est plus que présent, les huit descripteurs retenus sont continus[12]. Les algorithmes d'apprentissage doivent alors déterminer des valeurs charnières au sein de l'intervalle de définition pour découper les descripteurs. Ce découpage se fait généralement en cherchant à maximiser un critère de partitionnement. L'algorithme commence par déterminer le critère de gain d'incertitude défini comme un dérivé de l'entropie de Shannon, puis vérifie qu'il y a bien gain d'information en utilisant cette valeur charnière.

La méthode de segmentation la plus évidente consiste a définir des intervalles de taille égale.

[12] CATMInE utilise un réseau de neurones pour la modélisation du domaine, les valeurs continues sont donc nécessaires

Le nombre d'intervalles peut être spécifié par l'expert. Une variante de cette méthode consiste à segmenter le descripteur en intervalles de fréquences égales : soit n enregistrements pour k intervalles, alors les intervalles auront une fréquence de n/k valeurs adjacentes. De cette variante, en découle une nouvelle consistant à réduire l'entropie de chaque intervalle proposé en ajustant les frontières.

Un autre procédé, proposé au travers de l'algorithme 1R ([Holte, 1993]), tente d'établir des intervalles purs, chacun contenant une forte majorité d'exemples appartenant à une classe définie avec la contrainte que chaque intervalle doit inclure un minimum d'enregistrements.

Une méthode plus complexe, le ChiMerge, proposé par Kerber ([Kerber, 1992]) détermine les intervalles en observant les valeurs, puis vérifie si des intervalles adjacents ne doivent pas être fusionnés en utilisant la mesure de χ^2. Cette méthode vérifie l'hypothèse selon laquelle deux intervalles adjacents sont indépendants en faisant une mesure empirique de la fréquence attendue des classes représentant chaque intervalle.

Le principe d'entropie a aussi largement été utilisé pour répondre aux contraintes de la segmentation de descripteurs continus. [Fayyad & Irani, 1993] proposent une heuristique récursive pour minimiser l'entropie. Cette méthode est associée au critère de description de longueur minimum ([Rissanen, 1986]) afin de contrôler le nombre d'intervalles produits. Cette méthode est utilisée dans les processus de modélisation d'arbres d'induction, et est appliquée localement lors de la production de chaque nœud. [Ting, 1994] en propose une version globale.

Une des conclusions de [Zighed et al., 1999] est l'équivalence constatée entre les différentes méthodes possibles de segmentation de valeur continue. Ainsi, il n'existe pas de méthode meilleure qu'une autre pour traiter une telle problématique de façon automatique.

L'expert peut aussi intervenir en proposant des valeurs charnières, en fonction de son savoir-faire et de l'observation des descripteurs pour le jeu complet de données. Les valeurs de segmentations ne seront peut-être pas optimales comme elles pourraient l'être avec une méthode automatique. En revanche, les valeurs choisies seront fortement porteuses de sens : l'expert peut justifier ses choix même si celui-ci peut avoir de faux *a priori*. Cette intervention a été testée dans `CATMInE`. Les descripteurs ont été observés de manière à voir la répartition des valeurs en fonction des classes. Une segmentation naïve a été menée en segmentant en trois ou quatre intervalles les descripteurs. Un gain de prédiction a été observé, mais la sémantique des descripteurs a quand à elle été perdue (segmentation naïve, donc sans connaissance explicite).

L'exemple du descripteur relatif au nombre de caractères en commun entre la marque du plaignant et la marque du défendant est assez révélateur. Dans l'ensemble des décisions, il est possible par exemple d'observer celles qui ne sont contrefaisantes que pour un ensemble de 20% de caractères en communs. Ces décisions ayant probablement des caractéristiques particulières à observer (pourquoi y a-t-il contrefaçon avec si peu de caractères en commun ?). Or, si cette propriété n'est pas explicitement exprimée dans la base de données, il est difficile d'interpréter

ce comportement. Mettre en valeur ce dernier implique une opération de segmentation sur le descripteur, pour faire apparaître clairement les seuils significatifs. Cette information étant une mesure continue, il faut connaître les valeurs charnières pour lui accorder du sens.

5.2.4 Fusion de descripteurs

Une autre opération possible pour la réécriture des descripteurs est la fusion de ceux-ci. Cette démarche a pour objectif d'exprimer de nouvelles connaissances, par le biais de nouveaux descripteurs établis à partir d'existants. Cette opération n'a pas de trace dans la littérature en tant que procédé automatisable. Il incombe à l'expert de décider de fusionner ou non des descripteurs en fonction de son savoir-faire. L'objet d'une telle méthode permet selon J. Deer et P. Eklund ([Deer & Eklund, 1996]) d'éviter de devoir déterminer le meilleur descripteur parmi l'ensemble disponible. Pour cela, il faut pouvoir être en mesure de réaliser une disjonction de descripteurs équivalents avec un opérateur de type *ET* logique évolué, et une conjonction de descripteurs non équivalents avec un opérateur de type *OU* logique adapté. Ces types d'opérateurs sont détaillés dans [Cross & Sudcamp, 1991].

En procédant ainsi, l'expert réduit le nombre de descripteurs utilisés pour décrire le domaine et simplifie dans le même temps le processus de modélisation en réduisant l'espace de recherche et en introduisant un descripteur plus discriminant. Les nouveaux descripteurs issus de la fusion de deux ou plusieurs autres descripteurs héritent d'une valeur sémantique pouvant être considérée comme la composée des valeurs sources.

En revanche, fusionner des descripteurs ne veut pas dire mettre bout à bout l'intégralité des modalités disponibles. Bien au contraire, les modalités possibles deviennent la combinaison des modalités des descripteurs les uns par rapport aux autres.

Dans `CATMInE`, nous avons appliqué cette fusion de descripteurs afin de produire les descripteurs relatifs à la longueur des sous-chaînes graphémiques et phonémiques. Ces descripteurs combinent à la fois le rapport que représente la sous-chaîne par rapport à la longueur de la marque et la position de celle-ci dans cette marque. L'hypothèse formulée étant de renforcer les différences entre deux sous-chaînes de rapports identiques mais n'ayant pas la même importance. En effet, l'expert a défini qu'une sous-chaîne contrefaisant en début de marque à plus d'impact (et donc d'importance) dans l'esprit du consommateur que si cette même sous-chaîne contrefaisante était située en milieu ou en fin de marque.

5.3 Sélection des descripteurs

La motivation principale de la sélection de descripteurs dans l'apprentissage est le gain apporté sur un aspect conceptuel : décider quels attributs utiliser pour décrire le concept ou encore quelle combinaison d'attributs semble nécessaire pour obtenir la meilleure induction. De plus,

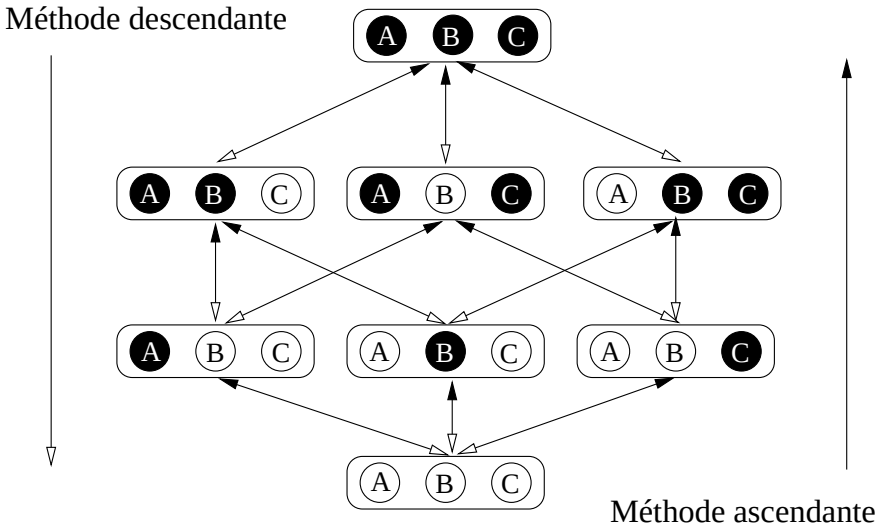


FIG. 5.2 – Principe ascendant ou descendant de sélection de descripteurs

cette opération a pour effet de réduire une fois de plus la combinatoire entre nombre d'enregistrements et nombre de descripteurs. La sélection de descripteurs a pour effet de chercher parmi l'ensemble des descripteurs possibles, un sous-ensemble représentatif avant tout processus de modélisation. Cela a évidement pour objectif de réduire les temps de calcul des différents algorithmes en réduisant l'espace de recherche, mais aussi d'assurer une qualité optimale des résultats. Le concept final induit est ainsi plus simple et plus compréhensible pour l'expert. Pour obtenir une méthode efficace de sélection de descripteurs, celle-ci doit suivre quatre règles incontournables ([Blum & Langley, 1997]) :

1. avoir un point de départ
2. une organisation de l'espace de recherche
3. une stratégie d'évaluation du sous-ensemble sélectionné
4. un critère d'arrêt

Pour garantir la sélection d'un sous-ensemble de descripteurs optimaux, les algorithmes utilisés doivent appliquer des méthodes prenant en compte l'interdépendance entre descripteurs. [Blum & Langley, 1997] présentent deux méthodes : une ascendante (où l'on ajoute les descripteurs) et une descendante (où l'on élimine les descripteurs). Ces deux méthodes répondent aux points 1 et 2 présentés précédemment. La figure 5.2 présente ce mécanisme de sélection de descripteurs. L'espace de recherche doit être vu comme un graphe où à chaque nœud correspond un sous-ensemble de descripteurs. L'état initial correspond à l'ensemble vide dans le cas des méthodes ascendantes (en bas de la figure 5.2), ou respectivement de l'ensemble complet des descripteurs pour les méthodes descendantes (en haut de la figure 5.2). Que cela soit pour l'ajout de descrip-

teurs (méthode ascendante) ou l'élimination (méthode descendante), l'action se fait en cherchant à maximiser le critère d'arrêt (qui peut être le score d'un apprentissage par exemple). Les descripteurs retenus par ces procédés le sont en fonction de leur pertinence. Se pose déjà un premier problème : pertinence par rapport à quoi ? Pour répondre à cette question, Avrim Blum et Pat Langley proposent les éléments de réponse suivants dans leur étude précédemment citée :

- pertinence par rapport au concept : le descripteur est fortement discriminant pour le concept étudié. Un descripteur est pertinent pour un concept s'il existe des exemples dans l'espace d'enregistrements pour lesquels modifier la valeur de ce descripteur affecte la classification fixée par la valeur de classe.
- faible pertinence : un descripteur d est faiblement pertinent si la suppression d'un sous-ensemble de descripteurs rend le descripteur d pertinent.
- pertinence comme mesure : pour déterminer le sous-ensemble de descripteurs le plus adéquat (dont l'erreur du modèle induit est la plus faible)
- utilité incrémentale : l'ajout du descripteur au sous-ensemble optimal offre une meilleure précision au modèle

Avec une telle définition, les méthodes de sélection doivent être étroitement liées à l'algorithme d'induction. De ce principe de parcours de l'espace de recherche, quatre grandes approches possibles de sélection de descripteurs ont émergé pour répondre à cette optimisation de pertinence du modèle. Chacune d'elle se distingue par l'application de l'algorithme d'induction et correspond au troisième point énoncé. Ces principes de sélections sont les suivants :

- Dans la première méthode, appelée enveloppante (*wrapper*), l'algorithme d'induction est intégré à la fonction d'évaluation.
- Pour la seconde méthode, dite de filtrage (*filter*), l'algorithme d'induction intervient après la sélection de descripteurs et filtre les sous-ensembles retenus.
- La troisième méthode, dite embarquée (*embedded*), inclue dans l'algorithme d'induction la sélection de descripteurs.
- La quatrième et dernière méthode dite de pondération des descripteurs (*feature weighting*) : la sélection des descripteurs est substituée par une procédure de pondération capable d'établir la pertinence des descripteurs.

Une bonne synthèse de ces différentes méthodes est présentée par Luigi Portinale et Lorenza Saitta dans [Portinale & Saitta, 2002].

5.3.1 La méthode *Wrapper*

La motivation première de la méthode *wrapper*, pour la recherche d'un sous-ensemble optimal de descripteurs, est liée à une maximisation des performances de reconnaissance du modèle induit. Les descripteurs retenus ne doivent pas uniquement dépendre des données mais aussi de

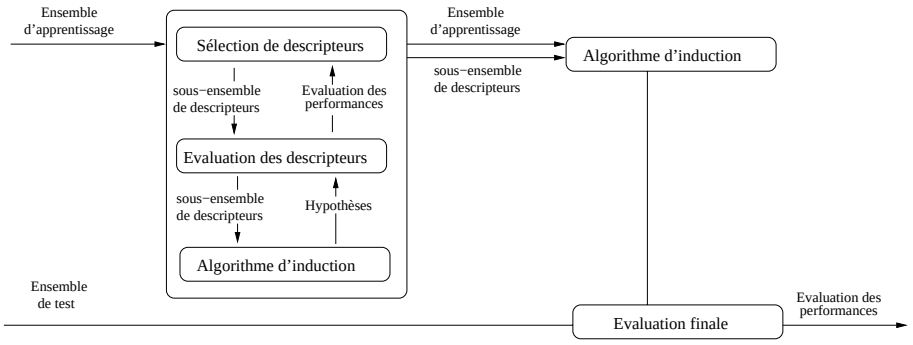


FIG. 5.3 – La méthode *wrapper* pour la sélection de descripteurs ([Kohavi & John, 1998]).

l'algorithme d'induction. En effet, [Kohavi & John, 1997] définit un sous-ensemble de descripteurs comme optimal, celui qui maximise la qualité du modèle induit sur l'ensemble des données. Le schéma 5.3 présente le principe de cette méthode.

La méthode *wrapper* garantit le respect de ces contraintes en intégrant l'algorithme d'induction de connaissances au sein du processus de sélection de descripteurs. Cependant, cette méthode nécessite d'importantes ressources de calcul : il faut pour être en mesure d'évaluer l'algorithme d'induction, le faire sur chaque sous-ensemble possible de descripteurs.

Cependant, Le principal désavantage de la méthode *wrapper* est qu'elle atteint ses limites avec l'augmentation du nombre de descripteurs et d'enregistrements ([Blum & Langley, 1997]). Cet inconvénient majeur en terme de complexité algorithmique (impliquant une modélisation pour chaque sous-ensemble de descripteurs testé) a conduit certains chercheurs à trouver des astuces de modélisation comme [Caruana & Freitag, 1994] qui proposent une stratégie pour garder en mémoire les arbres induits, permettant à l'algorithme de sélection de descripteurs de rechercher dans un espace de descripteurs plus large. Grâce à cette stratégie, les auteurs apportent une solution pour économiser la mémoire des ordinateurs afin d'effectuer plus de calculs.

5.3.2 La méthode *Filter*

Le concept de la méthode *filter* a été introduit pour pallier les limites, en terme de coûts de calcul, imposées par la méthode *wrapper*. Le principe est de tester les sous-ensembles de descripteurs en évaluant la qualité de ces ensembles, avant de les soumettre à l'algorithme d'induction. Pour cela le protocole consiste à prendre chaque descripteur indépendamment et de mesurer sa corrélation avec la valeur de classe. Ensuite, utiliser les k meilleurs descripteurs pour la modélisation, sachant qu'il faut pouvoir être en mesure de déterminer k.

La qualité est liée aux relations intrinsèques entre les descripteurs décrivant le domaine, relations qui doivent être les plus discriminantes possibles. Contrairement à la méthode *wrapper*, le filtrage des descripteurs avant toute opération d'induction de connaissances permet de traiter

des bases de données plus volumineuses.

En revanche, [John et al., 1994] démontrent que pour des bases où la méthode *wrapper* peut être appliquée, celle-ci se révélera bien supérieure à la méthode *filter*. Cependant, pour déterminer le meilleur sous-ensemble possible, la méthode de sélection de descripteurs *filter* doit tenir compte du parti pris de l'algorithme pour la construction du modèle afin de garantir une optimalité de celui-ci sur un jeu de données inconnu. Pour la méthode *wrapper* il est inutile de connaître cette information, le sous-ensemble est sélectionné par ajout ou suppression de descripteurs dans l'ensemble testé.

5.3.3 La méthode *Embedded*

Contrairement aux deux méthodes précédemment présentées, la méthode *embedded* (embarquée) correspond à la famille d'algorithmes d'induction réalisant dans un même temps la sélection et la construction du modèle. L'opération ne devient plus une étape à part entière où des contraintes précises sont appliquées, avant même de penser à modéliser le domaine, mais où la sélection et la modélisation ne font qu'un.

Chaque descripteur est évalué au fur et à mesure de la construction du modèle afin de partitionner au mieux les données en fonction de la valeur de classe. Ce partitionnement repose généralement sur des mesures d'entropies.

On retrouve dans cette catégorie les modélisations à partir d'arbres d'induction, et plus généralement tout algorithme établissant des conjonctions ou des disjonctions de descripteurs, sous forme de règles ou de branches, comme résultats.

5.3.4 La méthode de pondération

Cette dernière alternative, présentée par [Wettschereck & Aha, 1995], ordonne les descripteurs en fonction de leur pertinence. Ici l'intérêt porte sur une sélection explicite des descripteurs ce qui convient naturellement pour des résultats compréhensibles pour l'homme.

Il y a deux façon d'aborder cette méthode : une supervisée par l'expert et une seconde non supervisée. Quelque soit la méthode, le concept est de définir un poids (un nombre réel) à chaque descripteur : plus le poids est élevé, plus le descripteur est jugé pertinent. Dans le cas des méthodes supervisées, l'expert fixe les poids en fonction de ses croyances et de son savoir-faire. Dans le second cas de figure, les méthodes non supervisées ajustent, au fur et à mesure de leur apprentissage, le poids de chaque descripteur en fonction de l'erreur induite (tel que les réseaux de neurones). Tous les descripteurs sont initialisés à la même valeur puis, durant l'apprentissage, les exemples bien reconnus par l'algorithme d'induction entraînent l'augmentation du poids des descripteurs qui les composent, et dans le cas d'échec, le poids des descripteurs utilisés est réduit afin de moins favoriser ces descripteurs lors d'une prochaine décision.

Enfin, cette procédure implique d'utiliser des algorithmes de modélisation différents de ceux intervenant pour les trois précédentes méthodes. Ces algorithmes doivent pouvoir établir un modèle en favorisant les descripteurs de poids élevé dans la construction de celui-ci.

5.3.5 La sélection de descripteurs dans `CATMInE`

Dans `CATMInE`, les méthodes `wrapper` et `filter` ne sont pas nécessaires, le nombre de descripteurs est restreint à huit, plus la valeur de classe. L'espace de recherche est donc limité, et peut être traité dans des temps raisonnables. En revanche, en utilisant des algorithmes de modélisation comme les arbres d'induction, nous avons testé les principes de la méthode embarquée (*embedded*).

Dans l'évolution de notre projet (avec pour résultat `DeTTMInE`), nous avons démontré l'intérêt d'une pondération des descripteurs dans un processus décisionnel de type arbre de décision par l'indice de qualité du nœud auquel ils sont associés. Cet indice de qualité est défini par [Fournier, 2001]. Nous avons alors utilisé comme principe de sélection que le nœud de plus fort indice de qualité est le nœud le plus représentatif de l'arbre. Le descripteur qui y est associé est donc celui qui est le plus représentatif de l'arbre. Nous présenterons en détails ces travaux dans la section 9.2 page 130.

5.4 Conclusions

Comme nous l'avons vu tout au long de ce chapitre, la transformation des données permet l'expression d'un phénomène non explicité en l'état. Nous retrouvons des démarches intellectuelles sur les descripteurs présents, pour observer des propriétés intrinsèques. La segmentation des descripteurs continus, ou la fusion de descripteurs en sont de bonnes illustrations.

Mais, adapter les descripteurs à un savoir-faire donné n'est pas pour autant garant d'une bonne qualité de modélisation. Pour cela, il faut aussi déterminer quels sont les descripteurs les plus à même de répondre au problème, les plus pertinents pour le domaine étudié. C'est à ce moment qu'interviennent les différentes méthodes de sélection de descripteurs que nous venons de présenter. Celles-ci vont permettre à l'expert d'évaluer une certaine notion, ou mesure d'intérêt de ceux-ci. Adopter une méthode *filter* sera moins coûteuse qu'une méthode *wrapper*. Toutefois, la complexité algorithmique de cette dernière produira de meilleurs résultats, mais sera étroitement liée à l'algorithme utilisé, contrairement à la méthode *filter* applicable à tout algorithme produisant des conjonctions ou des disjonctions de descripteurs comme résultats. Les deux dernières méthodes *embedded* et de *pondération* sont intrinsèques à une classe d'algorithmes les mettant explicitement en jeu (de manière beaucoup moins flexible que l'approche *wrapper*).

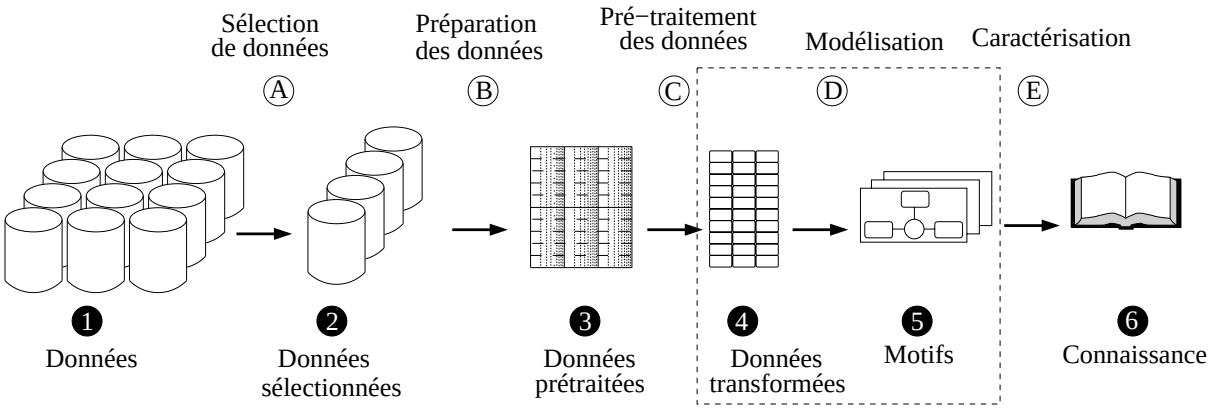
Cette étape de transformation des données a pour conséquence de valider les choix effectués dans les descripteurs établis. Elle permet aussi de corriger certains partis pris, et de déterminer

un sous-ensemble, jugé pertinent, de descripteurs.

Pour l'aide à la décision dans la contrefaçon des marques, nous avons mis en place des procédés de fusion de descripteurs et de sélection de descripteurs par pondération. Le procédé de fusion pour renforcer le pouvoir discriminant d'un descripteur et le second afin de maximiser le modèle de classification induit. Ce procédé repose sur l'utilisation de la notion d'indice de qualité de nœud, et est non supervisé. Cette information est ensuite utilisée dans le processus d'aide à la décision. Ce procédé a été comparé au même procédé d'induction n'intégrant pas cette qualité de nœud dans son processus d'induction.

Chapitre 6

Modélisation d'un phénomène exprimé par une base de données documentaire



Modélisation d'un phénomène : passage d'une base documentaire pré-traitée à des motifs caractérisant de la connaissance.

Sommaire

Dans le chapitre précèdent, nous avons abordé les principes de transformation de données à appliquer avant de modéliser un phénomène. Ces méthodes de pré-traitements sont effectuées une fois les données collectées et triées. Ces étapes permettent d'exprimer de nouveaux descripteurs à partir d'un sous-ensemble existant, de réduire les modalités d'un descripteur ou encore de segmenter des valeurs continues. Disposant d'une base de données préparée et orientée pour décrire un phénomène, l'exploiter afin de produire de nouveaux outils d'accès, d'archivage et de prédiction, implique d'en déduire un modèle. Ce dernier est généralement issu d'un processus d'induction cherchant à identifier des sous-ensembles d'enregistrements présentant des caractéristiques communes ou au contraire des caractéristiques divergentes. Dans ce chapitre nous présentons les motivations d'une telle démarche. Comme pour la sélection de données, la modélisation peut se faire de façon supervisée (section 6.4 page 87) ou non (section 6.5 page 89). Les stratégies essentielles à chacun de ces axes sont présentées. Ce chapitre du mémoire n'a toutefois pas la prétention de couvrir l'intégralité des méthodes disponibles, mais de ne retenir que celles qui, parmi les plus fréquemment utilisées, correspondent à notre projet. Ces méthodes ont fait l'objet de tests sur la problématique de `CATMInE` afin d'évaluer et de comparer leur pouvoir prédictif.

6.1 Motivations

A ce stade du processus de la découverte de connaissances, l'expert dispose d'une base de données exprimant le phénomène d'étude et issu d'en ensemble documentaire. Si les étapes précédentes ont été respectées, les descripteurs composant la base ne présentent aucune incohérence en terme de modalité. De cette base, nous souhaitons obtenir un modèle capable d'exprimer les relations intrinsèques entre descripteurs et donc les caractères communs de populations homogènes. Ces relations sont caractérisées par des modalités de descripteurs qui sont partagées entre les différents enregistrements du jeu de données.

Ces relations seront qualifiées de fortement discriminantes car utilisées pour caractériser le phénomène. Elles seront ensuite traduites en connaissances exploitables ou non par l'expert et peuvent revêtir plusieurs aspects tels des règles de classification ou encore des barycentres (appelés centroïdes) de regroupements. Ces connaissances serviront ensuite de support pour prédire de nouvelles instances et justifier les choix de ces nouvelles décisions. Ces connaissances peuvent aussi être utilisées pour classer et interroger une base de données dans le but d'accéder de façon évoluée à son contenu. Pour réaliser cette modélisation, nous retrouvons comme pour les précédentes étapes la dualité entre une voie explicative, la modélisation supervisée et une voie exploratoire par le biais d'une modélisation non supervisée. Ces deux orientations reposent sur un paramétrage qui diffère d'une méthode à une autre. Celui-ci influe sur le comportement de l'algorithme lors de la construction du modèle. La section suivante présente les détails de cet aspect.

6.2 Modèle d'une base de données documentaire : définition

La modélisation d'une base documentaire a pour objectif d'apporter de l'information et de déduire des connaissances à partir des documents électroniques composant cette base documentaire. Le résultat de cette opération est ce que nous définissons comme modèle. Ce modèle est déterminé par un algorithme d'apprentissage et est caractéristique de la combinaison de l'algorithme et de la représentation de la base documentaire.

Les algorithmes d'apprentissage sont répartis dans plusieurs familles algorithmiques en fonction de leur philosophie et donc de leur stratégie d'apprentissage. De plus, les modèles dépendent étroitement de la représentation des documents composant la base d'apprentissage. Cette représentation a été abordée dans le chapitre précédent et nous avons démontré la nécessité de valider cette phase avant toute tentative d'apprentissage. Nous entendons alors par modèle trois aspects :

1. la représentation des données sous forme de vecteurs de descripteurs (les données préparées)
2. la philosophie d'apprentissage

3. le résultat issu de l'apprentissage du jeu de données préparé.

Nous retiendrons ensuite les caractéristiques présentées ci-dessous pour définir un modèle et les attentes que l'on peut en avoir.

- Un des premiers aspects d'un modèle est sa capacité à déterminer des propriétés communes entre enregistrements et donc à les regrouper en fonction de ces propriétés communes, apportant une valeur sémantique à ces regroupements. Cela permet alors d'améliorer la conservation des documents en permettant un archivage et une consultation plus adaptés de ceux-ci. Les documents ne sont plus accessibles sur des propriétés de type : auteur, date, volume, journal, mais bien sur des valeurs plus intrinsèques. Dans `CATMInE`, cet accès peut être par exemple la sélection de tous les documents présentant des marques fortement similaires, ou ayant une sous-chaîne graphémique de longueur et de contenu précis.
- Une seconde motivation pour l'établissement d'un modèle est l'étude qui peut en être faite afin de comprendre le domaine étudié. Les regroupements d'enregistrements similaires peuvent être établis à partir d'un descripteur ou bien d'une combinaison de descripteurs. Cette information permet alors de mieux comprendre les regroupements et de faire apparaître des propriétés qui ne sont pas nécessairement connues de l'expert. Nous parlons alors de la connaissance induite d'un modèle. C'est ensuite à l'expert de juger la pertinence de cette connaissance en évaluant son intérêt dans la compréhension du domaine. Dans `CATMInE` et son évolution, cela permet de modéliser la jurisprudence. Cette jurisprudence caractérise le courant de pensée des décisions juridiques et permet à l'expert de comprendre l'évolution des décisions dans le temps.

De cette information déduite par la structure du modèle (les regroupements obtenus), nous disposons alors d'informations pouvant être utilisées pour trier des documents du domaine. Si l'apprentissage est réalisé sur un ensemble de documents précis, alors pour tout nouveau document du domaine, nous pouvons déterminer son appartenance à un des groupes établis. Nous parlons alors de prédiction. Cet aspect est la fonctionnalité première de `CATMInE`. Le système doit être en mesure de déterminer pour tout nouveau couple de marques s'il y a contrefaçon ou non et ce à partir de la connaissance induite pendant la phase d'apprentissage.

Cependant, il ne faut pas perdre de vue que le modèle est issu d'une vue subjective calculée par des procédés théoriquement objectifs. Par vue subjective, nous posons que la représentation des données, le phénomène étudié et les documents utilisés sont respectivement établis, définis et sélectionnés par un expert du domaine. Par calculs objectifs, nous faisons un rapprochement avec l'utilisation d'algorithmes et de méthodes plus ou moins supervisées pour établir le modèle.

Si nous tenons compte de ces contraintes, un modèle n'est jamais définitif, il évolue au fur et à mesure de la production des documents électroniques composant la base documentaire. Cela implique alors d'avoir des processus de mise à jour du modèle devant tenir compte de cet aspect. L'idéal étant bien entendu de ne pas recalculer la connaissance existante, mais de la compléter

et de la corriger afin de tenir compte de l'évolution du domaine.

Enfin, pour terminer l'énumération des propositions définissant un modèle, nous pensons que celui-ci ne doit jamais être considéré comme acquis et doit être comparé à d'autres modèles issus d'un apprentissage réalisé avec d'autres familles d'algorithmes. Pour répondre à cette proposition, certains auteurs ont réfléchi à des stratégies d'apprentissage prenant en compte différentes familles d'algorithmes afin de bénéficier de la combinaison des stratégies d'apprentissages possibles.

De là, lorsque l'expert a terminé de préparer les données, qu'il a défini les objectifs d'apprentissage, et qu'il a choisi le type de stratégies d'apprentissages pour réaliser sa modélisation, il va pouvoir agir d'une autre manière sur le modèle produit : les paramètres de ces stratégies.

6.3 Le paramétrage des méthodes de modélisation

L'ensemble des méthodes d'apprentissage, supervisées ou non, repose sur des paramètres permettant d'agir sur le processus d'obtention des connaissances. Ces paramètres vont chercher à maximiser une mesure dépendant du type de descripteur utilisé, ou permettant la segmentation des données en catégories optimales. Ce paramétrage correspond à des partis pris permettant d'agir sur la modélisation. Les méthodes d'apprentissage sont déjà composées de tel biais dans la définition même de l'algorithme utilisé. En fixant les paramètres, nous redéfinissons le comportement l'algorithme utilisé. La conséquence directe est alors d'augmenter le nombre de méthodes disponibles afin de traiter un problème.

Nous retrouvons ainsi comme type de paramètres le gain d'information ou encore le gain ratio, le second corrigeant le premier sur la sélection de descripteurs. Le gain d'information a tendance à ne retenir que les descripteurs ayant le plus de modalités. Selon Leo Breiman ([Breiman, 1996b]) cela a pour effet la dispersion des données. Pour limiter cet effet secondaire, la mesure de gain ratio a été introduite, maximisant cette fois le rapport entre gain d'information du descripteur et son entropie.

Pour les méthodes non supervisées (de regroupement essentiellement), des mesures comme celle du cosinus, du coefficient de corrélation de Pearson, du coefficient étendu de Jacquard ou encore de distance Euclidienne sont introduites pour répondre au problème de regroupement des données de façon pertinente. Les deux premières mesures font de deux enregistrements, des enregistrements similaires si ceux-ci vont dans la même direction[13]. Sur un autre principe, la mesure de distance euclidienne prend en compte la direction et la longueur des enregistrements. Enfin, la mesure de Jacquard étend la précédente en prenant en compte l'angle formé.

La gestion des variables continues et plus particulièrement de la recherche des valeurs charnières est la même problématique quel que soit l'algorithme appliqué. Il faut déterminer la valeur maximisant le critère de partition. Cette valeur sera la valeur charnière si elle maximise égale-

[13]Les descripteurs sont conservés dans un tableau, et peuvent donc être vus comme un vecteur

ment le critère de partition de l'ensemble des autres variables (principe retenu pour l'algorithme `C4.5`, [Quinlan, 1993]). Cependant cette méthode favorise les descripteurs très fragmentés, réputés peu fiables [Fournier, 2001]. Cet auteur résout ce problème de sélection de valeur pivot en les déterminant avant construction du modèle. Les pivots sont alors jugés plus fiables car ils reposent sur l'examen de l'ensemble des données et ne sont plus calculés à la volée. La méthode proposée détermine tous les intervalles de la variable dans lesquels il n'y a que des éléments d'une même classe ou inversement que des éléments différents.

Il existe aussi comme autres paramètres, les critères d'arrêts d'induction. Pour les algorithmes de regroupements, c'est généralement le nombre d'itérations nécessaires pour produire un modèle. Pour les arbres d'induction (méthode supervisée) nous retrouvons par exemple l'obtention de feuilles pures (non hétérogènes) ou contenant un nombre déterminé d'exemples ou encore l'épuisement des variables discriminantes pour la construction du modèle. Il est aussi possible d'agir sur des paramètres de généralisation des modèles obtenus. Pour les arbres d'induction, nous parlons alors d'élagage du modèle induit.

[Hilario, 2002] détaille l'impact de la variation des paramètres sur les performances de plusieurs algorithmes. Cette étude expérimentale compare neuf algorithmes d'apprentissage (dont des arbres de décision, des algorithmes producteurs de règles ou encore des réseaux de neurones) sur un ensemble de 70 bases de données préparées dite de références provenant de l'UCI[14]. Ce site regroupe une collection de bases de données, toutes reconnues par les acteurs du domaine, pour lesquelles des objectifs de fouilles de données sont proposés. Les conclusions issues de cette étude sont les suivantes :

1. l'incapacité de comparer des algorithmes en utilisant les paramètres par défaut, et que pour toute base de données réelles (a opposer à artificielle comme celle de l'UCI)

2. l'impossibilité de prendre le temps d'évaluer chaque algorithme avec chaque paramètre possible

3. et enfin, la nécessité d'améliorer les algorithmes d'apprentissage afin qu'ils évaluent au mieux leur stratégie d'élaboration de modèle en fonction de leur biais (ou parti pris)

La section suivante aborde les différentes méthodes possibles de construction de modèles. Il faut garder à l'esprit que les paramètres présentés sont peu nombreux au regard de toutes les méthodes disponibles. Chacune de ces méthodes met en place une stratégie particulière et toutes ont plus ou moins de paramètres qui leur sont propres. Toutefois, les paramètres présentés ci-dessus sont assez génériques et peuvent être mis en pratique dans nombre des méthodes.

[14] **U**niversity of **C**alifornia, **I**rvine, http ://kdd.ics.uci.edu/

6.4 Approche supervisée

Cette approche propose de mettre en relation des enregistrements étiquetés (par l'expert ou un procédé automatique) avec une valeur de classe (cf. section 2.4 page 31). L'étiquette des données correspond aux regroupements possibles des enregistrements. La modélisation est donc biaisée par un objectif à atteindre et dont l'algorithme a connaissance. Le procédé choisira alors les descripteurs permettant de distinguer le plus finement possible les classes établies pour le domaine. Cet objectif est équivalent à prédire la valeur de classe par une fonction établie avec les descripteurs disponibles. Cet but, si la base décrit bien le phénomène, est le seul caractérisable. Dans le cas de données mal préparées, l'objectif n'est peut être pas le phénomène le mieux caractérisé par les enregistrements et peut donc dégrader la qualité du modèle issu du processus d'induction. Si la valeur de classe est multi-valuée (avec des valeurs discrétisées) la littérature parle alors de classifications. Dans le cas contraire, le descripteur de classe est une valeur réelle (descripteur continu), alors il s'agira de résoudre un problème de régression. Dans la modélisation supervisée, deux courants se distinguent nettement : les procédés numériques qui reposent sur des fonctionnements de régression linéaire et d'analyse discriminante, et les procédés automatiques développés sur des approches statistiques ou symboliques.

6.4.1 Modèles numériques

Le principe des modèles numériques consiste à trouver des relations mathématiques entre différents descripteurs quantitatifs pour en expliquer un autre : la valeur de classe. Nous parlons alors de régression linéaire (c'est le cas par exemple avec les réseaux de neurones, [Dreyfus, 2002]). Cette méthode permet d'expliquer une variable par la mise en relation linéaire d'autres variables disponibles.

Ces relations permettent ensuite de prédire une valeur de classe, à partir des relations établies durant la modélisation. Il est important de garder en mémoire que le principe de régression linéaire ne peut s'appliquer qu'à des données quantitatives.

À partir des résultats obtenus, l'expert peut chercher à éliminer des descripteurs utilisés pour simplifier l'expression du domaine, à condition que ceux-ci ne soient pas pénalisant pour la qualité et la justesse du modèle.

Une deuxième approche de modèle numérique est l'analyse factorielle ([Institute, 1989]). À la différence de la régression linéaire, l'analyse factorielle permet de traiter des descripteurs quantitatifs caractérisant des données qualitatives sur un principe proche de la régression linéaire. Cette méthode compose de nouveaux descripteurs à partir de ceux décrivant les données avec pour objectifs de reproduire au mieux la géométrie de l'ensemble initial, dans un espace de dimensions inférieures. Nous obtenons alors un modèle d'approximation permettant d'attribuer une valeur de classe à un nouvel enregistrement.

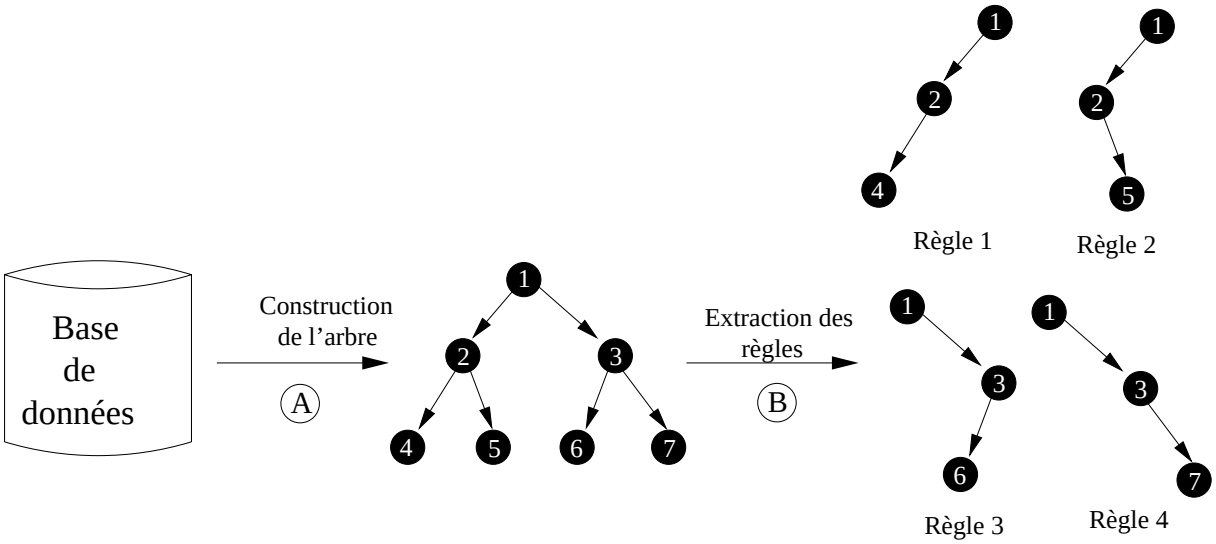


FIG. 6.1 – D'une base de données à des branches unaires.

Ces deux approches restent les stratégies les plus élémentaires des modèles numériques. Quelle que soit la méthode retenue, celle-ci nécessitera des coûts en terme de puissance de calcul, imposés par les calculs matriciels nécessaires à l'expression du domaine. Une étude détaillée de ces méthodes a été réalisée par [Auray et al., 1991].

6.4.2 Modèles d'apprentissage automatiques

Par modèles d'apprentissage automatiques, nous regroupons ici tous les procédés de modélisation reposant sur des analyses statistiques et symboliques avec pour stratégie l'accumulation des enregistrements ou bien le partitionnement successif de ceux-ci. Pour le premier cas de figure, la programmation logique inductive ([Lavrac & Dzeroski, 1994]) permet d'orienter l'exploration des données dans l'espace de recherche.

Toutefois, ce type de méthode repose sur un test de couverture appelé *thêta-subsumption* limitant l'algorithme par une complexité exponentielle en fonction de la taille de la solution (nombre de modalités possibles pour la valeur de classe). [Sebag & Gallinari, 2002] en font une étude détaillée.

Le partitionnement de l'espace de recherche, deuxième cas de figure des modèles automatiques, est représenté en très large partie par les modèles de type arbre d'induction, tel que *C4.5* présenté par [Quinlan, 1993]. La figure 6.1 présente succinctement le passage d'une base de données à un arbre d'induction et d'un arbre à des règles.

De la base de donnée, l'algorithme d'induction d'arbre établit un modèle (A). De cet arbre, le processus établit l'ensemble des règles possibles (B). Ces règles sont identifiées par le chemin allant de la racine de l'arbre (nœud 1) à chaque feuille (nœud 4, 5, 6 et 7).

FIG. 6.2 – Passage d'une branche à une règle de classification.

La représentation finale du modèle est une arborescence ou chaque nœud représente une sélection de descripteurs. Les feuilles de l'arbre sont caractérisées par la population identifiée par la branche conduisant de la racine de l'arbre à la feuille observée. Quelle que soit la méthode retenue, la connaissance issue de ces modèles est généralement formalisée par des règles. La figure 6.2 présente le passage d'une branche à une règle. À chaque nœud est associé une contrainte sur le descripteur, et la conjonction des nœuds identifiants la branche représente la règle.

Ces règles facilitent la compréhension du domaine par l'expert, à la condition qu'elles soient peu complexes et transcriptibles en langue naturelle ou en syntagmes explicites.

6.5 Modélisation non supervisée

La différence entre une modélisation supervisée et une modélisation non supervisée repose sur la connaissance des étiquettes (ou valeur de classe) de chaque enregistrement. Ces étiquettes biaisent la modélisation en orientant la recherche de relations entre descripteurs pour discriminer au mieux les enregistrements par rapport aux étiquettes possibles. Ne pas disposer de cette information permet à l'algorithme de modélisation non supervisée de déterminer des connaissances sur un phénomène non précisé, mais dont l'instance (par le biais des enregistrements) le décrit.

Si le domaine est bien représenté par les descripteurs, l'algorithme de modélisation devrait établir les bons regroupements. Dans le cas contraire, il est possible de voir émerger des sous-groupes non prévus dans la modélisation.

Les méthodes non supervisées vont chercher à identifier les enregistrements semblables et homogènes. Les connaissances ainsi déduites sont très fortes. En effet, comme nous n'orientons pas le regroupement d'exemples, une telle méthode est alors capable de déterminer des sous-ensembles dont l'expert n'avait pas supposé l'existence.

Dans la base documentaire de `CATMInE`, les contrefaçons de marques nominatives par d'autres marques nominatives déposées sont un sous-ensemble des contrefaçons de marques nominatives. Une analyse où chaque décision est étiquetée comme étant une contrefaçon de marques ne permet pas de faire apparaître les décisions faisant partie des contrefaçons de marques nominatives par

des marques déposées.

Sans *a priori* sur les faits retenus qui caractérisent la contrefaçon dans la décision, les jugements de contrefaçon entre marques nominatives déposées ont des descripteurs en communs avec les contrefaçons de marques par reproduction [15]. Ces descripteurs permettent des regroupements possibles entre les décisions avec des jugements issus de ces deux types de décisions et présentant des propriétés graphémiques et phonémiques similaires.

Les méthodes de modélisation non supervisées sont réparties au sein de trois grandes familles :

1. de partitionnement
2. hiérarchique
3. les règles d'associations

6.5.1 Les modèles de partitionnement

Les modèles de partitionnement construisent de façon directe des regroupements. Cette construction est à opposer à une construction incrémentale utilisée dans les méthodes hiérarchiques et présentée dans la sous-section suivante. Ces modèles de partitionnement se divisent en deux sous-catégories : un partitionnement par ré-allocation des enregistrements entre sous-ensembles, et un partitionnement par zones contenant beaucoup d'enregistrements (notion de densité).

Partitionnement par ré-allocation d'enregistrements

Les méthodes de partitionnement appliquant des principes de ré-allocation d'enregistrements entre sous-ensembles sont à nouveau caractérisées par trois grands courants :

1. les approches probabilistiques
2. les approches de type k-médoïdes
3. les approches de type k-means

Dans les approches probabilistiques, les données sont vues comme un modèle mixte, caractérisé par plusieurs probabilités de distribution. L'objectif est donc de déterminer des regroupements respectant ces probabilités, et de les raffiner par le biais de mesures intra et inter-regroupements. Dans cette famille d'algorithmes, nous retrouvons par exemple l'algorithme `EM` (Expectation Maximization) [McLachlan & Krishnan, 1997], `SNOB` [Wallace & Dowe, 1994], `AutoClass` [Cheeseman & Stutz, 1996] ou encore `MClust` [Fraley & Raferty, 1999]. Enfin, l'un des principaux avantages de cette technique est la clarté des résultats proposés.

[15] Ce type de décisions est donné à titre d'exemple. Ces décisions sont aussi présentes dans la base documentaire d'origine.

Pour les approches de type *k-médoïde*, un regroupement est représenté par un des enregistrements le constituant, le médoïde. La frontière du regroupement se détermine alors en ne retenant que les enregistrements les plus similaires. Nous retrouvons comme principaux algorithmes spécialisés dans ce type de modélisation, l'algorithme `PAM` (Partitioning Around Medoids) et l'algorithme `CLARA` (Clustering Large Application)[Kaufman & Rousseeuw, 1990].

Les approches de type *K-means* ([Hartigan, 1975] et [Hartigan & Wong, 1979]) ont pour principe de déterminer le centroïde (enregistrement représentant la moyenne de ceux retenus pour constituer le regroupement) de chaque regroupement par le calcul de leur barycentre.

Partitionnement par densité d'enregistrements

Le partitionnement par densité d'enregistrements correspond à la seconde sous-catégorie de partitionnement. Le raisonnement sur la densité consiste à déterminer les régions où la population est forte de celle où elle est faible. Les régions de faible densité sont considérées comme du bruit et les autres comme les regroupements a effectuer. L'expansion des regroupements suit donc la recherche des plus proches voisins et est guidée par la densité : les algorithmes mettant en place ce principe peuvent alors déterminer des regroupements de forme arbitraire. Les algorithmes les plus représentatifs incluent par exemple `DBSCAN` (Density Based Spatial Clustering of Application with Noise)[Ester et al., 1996], `OPTICS` (Ordering Points To Identify the Clustering Structure) [Ankerst et al., 1996] ou encore `DENCLUE` (DENsity-based CLUstering) [Hinneburg & Keim, 1998].

6.5.2 Les modèles hiérarchiques

Une autre façon de produire des modèles de manière non supervisée consiste à établir une hiérarchie entre les regroupements. Le principe se résume à fusionner des sous-ensembles présentant des caractéristiques communes afin d'obtenir un nouvel ensemble, ou bien de scinder un regroupement n'offrant pas les attentes en terme d'homogénéité. Le premier cas de figure est dit ascendant, contrairement au second, qualifié de descendant. La mesure d'homogénéité guide l'algorithme. Avec des mesures géométriques (de distance, tel que AGlomerative NESting, `AGNES`, [Kaufman & Rousseeuw, 1990]), la forme des clusters sera plutôt convexe, tandis que d'autres procédés de regroupement tel que Clustering Using REpresentatives (`CURE`) [Guha et al., 1998] la forme des clusters devient arbitraire : un ensemble représentatif d'enregistrements est utilisé pour caractériser un cluster (et non un seul point, centroïde du cluster). Les résultats issus d'une hiérarchisation des clusters produisent un dendogramme (un arbre) des clusters. Cela permet ainsi d'étudier l'ensemble solution en prenant en compte les différents niveaux de granularité produits.

6.5.3 Les modèles de règles d'association

La recherche de règles d'associations dans le *Machine Learning* est une discipline encore récente, très proche de la fouille de données. Cette recherche s'applique généralement à des données symboliques binarisées (chaque descripteur a un intervalle de définition se réduisant à deux modalités).

Le principe des règles d'association est d'isoler des motifs fréquents, parmi un ensemble de motifs candidats. Un motif fréquent se définit par un support (nombre d'occurrences), et pour être considéré comme fréquent, ce support doit être supérieur à un seuil fixé par l'expert. Les motifs candidats sont quant à eux des motifs dont le support n'est pas encore déterminé. Ceci est une propriété importante car pour n'importe quel sous-ensemble de motifs candidats si ce n'est pas un motif fréquent alors l'ensemble des motifs candidats ne le seront pas.

Les algorithmes actuels (`Apriori`, [Agrawal & Srikant, 1994] ou `mvminner` [Rioult, 2002]) reposent sur la propriété de monotonicité permettant d'élaguer le nombre de motifs à un niveau donné, optimisant alors les accès à la base de données. Ils offrent en plus la possibilité d'être exprimés par une représentation condensée sous forme équivalente ou rapprochée. Cette représentation permet de limiter le volume de connaissances fournies à l'expert pour analyse.

Pour trier l'information, des mesures d'intérêt subjectives (fournies par l'expert) ou objectives (fondées sur un rapport entre le support et le nombre total d'exemples) sont disponibles. Ces méthodes sont encore jeunes dans le domaine de la recherche de connaissances, mais commencent à y être intégrées de plus en plus. L'application dans l'extraction de regroupements pour la découverte de communautés d'intérêt [Durand, 2004] est un des exemples commençant à émerger avec ce type de représentations que sont les motifs fréquents.

6.6 Les méta-modèles

Les méta-modèles rendent possible ce qui ne l'était pas avec les modèles simples abordés précédemment. Ils permettent de combiner plusieurs algorithmes et donc de tenir compte de leurs spécificités pour modéliser un problème. En effet, les classifieurs, en fonction de leur comportement (ou parti pris) dans une modélisation ne peuvent couvrir l'intégralité des exemples présents dans une base documentaire ([Vilalta & Drissi, 2002]). L'objectif des méta-modèles est donc de bénéficier des spécificités de chaque algorithme utilisé pour modéliser un ensemble de données.

Le méta-modèle peut par exemple caractériser une base en ne conservant que l'algorithme le plus efficace parmi un ensemble disponible. Cette sélection s'effectue en comparant les résultats des différents algorithmes. L'algorithme ayant le meilleur pouvoir prédictif sera alors retenu.

Le méta-modèle peut aussi, à chaque expérience, utiliser tous les algorithmes dont il dispose. La prise de décision se fera alors en combinant l'ensemble des résultats obtenus avec les différents

algorithmes. Cela peut être une moyenne ou encore un choix par pondération des algorithmes.

Enfin, une autre approche consiste à diviser le jeu de données en sous-ensembles. Ceux-ci sont alors caractérisés par des algorithmes. Le schéma consiste à appliquer différents types d'algorithmes à chaque sous-ensemble et à ne conserver que l'algorithme ayant réalisé le meilleur apprentissage de ce sous-ensemble. Chaque ensemble est donc étroitement lié à l'algorithme le plus performant pour lui-même. Cette caractéristique s'explique par le fait que les algorithmes mettent en jeu différentes stratégies d'apprentissage. Ils ne vont donc pas apprendre le jeu de données de la même manière et certains vont se révéler meilleurs que d'autres pour certaines enregistrements et plus faibles sur d'autres.

Ensuite pour tout nouvel exemple, l'algorithme de prise de décision place le nouvel exemple dans l'espace de définition du domaine et applique l'algorithme dédié au sous-ensemble le plus proche du nouvel exemple. Toute la difficulté réside donc dans la combinaison des modèles.

Le principe même du méta-modèle impose donc à un moment de faire un choix sur la décision à retenir parmi l'ensemble disponible. Ce choix peut être de trois natures différentes : le vote, l'empilement et la proximité entre modèles. Ces trois notions sont détaillées dans la suite de cette section.

6.6.1 Le méta-modèle de type vote

La méthode de vote est la plus simple et la plus naturelle à mettre en place. Chaque algorithme intervenant dans le méta-modèle produit une solution. La solution majoritaire sera alors retenue pour le choix final du processus décisionnel. L'influence d'un algorithme sur la décision dépend bien évidement de la qualité du modèle produit par celui-ci. Une approche consiste à donner à chaque algorithme un poids lié à la précision du modèle. Dans [Littlestone & Warmuth, 1989], l'algorithme `WM` est proposé. Ici, l'intuition de pondération consiste à donner un poids identique à tous les algorithmes. À chaque prédiction le méta-modèle décrémente le poids des algorithmes ayant prédit la mauvaise issue par rapport à la décision collective.

6.6.2 Le méta-modèle de type empilement

Le méta-modèle de type empilement (dit de *stacking*) se propose de modéliser le domaine avec chaque algorithme retenu pour la modélisation. Ceux-ci sont alors modélisés (en tant que processus du `KDD`) afin que leurs résultats[16] soient exprimés à l'aide de descripteurs partagés. Ces descripteurs représentent alors un méta-niveau de modélisation du domaine. Un algorithme de décision est alors appliqué sur ce méta-modèle et apprend les caractéristiques et performances de chaque algorithme appliqué dans la première modélisation. Par exemple, dans [Todorovski & Dzeroski, 2000] de Todorovski et Dzeroski, le principe d'empilement mis en pratique repose

[16] La valeur de classe déduite par l'algorithme

sur un arbre de décision pour le méta-modèle. Cet arbre, à la différence des arbres classiques, recommande l'algorithme le plus adapté pour cet exemple plutôt que d'en prédire l'issue. L'apprentissage de cet arbre méta repose sur des caractéristiques de probabilité maximale d'une classe, d'entropie des distributions de classe pour chaque classifieur, et l'issue prédite par le classifieur en plus du descripteur de classe de l'exemple.

6.6.3 Le méta-modèle en cascade

Le méta-modèle dit en cascade repose sur le principe qu'un classifieur de base, A, bon en généralisation (réseau neuronal, par exemple) peut modéliser une large partie d'un jeu de données, mais qu'un sous-ensemble plus délicat d'enregistrements ne sera pas pris en compte. Ces exemples sont désignés comme des exceptions. Ce sous-ensemble sera alors traité par un algorithme B, plus coûteux en ressources de calculs, capable de produire un modèle plus complexe des données restantes ([Kaynak & Alpaydin, 2000]).

La principale difficulté porte sur la reconnaissance d'un nouvel exemple et donc de la couverture de l'algorithme A et de l'algorithme B. Pour cela, pendant la phase d'apprentissage, la confiance de l'algorithme A est mesurée. Cette mesure dépend de l'approximation de l'appartenance (distance au concept) de chaque exemple à apprendre, à l'ensemble c des classes possibles. Il y a une ambiguïté forte lorsque cette mesure est égale à $1/c$ (la différence entre valeur attendue et valeur calculée est forte). L'algorithme est réputé fiable lorsque cette mesure est proche de 1 (la différence entre valeur attendue et valeur calculée est faible). Pour tout nouvel exemple à apprendre, si cette mesure de confiance est supérieure à un seuil s fixé, l'algorithme A est appliqué, sinon, l'algorithme B est utilisé pour apprendre l'exemple et celui-ci est considéré comme exception.

Dans une phase de test, l'algorithme A est appliqué et une solution est calculée. Ensuite, il faut vérifier si A est suffisamment sûr. Si c'est le cas, A est appliqué, sinon, B est appliqué et sa décision retenue.

Pour affiner la méthode, il est possible de mettre en place plusieurs types d'algorithmes de complexité croissante à la place de l'algorithme A puis d'appliquer l'algorithme B. Par exemple, [Kaynak, 1997] démonte qu'enchaîner l'apprentissage de l'algorithme des k plus proches voisins à celui d'un réseau de neurones à simple couche permet d'améliorer la qualité de la décision.

6.6.4 Le *boosting*

Le boosting a été introduit par Schapire en 1990 ([Schapire, 1990]), et l'algorithme le plus connu est *AdaBoost* ([Freund & Schapire, 1996]). Le principe consiste à établir lors d'une modélisation plusieurs sous-ensembles de données issus de la base initiale. Pour chaque sous-ensemble, un classifieur est établi (choix parmi l'ensemble des classifieurs existant). Nous parlons alors de

classifieur de base.

Une fois les classifieurs établis, les sous-ensembles sont soumis à validation avec les classifieurs de base associés. Chaque instance de chaque sous-ensemble est alors pondérée en fonction de sa reconnaissance ou non par le classifieur de base. Le but est de forcer l'algorithme à minimiser l'erreur attendue ([Bauer & Kohavi, 1999]).

L'ensemble de ces informations - classifieurs de base, sous-ensemble et pondération des instances - représente le modèle du jeu de données étudié.

Pour une nouvelle décision, le principe consiste à établir le choix de l'algorithme de décision par une méthode de vote. Chaque algorithme de base est alors pondéré en fonction du poids des instances qui composent son échantillon d'apprentissage.

Cette méthode a été établie pour améliorer la performance des algorithmes peu efficace, ayant besoin d'être plus précis qu'une prédiction aléatoire ([Tsymbal & Puuronen, 2000]). Elle est proche du principe de méta-modèle en cascade, mais diffère sur les éléments suivants [Kaynak & Alpaydin, 2000] :

1. *AdaBoost* utilise un grand nombre d'algorithmes, les méta-modèles en cascade utilisent peu d'algorithmes afin de réduire la complexité
2. *AdaBoost* utilise le même algorithme et la même architecture pour tous les classifieurs. Une telle combinaison ne permet pas de bénéficier de la combinaison d'algorithmes différents comme le font les méta-modèles en cascade. Cette combinaison permet de réduire la corrélation entre les algorithmes en prenant en compte les différentes stratégies d'apprentissage.
3. *AdaBoost* est un système multi-experts (principe de vote entre algorithmes) alors que les méta-modèles en cascade utilisent un principe décisionnel échelonné n'impliquant pas de consulter tous les classifieurs.
4. *AdaBoost* utilise les erreurs de classification (bien classé, mal classé) plutôt qu'une mesure de l'erreur comme les méta-modèles en cascade. Cette mesure de l'erreur permet d'affiner la couverture de chaque algorithme.

6.6.5 Le *bagging*

Le *bagging* (pour `bootstrap aggregating`) génère à partir d'un jeu de données plusieurs sous-ensembles (les *boostraps*).

Puis, pour chaque sous-ensemble, la méthode produit un modèle dit de base (comme pour le *boosting*). Pour Bauer et Kohavi ([Bauer & Kohavi, 1999]) chaque classifieur de base est produit par un modèle d'apprentissage qualifié de *non stable* (réseau de neurones, arbres de décision).

Pour une nouvelle décision, le modèle applique alors la décision majoritaire obtenue à partir des prédictions de chaque classifieur de base, et offre un début de taxonomie utile.

[Breiman, 1996a] présente cette technique en détails et montre que ce procédé dégrade la qualité des modèles réputés stable comme les plus proches voisins. Ce phénomène est expliqué comme résultant de l'utilisation de petits sous-ensembles de données pour la modélisation de chaque classifieur de base.

Cette méthode est très proche des méta-modèles de type empilement par le fait d'appliquer la décision majoritaire. La différence repose sur l'utilisation d'une seule classe d'algorithmes et sur l'échantillonnage du jeu de données initial en plusieurs sous-ensembles. Enfin, comme `AdaBoost`, le *bagging* utilise l'erreur de classification (bien classé ou non) ce qui limite sa précision comparativement aux méta-modèles en cascade.

6.6.6 Bilan

Le diagramme 6.3 page suivante présente les différentes méthodes abordées dans ce chapitre. Ce schéma permet ainsi de mieux visualiser les relations entre familles d'algorithmes et d'avoir accès dans le même temps à des références sur les différentes philosophies de modélisation et leur mise en application. La nomenclature est la suivante :

Philosophie, Application (Références)

Il n'est évidement pas complet, mais il présente les grandes tendances de l'aide à la décision assistée par ordinateur.

6.7 La modélisation dans `CATMInE`

Ces différentes méthodes de modélisation et donc d'induction de connaissance ont été testées sur la base de données documentaire utilisée dans `CATMInE` afin de déterminer le meilleur procédé d'apprentissage.

Pour cela nous avons utilisé la plate-forme d'expérimentation `Weka` présentée dans [Witten & Eibe, 2005]. Cet outil permet d'établir des scénarios de modélisation en produisant des chaînes de traitement afin de préparer et sélectionner les descripteurs, puis d'appliquer des algorithmes de modélisation à partir de l'ensemble des classifieurs disponibles. Enfin, le `Weka` permet de tester ces algorithmes et de les évaluer avec différentes mesures de qualité.

Les résultats sont présentés au travers d'une interface permettant de contrôler l'affichage des résultats.

Nous avons donc comparé et évalué les algorithmes suivants en tenant compte de leur classe d'algorithme :

1. un réseau de neurones, modèle numérique supervisé
2. un arbre de décision, modèle automatique supervisé

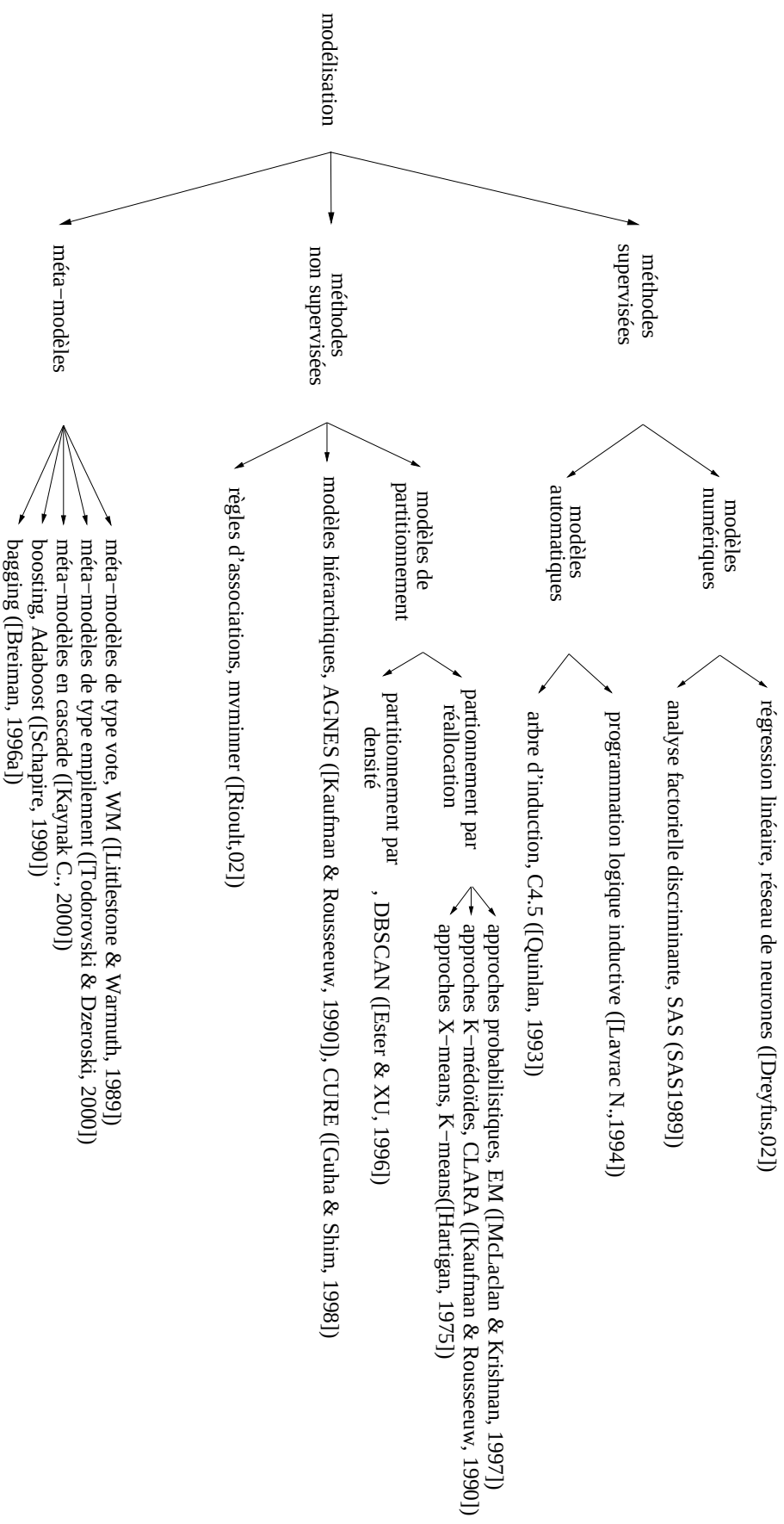


FIG. 6.3 – Schéma récapitulatif des algorithmes de modélisation et de leur réalisation

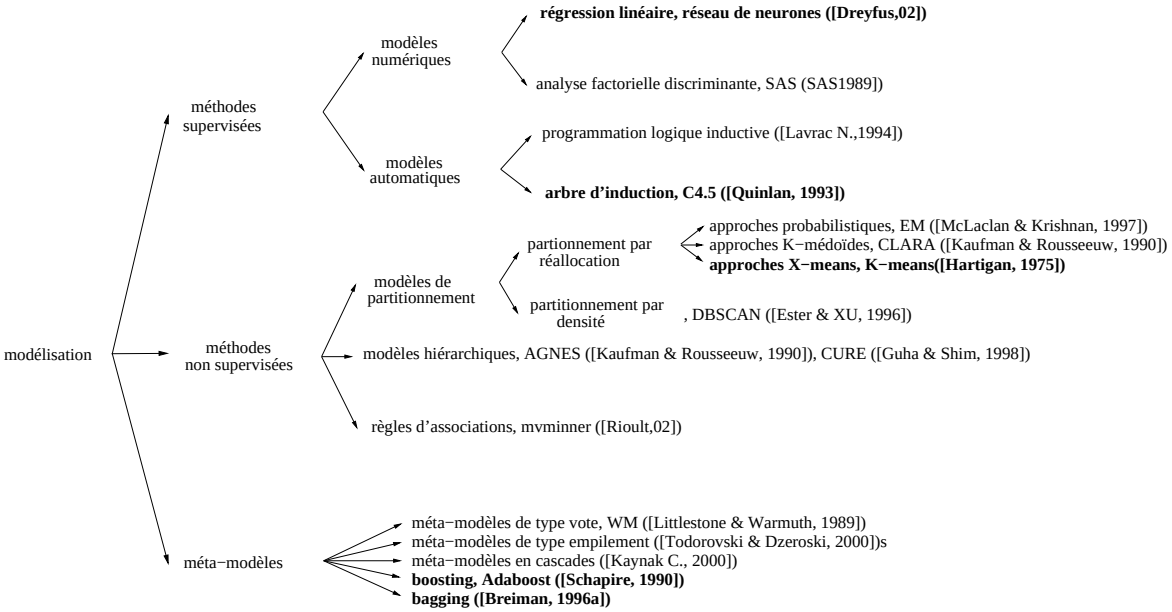


FIG. 6.4 – Algorithmes retenus pour la modélisation dans `CATMInE`.

3. un réseau bayésien,
4. un algorithme de regroupement, modèle de partitionnement non supervisé
5. des méta-modèles :
   - un méta-modèle de type *Bagging*, avec des arbres de décisions pour classifieurs de bases
   - un méta-modèle de type *Boosting*, avec des arbres de décisions pour classifieurs de bases

En procédant ainsi, nous avons validé notre recherche de connaissances à l'aide de l'ensemble des philosophies d'apprentissage disponibles, et nous nous sommes assuré d'une étude complète du meilleur procédé possible, ce que rappelle la figure 6.4. La seule stratégie qui n'a pas été utilisée de façon approfondie est relative aux règles d'associations. En effet, pour utiliser une telle méthode, il faut pouvoir être en mesure de binariser les données. Cette étape consiste à transformer un descripteur numérique ou symbolique (à plus de deux modalités) en autant de descripteurs binaires que nécessaire. Cette transformation est très délicate à réaliser dans notre cas de figure : tous les descripteurs utilisés sont numériques, les segmenter en descripteurs binaires représente une véritable difficulté. Des méthodes automatiques ont été testées sans réels succès.

Le détail de ces expériences est présenté dans le chapitre 9 page 127. Les résultats laissent apparaître, pour ce type de données et leur préparation associée, que les modèles prédictifs les plus probants sont les réseaux de neurones et les réseaux bayésiens. Les algorithmes de regroupement non supervisés ont permis de mettre en valeur un sous-ensemble de données présentant des caractéristiques communes pour des valeurs de classes différentes. Nous avons pu bénéficier de chacune des philosophies afin de mieux observer le jeu de données utilisé. Ces observations ont

ensuite un rôle important dans la compréhension de la connaissance induite et de la correction de la modélisation.

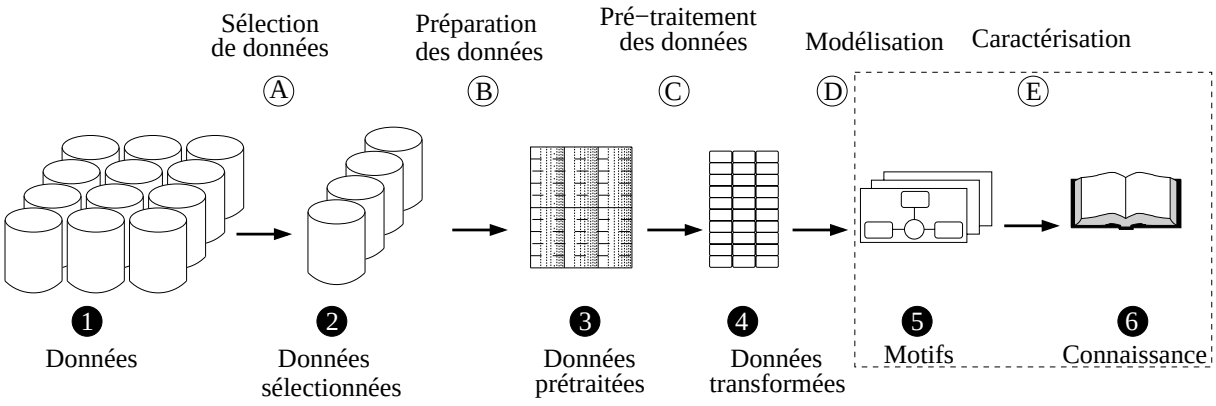
6.8 Conclusions

De manière générale, il n'existe pas d'algorithme plus performant que d'autres, ce que rappel l'étude de [Zenko et al., 2001]. Le tableau comparatif qu'ils proposent, présente dix algorithmes (de familles différentes) sur vingt et une bases extraites de l'UCI. Le taux d'erreur moyen observé est de 12,51% avec un écart de 3,02% au plus haut et 1,59% au plus bas. De nombreuses méthodes permettent de répondre aux problèmes de sources de données documentaires multiples, ou trop volumineuses. D'autres effectuent des recherches de regroupements avec ou sans *a priori* sur le jeu de données. Cela dépend alors des attentes de l'expert : sait-il ce qu'il cherche, comment veut-il que cela soit formalisé.

Dans une recherche d'optimalité, l'expert devrait pouvoir tester l'intégralité des méthodes disponibles par le biais d'une plate-forme dédiée afin de sélectionner l'outil le plus adapté à ses besoins. Cependant, établir une telle plate-forme reste délicat. Il faut pouvoir être en mesure de transposer les connaissances de l'informaticien en une matière exploitable par l'expert (définition des critères de paramétrages, des algorithmes par exemple). À cela vient s'ajouter la connaissance induite et son interprétation par l'expert. Est-ce qu'une connaissance avec une faible représentativité doit être écartée ou doit-elle être considérée comme stratégique. Cela ne peut être évalué sans intervention de l'expert et tous les outils de retour sur le processus de modélisation doivent être mis à disposition de l'expert afin qu'il détermine les enregistrements concernés par cette connaissance et puisse prendre des décisions sur la valeur à apporter à l'ensemble des résultats proposés.

Chapitre 7

Évaluation interactive des motifs : vers la formalisation de la connaissance



L'évaluation interactive des motifs : passage d'un ensemble de motifs à de la connaissance.

Sommaire

7.1 Généralités

Au cours des précédents chapitres, nous avons abordé une vue d'ensemble du principe de mis en place d'un processus décisionnel reposant sur l'apprentissage de l'existant. Nous avons présenté différentes méthodes pour caractériser un domaine, le modéliser et faire intervenir l'expert tout au long de ce processus. Si cette étape du processus de modélisation n'est pas réalisée de manière rigoureuse, l'expert perd alors tous les efforts investis dans les étapes précédentes du processus. Une telle démarche de modélisation a pour finalité l'aide à la décision, voire la prédiction en déterminant des connaissances nouvelles, valides et utiles [Fayyad et al., 1996a]. Or cette définition de la connaissance reste très subjective et est étroitement liée aux attentes de l'expert pour un tel système. Tout au long de ce chapitre nous allons utiliser les termes de motif, règle et connaissance. Les deux premiers, motif et règle, sont à prendre au même sens. Le dernier, connaissance, est à considérer comme un ensemble de motifs (et donc de règles). La connaissance n'est pas nécessairement l'intégralité des motifs calculés, mais peut exprimer un sous-ensemble de ceux-ci. Nous parlerons par exemple de connaissance stratégique. Celle-ci représente l'ensemble des motifs les plus pertinents (la définition de pertinence étant bien entendue fortement subjective) qui eux même sont un sous-ensemble du total des motifs possibles et calculés.

Une connaissance modélisée pour être ensuite traitée de façon automatique peut ne pas être exploitable en l'état par un expert. Une traduction en langue naturelle ou en langage logique s'impose pour permettre un accès à cette information. Par accès à l'information, nous faisons référence à la lecture, compréhension et interprétation de celle-ci. En effet, nous passons d'une représentation sous forme structurelle, liée au procédé de modélisation, à une représentation humainement interprétable. La représentation finale n'impose pas nécessairement d'avoir des phrases pour exprimer une information mais peut aussi être réalisée par le biais d'une structure hiérarchique (les dendogrammes), ou encore de réseaux sémantiques.

Quelque soit le procédé de modélisation mis en place, le volume de connaissance est lié au

volume de données et à leur préparation. Si les données sont représentées avec un seul descripteur symbolique composé de deux modalités, la distinction entre les données documentaire se fera alors en deux groupes, chacun caractérisé par l'une des deux modalités. En augmentant le nombre de descripteurs (numériques ou symboliques) l'expert augmente alors la combinatoire entre enregistrements et descripteurs et donc la finesse pouvant caractériser les données.

L'inconvénient majeur des principes de modélisation repose sur leur capacité à retrouver des motifs en masse et dont la majorité se traduit par des banalités. Ceux-ci serviront certes à classer de nouvelles instances issues du domaine, mais n'apporteront pas de nouveauté pour l'expert. Par exemple, dans `CATMInE` découvrir que si une marque est la copie conforme d'une autre, alors cette marque est contrefaisante n'est pas une conaissance hautement stratégique. Évidement ce cas est extrême mais exprime clairement ce que nous entendons par banalités.

Retrouver ce type de motif est tout de même réconfortant : le système a bien identifié ce pourquoi il a été mis en place et la formalisation des données permet bien l'expression du phénomène. Mais dans de tels procédés, la connaissance la plus motivante pour l'expert est celle relative aux enregistrements stratégiques caractérisant le savoir-faire implicite de l'expert. Il faut donc pouvoir être capable d'identifier les instances caractérisant cette connaissance afin de la retrouver plus facilement parmi l'ensemble produit par le modèle appliqué.

Comme énoncé précédemment, section 2.2 page 29, la qualification de cette connaissance est purement subjective. Deux orientations sont alors possibles : évaluer de façon automatique l'intérêt de la connaissance, ou bien permettre à l'expert d'explorer celle-ci par le biais d'interfaces adaptées autorisant la construction d'heuristiques de recherche en fonction des caractéristiques disponibles.

7.2 Évaluer la qualité des connaissances induites

L'évaluation de la qualité des connaissances permet de déterminer si la modélisation est un succès ou le cas contraire un échec. Cependant la connaissance induite par des procédés d'apprentissage automatiques produit beaucoup d'éléments dont le traitement par un humain peut se révéler ardu. Nous proposons dans cette section de présenter les principes de mesures automatiques puis la façon dont elles peuvent être conduites par un expert.

Il est nécessaire de garder à l'esprit que toutes ces mesures sont purement objectives (elles sont le résultat de calculs numériques) mais que leur définition est supervisée et donc subjective. En effet, ces mesures ont été mises en place afin de mettre en valeur un phénomène précis. De plus, l'objectif de ces mesures est de déterminer ce que nous appelons la connaissance stratégique ou encore pertinente (les termes sont équivalents). Or cette connaissance est qualifiée ainsi par l'expert qui à son tour à une vue subjectives du phénomène (il faut aussi garder à l'esprit que l'expert n'est pas nécessairement omniscient dans son domaine et qu'il peut aussi se tromper).

7.2.1 Évaluation par mesures automatiques

L'évaluation de la qualité des connaissances de façon automatique se traduit par la mise en place de mesures d'intérêt. Ces mesures sont liées à la représentation des connaissances, à leur principe de calcul, à leur portée, et si elles reposent sur des fondements objectifs ou subjectifs. La représentation des connaissances est donc la première des contraintes, et peut être de type règle de classification (issues des arbres de décision par exemple), règles d'association (liées aux procédés d'études des motifs fréquents), structure hiérarchique des concepts ou encore indépendante du format. [Hilderman & Hamilton, 1999] proposent une étude très complète des différentes mesures d'intérêt de la connaissance induite.

Le schéma 7.1 page suivante présente un classement de ces mesures. Celles-ci ne sont pas nommées mais numérotées afin d'en faciliter la lecture. Le tableau 7.1 page 106 permet de faire la correspondance entre l'index proposé par le schéma 7.1 page suivante et la mesure en elle-même.

Les mesures sont classées en fonction de leur catégorie : subjective ou objective. Le classement est ensuite affiné en fonction de la portée de la mesure : applicable à une règle ou à un ensemble de règles. Enfin, les feuilles de l'arborescence proposée reposent sur le type de calcul que représente la mesure : distance, probabilité ou syntactique.

L'ensemble des mesures présentées dans le tableau 7.1 page 106 regroupe l'index de la mesure auquel il est fait référence dans le schéma 7.1 page suivante, leur(s) auteur(s), la publication relative à la mesure et enfin, le type de règles auxquelles la mesure s'applique : classification, association ou de relation généralisée.

L'étude de [Hilderman & Hamilton, 1999] tient compte des différents procédés de classification : motifs fréquents, partitionnement, classification, ou encore sur les procédés de recherche de motifs dans les séries temporelles. Toutefois, ces auteurs ne proposent pas de mesure idéale. Le choix de l'une ou l'autre peut ensuite reposer sur :

- une comparaison des résultats proposés par chacune des mesures de même catégorie
- les attentes d'extraction de la connaissance la plus stratégique pour l'expert.

Dans le premier cas de figure, nous pouvons par exemple, pour un processus de modélisation produisant des règles de classification comme résultat, appliquer les mesures 1, 2, 3 et 4 et croiser les résultats obtenus. Cette comparaison de résultats permet de mettre en valeur la connaissance considérée comme pertinente par ces quatre mesures si il y en a.

Pour le second cas de figure, le choix par l'expert de la mesure lui semblant la plus adaptée, implique que celui-ci doit alors maîtriser chacune de ces mesures et leur procédé de calcul. C'est cette connaissance des parti-pris sur les mesures d'intérêt qui rendra les mesures, et donc la connaissance sélectionnée, plus pertinente qu'une autre.

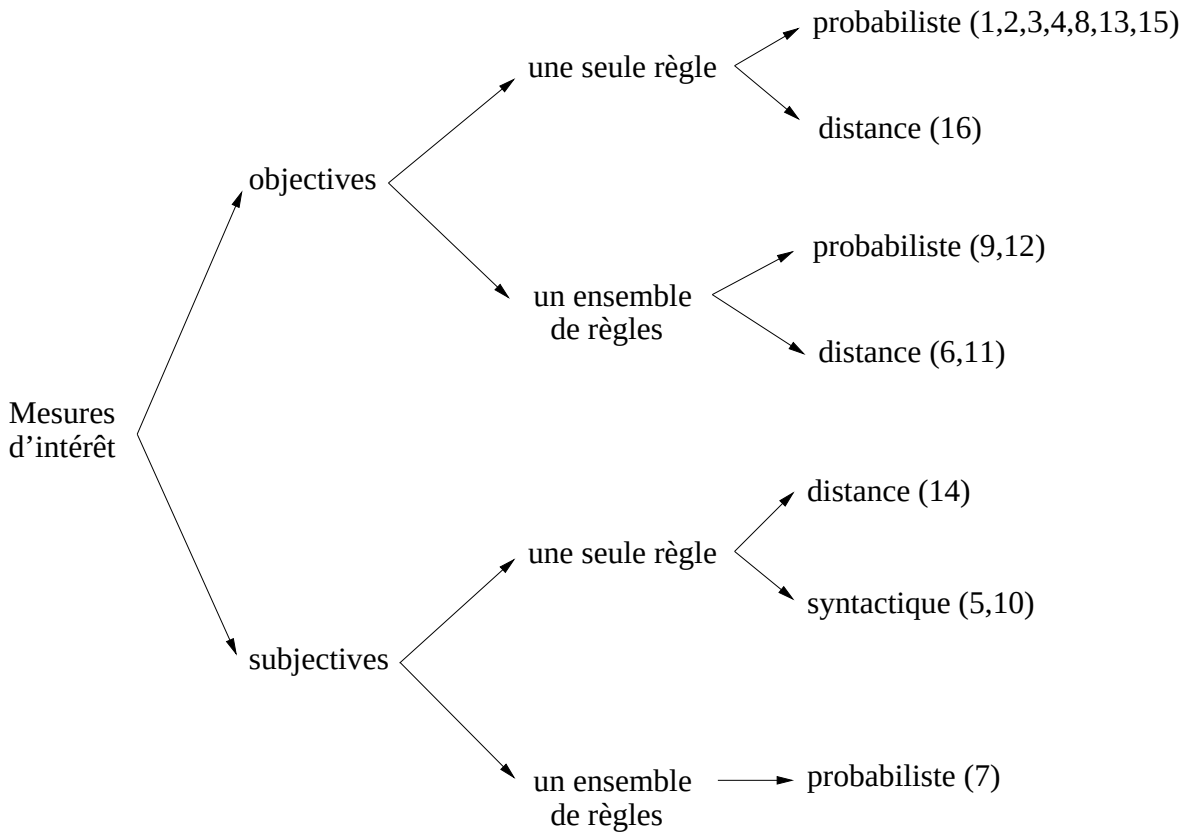


FIG. 7.1 – Classement des différentes mesures d'intérêt de règles caractérisant la connaissance induite par un processus de modélisation

indexe	Mesure d'intérêt	type de règle
1	Mesure Piatesky-Shapiro ([Piatesky-Shapiro, 1991])	classification
2	J-mesure de Smyth et Goodman ([Smyth & Goodman, 1991])	classification
3	Raffinement de Major et Mangano ([Major & Mangano, 1993])	classification
4	Itemset de Agrawal et Srikant ([Agrawal et al., 1993])	association
5	Modèle de Klemettinen ([Klemettinen et al., 1994])	association
6	I-mesure de Hamilton et Fudger ([Hamilton & Fudger, 1995])	relations généralisées
7	Intérêt de Silbershatz et Tuzhilin ([Silberschatz & Tuzhilin, 1995])	indépendant
8	Intérêt de Kamber et Shinghal ([Kamber & Shinghal, 1996])	classification
9	Crédibilité de Hamilton et al. ([Hamilton et al., 1997])	relations généralisées
10	Impressions générales de Liu et al. ([Liu et al., 1997])	classification
11	Distance métrique de Gago et Bento ([Gago & Bentos, 1998])	classification
12	Mesure de surprise de Freita ([Freitas, 1998])	independant
13	Intérêt de Gray et Orlowska ([Gray & Orlowska, 1998])	association
14	L'intérêt de Dong et Li ([Dong & Li, 1998])	association
15	Exception fiable de Liu et al. ([Liu et al., 1999])	association
16	Particularité de Zhong et al. ([Zhong et al., 1999])	association

TAB. 7.1 – Mesures d'intérêts classées dans le schéma 7.1 page précédente et présentées par [Hilderman & Hamilton, 1999].

7.2.2 Évaluation dirigée par l'expert

L'évaluation guidée par l'expert représente le second procédé pour mesurer la pertinence de la connaissance induite. [Sahar, 1999] propose d'orienter la sélection de la connaissance par le désintérêt de l'expert pour certains résultats afin d'orienter la recherche. La solution apportée par cet auteur s'applique aux règles d'associations qu'il classe en 4 catégories :

1. les règles vraies mais inintéressantes
2. les règles fausses mais intéressantes
3. Les règles fausses et inintéressantes
4. les règles vraies et intéressantes

La véracité d'une règle porte sur la sémantique de celle-ci. L'exemple utilisé par [Sahar, 1999] permet de bien saisir les distinctions entre chaque catégorie de règle :

1. Un mari implique d'être marié ➟ vraies mais inintéressantes
2. un homme implique d'être marié ➟ fausses mais intéressantes
3. un homme gagnant plus de 50.000 dollars par an implique d'être marié ➟ fausses et inintéressantes
4. enfin pour la dernière catégorie de règle, l'expert détermine que la règle est intéressante en fonction de son savoir-faire, de ses croyances et de ses attentes.

Bien évidement, cette méthode est facilement extensible à d'autres procédés d'apprentissage produisant des règles de classification (comme les arbres d'induction par exemple).

De cette qualification des règles, [Sahar, 1999] propose ensuite de placer l'expert dans un processus de classement des règles induites en quatre étapes :

1. sélection de la meilleure règle parmi l'ensemble possible par procédés automatiques : combinaison du support, de la confiance et de mesures d'intérêt. Puis recherche de l'ancêtre de la règle (si nécessaire) pour faciliter la compréhension de celle-ci par l'expert
2. interprétation du meilleur candidat par l'expert
3. choix par l'expert de continuer ou non : si la règle couvre peu de règles filles, est-il nécessaire de poursuivre pour éliminer peu de règles
4. tri des règles :
   - si la règle n'est pas retenue, suppression de la règle et de ses descendantes de l'ensemble possible de règles
   - si la règle convient, conservation de celle-ci dans la base de connaissances

Avec, un tel procédé, [Sahar, 1999] démontre qu'après trois itérations de l'algorithme proposé, il est possible d'éliminer 30% des règles d'un ensemble induit par un algorithme produisant des

règles d'association. Au delà de trois itérations, le nombre de règles tend vers un seuil propre à chaque base de données.

Une autre façon de procéder, abordée par [Venturini et al., 1997], consiste à explorer l'espace de recherche en interaction avec l'expert pour construire des requêtes de sélection par optimisations interactives de celles-ci. Ce procédé s'inscrit dans le cadre d'une fouille visuelle de la connaissance et est proche des travaux d'Interaction Homme Machine, sur l'analyse du comportement et des habitudes de l'expert. Ce même groupe d'auteurs s'est aussi penché sur la visualisation en trois dimensions de bases de données afin d'augmenter les interactions possibles entre expert et jeu de données.

Enfin, Une autre façon de procéder consiste à évaluer la surprise que peut avoir l'expert face à certains résultats. Pour cela, [Silberschatz & Tuzhilin, 1996] proposent d'évaluer la connaissance en fonction de deux facteurs :

1. le gain apporté pour traiter la tâche relative
2. la surprise de l'expert à la découverte des connaissances induites.

La notion de gain apporté pour traiter la tâche relative à l'apprentissage est, selon les auteurs, une importante mesure subjective de l'intérêt d'une règle. Cette mesure apporte en effet la notion que l'expert est généralement intéressé par la connaissance lui permettant de mieux pratiquer son domaine d'expertise (en agissant de façon spécifique dessus) en réponse à la nouvelle connaissance induite.

Pour ce qui est de la surprise (ou de l'inattendu), il s'agit de capturer la surprise ressentie par l'expert à la lecture de la règle. Si l'expert est surpris par cette connaissance, c'est que celle-ci vient contredire les croyances de ce dernier sur son domaine. Les croyances contredites sont alors de deux types :

1. croyances fortes : elles ne peuvent être contredites par un seul exemple. L'expert croit généralement à une erreur du procédé de modélisation et du jeu de données. De plus, ce type de croyances est particulièrement subjectif et varie d'un expert à un autre. Toutefois, si un motif contredit ce type de croyances, alors il est toujours intéressant pour l'expert, mais il ne changera en rien ses croyances.
2. croyances faibles : l'expert souhaite les confirmer ou non par des preuves. L'expert a plusieurs degrés de croyances faibles. Si une croyance vient à être changée, alors une croyance plus importante peut être remis en cause.

Pour cela les deux auteurs introduisent une nouvelle mesure, combinant ces deux aspects : utilité et surprise. Les motifs les plus surprenants sont généralement utiles. En créant la surprise, ils révèlent à l'expert un comportement dont il n'a pas conscience. Ils changent alors le comportement de l'expert. La plupart des motifs utiles sont inattendus. En effet, le comportement de

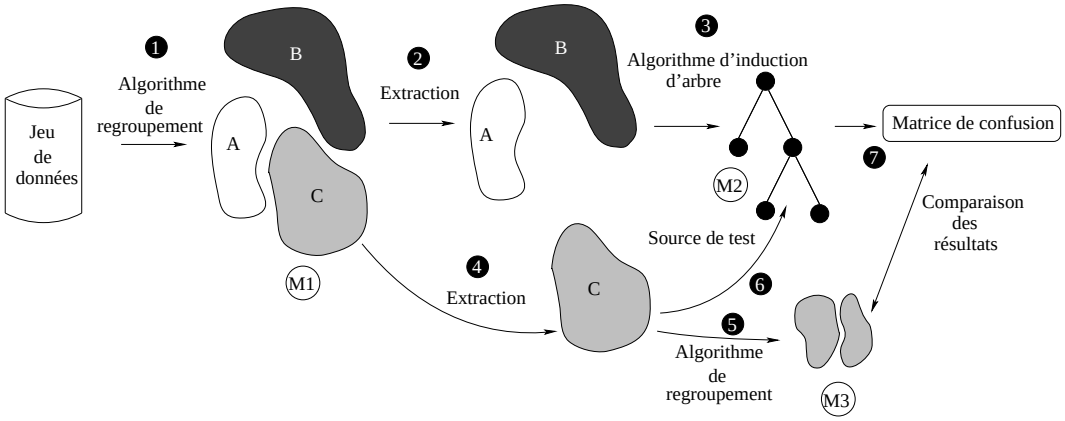


FIG. 7.2 – Processus d'évaluation de la connaissance dans CATMInE.

l'expert est guidé par ses croyances, donc découvrir un motif utile c'est ne pas en avoir conscience, et donc avoir des comportements qui n'incluent pas ces motifs utiles.

Tous ces procédés utilisés pour déterminer la pertinence des motifs font intervenir l'expert. De ce fait, et comme nous l'avons énoncé précédemment, la connaissance d'un expert, ce qu'il juge pertinent ou non, varie d'un expert à l'autre. La notion de mesure est alors fortement subjective et doit impliquer l'omniscience de l'expert.

7.3 Évaluation de la connaissance dans CATMInE

L'évaluation de la connaissance dans CATMInE s'est effectuée par des procédés statistiques d'évaluation des scores de classification et des procédés subjectifs : la pertinence des motifs induits.

L'évaluation objective de la connaissance a été réalisée par la comparaison d'algorithmes non supervisés et d'algorithmes supervisés. Cette évaluation a pour but de confronter les résultats obtenus afin de déterminer si les choix de modélisation par l'expert sont pertinents. Par choix, nous couvrons l'intégralité du processus décisionnel : choix des exemples, définition des descripteurs, choix des descripteurs et enfin choix de l'algorithme de modélisation.

Pour cela nous avons procédé selon le protocole détaillé par la figure 7.2. La démarche consiste à isoler les données pouvant être parfaitement caractérisées par un algorithme de regroupement, et donc un procédé non supervisé. Le nombre de regroupements cherchés correspondant au nombre de classes définies par l'expert. Nous avons alors observé comme résultat la production de deux regroupements purs et d'un regroupement hétérogène. Les deux regroupements purs correspondent bien à des jugements caractérisant la contrefaçon et la non-contrefaçon de marques nominatives. Le dernier regroupement correspond à un mélange des deux. Ce regroupement nous laisse supposer trois hypothèses :

1. les décisions inclues dans ce regroupement caractérisent la connaissance stratégique

2. les décisions sont à considérer comme du bruit : il y a eu erreur dans la sélection de données

3. les descripteurs utilisés ne permettent pas de finement différencier les décisions

Afin de raffiner les informations relatives à ces décisions, nous avons alors procédé à une nouvelle modélisation du domaine. Cette modélisation utilise comme source d'apprentissage les décisions issues des deux regroupements purs et comme source de test le troisième regroupement, qualifié d'hétérogène.

Les résultats ont permis d'isoler un sous-ensemble de seize exemples mettant en avant les faiblesses des descripteurs utilisés pour identifier les décisions. Cela a donc pour conséquence une connaissance que nous pouvons qualifier de peu fiable car inapte à répondre aux contraintes fixées par l'expert. Les détails des résultats de ce protocole de comparaison sont présentés dans la section 10 page 147.

L'évaluation subjective s'est faite par utilisation de l'outil. l'expert a pu expérimenter la notion de surprise présentée par [Silberschatz & Tuzhilin, 1996]. Ainsi, en réalisant une étude d'un nouveau cas, le système a produit une règle allant à l'encontre de ses croyances. La première réaction a été de critiquer le système en argumentant sur le manque de pertinence de la connaissance. Nous avions donc affaire a une croyance forte. La justification de la règle proposée par les exemples la caractérisant a permis de démontrer auprès de l'expert la pertinence de la connaissance et son préjugé face à la décision proposée. Nous pouvons alors corroborer ce que présente [Silberschatz & Tuzhilin, 1996], et nous soulevons l'intérêt de la surprise face à un motif de connaissance pour le qualifier.

Cette évaluation subjective a aussi permis de mettre en avant le crédit qu'apporte l'expert face aux résultats de la plate-forme. Dans l'exemple précédent, l'expert exprime le rejet de la justification à cause de ses croyances sur la question formulée, et remet en doute l'efficacité du système. Cependant, l'étude des résultats de proximité proposés par la plate-forme a permis à l'expert de comprendre la décision de la machine. Cette compréhension aborde sous un autre angle la question soumise à la machine et entraîne l'expert dans un processus d'analyse et de remise en cause de ses croyances. Si les justifications et exemples sont pertinents, et que l'expert les accepte, alors le crédit apporté à la plate-forme est renforcé.

Toutefois, et face à un volume important de connaissances, il faut pouvoir être en mesure de les isoler afin de ne pas solliciter l'expert trop longtemps pour l'évaluation de la connaissance. Nous réfléchissons alors à un procédé permettant de caractériser toutes les croyances dites fortes de l'expert pour les confronter ensuite aux connaissances induites et évaluer leur impact.

7.4 Pour conclure

Un des meilleurs principes pour l'évaluation et la sélection de la connaissance consiste à impliquer profondément l'expert dans la démarche de modélisation. Pour cela le processus de création du modèle et d'extraction de connaissances doit être hautement interactif et explicatif. La construction du modèle doit pourvoir être en mesure d'apporter des éléments de réponse sur les choix retenus.

Par exemple dans les arbres de décision, le choix d'un attribut plutôt qu'un autre peut se faire sur le gain d'information apporté par cet attribut par rapport aux autres (ce choix étant établit avant de construire le modèle et l'expert peut utiliser un autre critère de sélection). Nous savons que dans ce cas de figure le descripteur retenu est celui apportant le plus d'informations, mais il n'est pas possible par exemple d'obtenir un classement de l'ensemble des descripteurs, sur cette mesure à chaque sélection. Cela permettrait alors d'observer qu'un descripteur a été retenu de justesse par rapport à un autre (les gains étant très proches). Bien évidement, pour un ensemble important de descripteurs, il faut pouvoir filtrer et présenter cette information afin de mettre en valeur ceux répondant à ce phénomène. L'expert pourra alors s'interroger plus facilement sur les choix réalisés par l'algorithme de modélisation afin de déterminer les descripteurs critiques (en terme de gain d'information) participant à la construction du modèle. Si la différence est trop forte, la sélection de l'attribut ne peut être remise en cause (le gain étant significativement plus important pour le descripteur retenu que pour les autres). Nous pouvons ensuite proposer à l'expert de changer le choix réalisé par l'algorithme, entraînant la reconstruction du modèle en fonction de cette intervention, afin d'étudier le nouveau modèle.

Cette exemple n'est pas le seul, que cela soit pour un réseau de neurones, un réseau bayésien, ou un tout autre type d'algorithme de modélisation : la présentation des partis pris à l'expert et l'action sur ceux-ci sont plus que nécessaires. Grâce à cette implication, l'expert pourra introduire, tout au long du processus, sa connaissance, son expertise et donc son savoir-faire pour améliorer la précision du modèle. Il sera aussi en mesure de préciser les éléments clefs sur lesquels orienter et agir pour évaluer et extraire la connaissance stratégique.

La connaissance induite est étroitement liée aux interventions de l'expert au cours de la modélisation. Dans la sélection de documents, celui-ci oriente déjà la connaissance induite en se focalisant sur un problème qu'il juge nécessaire d'étudier. Lors de la création des descripteurs, la sémantique mise en place sert à la formalisation de cette connaissance. Enfin, lors de l'évaluation de la connaissance, l'expert saura définir les bons critères d'évaluation de la connaissance en tenant compte des objectifs fixés lors de la mise en place de l'étude et de la modélisation du domaine. Enfin, l'intégration de la connaissance au sein d'un processus décisionnel peut encore faire l'objet d'une intervention de l'expert en introduisant des connaissances ciblées du domaine pour faciliter la gestion et l'application des connaissances produites.

Deuxième partie

CATMInE : une plate forme d'aide à la décision

Chapitre 8

Dispositif de correction d'un processus décisionnel

Sommaire

Le processus de modélisation présenté dans les chapitres précédents implique de faire des choix (partis pris) dans :

- la sélection des données utilisées
- la sélection des descripteurs
- la représentation des descripteurs
- l'algorithme de modélisation

Cependant, il est assez évident de comprendre que de telles démarches de sélection et de représentation peuvent être sources d'erreurs. Même si les choix sont réalisés par l'expert, ce dernier n'est pas omniscient et il est guidé par son expertise, ses croyances, ses intuitions et ses intentions. Celles-ci ne sont pas infaillibles, c'est pourquoi nous proposons de mettre en place une plate-forme d'évaluation et de correction d'un processus décisionnel afin d'aider l'expert à mieux comprendre son savoir-faire. Il est donc nécessaire de pouvoir corriger les différents choix retenus si les résultats obtenus ne correspondent pas aux attentes de l'expert.

Pour répondre à ces besoins, nous présentons dans ce chapitre une démarche interactive et cyclique de correction de l'ensemble du processus de modélisation (8.1 page suivante) admise par la communauté de l'aide à la décision, ainsi que les détails des corrections possibles, étape par étape (8.2 page 118).

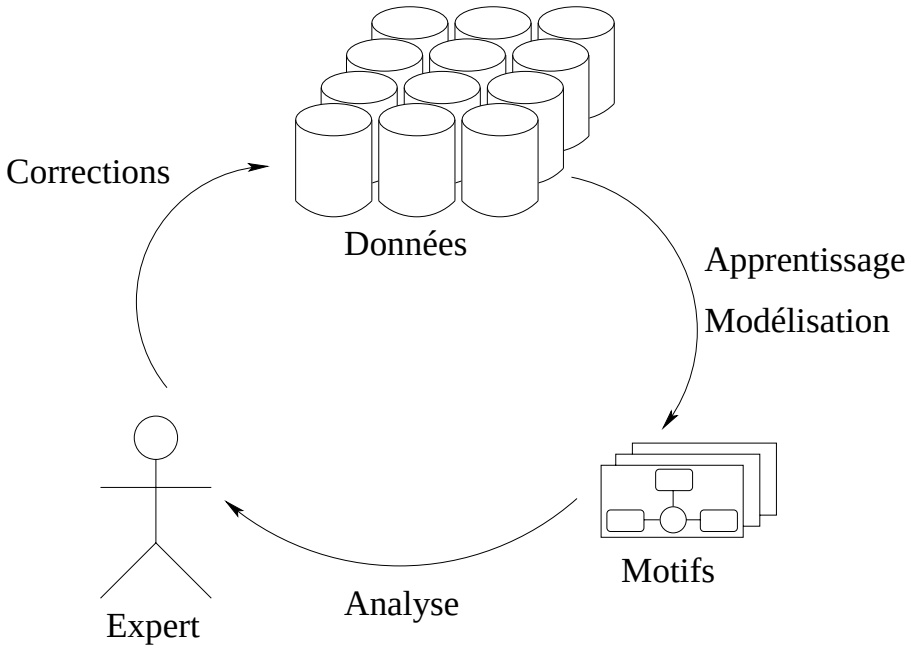


FIG. 8.1 – Boucle de correction ([Fournier, 2001]).

8.1 Correction de modèle : une approche interactive cyclique

L'approche de modélisation présentée dans ce mémoire est linaire : nous passons d'une étape à l'autre uniquement si la première est validée par l'expert, si elle est rendue *robuste*. Cependant, dans ce processus de modélisation, la correction pour l'amélioration des résultats est une tâche charnière. Il existe donc une très forte notion de retour au sein de ce processus linéaire pour en améliorer les résultats. Nombreux sont les auteurs ayant abordés ce point ([Fournier, 2001]), en apportant des solutions sur le processus de corrections. La tendance observée consiste à boucler entre la modélisation du domaine et la mise en place du jeu de données. La figure 8.1 rappelle cette notion d'interaction entre choix techniques pour la modélisation, processus de modélisation et intervention de l'expert pour adapter les corrections. Au cours de ce processus de correction, l'expert agit sur la modélisation du domaine en adaptant la représentation des données. Cela revient concrètement à modifier certains descripteurs en les fusionnant avec d'autres ou en créant de nouveaux descripteurs à partir des anciens. De plus, l'expert peut adapter les modalités des descripteurs, en effectuant un changement d'échelle ou en segmentant les valeurs continues ([Dougherty et al., 1995]) afin d'augmenter leur pouvoir discriminant. Ce processus permet d'explorer plus finement les données tout en profitant des interactions de l'expert. Ces interactions offrent aussi l'avantage d'impliquer plus fortement l'expert dans la modélisation et lui permettent de s'approprier l'application plus facilement tout en garantissant la capture de son savoir-faire. Le protocole de correction s'arrête lorsque l'expert considérera qu'il n'a plus rien à apporter au

modèle.

Cependant, un tel protocole de correction impose la mise en place d'interfaces dédiées à la correction/adaptation du jeu de données si l'on veut pouvoir bénéficier efficacement des interactions avec l'expert. Ces éléments sont étroitement liés au domaine et doivent donc faire l'objet d'une mise au point rigoureuse pour avoir un modèle de correction stable et générique, ou pour avoir un modèle étroitement lié au domaine, offrant un maximum de souplesse en fonction des données.

Quelles que soient les interactions avec l'expert, l'utilisation d'outils de visualisations statistiques reste nécessaire que ce soit pour les résultats produits, les descripteurs utilisés, la couverture de certaines connaissances ou la qualité des données utilisées. Pour autant, il faut que l'expert sache ce qu'il veut et comment il compte l'évaluer. Cela peut par exemple passer par la définition de motifs types correspondant aux attentes du modèle et permettant la comparaison avec l'ensemble des enregistrements présents.

L'expert peut aussi s'appuyer sur les résultats des tests réalisés sur le modèle, observer les exemples non reconnus et comprendre les erreurs. Plus le processus sera interactif (avec pour objectif de capturer un maximum d'interactions avec l'expert), plus les résultats correspondront aux attentes de l'expert. Ces solutions ayant été affinées au fur et à mesure des interactions, elles rendent le processus décisionnel dédié au domaine auquel il est rattaché.

Le danger d'une telle démarche de correction est la simplification de l'expression du domaine. Si le jeu de données n'est pas sélectionné avec beaucoup d'attention, les descripteurs ne permettront pas de bien distinguer les enregistrements. En corrigeant la représentation du domaine, l'expert risque de simplifier l'expression du domaine pour en permettre un meilleur apprentissage. Il est alors délicat de garantir la consistance du modèle. La connaissance induite sera précise mais ne correspondra pas aux attentes initiales : la réécriture des descripteurs entraîne un glissement vers une simplification de la sémantique qu'ils expriment.

Le processus de correction se termine lorsque le modèle induit correspond aux attentes de l'expert. Celles-ci sont de l'ordre de la qualité des connaissances et de l'efficacité de prédiction. Les deux ne sont pas antagonistes, mais il n'est pas évident de les concilier. Par exemple, les réseaux de neurones sont de bons outils de prédiction, mais l'exploitation de la connaissance produite reste délicate. Il est tout à fait possible de déterminer l'importance des descripteurs, mais quant à connaître les modalités exactes ou l'importance d'un descripteur par rapport à un autre, la tâche se révèle plus délicate, comme le présente Filipe Borges dans [Borges et al., 2003]. Dans le même principe, les arbres de décisions permettent une expression des connaissances sous forme de règles plus abordables pour l'expert, mais l'appréhension d'une structure arborescente très complexe peut être une difficulté dans la compréhension du modèle induit par l'algorithme.

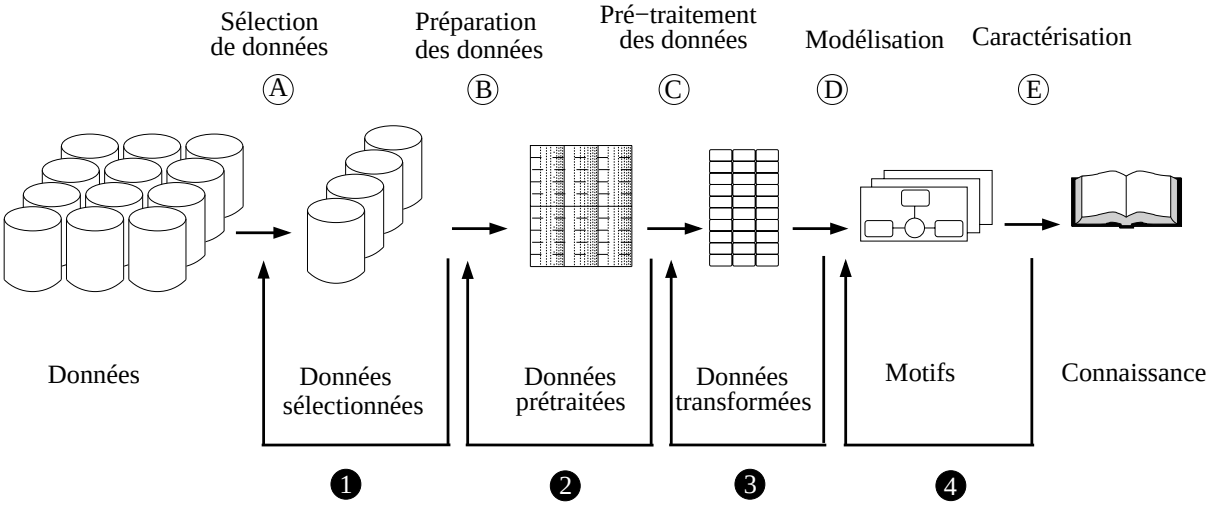


FIG. 8.2 – Principe de correction par retour sur le processus de fouille de données.

8.2 Correction de modèle : étape par étape

Nous venons de présenter dans la section précédente un principe de correction de modèle impliquant l'expert tout au long de ce processus. Ce processus boucle entre : adapter les données, établir un modèle, analyser les motifs de connaissance obtenus et corriger le jeu de données. Cette correction du jeu de données s'applique aux données transformées, résultant du processus de prétraitement des données.

Nous pouvons toutefois étendre la notion de correction/apprentissage abordée précédemment comme le présente la figure 8.2 :

1. de la caractérisation des motifs, adapter l'algorithme de modélisation (4)
2. de l'algorithme de modélisation, modifier les données (3) transformées, et donc le prétraitement de celles-ci
3. des données transformées, revoir la préparation des enregistrements (2)
4. de la préparation des enregistrements, adapter l'échantillonnage (1)

Cette fois, nous définissons plus finement ce que [Fournier, 2001] présente comme cycle de correction, se limitant au changement de paramètres de l'algorithme et à l'adaptation des données. Celui-ci ne se limite pas à corriger la représentation des données et à refaire une modélisation de domaine, mais bien au contraire, à adapter chaque étape d'élaboration du modèle, en nous rapprochant de ce que propose [Crémilleux, 2000]. Ainsi, par correction, nous pointons la correction du jeu de données, sa sélection, préparation et transformation, puis le choix du modèle (et implicitement des paramètres) et enfin la caractérisation des connaissances.

Ce protocole met alors en évidence une complexité abordée tout au long de ce premier chapitre : la multitude de méthodes disponibles pour chaque étape de modélisation d'un phénomène.

Cette complexité est présentée succinctement dans la figure 8.3 page suivante. Ici pour des besoins de lisibilité, nous présentons les méthodes sous leur aspect philosophique. Ainsi, pour la modélisation, lorsque nous faisons référence aux méta-modèles, il faut alors comprendre méta-modèle de type :

- vote
- empilement
- cascade
- boosting
- bagging

Il en va donc de même pour les autres étapes de la modélisation et des procédés présentés. Nous faisons alors référence à tous les chapitres précédents présentant chacune de ces méthodes.

Cela a pour effet d'augmenter considérablement l'espace de recherche d'une modélisation optimale. Cet espace de recherche est défini par la combinaison de l'intégralité des méthodes disponibles pour chaque étape du processus de modélisation. Il faut alors pouvoir assister l'expert dans cette démarche. Ainsi dans notre schéma, nous présentons les choix réalisés dans le projet `CATMInE` pour la première modélisation :

1. sélection de données par expertise
2. gestion du bruit
3. fusion de descripteurs
4. modélisation numériques : réseau de neurones
5. caractérisation de la connaissance par mesure automatique : score de classification

Une autre remarque sur cet espace de recherche : les étapes relatives au traitement des descripteurs s'appliquent autant de fois qu'il y a de descripteurs.

En pratiquant ensuite le protocole de correction présenté dans la figure 8.1 page 116, nous avons obtenu une nouvelle version de `CATMInE` : `DeTTMInE`. Cette fois, les choix sont les suivants :

1. sélection de données semi-automatique, validée par expertise
2. gestion du bruit et des valeurs manquantes
3. fusion de descripteurs
4. modélisation automatique : arbre de décision
5. caractérisation de la connaissance par mesure automatique : score de classification, mesure de qualité de nœud

Nous reviendrons sur cette correction de `CATMInE` dans la suite de ce mémoire. En revanche nous avançons qu'un bureau virtuel permettant de réaliser avec le plus d'interactivité possible l'ensemble des étapes de modélisation est nécessaire, la section 11.2 page 171 en présente le

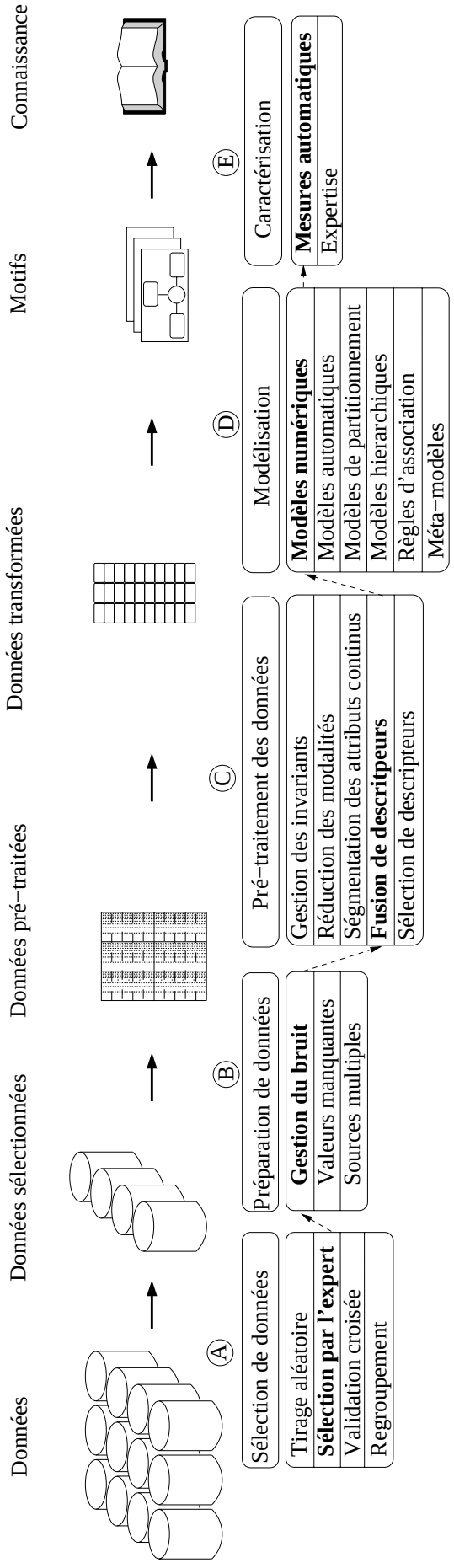


FIG. 8.3 – Complexité du processus KDD.

principe. Il peut ensuite être utile d'étudier le comportement des experts face à une telle panoplie d'outils afin de déterminer les méthodes les plus pertinentes, voire de proposer à tout expert cherchant à modéliser un domaine une suggestion des outils en fonction des choix réalisés. Par exemple, si l'apprentissage détermine qu'un ensemble d'experts préfère appliquer un arbre de décision lorsqu'ils ont segmenté les valeurs continues présentes dans leur base de données, alors nous pourrons être en mesure de le suggérer à tout nouvel expert ayant pratiqué le même prétraitement de données et ce avant qu'il ne choisisse un algorithme de modélisation.

8.3 Observations

Usuellement, la base de données brute n'est que très rarement remise en cause. Seule la base raffinée, où les descripteurs ont été normalisés, est sujette à corrections. Cela s'explique par le fait que la base brute est généralement fournie par l'expert du domaine, dont nous ne remettons pas en question l'intégrité. Toutefois, si cette base n'a pas été établie en collaboration avec des experts en apprentissage, celle-ci ne caractérise peut être pas le domaine et donc ne garantit pas la qualité d'un traitement d'extraction de connaissances. Le monde industriel a tendance à fournir à la communauté informatique une masse de documents que ceux-ci doivent admettre comme corrects car fournis par un expert du domaine.

Autre aspect important, nous nous sommes vite aperçu que quelque soit l'algorithme appliqué pour modéliser le jeu de données, les résultats varient peu. En revanche, changer le jeu de données fait varier considérablement les modèles (et implicitement leur pouvoir prédictif). Nous pensons alors que dans un processus de correction, la source de données est bien plus importante que l'algorithme utilisé et que l'accent doit être porté sur celle-ci pour améliorer la modélisation d'un domaine.

Dans notre contexte de travail, le document électronique juridique, l'extraction d'informations et la représentation des documents sont les principaux freins au bon fonctionnement des algorithmes de modélisation. Tout au long de ces travaux de recherche nous avons émis les hypothèses selon lesquelles :

1. le processus de modélisation était inadapté, et nous avons démontré par l'expérience le contraire.
2. le jeu de données était mal caractérisé, nous avons adapté les descripteurs, et démontré l'apport d'une telle démarche.
3. les données n'étaient pas pertinentes, nous avons remis en cause la base de données, développé des outils d'extraction et mis en place une nouvelle base, sans apport significatif, mais avec un jeu de données cette fois garanti.

Nous avons étudié en profondeur les outils de modélisation et de traitement des données car

nous n'avons pas cherché à remettre en cause l'expertise utilisée. Nous pensons que celle-ci était incomplète. Des informations supplémentaires sont nécessaires afin de répondre au problème posé : la contrefaçon de marques nominatives. Nous avons alors orienté ces travaux de recherche sur la manière de permettre à l'expert de s'approprier les outils d'aide à la décision. Pour cela nous nous sommes inspiré du procédé de correction d'un processus de modélisation que nous présentons dans la section 8.1 page 116.

Nous soulevons alors trois hypothèses :

1. remettre les descripteurs en question
2. remettre les données en question
3. remettre les documents en question

Nous proposons alors un découpage en trois parties de ce processus de correction avec un retour en arrière afin de satisfaire les trois hypothèses précédentes :

1. la modélisation et l'évaluation des résultats : remise en question des descripteurs et de leur pouvoir discriminant.
2. la transformation des données : remise en question des données utilisées pour la modélisation.
3. la sélection et la préparation des données : remise en question des documents et de leur interprétation.

La première partie n'est faisable que si des différences notables sont observées lors de la comparaison de différents algorithmes de modélisation. À cet effet, l'expert pourra alors influer sur les paramètres des modèles les plus pertinents en terme de qualité d'apprentissage.

La seconde partie, la transformation des données, est quant à elle particulièrement importante. C'est elle qui va permettre d'exprimer le phénomène à partir des descripteurs proposés par l'expert. Ainsi, en fonction des calculs appliqués sur les descripteurs, l'expert va pouvoir mettre l'accent sur un aspect particulier exprimé dans les données.

Enfin, la troisième partie n'est pas nécessairement cyclique. Une fois les données sélectionnées par l'expert, nous pouvons tout de même penser que celles-ci correspondent au domaine d'étude. Cependant, l'expérience de `CATMInE` nous a démontré le contraire : le premier jeu de données fourni par l'expert contenait des décisions pertinentes au domaine. En revanche, l'information retenue pour constituer les descripteurs caractérisant ces décisions était erronée (nom de la partie à la place du nom de la marque par exemple). Pouvoir visualiser à nouveau les documents sélectionnés ainsi que l'information retenue peut être utile afin que l'expert se rassure sur les données utilisées et puisse écarter définitivement la sélection et la préparation de données comme sources potentielles d'erreurs. Le retour sur cette étape peut être cyclique, mais nous pouvons considérer qu'un seul retour suffit afin de garantir la qualité des données.

8.4 Complexité des choix conceptuels dans un processus décisionnel

Dans les chapitres précédents nous avons présenté l'intégralité du protocole permettant d'établir un processus décisionnel. Nous avons relié ce protocole à notre problématique relative aux documents électroniques juridiques, et à notre domaine d'application précis : la contrefaçon de marques nominatives.

Nous avons ensuite démontré le principe d'une approche cyclique de corrections afin d'établir le plus finement possible un modèle caractéristique du domaine d'étude, et justifié cette approche relativement à l'ensemble des méthodes disponibles.

Afin d'avoir une vision plus réaliste des difficultés relatives à l'ensemble des méthodes disponibles, la figure 8.4 page suivante présente une vision plus précise de l'ensemble des méthodes disponibles pour chaque étape du protocole de modélisation où chaque option est un nœud et un ensemble de nœuds représente un chemin. Ce chemin caractérise un processus avec une qualité qui lui est propre.

En plus de devoir établir un parcours à travers toutes ces méthodes en sélectionnant une approche par étape, il y a aussi pour certaines étapes la possibilité de combiner plusieurs des approches disponibles dans le traitement de l'information. Par exemple dans la transformation des données, la gestion du bruit est complémentaire des valeurs manquantes (l'un n'empêchant pas l'autre). Chaque décision, chaque parti-pris, a des conséquences sur les étapes situées en aval. De plus, face à l'ensemble des choix possibles, à l'ensemble des méthodes disponibles permettant de traiter une même problématique, il est très délicat d'établir un traitement optimal. De ce constat, nous observons alors une complexité conceptuelle encore accrue de l'approche de modélisation. Cette compléxité ne permet pas à l'expert de sélectionner avec beaucoup de sûreté les outils les plus adaptés pour répondre à ses besoins sans l'utilisation d'une plate-forme adaptée. cette plate-forme doit favoriser les interactions avec l'expert tout au long de la construction du processus décisionnel.

A ce constat de complexité conceptuelle vient s'ajouter la non connaissance d'un chemin idéal en fonction des choix réalisés à chaque étape. Vouloir déterminer les erreurs en fonction des choix réalisés en amont du processus de modélisation ne peut être mené efficacement sans une approche exploratoire. Cette approche exploratoire permet de lever la difficulté et ne peut être mise en évidence que dans l'approche cyclique de correction.

Nous allons présenter dans le chapitre suivant (chapitre 9 page 127), au travers d'une approche pragmatique, les choix que nous avons retenus dans notre cadre applicatif `CATMInE`. Nous allons présenter la façon dont nous avons cherché à évaluer la validité de l'algorithme de modélisation. À partir des observations réalisées sur la validité du processus de modélisation, nous nous sommes intéressé à l'étude de la validité de la transformation des données appliquée dans

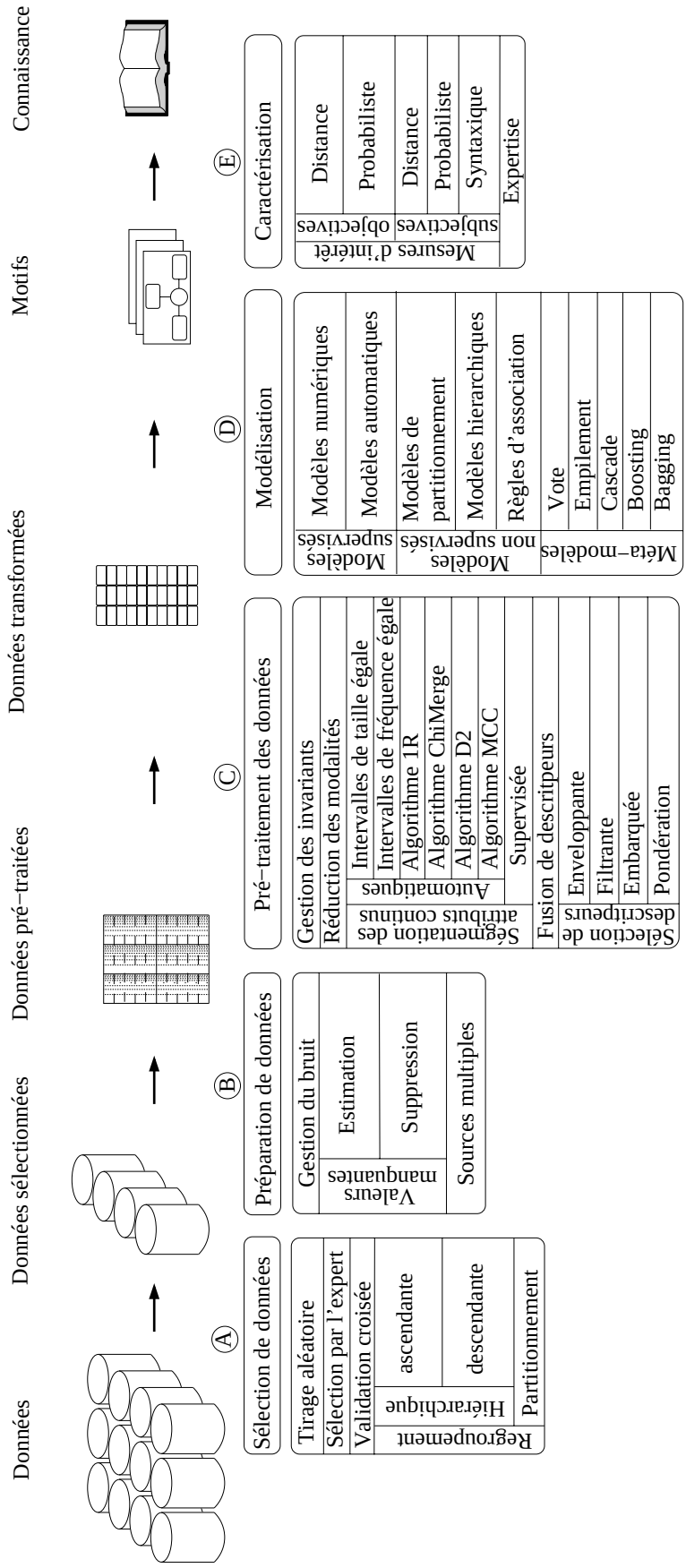


FIG. 8.4 – Complexité détaillée du processus KDD.

`CATMInE` (section 10 page 147). Cette étude nous permet d'identifier les enregistrements du jeu de données posant des difficultés pour la modélisation, et les conséquences qui peuvent être déduites de ces données : connaissance stratégique, erreurs de modélisation ou bruit.

Enfin, nous terminerons cette partie sur la nécessité d'établir de nouveaux chemins dans ce protocole de modélisation afin d'optimiser la modélisation et d'obtenir une meilleure qualité du modèle.

Chapitre 9

Étude de la validité du modèle

Sommaire

Dans `CATMInE`, le processus décisionnel proposé repose sur la mise en place d'un réseau de neurones[17]. Ce type d'application ne permet pas de mesurer efficacement la qualité des connaissances collectées au cours du processus d'apprentissage. Les connaissances correspondent à une fonction mathématico-logique simplifiée intégrant des heuristiques pour laquelle il est délicat de mesurer la pertinence. Une façon de procéder, commune à tout processus d'apprentissage, est

[17] L'algorithme a été spécifié dans la définition du contrat de recherche et développement établi en 2001 entre l'Université de Caen par l'intermédiaire de Thomas Lebarbé et le cabinet Breese Derambure Majerowicz

l'étude des résultats de classification sur un ensemble d'enregistrements de test. Nous avons complété le système d'aide à la décision par un procédé d'étude de proximité géométrique proposé permettant de déterminer des documents similaires pour l'expert afin de justifier la décision du réseau de neurones. Ce procédé étant complètement indépendant du processus de modélisation, il ne peut être utilisé en l'état afin d'évaluer la qualité des résultats du réseau mais pour les appréhender.

Nous nous limiterons donc à l'étude du score de classification du réseau de neurones pour déterminer si ce dernier modélise de façon adéquate la problématique étudiée.

Afin d'être rigoureux avec le protocole de correction présenté dans le chapitre 8 page 115, nous comparerons la décision produite par `CATMInE` à d'autres procédés de classification (cf. section 9.1). L'idée sous-jacente étant de s'assurer de la pertinence des décisions produites par `CATMInE`. Par pertinence, nous voulons surtout nous assurer que les modèles établis sont équivalents à d'autres, produits par d'autres algorithmes. Évidement, si d'autres algorithmes produisent des résultats nettement supérieurs à ceux obtenus avec `CATMInE`, alors il faudra remettre en question les choix établis pour le processus de modélisation. Inversement, si les autres modèles suivent les résultats observés pour `CATMInE`, alors l'algorithme de décision peut être considéré comme pertinent.

Nous comparons dans la section 9.2 page 130 l'algorithme de `CATMInE` à un ensemble d'algorithmes et vérifions si un algorithme adapté au domaine peut améliorer le modèle en comparant ses résultats à ceux précédemment obtenus avec les différents algorithmes. Nous validerons ainsi cette étape de modélisation en adaptant les partis pris de la construction du modèle. Cette adaptation consiste à prendre en compte des spécificités juridiques pour la construction du modèle.

9.1 Plate-forme d'expérimentation

L'objectif de ces expériences est de comparer le réseau de neurones mis en place auparavant aux résultats de `C4.5`, d'un réseau bayésien, d'un algorithme de `bagging` et d'un de `boosting`. Pour mener à bien ce protocole de comparaison, nous avons utilisé la plate-forme `Weka` [Witten & Eibe, 2005]. Celle-ci permet d'appliquer différents algorithmes d'apprentissage sur un même jeu de données afin d'en étudier les comportements. Cette plate-forme a deux modes d'utilisation : exploration de données et expérimentations. Chacun de ces deux modes dispose d'interfaces graphiques intuitives pour leur utilisation. Le mode d'exploration d'un jeu de données fournit un accès simple à toutes les méthodes de pré-traitement des données, d'apprentissage, de sélection d'attributs et de modules de visualisation dans un environnement adapté à l'exploration des données. Le second mode, dit d'expérimentation, permet de mener des études d'envergures sur des jeux de données. Les résultats sont conservés au sein d'une base de données pour un accès et des analyses adaptés.

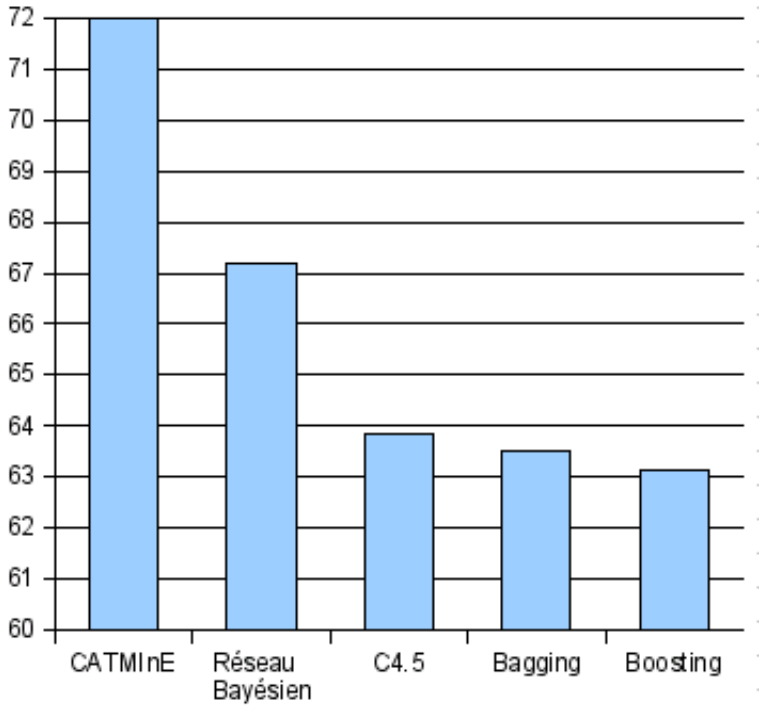


FIG. 9.1 – Taux de réussite sur une validation croisée

Chacun des algorithmes présentés a été utilisé avec ses paramètres par défaut. Une validation croisée (de 10 paquets) ([J.Petrak, 2000]) a été appliquée pour déterminer une moyenne des scores de classification. Ceux-ci sont présentés dans le graphique 9.1.

De manière générale, le réseau de neurones utilisé dans `CATMInE` semble être légèrement plus efficace (de 5-9%) que le principe du réseau bayésien ou d'arbre de décision. Les 4 algorithmes supplémentaires sont, quand à eux, à peu prés équivalents en terme de résultats.

Le tableau 9.1 page suivante présente le score et la déviation standard pour chacun des paquets de la validation croisée effectuée précédemment (par algorithme étudié).

Indépendamment des performances, les algorithmes se comportent tous de la même manière face aux différents jeux de données utilisés pour les mesurer. Nous observons par exemple que le score le plus médiocre est obtenu sur le sixième jeu de données. Tous les algorithmes testés suivent le même comportement, à savoir le score le plus bas de toute la série. Le score le plus élevé est obtenu sur le huitième jeu de test. De manière générale, quelque soit l'algorithme utilisés, les résultats sont similaires (les différences sont réalisées sur 1 ou 2 exemples de mieux reconnus). Le processus de modélisation peut alors être écarté des processus nécessitant une analyse approfondie. Ces résultats semblent indiquer que le jeu de données utilisé pour la modélisation n'est pas adapté pour établir des motifs de classification permettant une aide à la décision optimale.

Des 4 algorithmes mis en concurrence avec le réseau de neurones mis en place pour répondre

Jeu de données	CATMInE	Réseau bayésien	C4.5	Bagging	Boosting
marques 1	73.27 (±13.00)	67.27 (±10.67)	66.36 (±9.63)	66.36 (±8.62)	64.55 (±9.04)
marques 2	74.18 (±15.00)	68.18 (±7.73)	59.09 (±7.73)	64.45 (±11.97)	61.82 (±13.42)
marques 3	70.55 (±10.88)	69.09 (±14.97)	66.36 (±9.63)	63.55 (±10.00)	64.55 (±10.00)
marques 4	73.27 (±7.67)	66.36 (±12.16)	66.36 (±9.63)	65.45 (±11.97)	61.82 (±9.39)
marques 5	74.00 (±11.38)	70.00 (±6.14)	65.45 (±9.39)	63.55 (±12.46)	62.73 (±12.46)
marques 6	62.73 (±18.97)	58.00 (±12.29)	50.00 (±12.47)	55.00 (±16.50)	57.00 (±12.52)
marques 7	71.00 (±17.80)	61.00 (±11.97)	65.00 (±11.35)	60.00 (±15.63)	63.00 (±16.36)
marques 8	76.00 (±10.33)	75.00 (±12.69)	71.00 (±11.01)	71.00 (±12.29)	72.00 (±16.87)
marques 9	72.00 (±12.65)	65.00 (±25.06)	66.00 (±19.55)	62.00 (±21.11)	61.00 (±24.70)
marques 10	73.00 (±14.18)	72.00 (±13.17)	63.00 (±16.36)	64.00 (±20.66)	63.00 (±18.29)
moyenne	72 (±13.18)	67.19(±19.91)	63.86 (±11.67)	63.53 (±14.31)	63.14 (±14.31)

TAB. 9.1 – Détails de la validation croisée (entre parenthèses, la déviation standard)

aux besoins de `CATMInE`, seuls les réseaux bayésiens sont véritablement comparables. En effet, sur 3 paquets de validation (série 3, 8 et 10) les réseaux bayésiens sont à 1 point du réseau de neurones. Pour les autres paquets, le réseau de neurones est le plus efficace.

En revanche, le réseau de neurones est l'algorithme présentant les plus forts taux de déviation standard (quatre paquets de validation : 1, 2, 6, 7 sur les 10), laissant envisager une répartition des données plutôt large.

A la vue de ces résultats, l'algorithme de décision n'est pas à incriminer dans la qualité des modèles induis : les algorithmes de modélisation évalués sur le jeu de données ont des résultats et des comportements similaires. En revanche, la représentation du domaine (enregistrements et descripteurs) nécessite un retour correctif afin de les optimiser pour renforcer le pouvoir discriminatoire des descripteurs. Cependant, avant d'en arriver à ce retour, nous vérifions que la mise en place d'un algorithme adapté au domaine modélisé ne pouvait faire mieux. Nous proposons alors un algorithme capable de prendre en compte dans son processus de modélisation et de reconnaissance des spécificités du domaine juridique.

9.2 DeTTMInE : une approche adaptée au domaine juridique

Decision **T**ree for **T**rade **M**arks **In**fringement **E**valuation

9.2.1 Les motivations

Dans la section précédente nous avons comparé l'algorithme utilisé (un réseau de neurones) à 4 autres types d'algorithmes (issus de différents courant de l'apprentissage). Les résultats obtenus ont démontré que quelque soit l'algorithme utilisé, le modèle ne serait pas optimal. Nous

proposons alors de vérifier si un algorithme adoptant un paradigme de modélisation relatif au domaine ne serait pas plus efficace. Ce comportement consiste à prendre en compte l'atypicité des données juridiques utilisées, moyennant l'ajout d'un parti pris dans l'algorithme de modélisation. De plus, le pouvoir explicatif des décisions doit aussi être pris en compte afin d'assister plus finement l'expert en lui apportant des éléments de réponse concernant l'ensemble des décisions de références proposées.

9.2.2 L'intelligibilité des résultats

L'intelligibilité est le second point fort du procédé de classification à mettre en place. Dans `CATMInE`, l'aide à la décision est assurée par un réseau de neurones. Celui-ci étant opaque dans ses calculs, nous ne pouvons pas connaître ou déduire les caractéristiques qui ont mené aux résultats proposés. L'une des motivations de ce travail est de pouvoir apporter au juriste des éléments pour le guider dans son étude.

9.2.3 La gestion des cours décisionnelles, parti pris de la modélisation

Dans notre cas d'étude, un jugement peut être remis en cause si l'une des deux parties n'est pas satisfaite du jugement. Ainsi, lorsqu'un procès a lieu pour la première fois, le *Tribunal de Grande Instance* (TGI) observe les faits et la loi et rend un jugement. La partie lésée, mécontente ou insatisfaite par le verdict a alors la possibilité de porter le jugement devant la *Cour d'appel* (CA). Si cette dernière invalide le jugement rendu par le TGI, alors celui-ci n'est plus valable, et seul le jugement de la Cour d'Appel prévaut. De nouveau, si l'une des parties n'est toujours pas satisfaite par le jugement rendu en Cour d'Appel, il lui reste la possibilité de porter l'affaire en *Cour de Cassation* (CCASS) qui elle juge les points de droit et non les faits. Pour les mêmes raisons que la Cour d'Appel, si la Cour de Cassation infirme un point de droit relatif au jugement rendu en appel, l'affaire est alors renvoyée en Cour d'Appel.

Un autre aspect de cette atypicité des données est relatif au temps. Soit un jugement J_1 rendu à une date D_1 pour une décision V_1. Si il existe un jugement J_2, rendu à une date D_2 tel que $D_1 < D_2$ alors, le jugement J_2 fait jurisprudence sur le jugement J_1. Cet aspect temporel caractérise le courant de pensée juridique et de ceux rendant les décisions, avec la volonté d'être le plus fidèle au courant de pensée de la société. Il est de plus complètement en relation avec la notion de niveaux décisionnels : seule la cour d'Appel (et évidement Cour de Cassation) pourra contredire une décision de niveau équivalent pour une décision antérieure.

9.2.4 Les arbres de décisions et la qualité de nœuds

L'hypothèse de travail retenue consiste à classifier la majorité des exemples avec des règles simples et les exemples difficiles à associer à l'aide de règles plus complexes. L'algorithme retenu

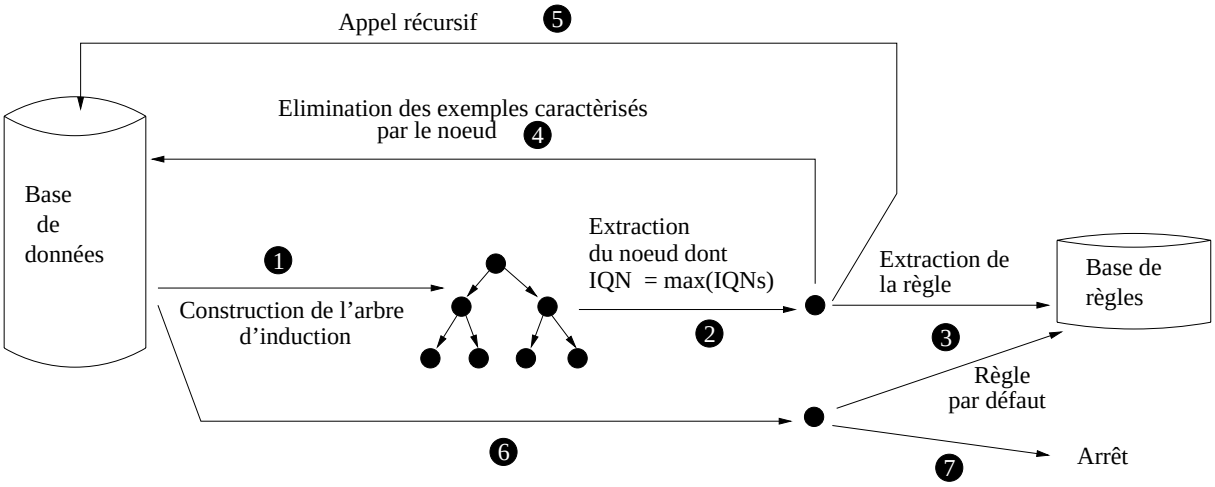


FIG. 9.2 – Principe récursif de DeTTMInE

pour produire ces règles est de type arbre de décision. L'intégration de mesures de qualité de nœud dans les arbres de décision [Fournier, 2001], permet de développer cette hypothèse. Cette qualité de nœud (`IQN`) repose sur la conjonction de ces trois notions :

- le poids du nœud dans l'arbre (représentativité du nœud)
- la mesure d'impureté, normalisée entre 0 et 1
- une fonction d'amortissement dépendant de la profondeur du nœud dans l'arbre

Cette mesure qualitative a été introduite dans le but d'évaluer les modèles induits en les étudiant localement (nœud, sous-arbre) ou globalement (arbre). À partir de la mesure locale fournie par cet indice de qualité de nœud, nous pouvons alors réaliser notre sélection de descripteur selon les contraintes évoquées ci-dessus. Le principe de l'algorithme est détaillé dans la figure 9.2.

Dans un premier temps, le processus décisionnel produit un arbre de décision à partir de la base de données (étape 1). Ensuite, l'algorithme cherche le nœud de qualité la plus élevée (étape 2) parmi l'ensemble des nœuds composant l'arbre. Cela traduit alors le fait que le nœud retenu correspond à beaucoup d'enregistrements (première propriété de l'indice de qualité de nœud). Une règle de classification en est extraite (étape 3). Cette sélection de nœud implique une notion de hiérarchie entre ceux-ci (le premier étant plus important que le second, etc...).

Les exemples associés à ce nœud sont alors effacés du jeu d'apprentissage (étape 4), après avoir déterminé le niveau décisionnel le plus fort présent parmi ces exemples. Cette information vient alors enrichir la règle de classification produite, offrant un moyen de pondération de la règle. Le procédé est appliqué de nouveau sur le reste du jeu de données (étape 5). Lorsque le procédé de modélisation produit un arbre se restreignant à une racine (étape 6), une règle dite *par défaut* est insérée dans l'ensemble des règles extraites (étape 3) et le procédé s'arrête (étape 7). Cette règle sera utilisée dans le processus décisionnel lorsque aucune autre règle ne pourra

FIG. 9.3 – Production d'une règle après sélection du meilleur nœud.

s'appliquer.

9.2.5 Principe prédictif

Ce procédé, novateur par son utilisation de l'indice de qualité de nœud permet, comme nous l'avons présenté précédemment, la mise en place d'une méthode de sélection de descripteurs, à partir d'arbres de décision. Celle-ci, appliquée de manière récursive, entraîne à chaque appel l'extraction d'un nœud, représentant la terminaison d'une branche. La figure 9.3 présente la transformation de cette branche en règle. La branche est formée par le chemin de nœuds allant de la racine de l'arbre au nœud sélectionné. L'ensemble des nœuds rattachés à cette branche représente des descripteurs auxquels il est associé un test pour une valuation particulière (valuation appartenant au domaine du descripteur). L'ensemble de ces tests représente alors une règle caractérisant le nœud sélectionné. De plus, l'ordre hiérarchique évoqué précédemment (et imposé par la notion d'indice de qualité de nœud) affecte directement les règles qui doivent le respecter. Cet ordonnancement intervient pour la reconnaissance d'un nouvel exemple : il faut appliquer à l'exemple la première règle trouvée qui le caractérise. Cet aspect étant motivé par la notion de hiérarchie de nœuds.

Pour mieux comprendre, si la première règle caractérise un exemple et si une ou plusieurs autres règles le caractérisent aussi, alors la première (correspondant au premier nœud extrait qui à le plus fort `IQN`, donc le plus représentatif des données) est considérée comme la plus représentative de toutes celles applicables. Il en va de même si la première règle applicable est la quatrième règle induite, et qu'il en existe d'autres après celle-ci (dont la création est postérieure à la quatrième).

9.2.6 Adaptation des règles au domaine juridique

Depuis le début de la présentation des spécifications de `DeTTMInE`, le caractère atypique des données a été posé et expliqué en détails. Ce phénomène ne peut pas être utilisé pour

l'apprentissage[18], mais il est nécessaire de le prendre en compte pour la reconnaissance. Pour y répondre, la méthode consiste à utiliser la cour décisionnelle pour inclure un point de sélection supplémentaire entre les règles lors de la recherche de la règle optimale pour un nouvel exemple.

Pour réaliser ce traitement, le prototype développé permet, via un paramètre d'appel, l'utilisation ou non de la cour décisionnelle et de la population associée.

L'intérêt d'un tel paramètre d'appel porte sur l'étude d'autres bases de données que celle utilisée dans ces travaux en se comportant alors comme un arbre de décision classique. Cela permet ainsi d'évaluer le prototype de création d'arbre par indice de qualité de nœud à d'autres algorithmes et de mesurer l'intérêt d'une telle gestion des règles induites. Ensuite, connaissant les performances du prototype, nous pouvons alors déterminer si la gestion des règles, via le procédé de gestion des cours décisionnelles, est une méthode apportant une valeur ajoutée au prototype.

Si le paramètre de gestion des cours décisionnelles est utilisé, alors le système va déterminer le meilleur nœud (présentant le plus fort `IQN`) puis il détermine le niveau juridique le plus haut atteint pour l'ensemble des exemples associés au nœud sélectionné. En plus de ce niveau, le nombre total d'exemples associés au niveau retenu est aussi calculé. Ces deux informations sont alors rajoutées à la fin des règles produites et entreront dans un processus décisionnel de sélection de règles pour la reconnaissance. Ce processus est expliqué dans la sous-section suivante.

9.2.7 La gestion des règles de classification par le niveau décisionnel

La gestion des règles découle directement des informations rajoutées (niveau décisionnel et population) pour répondre au problème de sélection intelligente de règle. Tout d'abord, dans la section 9.2 page 130, l'ordre hiérarchique des règles a été expliqué et est issu de la sélection de descripteurs, sélection obtenue par le biais de l'indice de qualité de nœud. Ici, l'ajout à la règle du rang de la cour décisionnelle et du nombre d'exemples associés est à mettre en valeur. Cette gestion consiste à déterminer l'ensemble des règles applicables pour un exemple. La figure 9.4 page suivante en présente les mécanismes, qui sont les suivants :

1. le système recherche l'ensemble des règles applicables (étape 1)
2. le système isole la règle de plus haut niveau juridique, dans cet ensemble (étape 2)
3. si plusieurs règles sont de même niveau juridique, alors le système repose son choix sur le nombre d'exemples associés à cette règle pour ce niveau (étape 3)
4. si, de nouveau, il existe plusieurs règles applicables à niveau égal et à population d'exemples égale, alors le système va rechercher la règle la plus longue (étape 4)
5. enfin, s'il existe encore plusieurs règles applicables, le système prendra alors la première (étape 5)

[18] Le niveau décisionnel final ne peut être déterminé à l'avance

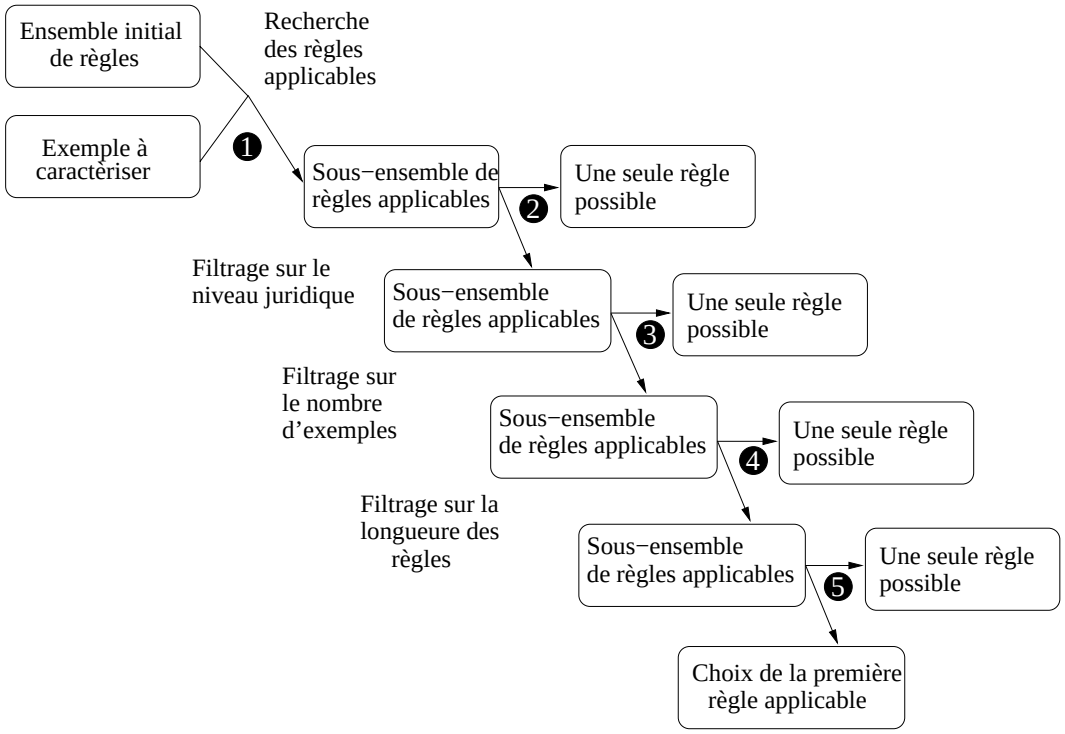


FIG. 9.4 – Procédé de gestion des règles.

Le principe de cette sélection de règle est double, il prend en compte :

1. les contraintes du domaine
2. les propriétés liées à l'indice de qualité de nœud

Les contraintes du domaine sont caractérisées par le niveau décisionnel et la population. Le niveau décisionnel rappelle que les décisions rendues en Cour de Cassation sont plus importantes que celles rendues en Cour d'Appel, elles-mêmes plus importantes que celles issues d'un Tribunal de Grande Instance. La population caractérisant une règle est liée à la prise en compte de la jurisprudence, qui tend à uniformiser les décisions futures en se rapportant à des décisions antérieures. Plus il y a de décisions pour une règle, et plus le motif de ces décisions a été admis par les experts pour l'ensemble des décisions partageant les mêmes propriétés.

Lorsqu'il n'est pas possible d'obtenir une règle unique pour identifier une nouvelle décision en appliquant les contraintes du domaine, le système choisit alors celle utilisant le plus de descripteurs pour expliquer la décision. Nous utilisons alors le fait que plus la règle est précise, et plus cette règle apporte de l'information pour comprendre la décision.

Enfin, s'il n'est toujours pas possible d'obtenir une règle à partir des sélections précédentes, nous retenons la première règle du sous-ensemble obtenu. Nous appliquons alors les propriétés issues de la sélection des règles et impliquant les principes d'indice de qualité de nœud comme

Nom	désignation	instances	classes	attributs
Balance Scale Database	bal	625	3	4
Breast Cancer Database	breast	286	2	9
Credit Screening Databases	crx	653	2	15
Solar Flare Databases	flare	1066	7	12
Housing Database (Boston)	housing	506	2	12
Lymphography Database	lympho	148	4	19
Pima Indians Diabetes Database	pima	768	3	8
Zoo Database	zoo	101	2	17

TAB. 9.2 – Bases de données retenues pour les expérimentations

critères de qualité : représentativité, pureté et profondeur dans l'arbre.

9.2.8 Appréciation des performances de `DeTTMInE` sur des bases d'évaluation

Pour permettre une évaluation pertinente du processus de classification, il est nécessaire de pouvoir le comparer à des valeurs de référence sur des bases de données classiques. Un tel procédé permet de ne pas remettre en cause les principes utilisés dans `DeTTMInE` : si l'algorithme se comporte de manière identique aux autres utilisés pour la comparaison sur ces données, alors l'algorithme de `DeTTMInE` peut être validé.

Les bases de données utilisées pour l'évaluation sont toutes issues du site de l'`UCI`[19]. Ce site regroupe en effet une collection de bases de données, toutes reconnues par les acteurs des domaines respectifs, pour lesquelles des objectifs de fouille de données sont proposés. Ici, seul le matériel brut - les bases de données - est utile pour la démarche de ce travail de recherche.

A partir de telles bases de données, il est alors possible de comparer différents classifieurs. Les bases de données ne pouvant pas être remises en cause dans l'interprétation des résultats. Les bases utilisées pour mener les expérimentations sont présentées dans le tableau 9.2, et détaillées en annexe A page 201.

L'ensemble des bases utilisées pour ce travail a subi un pré-traitement en vue de coder les données textuelles (par simple remplacement des valeurs d'attributs par une valeur numérique unique dont l'équivalence est conservée dans un fichier dit de dictionnaire.

A l'aide de tous les éléments déployés, il est maintenant possible de comparer le classifieur de `DeTTMInE` à un reconnu, `C4.5`, sur des bases de données, elles aussi admises. `C4.5` bénéficie en effet d'une longue expérience (la littérature en témoigne aisément), de plus il offre une certaine

[19] **U**niversity of **C**alifornia, **I**rvine, http ://kdd.ics.uci.edu/

base	classifieur	score
breast	DeTTMInE	74,04%
	C4.5	74,70%
	C4.5rules	68,62%
crx	DeTTMInE	64,63%
	C4.5	69,38%
	C4.5rules	69,30%
zoo	DeTTMInE	93,23%
	C4.5	90,18%
	C4.5rules	91,22%
pima	DeTTMInE	74,08%
	C4.5	74,40%
	C4.5rules	72,26%

base	classifieur	score
bal	DeTTMInE	77,26%
	C4.5	61,56%
	C4.5rules	72,56%
lympho	DeTTMInE	78,83%
	C4.5	78,28%
	C4.5rules	81,08%
housing	DeTTMInE	73,53%
	C4.5	81,40%
	C4.5rules	81,56%
flare	DeTTMInE	72,62%
	C4.5	73,38%
	C4.5rules	73,54%

TAB. 9.3 – Résultats expérimentaux comparant `DeTTMInE` à `C4.5` et `C4.5rules` sur des bases de données issues de l'`UCI`

aisance à l'utilisation, et des résultats complets sur les actions réalisées.

`C4.5` dispose de deux modes opératoires : le premier repose sur l'arbre de décision produit par l'intermédiaire de la base d'apprentissage. Le second produit un ensemble de règles, à partir desquelles `C4.5` essai de caractériser les exemples de reconnaissance.

Les résultats de ces comparaisons, sur l'ensemble des bases retenues, sont consignés dans le tableau 9.3 et synthétisés dans l'histogramme 9.5 page suivante. Ce tableau de résultats présente pour chaque base de données testée le calcul du taux de reconnaissance pour chaque classifieur comparé. Ce taux correspond à la moyenne[20] des scores réalisés sur l'ensemble des paires de fichiers générés par validation croisée (cf. 3.3.1 page 46).

Ces résultats sont dans l'ensemble satisfaisants. `DeTTMInE` dépasse `C4.5` et `C4.5rules` sur certaines bases (`bal` et `zoo`), et dans les cas où il ne les dépasse pas (`breast`, `pima`, `lympho` et `flare`), il reste à leur niveau (à plus ou moins 2-3%). Cependant, sur les bases `crx` et `housing` `DeTTMInE` est nettement moins bon (de l'ordre de 5 à 8%).

De manière générale, pour 6 des bases testées, `DeTTMInE` se révèle meilleur ou égal à `C4.5` et `C4.5rules`. Hormis un procédé de sélection de descripteurs à partir desquels sont déterminées des règles, il n'y a pas d'autre traitement pour en améliorer la qualité. Or il existe des traitements applicables qui augmentent la qualité des règles. `C4.5rules` améliore les règles produites par `C4.5` avec des mesures statistiques ou encore des procédés de généralisation. Pour des raisons de

[20] Le détail de cette moyenne par base est présenté en annexe B page 205

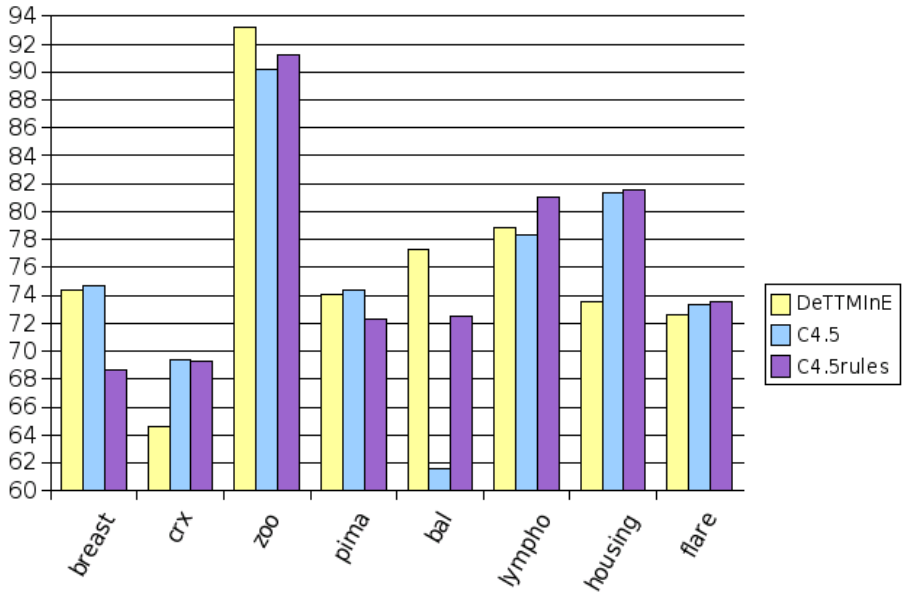


FIG. 9.5 – Synthèse des résultats obtenus pour la compraison de `DeTTMInE` à `C4.5` et `C4.5rules` sur les bases de l'`UCI`.

temps, de tels procédés n'ont pas été mis en place. Cependant, il est évident que dans un futur travail cela sera nécessaire. La reconnaissance mesurée sur `DeTTMInE` est donc appréciable, et les scores obtenus correspondent aux attentes que nous nous sommes fixées pour ce travail.

9.2.9 Évaluation des performances de `DeTTMInE` sur la base documentaire juridique

L'évaluation de `DeTTMInE` s'est faite comparativement à `C4.5` car tous deux produisent un système de règles pour classifier de nouveaux exemples et reposent tous deux sur un principe d'arbre de décision assez similaire. Le jeu de comparaison a été défini par une validation croisée proposant en test une dizaine d'exemples pour un apprentissage composé d'une centaine de cas.

La première information à observer est la comparaison du score de `C4.5` à celui de `DeTTMInE`. Le premier a un score de 63.86% et le second un score de 66%. Le premier point remarquable dans cette évaluation est la gestion des cours décisionnelles. Cette gestion apporte peu, en terme de score de reconnaissance, comparativement à un algorithme ne la prenant pas en compte.

Le second point remarquable est relatif aux résultats consignés dans les tableaux 9.4 et 9.5 page ci-contre. Ces deux tableaux présentent le nombre de règles induites, le nombre de règles utilisées pour valider les jeux de données test et le nombre d'utilisation de la règle par défaut durant la validation sur les jeux de test. Cette règle par défaut est appliquée quand aucune autre règle ne s'applique à l'exemple testé.

Jeu de données	Règles induites	Règles utilisées	Règle par défaut
marque 1	8	6	0
marque 2	9	4	1
marque 3	5	5	5
marque 4	4	3	2
marque 5	10	6	1
marque 6	5	2	1
marque 7	5	3	4
marque 8	6	5	3
marque 9	6	6	3
marque 10	5	3	7
moyenne	6	4	3

TAB. 9.4 – Observation de la gestion des règles induites par C4.5 sur une validation croisée (10 paquets)

Jeu de données	Règles induites	Règles utilisées	Règle par défaut
marque 1	13	6	0
marque 2	13	8	1
marque 3	11	6	0
marque 4	11	7	0
marque 5	10	6	0
marque 6	10	6	0
marque 7	10	4	0
marque 8	17	7	1
marque 9	12	6	0
marque 10	9	5	0
moyenne	12	6	0

TAB. 9.5 – Observation de la gestion des règles induites par DeTTMInE sur une validation croisée (10 paquets)

Un autre résultat important est, cette fois, relatif aux règles produites et à leur utilisation. Par rapport à `C4.5`, `DeTTMInE` produit environ une douzaine de règles pour le jeu de données testé et en utilise généralement 6 pour qualifier le jeu de test. En revanche, `C4.5` induit en moyenne 6 règles (soit deux fois moins de règles que `DeTTMInE`) et en applique généralement 4.

Notre attention se porte alors sur le comportement de l'algorithme, et l'utilisation des règles induites. Plus particulièrement, nous allons observer l'utilisation de la règle par défaut. Cette règle est à considérer quand aucune des autres règles possibles ne satisfait l'exemple à reconnaître. Cette interprétation est alors plus délicate car il faut avoir connaissance des autres règles pour comprendre la décision. Sur ces ensembles de règles appliquées, à une exception près (le premier jeu de données), `C4.5` utilise au moins une fois le cas par défaut avec un maximum observé à 7 exemples reconnus sur 10 par le cas par défaut. Contrairement à cela `DeTTMInE` n'a utilisé le cas par défaut qu'à deux reprises et sur deux jeux de test différents.

L'utilisation parcimonieuse de la règle par défaut par `DeTTMInE` apporte alors une valeur ajoutée non négligeable sur l'explication d'une décision, contrairement à `C4.5`.

9.2.10 Bilan de `DeTTMInE`

Le bilan de cette expérience, consistant à produire un algorithme de classification applicable aux documents électroniques juridiques caractérisant des jurisprudences, démontre que :

1. la construction de règles de classification reposant sur l'indice de qualité de nœud est aussi performante que les arbres d'induction de type `C4.5`
2. l'utilisation d'un algorithme orienté vers le domaine d'étude ne dégrade pas la qualité décisionnelle de l'algorithme par rapport aux algorithmes usuels
3. l'utilisation d'indice de qualité réduit l'utilisation d'une règle de classification dite par défaut
4. la sélection de règle par prise en compte des spécificités juridiques permet de réduire l'effet de bord de la sélection par indice de qualité de nœud : plusieurs règles possibles par enregistrement testé

Nous pouvons maintenant avancer le fait que la qualité des résultats des algorithmes testés n'est pas liée aux algorithmes de modélisation mais probablement à la qualité du jeu de données et de sa représentation. En effet, nous venons de démontrer avec `DeTTMInE` que l'utilisation d'un algorithme dédié au domaine d'étude n'était pas suffisante pour garantir la justesse des connaissances induites et que des algorithmes que l'on peut qualifier de génériques se comportaient de manière similaire. Ces deux observations conduisent alors à une remise en cause de la représentation des documents électroniques qui ne semble pas être optimale contrairement à celle des outils de modélisation.

Nombre de Regroupements	Entropie	Pureté
1	0.9994	0.5142
2	0.9438	0.638
3	**0.8898**	**0.6857**
4	0.9483	0.619
5	0.9245	0.6571
6	0.9178	0.6476
7	0.8768	0.6952
8	0.8123	0.6761

TAB. 9.6 – Mesures de l'entropie et de pureté en fonction du nombre de regroupements recherchés

9.3 Comparaison avec un algorithme de regroupements

La comparaison avec un algorithme de regroupements permet de déterminer la pertinence du jeu de données et de sa représentation sous forme de descripteurs. En effectuant une telle recherche, l'algorithme de regroupement va déterminer les données les plus corrélées entre elles. L'étude des résultats permet alors de déterminer si l'ensemble des documents correspond au problème étudié.

9.3.1 Approche exploratoire non supervisée

L'étude de regroupements de manière non supervisée permet de faire apparaître le nombre de regroupements idéal pour le jeu de données étudié. Nous avons alors réalisé une série d'expériences afin d'étudier l'entropie minimale et la pureté maximale lorsque l'on fait varier le nombre de regroupements cherchés. Les résultats sont consignés dans le tableau 9.6.

Pour cette expérience, nous avons fait varier le nombre de regroupements cherchés de 1 à 7. La comparaison des solutions se fait sur la base de deux mesures présentées dans [Zhao & Karypis, 2001] et [Zhao & Karypis, 2004] :

1. l'entropie qui mesure la distribution des documents au travers des regroupements
2. la pureté : évaluant la représentativité de la classe la plus présente dans le regroupement calculé.

Une entropie locale est calculée à partir de l'équation 9.1 où C est un regroupement de taille n_c, n le nombre de classes dans le jeu de données, et n_c^i le nombre de documents de la *i-ème* classe assignés au *c-ième* regroupement.

$$Entropie(C) = -\frac{1}{\log n} \sum_{i=1}^{n} \frac{n_c^i}{n_c} \log \frac{n_c^i}{n_c} \tag{9.1}$$

De cette définition, nous pouvons alors introduire l'entropie globale (équation 9.2) comme la somme de toutes les entropies locales, pondérée en fonction de la taille du regroupement (en nombre de documents)

$$EntropieGlobale = \sum_{i=1}^{n} \frac{n_i}{n} Entropie(C) \tag{9.2}$$

D'une façon similaire, la mesure de la pureté est définie par l'équation 9.3, où C est un regroupement de taille n_c et n_c^i le nombre de documents de la *i-ème* classe assignés au *c-ième* regroupement.

$$Pureté(C) = \frac{1}{n_c} max(n_c^i) \tag{9.3}$$

La pureté représente la fraction de l'ensemble du regroupement correspondant à la plus large classe présente. Comme pour l'entropie, la pureté peut être généralisée comme la somme des puretés de chaque regroupement, pondérée, là encore, par la taille du regroupement (équation 9.4).

$$PuretéGlobale = \sum_{i=1}^{n} \frac{n_c}{n} Pureté(C) \tag{9.4}$$

En générale, une bonne solution est celle dont les regroupements ne contiennent que les documents d'une seule classe. Dans ces cas là, l'entropie vaut 0 (en générale plus l'entropie tend vers 0, meilleure est la solution) et la pureté à 1 (chacune définie sur l'intervalle [0 ;1]). Ces mesures sont bien adaptées pour ces algorithmes qui établissent les regroupements en fonction de mesures de distances euclidiennes.

Nous observons (tableau 9.6 page précédente) alors que pour un ensemble de trois regroupements, la pureté et l'entropie sont pour la première la plus forte et pour la seconde la plus faible. Toutefois, pour sept regroupements, nous obtenons des résultats similaires à ceux obtenus pour trois regroupements. Cet effet est expliqué par une sur-segmentation de l'espace de recherche en un ensemble de sous-ensembles d'enregistrements très fins. Ce principe de sur-segmentation ou encore de sur-apprentissage est défini dans l'exemple de la figure 9.6 page suivante. Lorsque l'algorithme doit modéliser deux classes, il scinde alors l'espace de données en deux (cas A). Lorsque ce nombre de classes augmente, alors l'espace de recherche est segmenté plus finement, conduisant à une spécialisation très fine des regroupements ainsi proposés afin de minimiser l'entropie des regroupements calculés. De plus, ces frontières peuvent ne pas correspondre aux données réelles.

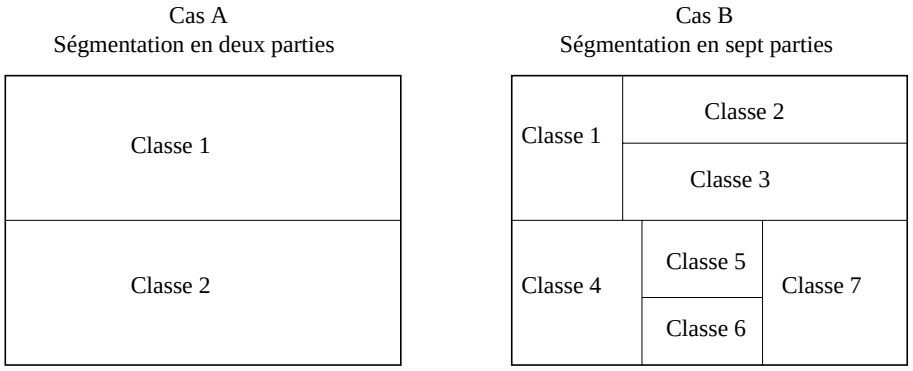


FIG. 9.6 – Passage d'un découpage grossier en sur-aprentissage par augmentation du nombre de sous-ensembles recherchés.

9.3.2 Approche exploratoire supervisée

Une autre approche exploratoire, cette fois de manière supervisée, permet d'observer la définition du domaine. L'idée est que la valeur de classe des enregistrements est un descripteur fortement discriminatoire. Dans `CATMInE` ce descripteur est binaire : contrefaçon ou non. En regroupant les données sur un tel descripteur, l'algorithme de modélisation va regrouper les données en deux sous-ensembles (un par valeur de classe). Si en revanche, l'algorithme utilise d'autres descripteurs pour produire 1 ou l'ensemble des regroupements, cela démontre alors qu'il existe une association de descripteurs plus pertinente en terme de gain d'information que l'utilisation de la valeur de classe. Cette démarche permet de mettre en valeur des erreurs de représentation de données. De plus, si l'algorithme offre plus de regroupements que de valeurs de classe, l'étude des regroupements peut faire apparaître des erreurs de documents et de représentation des données.

Afin d'évaluer les regroupements obtenus, nous utiliserons à nouveau les mesures de pureté et d'entropie définies dans la section précédente. Nous mesurons alors une recherche, de deux, puis trois et enfin quatre regroupements (les résultats révèlent qu'il est inutile d'aller au delà de quatre regroupements). Ces résultats sont présentés dans les tableaux 9.7, 9.8 et 9.9 page suivante.

	Cluster 1	Cluster 2	Global
Entropie	0.9288	0.9023	0.9177
Pureté	0.6557	0.6818	0.6666

TAB. 9.7 – Entropie et pureté pour une recherche supervisée de deux regroupements

Les résultats de la recherche déterminant deux regroupements sont obtenus sans l'utilisation de la valeur de classe. L'algorithme privilégiant une combinaison de descripteurs. Cette

	Cluster 1	Cluster 2	Cluster 3	Global
Entropie	0	0.934	0	0.356
Pureté	1	0.65	1	0.867

TAB. 9.8 – Entropie et pureté pour une recherche supervisée de trois regroupements

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Global
Entropie	0	0	0	0	0
Pureté	1	1	1	1	1

TAB. 9.9 – Entropie et pureté pour une recherche supervisée de quatre regroupements

combinaison a un pouvoir discriminatoire plus fort que la valeur de classe et maximise le gain d'information. Ces résultats sont surprenant sur deux aspects :

1. La valeur de classe n'est pas utilisée pour trier les données
2. Le nombre de regroupements cherché correspond à la définition du domaine.

Pour la recherche de trois regroupements, la valeur de classe est cette fois utilisée pour 2 sous-ensembles, le troisième étant défini par une combinaison de descripteurs. Ici, nous interprétons ces résultats comme le fait que deux sous-ensembles du jeu de données partagent de fortes similitudes et un troisième sous-ensemble regroupe des données dont la caractérisation par une combinaison de descripteurs est meilleure que l'utilisation de la valeur de classes. Nous pouvons alors formuler l'hypothèse suivante : ces documents peuvent être du bruit, des données dont la caractérisation par les descripteur n'est pas optimale, ou encore de la connaissance stratégique.

Enfin, pour une étude portant sur quatre regroupements, les résultats mettent en évidence qu'il existe bien une frontière dans le troisième regroupement observé dans l'expérience précédente. Le fait de rajouter un centroïde dans la recherche permet d'affiner la segmentation de l'espace de recherche et de mettre en évidence la frontière entre les données de ce sous-ensemble.

9.4 Observations

Les résultats des recherches réalisées, et plus particulièrement l'étude de regroupements, montrent que pour trois catégories (tableau 9.8), il en existe une dont l'entropie et la pureté sont médiocres (dénommée C par la suite), relativement aux deux autres catégories proposées. Il faut savoir que pour ce type d'algorithme, l'expert fixe le nombre de regroupements à déterminer. Toutefois, s'il existe une solution optimale faisant intervenir moins de regroupements, alors l'algorithme choisira cette solution. Dans cette étude ce n'est pas le cas et les solutions proposées correspondent aux paramètres fixés.

La conséquence directe de ces observations est la remise en cause de la qualité des modèles des données utilisées : les différents algorithmes expérimentés ont un comportement similaire et les algorithmes de regroupements mettent en évidence un sous-ensemble de données litigieux. La transformation des données ne semble pas adaptées, et certains exemples sont proches avec la définition actuelle des descripteurs existants mais ont une issue différente dans la décision.

Nous allons donc orienter notre exploration du domaine (chapitre 10 page 147) sur le prétraitement des données et la recherche des exemples proches. Ces exemples peuvent caractériser une connaissance critique (la majorité des exemples sont reconnus) où le savoir-faire s'exprime plus profondément, ou bien des données inadaptées à la définition du domaine d'étude.

Chapitre 10

Étude de la validité de la transformation des données

Sommaire

Pour étudier la transformation des données et plus particulièrement le pré-traitement de celles-ci, nous allons, dans un premier temps, appliquer un protocole de sélection des données litigieuses, puis, dans un second temps, nous étudierons ce sous-ensemble afin de valider ou invalider les choix réalisés dans la modélisation de `CATMInE`, et de déterminer des améliorations possibles.

10.1 Déterminer les enregistrements litigieux

Pour déterminer les enregistrements litigieux, nous allons suivre un protocole expérimental afin de mettre en valeur ces données ([Renaux, 2005a]). Ce protocole est présenté dans la figure 10.1 page suivante.

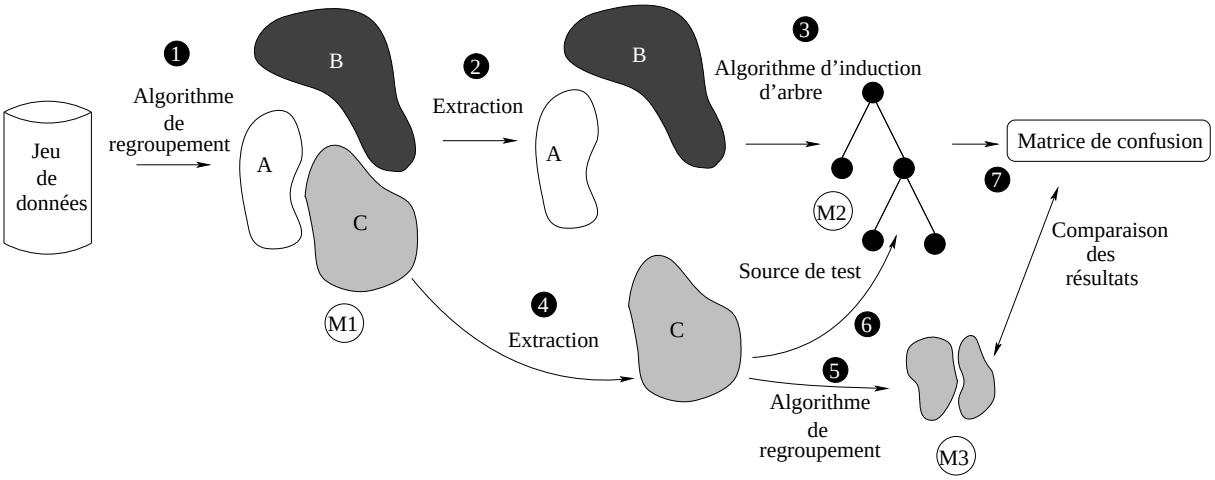


FIG. 10.1 – Protocole de sélection des données ambiguës.

Le principe consiste à établir par le biais d'un algorithme de regroupement, ici `K-means`, un modèle des données en prenant en compte la valeur de classe (étape 1). Nous avons démontré dans la section précédente que cette expérience entraîne la production de trois regroupements (modèle $M1$) : deux parfaits (regroupements A et B) et un hétérogène (le regroupement C). De ces résultats, nous isolons les données des regroupements A et B (étape 2) et nous établissons un modèle $M2$ avec un algorithme d'arbre d'apprentissage (étape 3) de type `C4.5`. Cette démarche s'inspire du méta-modèle de type en cascade (présenté dans la section 6.6.3 page 94). Le sous-ensemble d'algorithmes étant restreint à `K-means` et `C4.5`.

Nous établissons ensuite à partir des données du regroupement C (étape 4), un modèle $M3$ avec `K-means`, paramétré pour ne pas prendre en compte la valeur de classe, et pour établir 2 regroupements (étape 5).

Le jeu de données C, déterminé dans l'étape 4, est aussi appliqué en tant que jeu test pour le modèle $M2$ (étape 6).

Enfin, nous comparons les résultats de l'étape 6 et de l'étape 5 en étudiant les matrices de confusions obtenues et en déterminant l'intersection des exemples mal identifiés par chacun des algorithmes (étape 7).

Le jeu de données est constitué de : 105 décisions, 54 contrefaçons et 51 non-contrefaçons. Cet ensemble de décisions est issu d'une sélection de documents présentée dans les section 3.4.3 page 50 où nous justifions le passage de 8450 décisions à 2576 (modification du droit des marques en 1977, document enrichi du texte complet de la décision en 1997), et de 2576 à 800 (sélection par l'expert). Enfin, dans la conclusion du chapitre 4 page 54, nous expliquons que nous avons uniquement conservé les jugements pour lesquels l'information recherchée était bien étiquetée (ce qui a été le cas de ces 105 décisions).

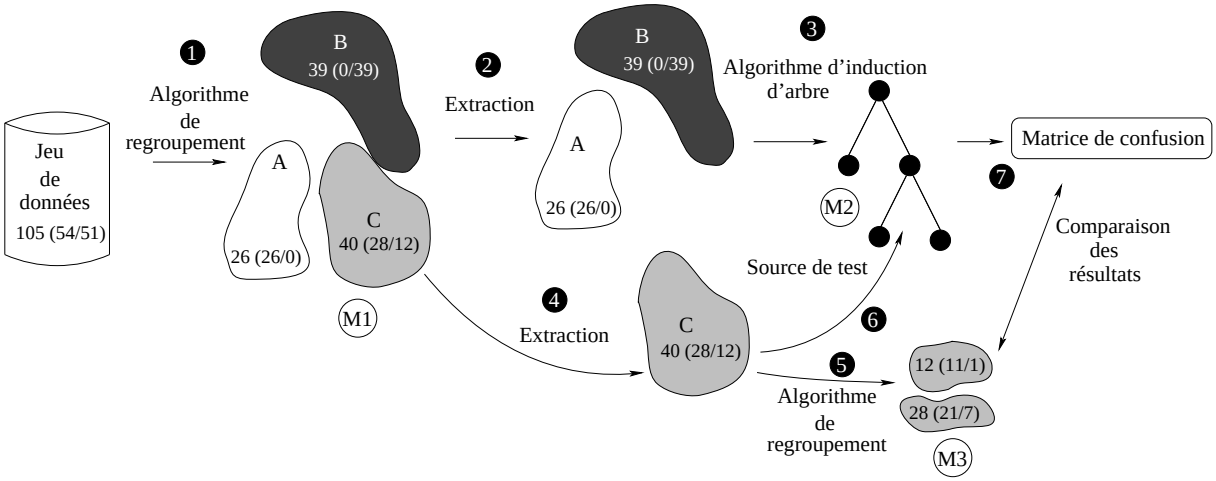


FIG. 10.2 – Protocole de sélection des données ambiguës (protocole valué).

C4.5	C	$\neg C$
C	12	23
$\neg C$	0	5

K-means	C	$\neg C$
C	11	21
$\neg C$	1	7

TAB. 10.1 – Matrices de confusion de l'arbre d'induction et des regroupements

La figure 10.2 synthétise les résultats de cette expérience. Nous avons constaté lors de la réalisation de ce protocole que :

- étape 1 : trois regroupements sont établis : deux regroupements purs : A et B (65 enregistrements : 26 contrefaçons et 39 non-contrefaçons) et un regroupement hétérogène, C (40 enregistrements : 28 contrefaçons pour 12 non-contrefaçons).
- le modèle $M2$ n'a que des feuilles pures, même après élagage.
- étape 6 : le modèle $M3$ est caractérisé par deux regroupements hétérogènes.
- étape 6 *bis* : un modèle $M3'$ est établi en prenant en compte la valeur de classe avec pour résultat deux regroupements parfait. Cela permet ainsi de démontrer qu'il existe bien une frontière au sein des enregistrements du regroupement C.

Les résultats observés à l'étape 7 sont consignés dans les matrices de confusion présentées dans les tableaux 10.1.

Ces deux matrices sont très proches l'une de l'autre en terme de classement. Cette proximité est expliquée par :

1. 12 contrefaçons étiquetées comme contrefaçons avec `C4.5`, pour 11 par `K-means`. L'intersection de ces deux sous-ensembles regroupe les mêmes enregistrements
2. 23 contrefaçons reconnues par `C4.5` comme non-contrefaçon, pour 21 par `K-means`. L'inter-

section des deux sous-ensembles représente 16 décisions

3. 5 non-contrefaçons étiquetées comme non-contrefaçon par `C4.5`, pour 7 par `K-means`. L'intersection de ces deux sous-ensemble est nulle
4. 1 non-contrefaçon classée comme contrefaçon par `K-means`, 0 pour `C4.5`

Cette proximité permet de poser comme hypothèse que tous les exemples reconnus par l'un ou l'autre des modèles sont considérés comme appartenant au domaine et identifiables par l'un ou l'autre des procédés de modélisation. En effet, les algorithmes sont mis en faute à ce niveau : ce que l'on ne peut pas identifier avec `C4.5` peut être caractérisé avec `K-means` et vice versa. Ainsi les points 1, 3 et 4 peuvent être résolus par cette hypothèse.

Le point 1 met en avant le fait que `C4.5` trouve un enregistrement de plus que `K-means`. Le bénéfice est double : tout d'abord `C4.5` est capable d'identifier un exemple de mieux et `C4.5` ne commet pas l'erreur de `K-means`. Cette particularité est valable à chaque fois que l'un des deux algorithmes est meilleur que l'autre.

Le point 3 valide aussi notre hypothèse. L'intersection des deux sous-ensembles étant nulle, les cinq enregistrements trouvés par `C4.5` ne le sont pas avec `K-means` et les sept enregistrements caractérisés par `K-means` ne sont pas identifiés par `C4.5`.

Pour le point 4, il s'agit de l'enregistrement supplémentaire trouvé par `C4.5`, et qui valide ainsi notre hypothèse.

Le bilan de cette expérience est que 11 enregistrements caractérisant des non-contrefaçons sont bien identifiés par `C4.5` et `K-means`, 13 le sont par l'un ou l'autre (les 7 contrefaçons bien identifiées par `K-means`, et les 5 par `C4.5`) et que 16 enregistrements sont des erreurs communes. En effet, avec les propriétés observées aux points trois et quatre, le sous-ensemble de contrefaçons étiquetées comme non-contrefaçons par `K-means`, contient les 5 enregistrements bien identifiés par `C4.5` comme non-contrefaçon : l'intersection des contrefaçons bien identifiées par les deux algorithmes est nulle, et la non-contrefaçon mal identifiée par `K-means`, l'est par `C4.5`. Si l'on soustrait ces 5 enregistrements aux 21 erreurs, il en reste 16. Il en va de même avec les 7 enregistrements de contrefaçon bien identifiés par `K-means` sur le sous-ensemble de non-contrefaçons mal identifié par `C4.5`.

10.2 Étude des enregistrements mal identifiés

10.2.1 Caractérisation des exemples

Pour permettre d'étudier les enregistrements mal identifiés par le protocole mis en place dans la section précédente, nous avons mesuré les coordonnées du centroïde formé par ces 16 enregistrements, ainsi que la déviation standard (ou écart-type moyen) des enregistrements par rapport à ce centroïde. Le tableau 10.2 page ci-contre consigne les mesures obtenues.

Descripteur	Écart-type moyen inférieur	Valeur moyenne	Écart-type moyen supérieur
descripteur 1	**0,104 (12,55%)**	0,830341	**0,081 (9,76%)**
descripteur 2	0,169 (28,07%)	0,602984	**0,101 (16,84%)**
descripteur 3	0,282 (78,35%)	0,360938	0,169 (47,01%)
descripteur 4	0,236 (93,00%)	0,254471	0,141 (55,80%)
descripteur 5	**0,155 (19,61%)**	0,792950	**0,093 (11,76%)**
descripteur 6	**0,067 (13,67%)**	0,494485	0,148 (30,08%)
descripteur 7	0,188 (70,38%)	0,267807	0,113 (42,23%)
descripteur 8	0,130 (76,24%)	0,170863	0,078 (45,75%)

TAB. 10.2 – Distances euclidiennes des 16 décisions isolées par rapport au centroïde qu'elles caractérisent (entre parenthèses est rapporté l'écart-type)

Descripteur	Centroïde1	Distance moyenne	Centroïde2	Distance moyenne
descripteur 1	0.5875	**0.1169 (19%)**	0.8747	**0.0955 (10%)**
descripteur 2	0.7909	**0.1207 (15%)**	0.6558	0.1506 (23%)
descripteur 3	0.3818	0.1667 (44%)	0.3735	0.1905 (51%)
descripteur 4	0.6032	0.1259 (21%)	0.2669	0.1431 (54%)
descripteur 5	0.4362	**0.0705 (16%)**	0.8064	***0.1202 (15%)***
descripteur 6	0.6898	**0.1254 (18%)**	0.5621	0.1553 (28%)
descripteur 7	0.3337	0.1243 (37%)	0.3129	0.1476 (47%)
descripteur 8	0.5179	**0.0884 (17%)**	0.221	0.1083 (49%)

TAB. 10.3 – Identification des descripteurs caractérisant les regroupements obtenus avec K-means

Dans ce tableau, nous avons relevé l'écart-type moyen inférieur et l'écart-type moyen supérieur observé pour chaque coordonnée du centroïde. Entre parenthèses, nous avons calculé le rapport entre l'écart-type et la coordonnée du centroïde. Nous avons mis en évidence (en gras) les résultats pour lesquels l'écart-type observé pour chaque descripteur est très faible (moins de 20%). Cette propriété révèle alors les descripteurs caractérisant le regroupement : descripteurs pour lesquels les données sont concentrées et proches autour des coordonnées du centroïde.

Nous observons que les enregistrements sont peu épars pour les descripteurs 1, 2, 5 et 6. Nous allons maintenant faire les mêmes opérations mais cette fois avec toutes les données issues du regroupement C obtenu à l'étape 1 par le protocole présenté dans la figure 10.1 page 148. Les résultats sont consignés dans le tableau 10.3.

Nous observons pour le centroïde 1 que les données sont peu éparpillées avec les descripteurs

	C			$C/\{16\}$			C			$C/\{16\}$		
	Cluster A (x,y) ent			Cluster A (x,y) ent			Cluster B (x,y) ent			Cluster B (x,y) ent		
	Coord.	σ	Rapport	Coord.	σ	Rapport	Coord.	σ	Rapport	Coord.	σ	Rapport
Desc 1	0,8747	0,0955	**10,92%**	0,8893	0,1425	**16,02%**	0,5875	0,1169	**19,90%**	0,6470	0,0981	**15,16%**
Desc 2	0,6658	0,1506	22,62%	0,7128	0,1290	**18,10%**	0,7909	0,1207	**15,26%**	0,7826	0,1427	**18,23%**
Desc 3	0,3735	0,1906	51,03%	0,3325	0,2082	62,62%	0,3818	0,1667	43,66%	0,4949	0,1356	27,40%
Desc 4	0,2669	0,1431	53,62%	0,2557	0,0806	31,52%	0,6032	0,1259	20,87%	0,6504	0,0852	**13,10%**
Desc 5	0,8064	0,1202	**14,91%**	0,7990	0,0865	**10,83%**	0,4362	0,0705	**16,16%**	0,4778	0,1416	29,64%
Desc 6	0,5621	0,1553	27,63%	0,6423	0,1442	22,45%	0,6898	0,1254	**18,18%**	0,6648	0,1637	24,62%
Desc 7	0,3129	0,1476	47,17%	0,3309	0,0846	25,57%	0,3337	0,1243	37,25%	0,3879	0,1148	29,60%
Desc 8	0,221	0,1083	49,00%	0,2648	0,0731	27,61%	0,5179	0,0884	**17,07%**	0,5304	0,0942	**17,76%**

TAB. 10.4 – Comparaison des déviations standard sur le regroupement C et sur ce même regroupement privé des exemples litigieux.

1, 2, 4, 6 et 8. Pour le second centroïde, les descripteurs ayant le même type de propriétés sont les descripteurs 1 et 5.

En rapportant ces résultats à ceux obtenus pour l'étude des 16 décisions sélectionnées (tableau 10.2 page précédente), nous observons que sur les six descripteurs utilisés par `K-means` pour traiter le regroupement C, quatre le sont pour définir le centroïde des 16 décisions.

Nous pouvons aussi constater que dans la définition du centroïde des 16 décisions sélectionnées, les descripteurs 2 et 6 sont utilisés dans la définition du premier centroïde du sous-ensemble C et que les descripteurs 1 et 5 le sont dans la définition du second centroïde proposé. Ces 16 décisions, sur les 40, participent donc à la production de bruit dans les regroupements établis à partir de C en partageant des propriétés communes aux 2 regroupements calculés.

Enfin, afin de valider ces observations, nous avons réalisé une dernière expérience consistant à établir les regroupements sur le sous-ensemble C et les mêmes regroupements, calculés avec les données du sous-ensemble C, privé des 16 enregistrements. Cette expérience a pour but de démontrer qu'avec ou sans ces exemples, l'influence des descripteurs dans la construction des regroupements ne change pas (ou peu). Ces résultats sont consignés dans le tableau 10.4

Ce tableau confronte les coordonnées des centroïdes A et B obtenus à partir du jeu de données C établit dans le protocole 10.1 page 148, aux mêmes centroïdes, issus du même jeu de données pour lequel nous avons préalablement enlevé les exemples contraignants pour l'apprentissage. Dans la première colonne, nous avons rapporté les coordonnées du centroïde A issu d'une modélisation sur l'intégralité des données caractérisant le regroupement C. La deuxième colonne indique l'écart-type observé sur ces coordonnées et la troisième colonne le rapport que représente cette écart-type par rapport aux coordonnées. Nous avons fait apparaître en gras, les rapports entre écart-type et coordonnées inférieures à 20%. Les 4^e, 5^e et 6^e colonnes fournissent les mêmes informations, mais cette fois-ci pour le jeu de données privé des enregistrements contraignants. Les colonnes 7, 8 et 9 fournissent les informations sur le second centroïde observé avec l'intégralité des données. Enfin les 3 dernières colonnes présentent les résultats du second centroïde pour

Descripteur	Centroïde1	Ecart-type
descripteur 1	0.8162	**0.1268 (15,54%)**
descripteur 2	0.6828	**0.1479 (21,66%)**
descripteur 3	0.3751	0.1856 (49,48%)
descripteur 4	0.3341	0.1898 (56,81%)
descripteur 5	0.7323	**0.1606 (21,93%)**
descripteur 6	0.5876	**0.1516 (25,80%)**
descripteur 7	0.317	0.1442 (45,49%)
descripteur 8	0.2803	0.1477 (52,69%)

TAB. 10.5 – Coordonnées et écarts-type du regroupement C obtenu à l'étape 1 du protocole de modélisation

le jeu de données épuré.

Les descripteurs 1, 2, 5, 6 et 8 sont principalement utilisés pour la définition des centroïdes. Or nous avons précédemment observé que les descripteurs 1, 2, 5 et 6 interviennent dans la caractérisation des seize exemples posant problème. Ces descripteurs sont donc utilisés pour établir le modèle quelque soit le jeu de données

Enfin, la dernière étude à réaliser consiste à observer si les coordonnées du centroïde caractérisant le regroupement C, obtenues à l'étape 1 (et donc en tenant compte de l'influence des regroupements A et B sur C), reposent sur les mêmes descripteurs que ceux observés jusque là pour cet ensemble de données. Le tableau 10.5 présente ces mesures (coordonnées du centroïde, distance moyenne des exemples aux coordonnées, rapport de cette distance par rapport aux coordonnées).

Nous avons mis en gras les écarts-type relativement faibles. Nous observons alors qu'à nouveau, les descripteurs 1, 2, 5 et 6, sont utilisés pour établir ce regroupement. Cela va alors dans le sens des observations précédentes où ces mêmes descripteurs étaient utilisés pour la définition du centroïde caractérisant les données du regroupement C.

Dans cette caractérisation des exemples par le biais d'une approche exploratoire, nous avons observé les descripteurs intervenant dans la modélisation du sous-ensemble de données hétérogène C. Nous avons ensuite procédé de même sur des exemples considérés comme litigieux, issus de C et sur les exemples valides de ce même sous-ensemble. Les exemples litigieux ont été déterminés en confrontant les résultats de deux algorithmes de modélisation. Enfin, nous avons comparé les fondements sur lesquels reposent les résultats issus de l'étape 1, à ceux précédemment observés dans les autres expériences.

De cette étude et des résultats observés, nous pouvons conclure qu'avec ou sans les seize

exemples posant des difficultés de modélisation, les regroupements sont obtenus sur les mêmes bases. Ces données entraînent les contraintes sur le regroupement C. En effet, sans les exemples le regroupement C peut être validé par l'un ou l'autre des algorithmes. Il est à noter que les regroupements obtenus par `K-means` pour le sous-ensemble C ne sont pas purs.

10.2.2 Étude des exemples litigieux

Nous avons précédemment constaté qu'un sous-ensemble d'enregistrements représentait des difficultés de modélisation. Nous avons isolé ces exemples et étudié sur quelles bases ils sont regroupés. Nous avons alors démontré que les fondements sur lesquels ils reposaient, correspondaient aux mêmes spécificités que l'intégralité des données modélisées.

Nous allons maintenant étudier en détails ces seize exemples afin de déterminer si ceux-ci appartiennent bien au domaine de définition. S'il s'agit bien d'exemples appartenant au domaine, alors nous allons étudier leur définition de manière générale puis plus particulièrement sur les descripteurs identifiés comme coordonnées influentes des centroïdes. Le tableau 10.6 page suivante énumère ces seize exemples.

Ces seize décisions sont toutes des contrefaçons identifiées comme non-contrefaçons. Il est à noter que l'enregistrement 4 est identique au 14, il s'agit du même procès : dans le premier cas de figure il s'agit du verdict d'un Tribunal de Grande Instance, dans le second cas une décision rendue par une Cour d'Appel. Cet appariement est relatif uniquement aux données issues du troisième regroupement (le sous-ensemble C) isolé par notre protocole de sélection de documents électroniques (10.1 page 148). Cette association à des non-contrefaçons est donc relative aux propriétés dégagées dans ce sous-ensemble et non à l'intégralité du jeu de données.

En observant plus finement les décisions, Nous pouvons les regrouper en trois catégories :

1. la marque du plaignant est plus petite que la marque du défendant et est intégralement reprise (décisions 3, 8 et 11)
2. la marque du plaignant ayant un mot (ou plus) repris par la marque du défendant. Les deux marques étant de taille équivalente (décisions 1, 5, 6, 7, 10, 12, 15 et 16 ; la décision 5 étant un anagramme)
3. la marque du plaignant à moins d'un mot repris par la marque du défendant (décisions 2, 4, 9, 13 et 14)

Quelle que soit la décision, si l'on observe les descripteurs identifiés dans notre étude, nous observons alors :

– descripteur 1, nombre de caractères en commun entre la marque du plaignant et la marque du défendant : deux décisions (décision 14 et 16) sont au dessus de 60% en communs, les autres sont toutes à plus de 70%. Cela se traduit juste par une lettre non présente, rapporté au nombre de caractères composant la marque.

Indexe	Marque plaignant	Marque défendant
1	la poste	posteasy
2	vachon	agenda vacher
3	prince	tek sun princess
4 (TGI)	belle color	loreen paris
5	san marina	ana s. marin
6	soleil voyage	sous le soleil tour operator
7	intel inside	sound cinema inside
8	élite	élite, biosthéticien élite cm partenaire des laboratoires la biosthétique marcel contier francine frantin
9	biotherm	bioderma ph6
10	scott & fox	american scott
11	elle	photo lab'elle accueil qualité service
12	plein sud	latitude sud
13 (CA)	belle color	loreen paris
14	hom	xom
15	bottin mondain	l'annuaire mondain le plus ancien de france
16	mag 2	oromag

TAB. 10.6 – Liste des enregistrements, source d'erreur de modélisation

- descripteur 2, nombre de caractères en communs entre la marque du défendant et la marque du plaignant : généralement, la moitié ou plus des caractères sont en communs. Nous observons que cette propriété est due au fait que la marque du défendant est généralement plus longue que celle du plaignant (décisions 3, 8 et 11).
- descripteur 5, même information que le descripteur 1, mais dans une approche phonémique : comme ce descripteur est identique au premier avec une transformation phonétique, les propriétés trouvées sont nécessairement proches.
- descripteur 6, même information que le descripteur 2, mais dans une approche phonémique : nous observons ici des propriétés similaires au descripteur 2, pour les même raisons que le descripteur expliqué précédemment.

Les descripteurs incriminés sont symétriques (graphémiques / phonémiques), et sont utilisés par combinaison pour exprimer une partie du phénomène. Les propriétés graphémiques représentent des contrefaçons alors que la partie phonétique caractérise des non-contrefaçons.

Il est utile d'observer que les descripteurs 1 et 5 caractérisent, pour ce sous-ensemble, une contrefaçon et que les descripteurs deux et six, identifient une non-contrefaçon.

Enfin, les exemples que nous avons isolés correspondent bien au domaine d'étude et ne sont donc pas à considérer comme n'appartenant pas au domaine de définition. Nous pouvons écarter la sélection des documents comme source d'erreur potentielle dans cette modélisation.

Ces informations et particularités mettent en avant un problème dans la modélisation des données. Les descripteurs mis en place ne répondent pas suffisamment à la discrimination des enregistrements pour les exemples litigieux. Il faut donc reprendre la définition des descripteurs afin d'en établir de nouveaux. Ces nouveaux descripteurs devront avoir une sémantique proche, mais une interprétation plus adéquate.

10.3 Reconsidération des choix du protocole de modélisation

Nous savons, grâce aux expériences précédentes, que les outils de modélisation ne sont pas responsables de la qualité des modèles et que le jeu de données semble porter préjudice à l'ensemble. Nous avons alors déterminé par le biais d'une approche exploratoire les enregistrements représentant une connaissance stratégique pour le domaine. Ces exemples peuvent être considérés ainsi car, avec les descripteurs mis en place, ils se retrouvent à la frontière entre contrefaçons et non-contrefaçons.

Nous connaissons aussi les limites de la définition du domaine, conduisant ces exemples à être considérés comme une connaissance stratégique. Ces limites sont essentiellement liées aux propriétés des marques mettant en faute le processus de modélisation.

Afin d'améliorer le modèle et donc de prendre en compte ces exemples de façon optimale, nous nous intéressons à la définition du domaine des marques nominatives et la façon d'améliorer les chemins au sein des choix possibles dans le processus de modélisation présenté dans la figure 8.4 page 124. Cette opération correspond à un cycle de la démarche de correction d'un modèle (démarche présentée dans la figure 8.1 page 116). Nous allons donc étudier les procédés à affiner pour atteindre les objectifs fixés et la possibilité d'améliorer le processus de modélisation.

10.4 Retour correctif sur les descripteurs

Les différentes expériences réalisées dans ce chapitre ont démontré que le jeu de données correspond bien au domaine de définition. En revanche la définition des descripteurs semble léser le modèle. Tout d'abord, nous utilisons des descripteurs continus qui pénalisent les algorithmes de modélisation. Ce choix est lié à l'utilisation d'un réseau de neurones dans la première version de `CATMINE` qui impose des descripteurs continus. De plus, certains descripteurs correspondent à une fusion de descripteurs et donc de propriétés. Cette fusion n'est pas nécessairement adéquate

pour répondre au phénomène. Enfin, ces descripteurs ne permettent pas de prendre en compte certaines données spécifiques. Cette spécificité est généralement liée à la longueur, en terme de caractères, des marques impliquées dans une décision.

10.4.1 Vers une nouvelle définition des descripteurs

Nous nous intéressons dans cette section aux descripteurs résultant d'une fusion de descripteurs. Ces descripteurs sont ceux liés à la longueur de la sous-chaîne graphémique ou phonémique d'une marque dans l'autre, pondérée par sa position dans cette marque. Les autres descripteurs expriment une information unique, et ne font donc pas l'objet d'une telle étude.

Les descripteurs caractérisant la longueur de la plus longue sous-chaîne commune d'une marque A dans une marque B relativement à la position de cette sous-chaîne dans la marque B sont définis comme suit :

1. la longueur en terme de caractères de la plus longue sous-chaîne commune
2. position de cette sous-chaîne commune dans la marque B :
   - 1,5 si cette chaîne est au début
   - 1 si elle est au milieu
   - 0,5 si elle est à la fin

L'information de position a été définie afin de différencier les similitudes entre marques en accordant plus d'importance pour une similitude au début de marque par rapport à une présence en fin de marque. Cette information peut être représentée sous la forme d'un graphique selon les équations affines suivantes : $a*x+b$ ou a prend la valeur de la position, x représente le nombre de caractères et b est nul. Nous obtenons alors le graphique 10.3 page suivante présentant ces trois équations (une par position possible de la sous-chaîne). Nous pouvons alors observer ce type de propriétés : si la plus longue sous-chaîne commune vaut 80% de la marque du plaignant, et que cette chaîne est présente en fin de marque du défendant, alors cela est équivalent à une chaîne représentant 40% en milieu de marque et une chaîne de 26,6% en début de marque.

$$\begin{cases} \frac{3}{2} * x = 0,4 & x = 0,2666 & \text{(début)} \\ x = 0,4 & x = 0,4 & \text{(milieu)} \\ \frac{1}{2} * x = 0,4 & x = 0,8 & \text{(fin)} \end{cases}$$

Ce type d'équivalence se retrouve bien évidement pour d'autres proportions et d'autres placements. En revanche l'information véhiculée par ce genre de descripteur pose un problème : cela revient à accorder la même importance à des phénomènes bien distincts. Une sous-chaîne commune très faible en début de marque est tout aussi importante qu'une marque quasiment reprise intégralement en fin de marque. Nous pensons alors que cette information n'est pas sémantiquement valable. La définition de ce descripteur par une fusion de deux informations a été introduite

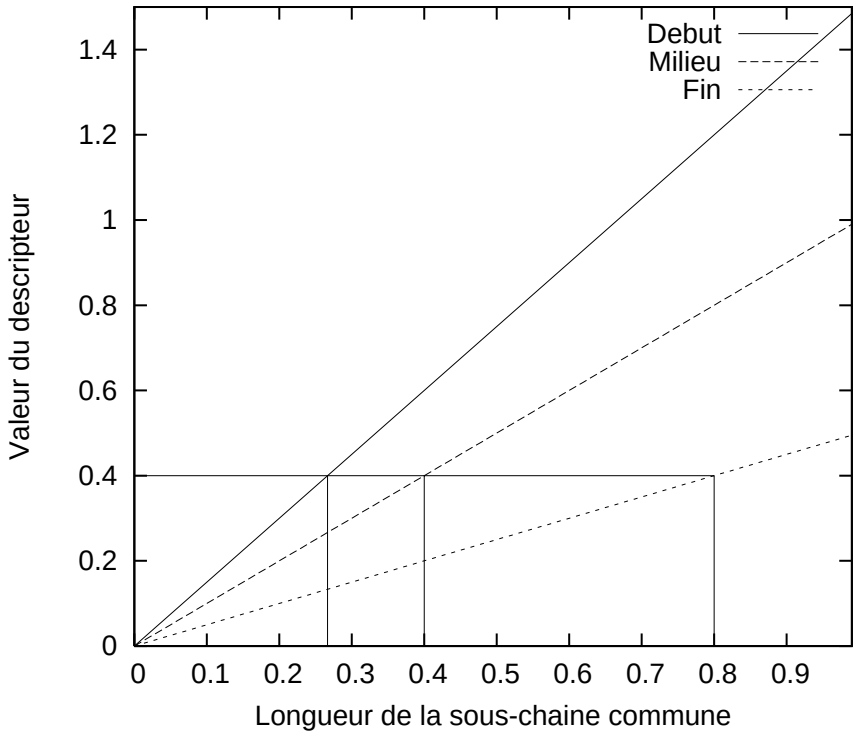


FIG. 10.3 – courbes représentatives des descripteurs présentés

pour illustrer qu'une chaîne reprise en début de marque a un impact plus important dans l'esprit du consommateur que lorsqu'elle est reprise en fin de marque. L'effet de bord observé n'était donc pas souhaité. Ce descripteur ne peut donc exister en l'état, et doit être reformulé afin de n'exprimer qu'une et une seule information.

La conséquence dans la modélisation est la création de quatre descripteurs supplémentaires afin de caractériser la position de la plus longue sous-chaîne commune entre la marque du plaignant et la marque du défendant, entre la marque du défendant et la marque du plaignant, ainsi que les mêmes calculs mais portant cette fois sur les transcriptions phonémiques des marques.

10.4.2 La segmentation des descripteurs continus

Dans la définition des descripteurs de `CATMInE` nous avons utilisé des descripteurs continus. Nous avons aussi rapporté que, dans la définition du domaine, de tels descripteurs sont contraignants pour les algorithmes de modélisation (autres que les réseaux de neurones). Nous nous sommes donc intéressé à l'étude d'une segmentation de ces descripteurs.

La disponibilité de l'expert pour cette étude n'ayant pu être assurée, nous avons alors étudié la répartition des enregistrements au travers de chaque descripteur afin d'observer s'il existe des seuils où il serait possible de les segmenter. Cette démarche n'a pas permis d'observer des seuils

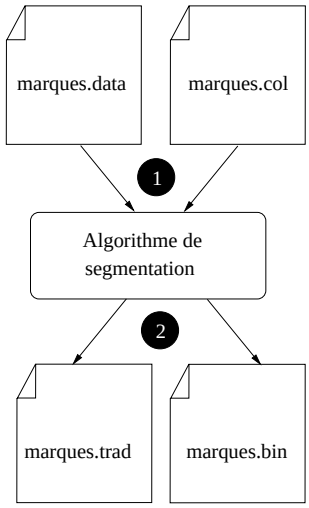


FIG. 10.4 – Principe de la segmentation des descripteurs d'une base de données.

`descripteur1` $<= 0.8 \rightarrow 1$
`descripteur1` $<= 0.9375 \rightarrow 2$
`descripteur1` $<= 1 \rightarrow 3$

TAB. 10.7 – Extrait du fichier `marques.trad` pour le descripteur 1

significatifs de segmentation. L'expression de la contrefaçon ne semble donc pas se traduire avec un seul descripteur, mais est le résultat d'une conjonction de descripteurs.

Nous avons alors orienté cette démarche de segmentation des descripteurs continus vers une approche plus naïve, en nous reposant sur les conclusions de [Zighed et al., 1999] rapportant le fait qu'aucune méthode de segmentation d'attributs continus (à l'exception prés du *ChiMerge*) est meilleure que l'autre. Notre approche a donc consisté à segmenter ces attributs en fonction de la répartition des enregistrements. Chaque descripteur est alors découpé en intervalles dont le nombre peut être prédéfini. Cette segmentation suit le principe du schéma 10.4.

A partir de la base de données brute (le fichier `marques.data`) et d'un fichier spécifiant pour chaque attribut, le nombre d'intervalles à calculer (le fichier `marques.col`), étape 1, l'algorithme de segmentation va produire deux fichiers, étape 2. Le premier, le fichier `marques.trad` présente les valeurs utilisées pour traduire chaque descripteur en fonction des seuils calculés ainsi que les contraintes exprimées par chaque seuil calculé. Par exemple, le fichier `marques.trad` exprimera les contraintes présentées dans le tableau 10.7 pour le descripteur 1. Pour tous les enregistrements du jeu de données, ceux dont la valeur du descripteur 1 est inférieure ou égale à 0,8, seront remplacés par 1, ceux pour lesquels la valeur sera comprise entre 0,8 (exclus) et 0,9375 seront remplacés par la valeur 2, et enfin tout ceux compris entre 0,9375 exclus et 1 seront remplacés

Descripteur	Premier seuil	Deuxième seuil	Troisième seuil
Descripteur 1	<=0.8	<=0.9375	<=1
Descripteur 2	0.666667	0.818182	1
Descripteur 3	0.545455	0.9	1.5
Descripteur 4	0.428571	0.818182	1.5
Descripteur 5	0.727273	0.916667	1
Descripteur 6	0.571429	0.75	1
Descripteur 7	0.444444	0.9375	1.5
Descripteur 8	0.4	0.75	1.5

TAB. 10.8 – Seuil déterminé par une approche naïve.

par 3.

Le second fichier, issu du processus de segmentation est le fichier `marques.bin`. Il correspond à la transformation de la base de données de départ en fonction du fichier `marques.trad`.

Nous avons fixé le découpage des descripteurs à trois intervalles, chacun contenant un tiers des données (quelque soit la valeur de classe des données). Cette approche naïve a alors calculé les seuils présentés dans le tableau 10.8.

Une expérimentation de ces transformations a permis d'augmenter le pouvoir décisionnel de notre modèle de 4%. Ce gain est tout de même significatif, faisant passer le taux de reconnaissance de `DeTTMInE` de 63.86% à 67.70%, tout en allégeant le calcul des modèles.

L'approche exploratoire que représente ce travail nécessite d'étudier pour chaque descripteur le nombre d'intervalles, en fonction des autres descripteurs. Ainsi il faudrait exécuter un processus calculant le nombre optimum d'intervalles pour chaque descripteur en tenant compte de l'ensemble des descripteurs.

Dans cette section nous avons vérifié si une telle démarche est possible pour notre cas d'étude et ce que nous pouvons en obtenir. Nous avons alors démontré que la segmentation des valeurs continues améliore le score de classification, mais que déterminer l'intégralité des modalités par valeur continues restait une tâche délicate à réaliser. Cela est essentiellement du à la complexité de l'algorithme à mettre en place. Intégrer l'expertise afin de déterminer les valeurs charnières permettrait d'optimiser ce traitement et de compléter le choix des valeurs par une information sémantique supplémentaire. Ce choix pourrait être établi en étudiant la répartition des exemples en fonction de la valeur de classe à travers les valeurs possibles du descripteur. Cette hypothèse est présentée dans l'histogramme 10.5 page ci-contre pour le descripteur 5, caractérisant le pourcentage de phonèmes de la marque du plaignant présents dans la marque du défendant. L'échelle est logarithmique, en rouge sont identifiées les contrefaçons et en bleue les non-contrefaçons. Nous

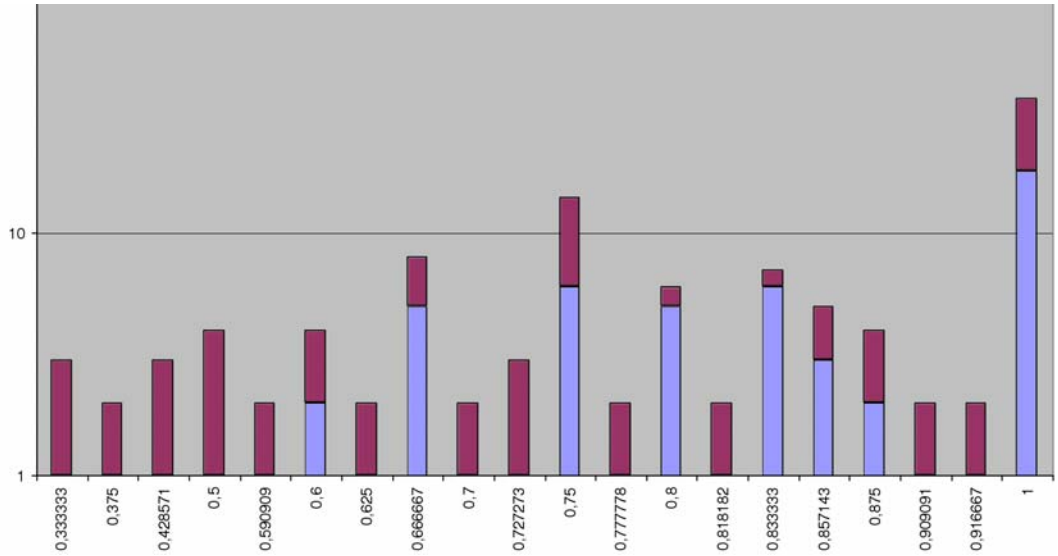


FIG. 10.5 – Répartition des exemples en fonction de la valeur de classe à travers les valeurs possibles du descripteur caractérisant le pourcentage de phonèmes de la marque du plaignant présents dans la marque du défendant.

pouvons alors établir le domaine de définition des contrefaçons comme la réunion des intervalles $[0; 0.5909] \cup [0.625; 0.666[\cup[0.7; 0.7272]\cup]0.75; 0.77] \cup [0.81; 0.83[\cup[0.90; 1[$.

10.5 Retour sur les documents en vue d'établir de nouveaux descripteurs

Dans cette section nous allons déterminer s'il existe des informations supplémentaires, dont il serait possible de tenir compte pour améliorer le processus décisionnel. Ces éléments ne peuvent être issus que de l'expertise sur le document, et plus particulièrement des éléments pris en compte pour la décision.

Dans de nombreuses décisions (mais pas l'intégralité), nous avons observé que le tribunal s'interroge sur la légitimité d'une partie à détenir une marque. Par légitimité nous sous-entendons le fait que la dénomination de la partie est en rapport avec la marque qu'elle détient. Par exemple, *The Coca Cola Company* a toute légitimité à détenir la marque *Coca Cola*®.

Nous avons donc observé à travers les décisions que si une entreprise n'a aucune légitimité à posséder une marque particulière, alors cette entreprise est déchue de sa marque au profit de l'autre partie.

Ainsi, il semble nécessaire d'introduire dans notre processus décisionnel la notion de légitimité d'une partie à détenir une marque. Le traitement consiste alors à vérifier que la marque est plus

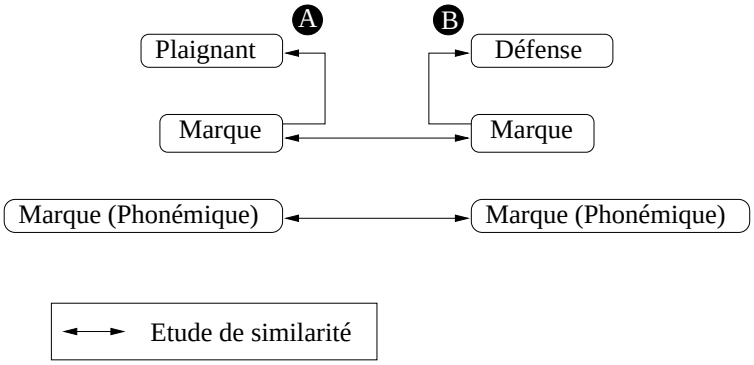


FIG. 10.6 – Relations de légitimité et de comparaison.

ou moins similaire au nom de la partie détentrice de celle-ci. Ce calcul peut être établi de la même manière que l'étude de similarité entre les marques constituant la décision.

La figure 10.6 présente les différents calculs de similarité à établir afin de déterminer si la marque du plaignant est détenue de façon légitime et si celle-ci reprend la marque du défendant. Ainsi, il sera utile de calculer pour chacune des parties si les marques sont détenues de manière légitime en établissant les comparaisons graphémique et phonémique de celles-ci avec leur partie respective (A, B). Cependant, certaines parties sont constituées de plusieurs personnes morales ou physiques. Il est nécessaire de prendre en compte cet aspect dans notre calcul. Nous proposons alors de ne vérifier cette propriété que dans un sens : celui de la marque dans la partie. Déterminer si le nom d'une partie est contenu dans une marque ne sera pas révélateur si celle-ci est composée de plusieurs personnes morales ou physiques. Il en est de même si la partie est un consortium tel que *Unilever* détenant plusieurs centaines (voir milliers) de marques.

Nous pouvons alors poser comme descripteurs supplémentaires, quelque soit la partie :

- la longueur de la plus longue sous-chaîne graphémique commune entre la marque du plaignant et sa partie, rapportée à la longueur de la marque
- la longueur de la plus longue sous-chaîne graphémique commune entre la marque du défendant et sa partie, rapportée à la longueur de la marque

Déterminer le nombre de caractères comme il est procédé dans la comparaison d'une marque est hasardeux, plus il y a de parties à détenir une marque et plus la probabilité de retrouver tous les caractères de cette marque dans les entités composant la partie sera forte. La comparaison entre marque reste inchangée, mais elle est complétée du descripteur représentant cette légitimité.

Cette propriété de légitimité n'intervenant pas dans toutes les décisions, celle-ci permet uniquement d'obtenir deux descripteurs supplémentaires pour regrouper les décisions et lever l'ambiguïté sur certaines. Cette propriété correspond bien à une des contraintes de la représentation du domaine, à savoir de caractériser les faits, et de ne pas être issue de la décision du juge (élément qui ne pourrait être déterminé à la lecture des faits).

Expériences	Position	Segmentation	Légitimité	score
Expérience 1	●	●	●	65.74%
Expérience 2	●	●	○	69.62%
Expérience 3	●	○	●	55.69%
Expérience 4	●	○	○	68.10%
Expérience 5	○	●	●	**71.63%**
Expérience 6	○	●	○	67.70%
Expérience 7	○	○	●	61.55%
Expérience 8	○	○	○	63.86%

TAB. 10.9 – Hypothèses prisent en compte et score obtenu pour chacune des modélisations réalisées (● hypothèse présente, ○ hypothèse absente)

10.6 Nouvelle validation de la modélisation établie

A l'aide des résultats observés et des hypothèses formulées tout au long de cette approche exploratoire, nous avons réalisé une nouvelle série de modélisations du domaine d'étude. Cette série de modélisation propose de prendre en compte les hypothèse suivantes pour caractériser le domaine :

- de segmenter les valeurs continues en trois ensembles
- d'intégrer la notion de légitimité entre les parties et leurs marques respectives
- de dissocier la position de la sous-chaîne (graphémique ou phonémique) relativement à sa taille.

Pour respecter le principe exploratoire de la création du processus décisionnel, nous allons comparer les différentes combinaisons possibles entre ces 3 hypothèses à la non prise en compte de celles-ci. Nous procédons ainsi à une sélection de descripteurs de type *wrapper* (cf. 5.3.1 page 76). Les expériences sont présentées dans la table de vérité 10.9, avec le score obtenu par l'algorithme `C4.5`, retenu pour réaliser cette série d'expériences.

Le graphique 10.7 page suivante synthétise les résultats obtenus pour chacune de ces expériences. Nous observons que les expériences 1, 2, 4, 5, 6 améliorent la qualité du modèle initial (expérience 8). Quatre de ces expériences (1, 2, 5, 6) sur les 5 impliquent la segmentation des données. Le meilleur résultat est obtenu avec la combinaison de la légitimité et de la segmentation des données (expérience 5). Les deuxièmes et troisièmes meilleurs modèles sont obtenus avec la position seule ou avec la combinaison de celle-ci et la segmentation des données.

L'expérience 5 permet de rapprocher le score de classification de `C4.5` à celui du réseau de neurones de `CATMInE` (modèle le plus efficace en prédiction) avec des scores respectifs de 71.69 % et de 72%. Pour les autres hypothèses, nous nous apercevons qu'il est délicat de se

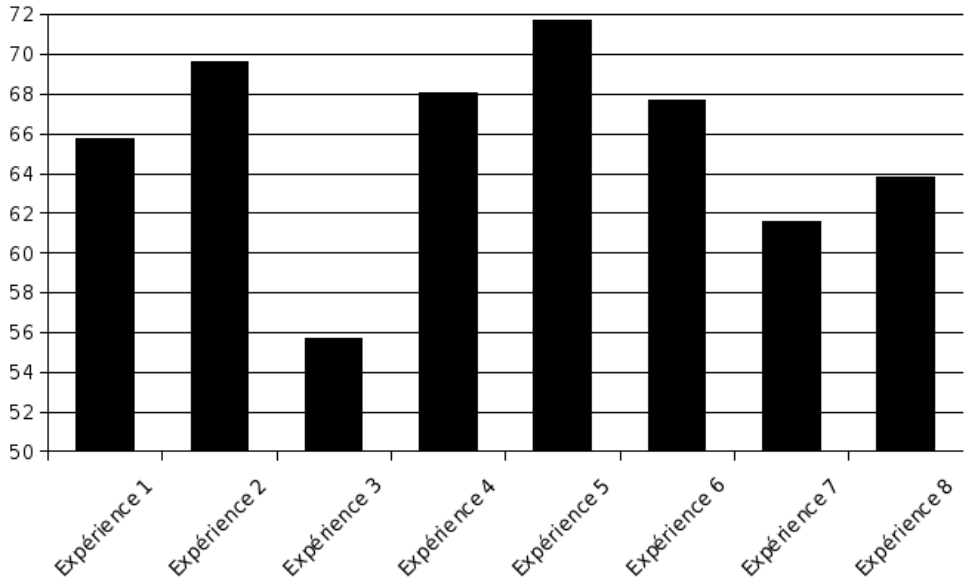


FIG. 10.7 – Résultats des expériences pour les nouveaux modèles possibles.

substituer à l'expert. Le descripteur caractérisant la longueur de la plus longue sous-chaîne commune, pondérée par sa position dans la marque ne semble plus poser problème. Pourtant nous démontrons auparavant que dans la recherche des descripteurs ne permettant pas de regrouper efficacement les données son rôle contribuait à baisser la qualité du modèle. Seule une approche exploratoire, menée par l'expert, et la définition de nouveaux descripteurs émanant de celui-ci permettraient d'améliorer la qualité du modèle recherché tout en prenant en compte la complexité que représente cette tâche.

Chapitre 11

CATMInE, une plate-forme d'aide à la décision en droit des marques

Sommaire

Nous avons présenté dans la partie précédente les fondements théoriques nécessaires à la réalisation d'un modèle issu d'une base de données. Nous avons alors réalisé un travail d'analyse approfondie afin d'identifier les problèmes inhérents à ce genre de processus : complexité des choix, maîtrise des outils de modélisation et d'évaluation de la connaissance. De l'ensemble de ces travaux, nous proposons alors dans ce chapitre le développement d'une plate-forme d'aide à la décision, permettant d'intégrer le savoir-faire de l'expert.

11.1 CATMInE : une première approche

Computer **A**ssisted **T**rade **M**arks **In**fringement **E**valuation

11.1.1 Les besoins de l'expert

Le thème général de ce travail de recherche a trouvé sa problématique et son originalité dans le domaine juridique. Ce projet, résultat d'une collaboration entre le cabinet conseil en propriété industrielle Breese Derambure Majerowicz (Paris) et l'Université de Caen[21], consistait à la mise en place d'un système de calcul du risque de contrefaçon entre des marques nominatives ([Renaux, 2003]).

En effet, ce genre de calcul représente une part de travail conséquente pour les juristes : le volume de données traitées pour mener une affaire est considérable. Ce volume se traduit par un ensemble de jurisprudences (jugements de tribunaux pour un litige), et n'est pas une représentation figée d'un concept, mais bien au contraire une évolution temporelle du monde juridique. Cette évolution est issue de celle des juristes qui à son tour est engendrée par l'évolution de la communauté. Les données décrivent des jugements qui présentent à certains niveaux des similitudes, mais dont les décisions associées peuvent différer.

Ces aspects d'évolution du monde entraînent la mise en place de nouveaux procédés d'apprentissage, capables de tenir compte de certains paramètres décrivant les décisions rendues. La mise en valeur de ces paramètres ne peut intervenir pour la prédiction, car ils sont non déterminés pour un nouvel exemple. Mais, ils sont indispensables pour le calcul du risque de contrefaçon, et doivent donc reposer sur les connaissances acquises.

Les besoins exprimés sont simples, mais leur réalisation un peu plus complexe. Pour chaque marque, nous ne nous intéressons qu'à son aspect nominatif. L'aspect graphique (logotype) est mis de côté, mais peut très bien faire l'objet de traitements similaires ultérieurement.

Lorsque qu'un industriel décide de déposer une marque pour valoriser ses produits, celui-ci consulte un cabinet d'experts en propriété industrielle. Ceux-ci ont pour fonction de déterminer si cette nouvelle marque risque de contrefaire une des 1.7 millions de marques enregistrées à ce jour en France. Si tel est le cas, lesquelles, et pourquoi ? Si ce n'est pas le cas, le dépôt peut alors être effectué dans les règles de l'art.

Pour réaliser cette tâche d'aide à la décision, l'expert a fourni une base de données (`JURINPI`), servant de support pour déterminer ce qui rend une marque potentiellement contrefaisante pour une autre. Ensuite, pour chaque nouveau dépôt, ou pour tout litige de marques déposées, le système doit pouvoir calculer un risque de contrefaçon et produire des jugements de référence.

Le deuxième cas de figure n'est pas improbable. Cette communauté d'experts travaille essentiellement à partir de documents papier, l'intégralité des marques disponibles leur étant fournie

[21] Contrat de recherche et développement, `CATMInE`, 2001, sous la direction scientifique de Thomas Lebarbé.

sur un support papier. La recherche d'une marque potentiellement contrefaite par une autre s'appelle la recherche d'antécédent. Le nombre de marques enregistrées est colossal, et ce travail de recherche se fait uniquement sur l'intuition de l'expert. Celui-ci n'est pas à l'abri d'une erreur et lui proposer des outils d'aide à la décision n'est pas pour lui déplaire.

11.1.2 Quelques principes

La réalisation d'une première étude du système d'aide à la décision a été réalisée par Thomas Lebarbé, et l'implantation de cette étude par nous-même en 2001. L'expert a été peu impliqué dans cette réalisation, mais les étapes clef lui ont été soumises et il les a validées. Nous entendons par étapes clefs, la production d'une base de données, la validation des descripteurs mis en place pour répondre à la problématique ainsi que la validation du modèle.

Les descripteurs mis en place, et validés par les experts ont été proposé par Thomas Lebarbé et sont au nombre de huit, plus une valeur de classe. Pour permettre une comparaison de marques, celles-ci doivent être comparées dans leur globalité. Pour ce faire, le système réalise une comparaison sur leur aspect graphémique (la façon dont elles sont écrites, indépendamment de la typographie mise en place) et leur aspect phonétique (la façon dont elles sont prononcées). Les descripteurs ont été conçus en fonction des critères des cours, consistant à vérifier qu'une marque ne crée pas de confusion (visuelle et sonore) dans l'esprit du consommateur.

L'obtention de la transcription phonétique se fait par un traducteur à base de règles contextuelles. Nous pouvons alors poser : *Soit `m1` la première marque et `m2` la seconde, à chacune d'elles est associée leur transcription phonémique respective, `p1` et `p2` (ce qui implique une traduction selon l'alphabet phonétique international (API), où un phonème est une unité minimale de prononciation),*

Les données seront alors représentées par les descripteurs suivants :

1. le pourcentage de caractères de `m1` présents dans `m2`
2. le pourcentage de caractères de `m2` présents dans `m1`
3. le pourcentage de la plus grande sous-chaîne graphémique de `m1` dans `m2` pondéré par un indice de position de la plus longue sous-chaîne graphémique de `m1` dans `m2`
4. le pourcentage de la plus grande sous-chaîne graphémique de `m2` dans `m1` pondéré par un indice de position de la plus longue sous-chaîne graphémique de `m2` dans `m1`
5. le pourcentage de phonèmes de `p1` dans `p2`
6. le pourcentage de phonèmes de `p2` dans `p1` `p1` dans `p2`
7. le pourcentage de la plus grande sous-chaîne phonémique de `p1` dans `p2` pondéré par un indice de position de la plus longue sous-chaîne phonémique de `p1` dans `p2`
8. le pourcentage de la plus grande sous-chaîne phonémique de `p2` dans `p1` pondéré par un indice de position de la plus longue sous-chaîne phonémique de `p2` dans `p1`

9. le jugement, 0 si non-contrefaçon, 1 si contrefaçon

Pour exemple, si l'on étudie la paire de marques (GOLF PLUS, GOLF'US)[22] au niveau graphémique nous pouvons observer :

- *GOLF PLUS* contient l'ensemble des lettres de *GOLF'US*
- *GOLF'US* contient 6 lettres sur 8 de *GOLF PLUS*
- *GOLF* est la plus longue sous-chaîne commune entre les deux marques
- La sous-chaîne *GOLF* correspond au mot d'attaque des deux marques

Au niveau phonémique, les traitements sont identiques, mais appliqués sur la traduction phonétique de la paire de marques. Cette traduction repose sur un système à base de règles contextuelles ([Renaux & Zreik, 2004], [Morel & Lacheret-Dujour, 1998], [Morel & Lacheret-Dujour, 2001]). En effet, l'utilisation d'un lexique phonétique ne permet pas la traduction de l'ensemble des marques qui peuvent être des néologismes (qui par définition ne peuvent être présents dans un lexique). Ainsi, un système à base de règles semble le plus adéquat : maintenance aisée, possibilité de représentation des néologismes et intégration des sigles en font ses principaux atouts. La table 11.1 page suivante présente un extrait de ce fichier de règles contextuelles pour la lettre *i*.

Ces règles sont conservées sous forme arborescente. Cette arborescence est symbolisée par le caractère `@` : plus le niveau de profondeur est important, et plus le graphème recherché sera précis. Nous avons mis en gras un exemple caractérisant les mots de type *ralliement, maniement*. Il s'agit ici du cas particulier d'un *i* en fin de mot suivi du graphème *ement*. Le caractère `#` précise l'absence du graphème recherché en début (avant le graphème) ou en fin de mot (après le graphème), ou les deux. Le graphème à traduire se trouve entre parenthèses. Nous précisons, lorsqu'il s'agit d'un cas particulier, le contexte dans lequel le graphème recherché doit se trouver pour que la règle soit applicable. Dans notre exemple, le graphème ne doit pas être présent en début de mot (présence du $\#$) et doit être suivi du graphème *ement*. La règle est ensuite complétée du phonème correspondant au graphème (i) et enfin une série d'exemples se rapportant au graphème sélectionné (*ralliement, maniement...*) est présentée pour caractériser la règle.

11.1.3 Premier cas de figure : évaluation de la contrefaçon

L'évaluation de la contrefaçon entre deux marques nominatives s'effectue par le calcul des descripteurs linguistiques entre les deux marques mises en opposition. Ces calculs reposent à la fois sur la représentation graphémique de la paire de marques, mais aussi sur leurs transcriptions phonétiques. Ces calculs dépendent aussi des pays mis en jeu pour la comparaison. Les traductions phonémiques et les bases d'apprentissage changent en fonction des pays ciblés. Le résultat de ces calculs est alors conservé dans un vecteur de similitudes linguistiques, un vecteur par pays.

[22] Golf House (ste, Italie) contre P. Schmidlin, contrefaçon par imitation, Tribunal de Grande Instance de Paris (Ch.03), 01/09/1999

|#(i)#	i	(virage, ...)
@|#(i)́#	i	(Abréviation i.)
@|#(i)voy#	y	(fiable, tiens, lion, ...)
@@|#(ie)	i	(amie, pie, vie, ...)
@@|#(i)ement	**i**	**(ralliement, maniement, ...)**
@@|#(ient)	i	(lient, rallient, rient, ...)
@|#voy(il)	y	(bail, réveil, seuil, fauteuil, ...)
@|#voy(ill)#	y	(paille, vieille, feuille, ...)
@@|#v(il)l#	i	(village, ville, ...)
@@|(il)l#	il	(illétré, illusion, ...)
@|#(im)cns#	î	(impôts, ...)
@@|#(im)m#	im	(immense, immeuble, immédiat, ...)
@|#(in)	î	(malin, ...)
@|#(in)cns#	î	(mince, ...)
|#(î)#	i	(abîme, dîme, ...)
|#(ï)#	i	(naïf, héroïne, ...)

TAB. 11.1 – Extrait du fichier de règles contextuelles pour la transcription phonémique dans `CATMInE`.

Une fois les descripteurs déterminés, il est alors possible de comparer le cas d'étude à l'ensemble de la jurisprudence par le biais du réseau de neurones préalablement entraîné sur la jurisprudence. Le résultat d'une telle comparaison se traduit par une score de contrefaçon, par pays d'étude, score compris entre 0 et 100 (proche de 100, le risque est important, proche de 0 le risque est faible). La figure 11.1 page suivante présente l'application CATMInE et la disposition de l'information. Le score de classification étant présenté dans la partie supérieure de la fenêtre. De plus à l'aide des vecteurs de similitudes linguistiques, il est possible de représenter le cas d'étude, comparativement à l'ensemble de la jurisprudence. Cette représentation est établie par trois graphiques caractérisant les proximités entre jugements (partie inférieure de l'application présentée dans la figure suivante). Ces graphiques sont calculés par le principe de la projection de Sammon ([Sammon, 1969], [Biswas et al., 1981]), permettant de projeter un ensemble de points (un par jugement) dans un espace à n dimensions vers un espace à m dimensions avec $m < n$, tout en conservant les distances entre les points. Nous obtenons alors une cartographie de la jurisprudence, permettant l'étude des jugements proches à la paire de marques en cours de comparaison.

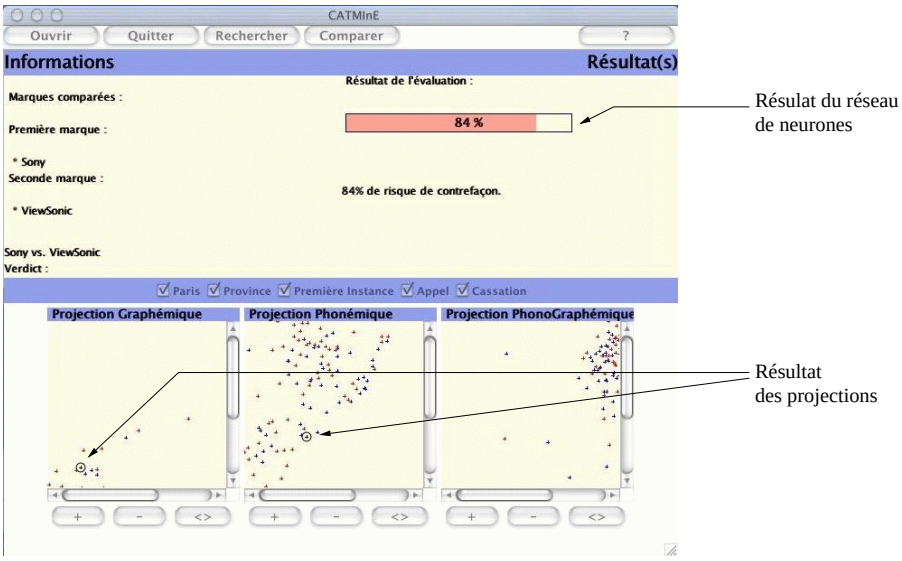


FIG. 11.1 – Résultats détaillés de la comparaison entre les marques Sony et Viewsonic pour la jurisprudence française

11.1.4 Deuxième cas de figure : la recherche d'antécédents

La recherche d'antécédents consiste à déterminer si une nouvelle marque, en cours de dépôt, risque de contrefaire une marque existante pour une série de classes de produits et services ciblées. Cette recherche représente un long travail d'investigation pour les professionnels et, est simplifiée par la base de données établie pour CATMInE.

En interrogeant la base de données, la liste des marques pour une série de classes de produits et services ciblés peut être obtenue pour être ensuite comparée une à une à celle en cours de dépôt.

Un vecteur de similitudes linguistiques est calculé pour chaque paire de marques et le risque de contrefaçon évalué (même procédure que pour la comparaison de marques). Le résultat d'une telle recherche est la liste, ordonnée par ordre décroissant du score de reconnaissance, de l'ensemble des marques obtenues pour une série de classes de produits et services ciblée.

Le juriste pourra ensuite interpréter les résultats et déterminer une stratégie pour les cas dont le risque de contrefaçon est élevé.

11.1.5 Troisième cas de figure : évaluation de la doctrine

Le premier point intéressant de ce troisième cas d'utilisation de CATMInE est l'évaluation d'une décision récente. En effet, le réseau de neurones permet de déterminer le degré de similarité avec la jurisprudence et les projections indiquent les proximités graphiques de cette nouvelle décision par rapport à cette jurisprudence. Ce cas d'utilisation permet alors au juriste d'avoir de

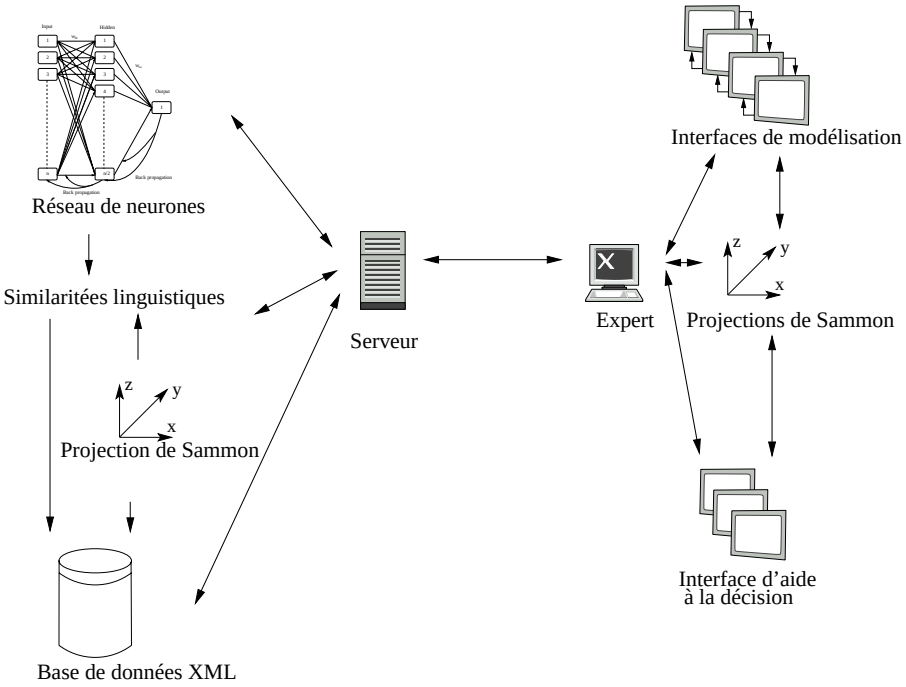


FIG. 11.2 – Architecture client serveur de CATMInE

nouveaux outils pour l'analyse et l'explication d'une nouvelle décision.

Le second point est le suivi de la jurisprudence dans le temps. En effet, grâce à l'ensemble des jugements, mis à jour régulièrement, il est possible d'étudier l'évolution de la jurisprudence comme par exemple l'étude des verdicts rendus par une cours donnée. Il est aussi possible d'étudier les revirements de jurisprudence, voire d'observer des signes avant-coureurs de ces revirements afin de mieux les anticiper.

11.2 La seconde version de CATMInE

Une seconde version de CATMInE a été produite ([Renaux, 2005b]). Celle-ci se base sur une architecture de type client serveur, présentée dans la figure 11.2. L'interface a été retravaillée pour une meilleure ergonomie et pour être intégrée directement à l'intranet du cabinet Breese. Les fondements restent les mêmes et aucune amélioration sur la formalisation du problème n'a été intégrée (contraintes contractuelles). Seule la maintenance et l'architecture logicielle ont été améliorées.

11.2.1 Une nouvelle architecture

L'expert retrouve bien évidement les trois applications principales de la première version :

1. comparaison de marques nominatives

2. recherche d'antécédents
3. étude de la jurisprudence disponible

La figure 11.3, présente de tels résultats pour la comparaison entre les marques Velux et Faelux[23] avec la nouvelle interface. Sur le principe de la projection de Sammon, et plus particulièrement sur

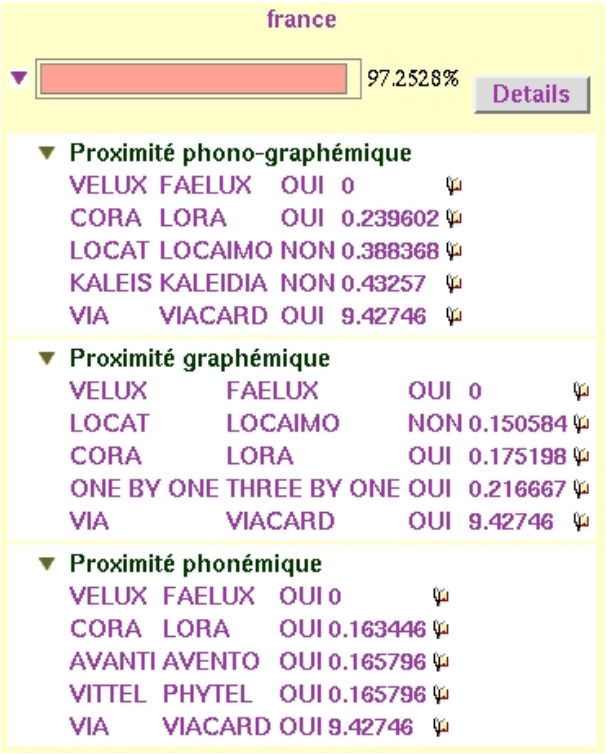


FIG. 11.3 – Résultats détaillés de la comparaison entre les marques Velux et Faelux pour la jurisprudence française

la notion de distance, CATMInE propose pour chaque résultat de comparaison produit par le réseau de neurones, la liste des cinq jugements les plus proches de la paire de marques en cours d'étude, par type de projection possible. De cette liste le juriste a accès au texte complet du jugement, lui permettant ainsi d'orienter ses recherches pour défendre au mieux les intérêts de ses clients. La figure 11.4 page suivante présente le résultat d'une comparaison phono-graphémique (mais il y a aussi une cartographie graphémique et une phonémique) sur la jurisprudence française. Les contrefaçons sont symbolisées par l'utilisation de la couleur rouge, la couleur bleu étant associée aux cas non contrefaisants.

L'expert analyse d'abord la page de résultats, présentée dans la figure 11.3 en prenant connaissance des 5 jugements de référence proposés par projection. Ensuite, par une étude de la cartogra-

[23] V KANN RASMUSSEN INDUSTRI A/S (Ste, Danemark) et VELUX FRANCE (SA) contre S.N.C. FAELUX DI FANTINI SERGIO EC (Ste, Italie), COUR D'APPEL DE PARIS (CH.04), 25/02/2002

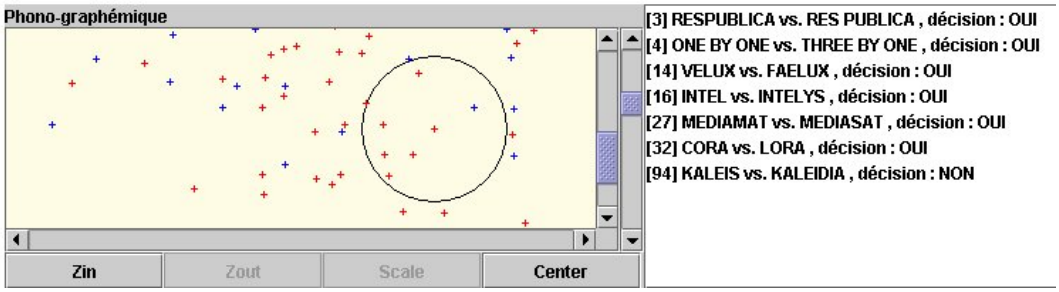


FIG. 11.4 – Projection Phono-graphémique pour la comparaison des marques Velux contre Faelux sur la jurisprudence française

phie associée, il retrouve le procédé de proximité proposé, complété d'une notion de disposition spatiale. Le fait de pouvoir obtenir une telle représentation permet à l'expert de mieux comprendre les résultats suggérés : il peut à la fois visualiser les jugements de référence par rapport à son cas d'étude et voir comment ces jurisprudences s'agencent dans l'espace de projection. Ce principe, mis en place pour compléter la compréhension des résultats, contribue à accréditer le système d'aide à la décision construisant son argumentation autour de ces décisions.

Si seul le risque de contrefaçon était proposé, l'expert n'aurait alors aucun moyen de comprendre les choix proposés et d'en retenir des connaissances pour le futur, permettant d'apporter du crédit au système.

Enfin, pour la recherche d'antécédents, l'innovation porte sur le filtrage des résultats en fonction d'un seuil fixé par l'utilisateur. La figure 11.5 page suivante présente le résultat d'une telle recherche.

11.2.2 Une plate-forme de modélisation aboutie

Cette plate-forme de modélisation, présentée dans [Renaux, 2006], permet de répondre à l'ensemble des éléments proposés dans le chapitre 2.3 page 30. La figure 11.6 page suivante présente une vue globale de l'intégralité des fonctionnalités de cette plate-forme.

La racine de cette arborescence représente le point d'entrée de la plate-forme CATMInE. De cette page d'accueil l'expert a la possibilité d'utiliser l'outil d'aide à la décision présenté précédemment (11.2 page 171, comparaison ou recherche d'antécédents). Le troisième choix disponible est relatif à l'administration du système. Cette partie administrative peut être restreinte à certains utilisateurs par le biais d'un accès contrôlé.

Dans la partie administrative, l'expert peut alors consulter les jurisprudences intégrées au système, ajouter de nouvelles jurisprudences (selon un formalisme précis) et gérer une jurisprudence particulière.

Ce troisième aspect est le plus intéressant dans le cadre de notre plate-forme d'aide à la déci-

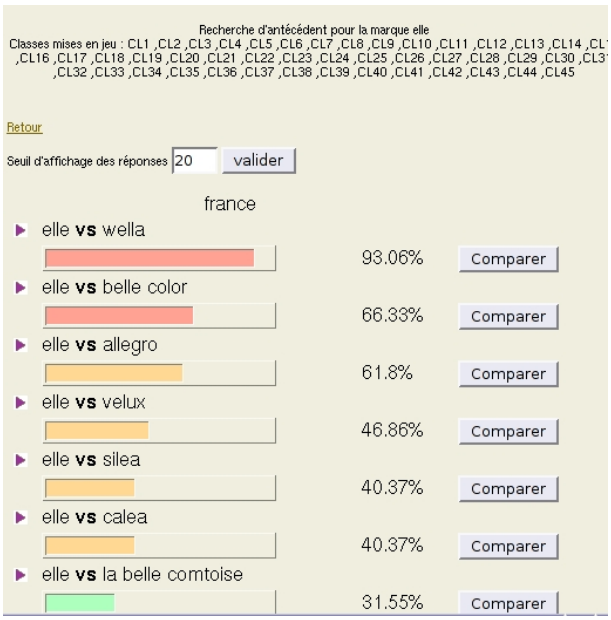


FIG. 11.5 – Recherche d'antécédents français pour la marque Velux dans l'intégralité des classes disponibles (45 classes de produits et services)

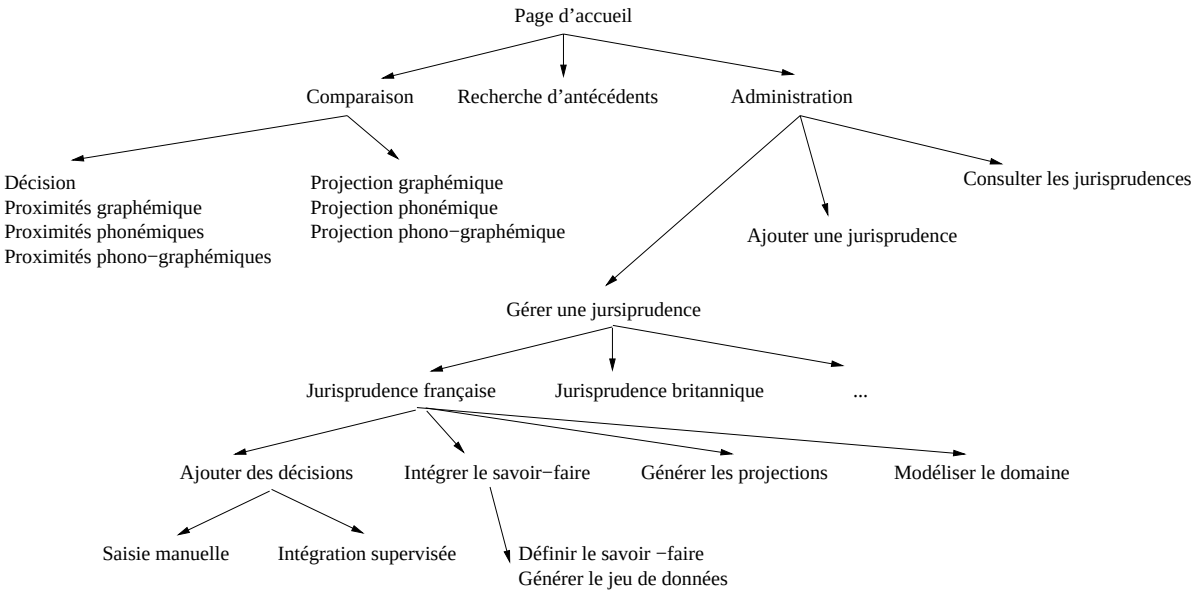


FIG. 11.6 – Vue d'ensemble de la plate-forme de modélisation CATMInE.

sion. Ainsi, en sélectionnant la jurisprudence à administrer, l'expert a la possibilité de compléter l'ensemble des décisions de deux manières :

1. ajout d'une décision manuellement
2. ajout d'un ensemble de décisions de façon supervisée.

Pour le premier cas de figure, l'expert a jugé la décision pertinente et c'est une décision qu'il a isolée de lui même. Dans le second cas de figure, nous faisons référence à des décisions issues de la base `JURINPI`. L'expert dispose alors d'un volume conséquent de décisions juridiques électroniques à intégrer au système, et la plate-forme lui propose de les traiter de façon semi-automatique : l'information utile est isolée et l'expert doit la valider avant de l'intégrer à l'ensemble des décisions constituant la jurisprudence.

Outre l'aspect d'alimentation en décisions du système, l'expert peut aussi intégrer son savoir-faire en définissant des opérations sur les éléments des décisions et produire une nouvelle base de données de modélisation. Lorsque cette base est définie, l'expert peut alors générer des projections ainsi que le modèle découlant du jeu de données. Les projections peuvent être utilisées dans l'aide à la décision ainsi que dans le cadre de la compréhension des résultats de la modélisation obtenue.

L'extraction d'informations

Pour palier le problème d'extraction d'informations présenté dans le chapitre 4 page 54, l'une des meilleures solutions reste encore de laisser à l'expert le soin de contrôler les informations retenues. Pour cela, des outils d'aide à l'extraction interactive d'informations doivent être mis en place afin d'assister l'expert dans cette tâche.

L'un des avantages de la base `JURINPI` reste sa structure `XML` (et ce même si celle-ci est plus que légère). Moyennant quelques règles d'extraction d'informations, propres à chaque balise, nous pouvons proposer à l'expert une information relativement pertinente.

Cependant, présenter l'information est une chose, être sûr qu'elle est bonne en est une autre. L'expert évaluera la pertinence de l'information extraite, et devra être en mesure de la corriger le cas échéant. Cette correction doit se faire de la façon la plus simple et la plus intuitive possible.

Afin d'assister l'expert dans une démarche d'extraction d'informations, une application fenêtrée, présentée dans la figure 11.7 page suivante, a été développée.

Cette application se décompose en deux parties : une partie supérieure présentant les informations extraites et une partie inférieure, contenant le document, épuré du cartouche de présentation (cf. 4.3 page 61).

La partie supérieure est composée de champs présentant l'information isolée automatiquement à partir du cartouche de présentation (figure 1.1 page 21). Devant chaque information présentée, un bouton rappelle ce à quoi celle-ci se rapporte (comme par exemple dans la figure 11.7 page suivante la valeur `M19970074` étiquetée par le label `Ref` pour référence). Le choix

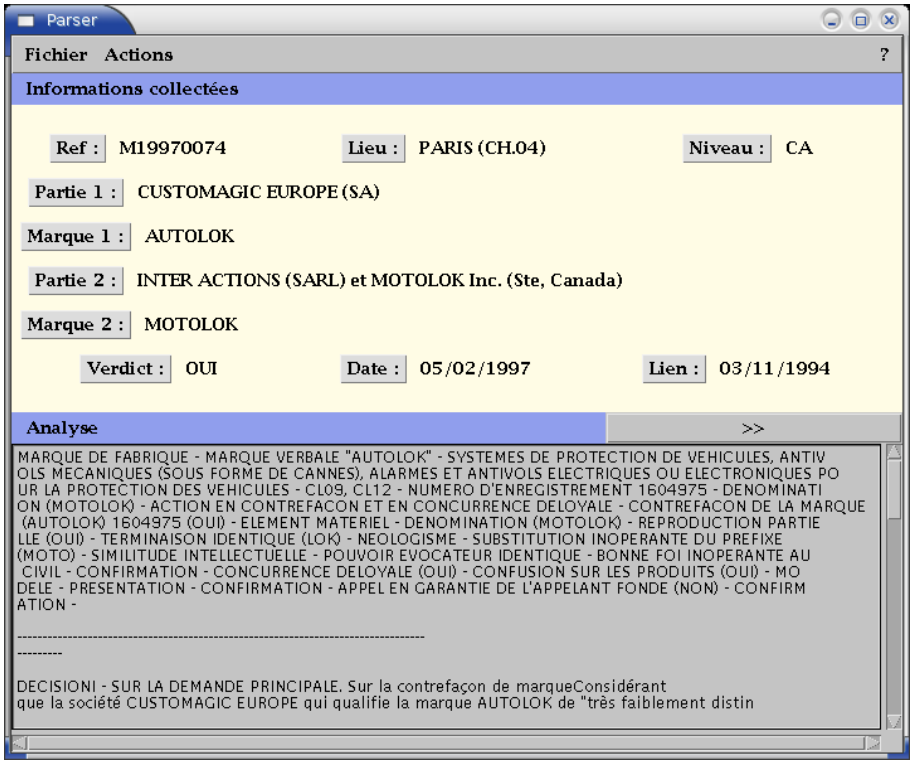


FIG. 11.7 – Outil d'extraction d'informations supervisée

d'utiliser un bouton n'est pas anodin. Lorsque l'information présentée n'est pas pertinente aux yeux de l'expert, celui-ci peut, à l'aide de la souris, surligner la zone de texte correspondant, dans la partie inférieure de l'application, et modifier l'information retenue en cliquant le bouton associé à celle-ci.

La figure 11.8 page ci-contre explique comment le système procède pour extraire l'information de la date du jugement. Il y a deux alternatives. La première, la balise `DAT` est renseignée. Le système se contente alors d'extraire l'intégralité du contenu et de remettre la date dans un style d'écriture française. Dans la seconde alternative, la balise `DAT` n'est pas renseignée. L'information reste toutefois disponible dans la balise `DOC` décrivant le lieu du jugement. L'extraction est un peu plus complexe, il faut isoler l'information présente dans le motif suivant :

– <DOC>w+(w+),*la date a extraire*</DOC>

Ce motif se lit comme l'information à extraire se situe entre :

1. une balise *DOC* ouvrante suivie de plusieurs mots (w+) suivi d'un parenthésage contenant un ou plusieurs mots ((w+))
2. une balise *DOC* fermante.

Une fois isolée, la date peut être réécrite et conservée au sein du modèle relationnel. A l'aide de ces mécanismes, nous avons pu établir une nouvelle banque de données qualifiée de rigoureuse. Les

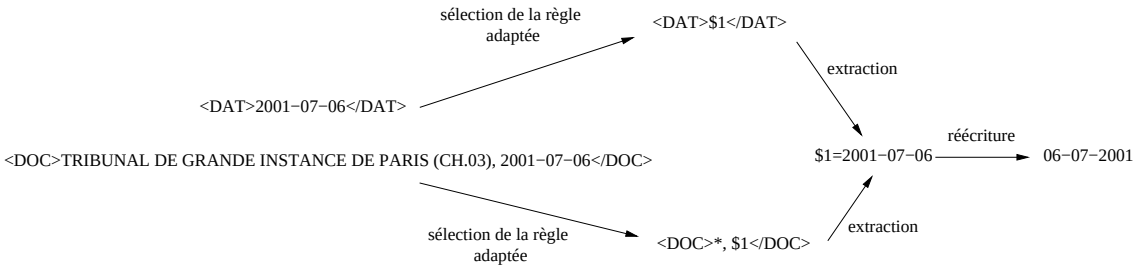


FIG. 11.8 – Principe d'extraction automatique d'informations : exemple de la date du jugement.

jugements insérés ont tous été contrôlés et l'information proposée jugée pertinente. Cette banque de données comptabilise 105 cas allant de 1997 à 2002. Cette correction du jeu de données n'a pas pour autant augmenté la qualité de l'apprentissage, cependant, elle a permis de s'assurer de la cohérence des exemples et de l'élimination des sources potentielles d'erreurs pour la modélisation.

Principe de l'intégration de l'expertise au jeu de données

Dans le but de formaliser le savoir-faire de l'expert et plus généralement son expertise, nous proposons un ensemble d'interfaces adaptées à cette tâche. Cette solution de formalisation d'un domaine est définie par l'intermédiaire d'un ensemble de pages internet, intégrées à la plate-forme, permettant à l'expert de définir sa connaissance en adaptant par l'intermédiaire d'opérateurs la représentation du domaine. Dans la figure 11.9 page suivante, nous présentons l'approche retenue afin d'aider un expert à formaliser son savoir-faire. La notion d'opérateur est introduite par le fait que l'expert dispose de l'information (les documents juridiques électroniques) mais que son savoir-faire implique une transformation de cette information pour expliciter un phénomène, dérivé de l'information. Cette nouvelle information représente alors la manière dont l'expert interprète l'information existante pour résoudre une tâche.

Dans la première étape de la figure 11.9 page suivante (A), l'expert a accès à sa base de données de jugements, relatives à ses habitudes de travail quotidiennes. Il peut alors ajouter des transformations sur certains attributs en interagissant avec le bouton *Ajouter*. Dans l'étape (B), l'expert peut choisir certains attributs parmi l'ensemble des attributs disponibles et leur appliquer des opérateurs ciblés afin d'établir de nouveaux documents, dérivés des jugements initiaux. De nouveaux opérateurs peuvent être créés en appuyant sur le bouton *Créer*. Dans l'étape (C), une autre page internet est affichée pour définir un nouvel opérateur. Dans l'étape (D), l'expert peut vérifier le bon fonctionnement de l'opérateur. Si c'est le cas, l'opérateur est alors ajouté à la liste des opérateurs disponibles, et dans le cas contraire, l'expert peut le corriger afin de le valider à nouveau.

Lorsque toutes les transformations sont définies, l'expert peut alors cliquer sur le bouton *Générer* de la première page afin d'appliquer toutes les transformations à la base documentaire.

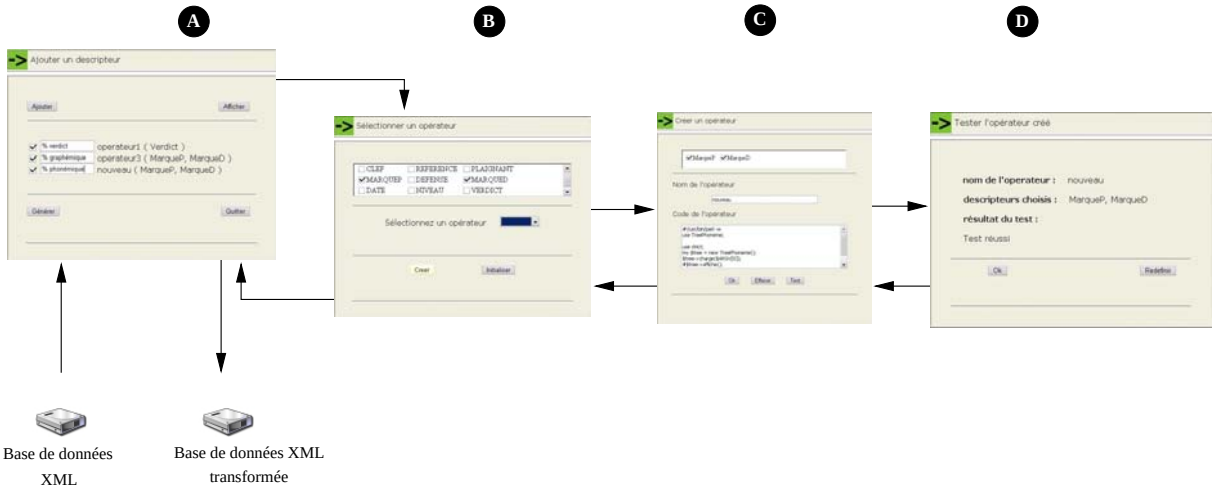


FIG. 11.9 – Processus de création d'une base de données intégrant le savoir-faire de l'expert.

L'expert dispose alors d'une nouvelle base de données décrivant toujours le domaine d'étude et intégrant en plus son expertise. Cette nouvelle base pourra alors être utilisée afin de modéliser l'activité de l'expert.

Intégration de l'expertise étape par étape

Nous allons maintenant revenir en détail sur chacune des pages développées et présentées précédemment pour permettre à l'expert d'intégrer son savoir-faire dans une base de données décrivant des documents électroniques juridiques. Nous avons présenté un principe reposant sur un ensemble de quatre pages. La première page, détaillée dans la figure 11.10 page suivante, permet à l'expert de visualiser les transformations qu'il va appliquer au contenu d'une base de données. A l'ouverture de l'application, cette page est bien évidement vide de toute transformation.

L'expert peut alors choisir d'ajouter de nouvelles transformations aux données en choisissant le bouton *Ajouter*. Une fois que l'expert a défini l'ensemble des transformations qu'il souhaite appliquer (dans notre exemple l'expert a déterminé trois transformations), il peut appuyer sur le bouton *Générer* afin de créer la nouvelle base de données issue de son expertise. Il peut le cas échéant désélectionner un opérateur s'il l'estime finalement non pertinent. Le bouton *Afficher* permet à l'expert de consulter la base de données sous forme tabulaire. Enfin, le bouton *Quitter* quitte l'application, sans appliquer les transformations sélectionnées.

La figure 11.11 page ci-contre présente la page permettant à l'expert de choisir un opérateur ainsi que les descripteurs sur lesquels appliquer l'opérateur afin d'intégrer à la base de données son savoir-faire.

Ce choix de l'opérateur est réalisé en sélectionnant les descripteurs (par l'intermédiaire des cases à cocher associée à chacun d'entre eux). La liste des opérateurs est alors recalculée pour ne

FIG. 11.10 – Interface de gestion de la transformation d'une base de données

FIG. 11.11 – Interface de définition d'une transformation

proposer que les opérateurs d'aritée correspondant au nombre de descripteurs sélectionnés. Une fois l'opérateur sélectionné, l'expert peut alors valider son choix par l'intermédiaire du bouton *Initialiser*. Cette action ramène alors l'expert sur la fenêtre de la figure 11.10 page précédente avec la liste des opérations mises à jour, en tenant compte de son nouveau choix.

Si l'expert ne trouve pas d'opérateur adapté à ses besoins, il a alors la possibilité d'en créer un afin de répondre à son expertise sur l'information retenue (bouton *Créer*). Ce choix entraîne alors l'expert sur la fenêtre présentée dans la figure 11.12.

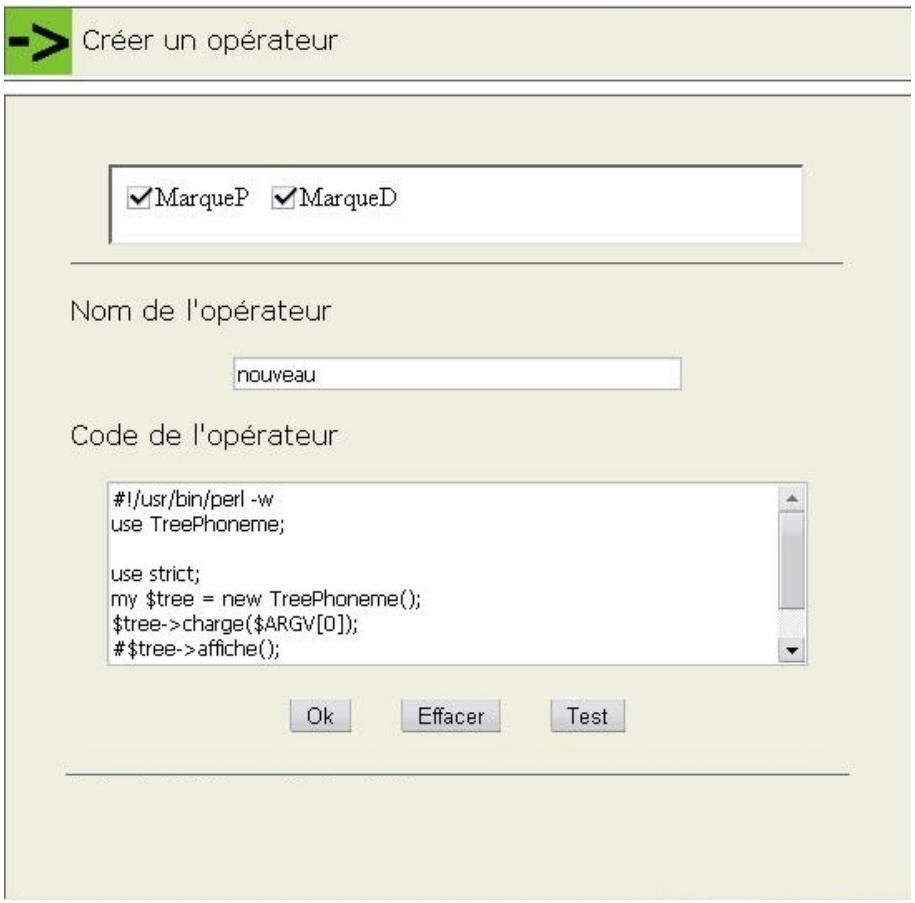


FIG. 11.12 – Interface de création d'une transformation

A ce stade de la modélisation du domaine, l'expert a la possibilité de créer un nouvel opérateur caractérisant son savoir-faire et plus particulièrement la façon dont il traite l'information qu'il a sélectionnée. Cela peut par exemple correspondre à isoler une information précise, ou encore à appliquer un calcul numérique ou linguistique sur cette information.

Ainsi, afin d'y parvenir, cette interface permet la saisie de l'opérateur via une fenêtre de saisie. Le haut de l'interface rappelle les descripteurs contenant l'information ciblée par ce traitement. La saisie de l'algorithme d'un nouvel opérateur se fait sur le même principe que l'écriture d'un

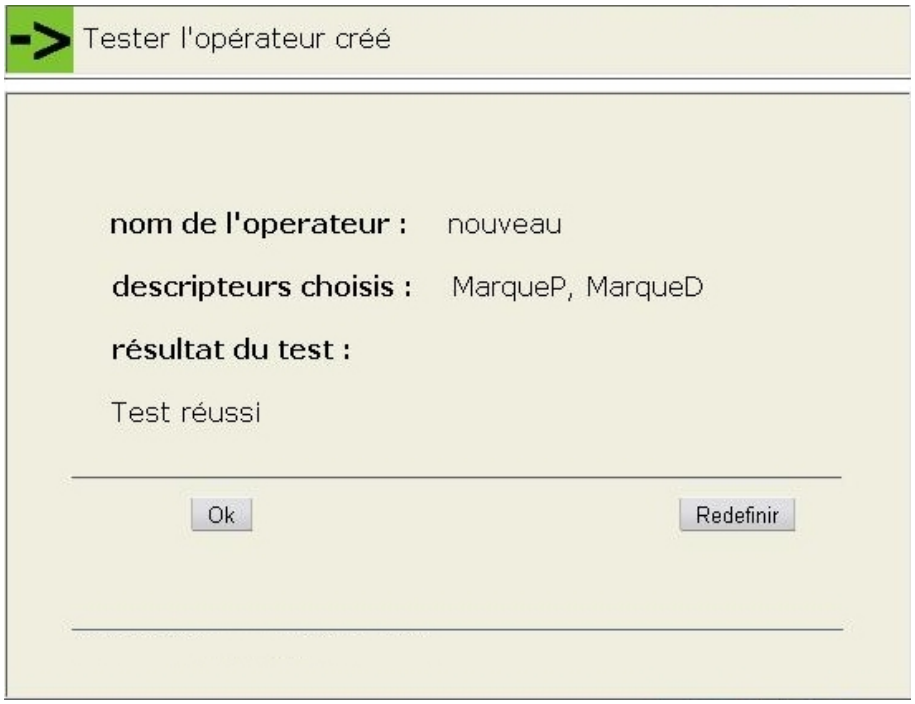


FIG. 11.13 – Interface de validation d'une transformation

script :

- la première ligne spécifie le langage de programmation comme par exemple :
  1. `#!/bin/sh` pour le `shell`
  2. `#!/bin/perl` pour le langage `perl`
- le corps du programme contenant l'algorithme de l'opérateur

Une fois l'opérateur saisi, l'expert peut alors choisir de le valider (bouton *Test*) sur l'intégralité du jeu de données afin d'observer le comportement de l'opérateur. Les résultats sont alors présentés sur le principe de la figure 11.13. Nous n'observons ici que le succès ou non de l'opérateur, il n'est pas possible d'afficher les résultats pour chaque enregistrement.

L'écriture de l'algorithme ne peut se faire sans connaissance des langages de programmation. Il est en effet difficile d'envisager des outils d'écriture intuitive de programme par le biais de page internet. Les langages de script semblent donc particulièrement adaptés pour le traitement des bases de données. Ils regroupent généralement des outils adaptés pour ce type de traitement : de nombreuses informations sont conservées par les systèmes d'exploitation (surtout vrai pour les systèmes de type `Linux` et `Mac OSX`) sous forme de bases de données, nécessitant certains post-traitements pour leur gestion.

Si l'expert ne souhaite pas valider ce nouvel opérateur, il peut alors le conserver directement dans l'ensemble des opérateurs disponibles en validant son algorithme à l'aide du bouton *Ok*.

Enfin, Le bouton *Effacer* permet d'effacer le contenu des zones de saisie pour en effectuer une nouvelle. Une telle action ne conserve pas ce qui a été saisi et non validé auparavant.

Une fois tous les opérateurs créés (si nécessaire) et sélectionnés, l'expert revient sur la fenêtre présentée dans le figure 11.10 page 179. Pour produire le nouveau jeu de données intégrant son savoir-faire, il lui suffit alors d'appuyer sur le bouton *Générer*. L'expert pourra ensuite visualiser le jeu de données produit en utilisant le bouton *Afficher* de cette même fenêtre.

11.2.3 Notes techniques sur la plate-forme CATMInE

Sans entrer dans les détails de la programmation, nous voulons présenter brièvement dans cette section les choix techniques réalisés lors de la mise en place de l'application CATMInE, répondant au schéma 11.2 page 171.

Le traducteur phonétique

Le principe du traducteur phonétique à base de règles contextuelles a été réalisé et modélisé par une structure objet avec le langage Perl. Le principe de structure objet intervient pour la représentation des règles en mémoire. Le tableau 11.1 page 169 présente l'exemple de la lettre i. L'ensemble du dictionnaire est donc réalisé sous la forme d'un arbre où chaque branche caractérise une lettre de l'alphabet, les chiffres et les caractères spéciaux.

Le choix du langage Perl repose sur son implémentation de la substitution d'un motif dans un flux textuel par un autre. Le programme prend donc en entrée une chaîne de caractères correspondant à une marque, applique les règles possibles et retourne la chaîne phonétique caractérisant la marque proposée.

La gestion du savoir faire

Historiquement, dans la première version de CATMInE, le savoir-faire était un calcul figé écrit en C++. Dans la nouvelle version cette intégration se fait directement par le biais de pages internet (en PHP, cf. figure 11.10 page 179 et suivantes) entraînant la création des données transformées sur le même principe que dans la première version, mais permettant des adaptations par l'expert.

Les projections de Sammon

Les projections de Sammon sont des cartes dans lesquels l'expert doit pouvoir naviguer, sélectionner et obtenir des informations sur les éléments caractérisés (cf. 11.4 page 173). Ces cartes sont déterminées par avance et complétées de l'étude en cours lors d'une comparaison.

L'algorithme de Sammon a été écrit en C++ afin de pouvoir exécuter la création des graphiques initiaux sur le serveur en utilisant l'hyperlien dédié à cela sur l'interface d'administration.

L'application étant orientée internet, nous avons utilisé le principe d'applet Java afin de répondre aux spécifications de présentation de ces outils. Cette applet a accès à la structure de données mise en place afin de pouvoir afficher les marques intervenant dans les décisions utilisées lors d'une sélection ou d'un pointage sur le graphique.

L'interface de modélisation

Dans la première version de CATMInE l'algorithme de modélisation utilisé est un réseau de neurones. Cette famille d'algorithmes est consommatrice de ressources (processeur et mémoire) pour établir un modèle. Nous avons donc écrit l'algorithme en C++ afin de se garantir une bonne rapidité d'exécution.

Lorsque l'expert effectue une comparaison, le réseau de neurones est alors chargé avec les différents poids le caractérisant et retenus après un apprentissage. Le retour est un score de classification exploité par la page internet appelante.

Dans le cas de DeTTMInE le programme charge les règles obtenues par un algorithme écrit en Java et retourne la règle sélectionnée pour un exemple proposé (celle-ci contient les descripteurs, la décision, la population caractérisée par la règle et le niveau décisionnel le plus important).

La gestion des données

Les données sont conservées de deux manières. La première par l'intermédiaire d'un modèle relationnel, mis en pratique avec MySQL, caractérisant l'essentiel d'une décision. La figure 11.14 page suivante présente le modèle relationnel appliqué. La seconde conserve intégralement les décisions sous forme de document XML. Celles-ci sont identifiées par un nom de fichier formulé à partir de la référence de la décision (la balise $< REF >$, cf. 4.3 page 65). Cette référence est utilisée comme clef primaire pour chaque décision dans la base de données MySQL. Il est alors possible d'avoir un résumé de l'information nécessaire à la modélisation du domaine et le texte complet de la décision dans le cas d'une consultation.

L'interface de consultation de la base de données permet d'écrire une requête relationnelle et d'obtenir la liste des documents ciblés pour les consulter.

11.2.4 Réutilisation effective de la plate-forme pour d'autres domaines d'étude

Comme nous l'avons énoncé précédemment, la plate-forme CATMInE est un outil pouvant être utilisé dans d'autres domaines d'étude, dès l'instant où l'expert dispose d'une base de données au format XML caractérisant le domaine ciblé. Toutefois, l'intégralité de la plate-forme ne peut être réutilisée en tant que telle et certains éléments nécessitent d'être adaptés. La figure 11.15 page suivante reprend le schéma 11.6 page 174 et propose une vue d'ensemble de la plate-forme en mettant en avant les éléments réutilisables de ceux dépendant du domaine d'étude.

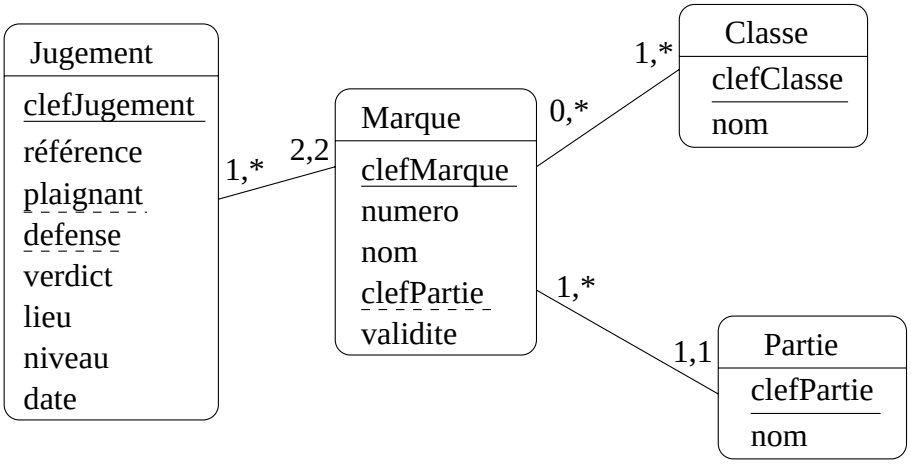


FIG. 11.14 – Modèle relationnel appliqué dans CATMInE

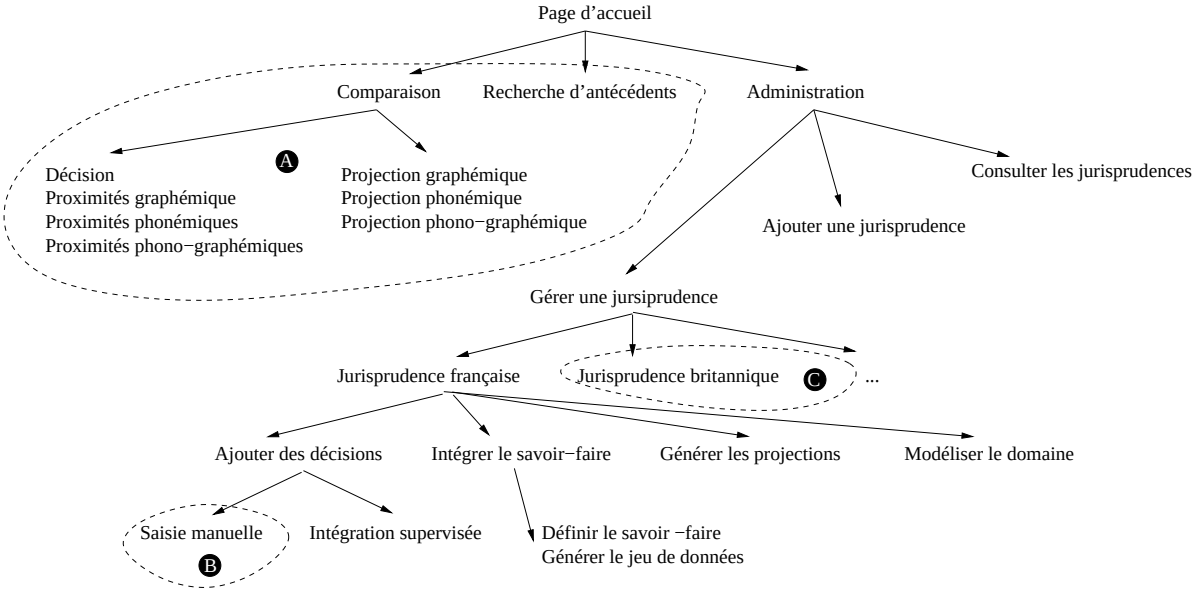


FIG. 11.15 – Distinction entre les éléments génériques ou non de la plate-forme de modélisation CATMInE.

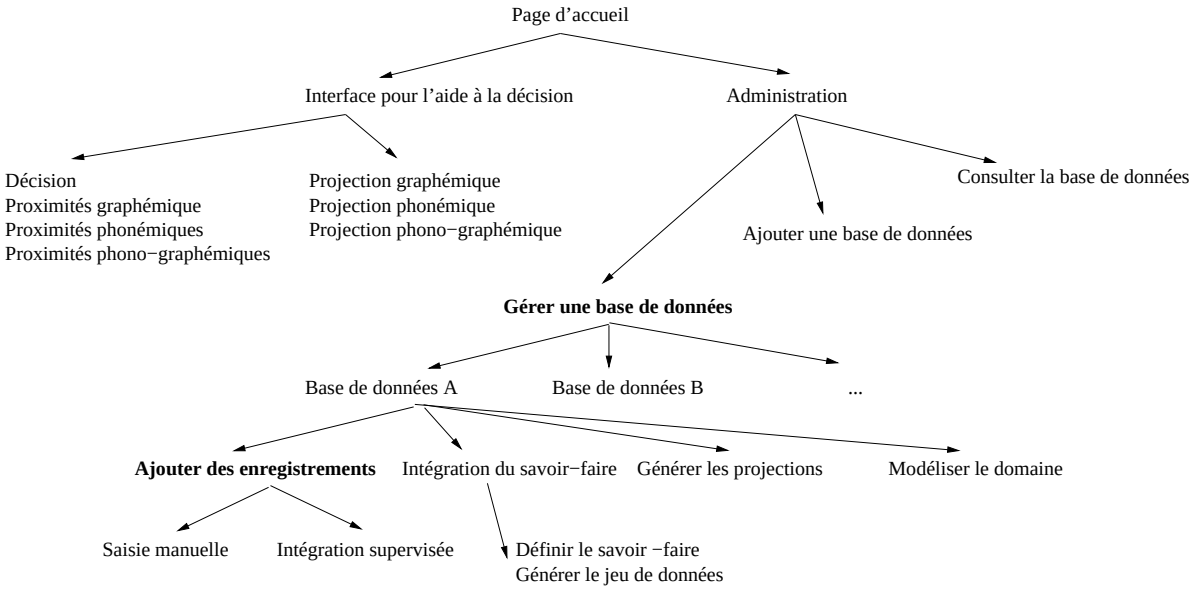


FIG. 11.16 – Vue générique de la plate-forme de modélisation CATMInE.

La première partie, spécifique au domaine, est la zone A (délimitée par des pointillés) qui regroupe les deux fonctionnalités de l'aide à la décision pour l'évaluation du risque de contrefaçon entre marques nominatives. En fonction de ce qui est demandé pour l'aide à la décision, ces interfaces doivent être adaptées à la problématique soulevée.

La fonctionnalité présente dans la partie B (saisie manuelle d'une décision) se présente sous la forme d'un formulaire où l'expert doit renseigner les champs obligatoires pour conserver une décision dans le système d'information. Cette fonctionnalité est contrainte par l'information manipulée dans la base de données et ne peut donc être utilisée en l'état pour un autre domaine.

Enfin, la fonctionnalité C est là aussi propre à la spécificité du domaine. Le dépôt d'une marque pouvant se faire à différentes échelles géographiques, il faut pouvoir avoir des décisions de références pour chaque pays impliqué dans le processus de dépôt. Toutefois, il est possible d'avoir des applications où le phénomène peut être traduit par le biais de plusieurs ensembles de données, distincts les uns des autres par leur sémantique mais identique par leur structure (ce qui est le cas entre les décisions de justices françaises et britanniques dans l'exemple).

La figure 11.16 présente une vue générique de la plate-forme. Si l'on superpose cette figure à la précédente, le seul changement fondamental est la suppression de la recherche d'antécédents. Cette fonctionnalité est un besoin spécifique pour le cabinet Breese Derambure Majerowicz, et ne peut être utilisée dans d'autres domaines.

La figure met alors en avant une architecture simple permettant de relier une base de données aux outils de gestion de celle-ci, ainsi que les fonctionnalités permettant l'exploitation de l'information contenue.

L'application se découpe en trois parties :

1. une partie utilisation
2. une partie administrative simplifiée
3. une partie administrative évoluée.

La première partie correspond à l'exploitation de la connaissance issue du traitement de la (ou des) bases disponibles. L'expert se trouve alors en position d'évaluation et d'interprétation d'une information relative à l'aide à la décision.

La seconde partie consiste en une administration simplifiée de la plate-forme. L'expert dispose alors de privilèges supplémentaires pour l'utilisation des outils d'administration du système d'information. Ces privilèges se résument simplement à un accès à ces éléments. À ce niveau, l'expert peut consulter la base de données disponible ainsi qu'ajouter des enregistrements supplémentaires à l'ensemble par l'intermédiaire de l'outil d'insertion supervisée, ou du formulaire de saisie, adapté aux enregistrements utilisés.

Enfin, la dernière partie propose à l'expert une administration évoluée de la plate-forme. Cette administration complète la précédente avec les outils d'intégration du savoir-faire, de gestion des projections et de modélisation du domaine. L'intégration du savoir-faire se fait par l'intermédiaire de formulaires. La gestion des projections s'effectue avec des hyperliens exécutant sur le serveur la création des projections. Enfin, pour la modélisation, soit l'expert l'effectue avec les paramètres par défaut, et il valide ses choix avec un hyperlien, soit il saisit les paramètres désirés au travers d'un formulaire et la création du modèle se fait à la validation du formulaire.

Le développement des éléments spécifiques à un domaine (tel que la page de comparaison ou la page de recherche d'antécédents dans `CATMInE`) ne peut être réalisé au travers de l'application. Dans notre cas de figure l'aspect générique est porté sur l'architecture de la plate-forme. Les fonctionnalités sont quant à elles spécifiques au domaine d'application et nécessitent donc d'être développées en conséquence.

11.2.5 Bilan de la plate-forme `CATMInE`

La plate-forme `CATMInE` présentée permet de répondre aux différentes étapes présentées dans le schéma 11.17 page ci-contre.

L'intégration d'un outil de sélection et de préparation des documents (cf. la figure 11.7 page 176) répond aux étapes A et B. Cet outil permet de sélectionner les documents (en choisissant de les conserver ou de les écarter du traitement). Il permet aussi de préparer partiellement les documents en éliminant tous les documents comportant des valeurs manquantes nécessaires à la modélisation du phénomène.

Les outils d'intégration du savoir-faire présentés dans les figures 11.10, 11.11 page 179, 11.12 page 180 et 11.13 page 181 permettent de répondre aussi aux procédés de pré-traitements des

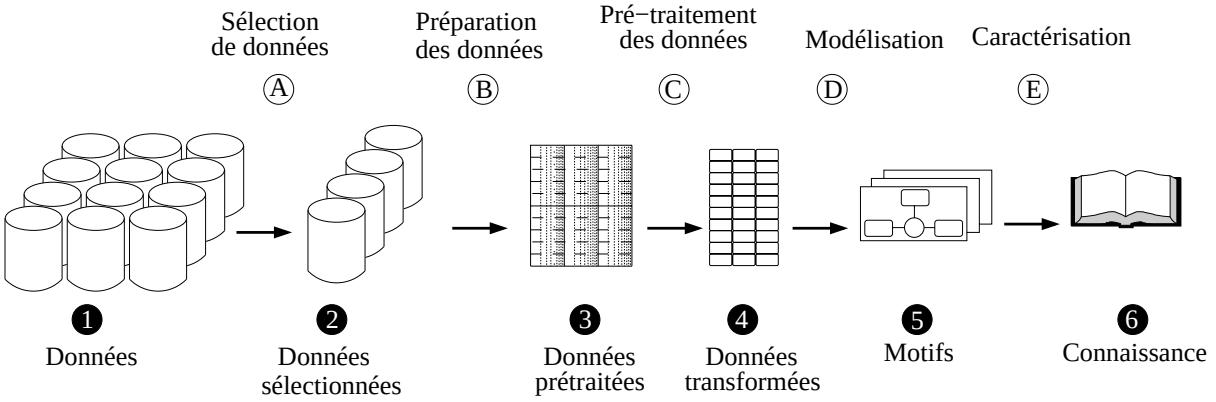


FIG. 11.17 – Vue d'ensemble des différentes étapes constituant la processus de fouille de données et d'extraction de connaissances.

données (étape C). L'expert peut en effet appliquer des opérateurs permettant la fusion, la segmentation ou tout autre type de traitements possibles sur les données conservées dans les étapes précédentes.

Les outils de validation et d'étude de la connaissance induite, représentés par les systèmes de projections des données (figure 11.4 page 173) et de mesure de qualité, permettent de répondre à l'étape E.

Toutefois, l'intégration de différents algorithmes de modélisation au sein de la plate-forme pour répondre à l'étape de modélisation (étape D) reste à mettre en place. `DeTTMInE`, l'évolution de `CATMInE`, offre une alternative au réseau de neurones, mais ne représente en aucun cas une souplesse pour l'intégration d'autres types d'algorithmes.

Il sera donc nécessaire dans la poursuite de ce travail d'axer nos efforts sur la définition d'un principe d'intégration des résultats d'un apprentissage. Cela permettra d'ajouter de nouveaux algorithmes de modélisation avec la définition simple d'un module convertissant les résultats de ces algorithmes vers ce module d'intégration, tel que le présente la figure 11.18 page suivante. Le module permettra alors d'effectuer la liaison entre l'algorithme de modélisation et la visualisation des résultats de prédiction de manière complètement transparente.

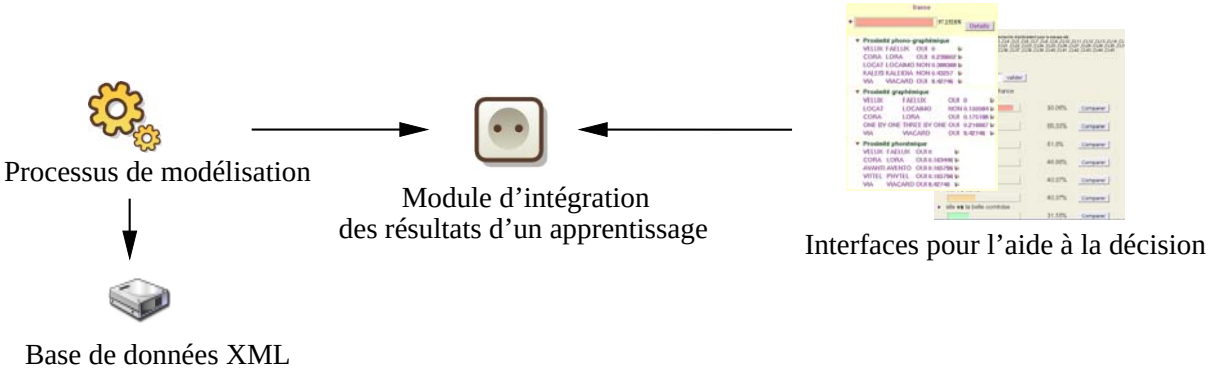


FIG. 11.18 – principe d'intégration à la plate-forme CATMInE des résultats d'un apprentissage réalisé avec n'importe quel algorithme de modélisation.

Bilan et perspectives

Chapitre 12

Bilan

La problématique de ces travaux de recherche est axée sur la classification de documents électroniques juridiques, relatifs à la contrefaçon de marques nominatives. Au cours de ce travail, nous nous sommes plus particulièrement interrogés sur les différents procédés et transformations permettant d'établir une modélisation de ce cas d'étude, ainsi que les moyens d'observer, de contrôler et d'améliorer les modèles obtenus.

La modélisation du domaine

La première contribution de ces travaux est de proposer une étude complète du processus de modélisation admis dans la découverte de connaissances à partir de bases de données et de l'apprentissage automatique. Nous avons alors présenté les différentes étapes clefs permettant une telle réalisation ainsi que la diversité des méthodes à chacune de ces étapes. La complexité de cette démarche réside alors dans l'utilisation optimale des méthodes et du choix de celles-ci en fonction des précédents choix réalisés.

Nous avons aussi démontré qu'une connaissance statistique des données ne suffit pas à résoudre ce probléme et que seul l'intervention de l'expert du domaine tout au long du processus permet de lever certaines ambiguïtés. Cependant, une expertise même précise ne peut être à même de lever l'indetermination liée à la multitude de choix disponibles car cette expertise représente des choix stratégiques et techniques guidées par un savoir implicite et non verbalisé. Ce savoir s'appuie sur des connaissances clairement articulées au niveau du document électronique juridique, et qui sont facilement transférables, car exprimées sous une forme tangible.

Nous avons alors soulevé la nécessité d'avoir une démarche exploratoire afin d'affiner le modèle et les connaissances découlant de ce modèle. Nous suggérons aussi de confronter constamment les décisions et les choix par comparaison avec des méthodes que nous pouvons qualifier de référence, et ce afin de garantir la justesse de la démarche.

Au travers de ce principe de modélisation, nous voulons permettre à l'expert de comprendre le processus décisionnel qu'il réalise, ainsi que de lui donner une objectivation de son savoir faire, en l'intégrant en tant qu'acteur central du processus.

La réalisation technique

La réalisation technique mise en forme par `CATMInE` et `DeTTMInE` représente une mise en œuvre du processus de découverte de connaissance à partir de bases de données. La plate-forme principale (`CATMInE`) permet d'intégrer les étapes de sélection, préparation et de pré-traitement des données. L'expert peut alors définir et réaliser des choix de conception et de modélisation aisément. Ces choix sont pratiqués au travers d'une approche cyclique favorisant l'expression et l'intégration de l'expertise.

Il est ensuite possible d'ajouter un algorithme (ou un ensemble d'algorithmes) de modélisation au choix. Dans notre solution, nous proposons un réseau de neurones (`CATMInE`) et un arbre de décision adaptés au domaine (`DeTTMInE`), mais tout autre type d'algorithme peut être déployé sur cette plate-forme.

L'interprétation des connaissances, dans le cas de `DeTTMInE`, peut se faire par lecture des règles, et observation des décisions correspondantes. Les règles apportent les explications sur la décision en elle même, les décisions permettant d'avoir une base solide de jurisprudences à faire valoir devant une Cour de justice pour défendre un dossier.

Pour `CATMInE` cette étape explicative de la décision nécessite l'intégration de procédés d'analyse des réseaux de neurones tels que ceux présentés dans [Borges et al., 2003] et la lecture des projections de Sammon, permettant de déterminer les jurisprudences de références et de comprendre les fondements de la décision du réseau de neurones.

Enfin, les résultats proposés à l'expert par la plate-forme `CATMInE`, et l'analyse qu'en fait celui-ci, ont permis d'en apporter une bonne accréditation. Le système, en proposant des jugements de référence, permet de renforcer la décision proposée. L'utilisation d'outils de visualisation tel que les projections de Sammon, augmente se pouvoir. L'expert peut juger visuellement (et autrement que par une valeur numérique) la notion de proximité avec en plus une disposition spatiale afin de mieux comprendre la décision proposée. Ce principe d'analyse permet d'éviter le rejet par l'expert de la décision dans la mesure où celle-ci est argumentée par des exemples proches du cas étudié par l'expert.

Note finale

La plate-forme présentée ainsi que le raisonnement pragmatique appliqué n'est pas spécifique au domaine de la contrefaçon de marques nominatives. Il peut en effet être étendu à l'ensemble

des questions dont la source de données est au format `XML` : chaque document correspond à une donnée et les balises `XML` utilisées dans la structure du document sont associés aux descritpeurs le caractérisant.

Le principe de cette plate-forme favorise les interactions entre l'expert et la machine dans un domaine précis afin d'améliorer les partis pris et l'intégration du savoir-faire de l'expert dans le modèle qu'il cherche à établir.

Les opérations sont alors réalisées sans faire intervenir un informaticien pour répondre à la question posée par l'expert et celui-ci peut bénéficier des interactions précédentes des autres experts avec le système.

Chapitre 13

Perspectives

Nos perspectives de recherches sont à envisager principalement sur deux axes. Le premier axe porte sur une vision à court terme de ces travaux de recherche : valider ces travaux sur d'autres types de données. Le second axe porte sur une approche à moyen terme : compléter la plate-forme avec de nouveaux outils afin d'optimiser l'intégration de l'expertise au jeu de données.

Évaluer la robustesse de la plate-forme

Nous avons présenté tout au long de ces travaux le principe de modélisation d'un ensemble documentaire, et nous avons validé cette approche à l'aide d'une base de documents électroniques juridiques. L'objectif était de mettre en valeur les difficultés conceptuelles inhérentes à une telle démarche de modélisation. Toutefois, afin de garantir que nous nous engageons sur la bonne voie, avec les bons outils, il serait judicieux de notre part d'envisager une évaluation de la plate-forme `CATMInE` sur d'autres types de données documentaires afin de juger sa robustesse.

Cette évaluation entraînera alors le passage d'une plate-forme dédiée, à une plate-forme ouverte à tout type de documents électroniques et donc de modélisation.

Compléter la plate-forme avec de nouveaux outils d'extraction d'informations

Pour la vision à moyen terme de l'évolution de ces travaux, nous pensons intégrer de nouveaux outils permettant de faciliter certains processus pour l'expert. Nous pensons essentiellement à l'extraction d'informations à partir des documents électroniques. Ce processus est pour l'instant réalisé de façon semi-supervisée (le système détermine l'information, l'expert la valide ou non) par le biais d'une interface `Java` où les règles d'extraction d'informations sont conservées dans un fichier précis. Or ces règles n'ont pas été définies afin d'être interprétées par un autre langage de programmation (nous faisons notamment référence à `Perl`). Nous envisageons pour cela de pro-

duire de nouvelles interfaces orientées vers l'Internet, permettant de consulter un sous-ensemble très restreint de documents. L'expert pourra alors définir pour chaque document :

- la région du document où il cherche l'information
- l'information précise au sein de cette région

Nous serons alors en mesure de produire des règles d'extraction d'informations respectant le formalisme des *expressions régulières* pouvant être intégrées à tout système capable de l'interpréter. De cette définition et interprétation, nous pensons qu'il reste cependant nécessaire de laisser à l'expert un contrôle sur les données isolées comme le permet déjà l'existant.

Mettre en valeur l'aspect collaboratif de la plate-forme

Une autre adaptation envisagée sur le moyen terme est relative à la définition d'un chemin optimal dans le processus de modélisation d'un jeu de données. Comme nous l'avons expliqué dans ce mémoire, chaque expert a sa propre vision du domaine qu'il définit. Cette vision est subjective et dépend du savoir-faire de celui-ci. De ce constat, nous imaginons qu'au sein d'un pool d'experts, aucun n'aura la même démarche pour une même tâche : les outils de préparation, de pré-traitement des données et de modélisation seront différents.

Étant donné que la plate-forme repose sur l'utilisation d'une architecture client/serveur, cela nous permet déjà de centraliser les différentes démarches appliquées par un groupe d'experts pour la préparation des données (les opérateurs de transformation sont partagés par le groupe d'expert utilisant la plate-forme (11.2.2 page 173). Cette centralisation permettra aussi de confronter, partager voire d'harmoniser les différentes expertises possibles au sein d'un même groupe d'experts. En combinant l'ensemble des ces expertises, le système peut être affiné en le faisant glisser vers un méta-modèle de décision, ou chaque classifieur reflète une expertise.

Une autre façon de procéder pourrait aussi être de déterminer un chemin optimal, prenant en compte l'expertise de chacun des intervenants. Le chapitre 9 page 127 se termine sur cette difficulté à déterminer le bon chemin parmi l'ensemble des processus disponibles. Nous pouvons alors envisager d'établir à l'aide d'un algorithme d'apprentissage des règles de sélection d'une méthode plutôt qu'une autre en fonction des habitudes des utilisateurs de la plate-forme et des choix retenus en amont de la méthode sélectionnée.

Amélioration de l'évaluation de la doctrine

Nous avons présenté dans la section 11.1.5 page 170, une utilisation de la plate-forme `CATMInE` permettant d'analyser l'ensemble de la jurisprudence afin d'en comprendre les fondements. Nous pensons que cette fonctionnalité peut être améliorée par la mise en place d'un procédé d'analyse temporel. Ainsi, en construisant une cartographie pour chaque nouvelle décision insérée, nous

aurons une vue animée de la jurisprudence. La cartographie de départ correspondra à la plus ancienne des décisions disponibles et celle d'arrivée, à l'intégralité des décisions. Cette vue mettra alors en évidence l'influence de certaines décisions par rapport à d'autres. Nous pourrons aussi matérialiser les décisions d'une information supplémentaires en établissant les vecteurs de déplacement de chacune d'entre elles.

Cette amélioration permettra à l'expert d'augmenter l'étude de la doctrine en observant la réorganisation de la jurisprudence à travers le temps, et permettra une meilleure anticipation des décisions futures.

Transfert de technologies

Le point de départ de ces travaux repose sur un contrat entre professionnels du droit et universitaires, ce qui met en avant le besoin de cette catégorie d'experts à envisager de nouvelles méthodes de travail faisant intervenir des outils adaptés pour répondre à des besoins précis.

Même si `CATMInE` a déjà fait l'objet de contrats, et même si cet applicatif bénéficie d'une validation de son installation et de son fonctionnement sur les 3 grands systèmes d'exploitations : Windows, MacOSX et Linux, nous pensons qu'il faut garder à l'esprit que d'autres institutions peuvent être demandeuses de tels systèmes. Nous faisons essentiellement allusions à d'autres cabinets juridiques ou encore à l'INPI, dont la base `JURINPI` provient.

Annexes

Annexe A

Les bases de données de l'UCI

Les bases de données sont issues du site de l'UCI[24]. Ce site regroupe une collection de bases de données reconnues par les acteurs du domaine, pour lesquelles des objectifs de fouilles de données sont proposés.

Balance Scale Database (bal)

- fournie par Tim Hume
- 625 instances, 4 attributs numériques
- 3 classes
- Pas de valeurs manquantes

Breast Cancer Database (breast)

- fournie par Ljubljana Oncology Institute
- base de données très utilisée
- 2 classes
- 286 instances, 9 attributs

Credit Screening Databases (crx)

- fournie par Japanese Credit Screening Database
- 2 classes
- 690 instances, 15 attributs
- contient quelques valeurs manquantes
- 653 instances réellement utilisées

[24] **U**niversity of **C**alifornia, **I**rvine, http ://kdd.ics.uci.edu/

Solar Flare Databases (flare)

- fournie par Gary Bradshaw
- 3 classes
- 1066 instances, 13 attributs
- prédiction d'attributs nominal
- pas de valeur manquante

Housing Database (housing)

- fournie par CMU StatLib Library
- loyer des logements dans la banlieue de Boston
- 2 classes
- 506 instances, 11 attributs continus et 1 binaire

Lymphography Database (lympho)

- fournie par Ljubljana Oncology Institute
- 4 classes
- 148 instances, 19 attributs
- pas de valeur manquante

Pima Indians Diabetes Database (pima)

- fournie par National Institute of Diabetes and Digestive and Kidney Diseases
- 3 classes binaires
- 8 attributs numériques
- 768 instances

Zoo Database (zoo)

- fournie par Richard Forsyth
- Artificielle
- 2 classes
- 101 instances, 17 attributs, 15 booléens et 2 numériques
- pas de valeur manquante

Mushrooms Database (Mushroom)

- fournie par Audobon Society Field Guide
- 2 classes

- 8124 instances
- 2480 valeurs manquantes pour le douzième attribut

Annexe B

Résultats expérimentaux

Ces résultats illustrent la comparaison, au travers de 5 séries d'apprentissage réalisées pour chaque base de données, les deux classifieurs utilisés : `C4.5` et `DeTTMInE` (section 9.2.8 page 136).

classifieur	série 1	série 2	série 3	série 4	série 5	moyenne
DeTTMInE	61,81%	74,54%	75,43%	75,43%	83,01%	**74,04%**
C4.5	65,50%	74,54%	78,90%	75,40%	79,20%	*74,70%*
C4.5rules	63,60%	58,20%	73,70%	68,40%	79,20%	68,62%

TAB. B.1 – Résultats par série pour la base `breast`

classifieur	série 1	série 2	série 3	série 4	série 5	moyenne
DeTTMInE	66,15%	66,15%	66,41%	61,83%	62,59%	**64,63%**
C4.5	73,10%	69,20%	69,50%	67,90%	67,20%	*69,38%*
C4.5rules	68,50%	73,40%	68,70%	67,20%	68,70%	69,30%

TAB. B.2 – Résultats par série pour la base `crx`

classifieur	série 1	série 2	série 3	série 4	série 5	moyenne
DeTTMInE	94,73%	100.00%	90,47%	85,71%	95,23%	**93,23%**
C4.5	89,50%	94,70%	100.00%	81.00%	85,70%	90,18%
C4.5rules	94,70%	94,70%	100.00%	81.00%	85,70%	*91,22%*

TAB. B.3 – Résultats par série pour la base `zoo`

classifieur	série 1	série 2	série 3	série 4	série 5	moyenne
DeTTMInE	72,36%	75.00%	72,72%	75,32%	75.00%	**74,08%**
C4.5	75.00%	71,70%	74,70%	76.00%	74,60%	*74,40%*
C4.5rules	71,70%	67,80%	77,90%	73,40%	70,50%	72,26%

TAB. B.4 – Résultats par série pour la base `pima`

classifieur	série 1	série 2	série 3	série 4	série 5	moyenne
DeTTMInE	68,29%	82,92%	73,80%	84,92%	76,37%	**77,26%**
C4.5	61,50%	59,30%	57,10%	69,30%	60,60%	61,56%
C4.5rules	80,50%	70,70%	65,90%	69,30%	76,40%	*72,56%*

TAB. B.5 – Résultats par série pour la base `bal`

classifieur	série 1	série 2	série 3	série 4	série 5	moyenne
DeTTMInE	71,42%	71,42%	84,37%	81,25%	85,71%	**78,83%**
C4.5	57,10%	85,70%	81,20%	78,10%	89,30%	78,28%
C4.5rules	64,30%	85,70%	78,10%	84,40%	92,90%	*81,08%*

TAB. B.6 – Résultats par série pour la base `lympho`

classifieur	série 1	série 2	série 3	série 4	série 5	moyenne
DeTTMInE	73,46%	74,48%	78,64%	72,81%	68,26%	**73,53%**
C4.5	77,60%	81,60%	86,40%	81,60%	79,80%	81,40%
C4.5rules	78,60%	78,60%	85,40%	82,50%	82,70%	*81,56%*

TAB. B.7 – Résultats par série pour la base `housing`

classifieur	série 1	série 2	série 3	série 4	série 5	moyenne
DeTTMInE	77,25%	72,98%	70,37%	71,75%	70,75%	**72,62%**
C4.5	77,30%	73,90%	74,10%	70,40%	71,20%	73,38%
C4.5rules	74,40%	73,90%	74,10%	74,10%	71,20%	*73,54%*

TAB. B.8 – Résultats par série pour la base `flare`

Annexe C

Extrait du texte complet d'une jurisprudence issu de la base documentaire JURINPI

FAITS ET PROCEDURE La société EXACOD, est inscrite au registre du commerce de Paris depuis le 17 avril 1998. Elle a pour activité la réalisation d'inventaires, la gestion de stocks, la rédaction de documents comptables et commerciaux, la commercialisation de banques de données informatiques et de toutes statistiques liées à la structure de stocks et aux achats de la clientèle.Elle indique faire usage du nom commercial EXACOD.Elle est titulaire de la marque semi-figurative EXACOD déposée le 3 juin 1999 et enregistrée sous le n° 99.796.368 pour désigner les produits et services des classes 9, 16, 35, 36 et 41 dont les appareils pour le traitement de l'information, la gestion des affaires commerciales, les conseils, l'information et le renseignement d'affaires, la gestion de dossier informatique, la programmation pour ordinateurs et la location de temps d'accès à un service serveur de bases de données.Le 7 décembre 1999, LA POSTE informait la société EXACOD de son intention de déposer la marque HEXACODES pour désigner en classe 9, 16 et 35 un fichier relatif aux voies, boîtes postales et localités. Puis, le 23 février 2000 elle a, malgré l'opposition de la société EXACOD, procédé au dépôt de la marque dénominative HEXACODES désignant en classes 9, 16 et 35 les logiciels de gestions de fichiers d'adresses destinés aux routeurs et aux grands émetteurs de courriers, les fiches, la constitution, la centralisation, la tenue et la mise à jour de fichiers d'adresses pour validation du contenu de bases de données (voies, boites postales, localités et Cedex). Cette marque est enregistrée sous le n° 00.3009.533.Invoquant le caractère contrefaisant de la marque HEXACODES et les atteintes portées à sa dénomination sociale et à son nom commercial, la société EXACOD a, par acte du 14 septembre 2000, fait assigner LA POSTE devant le tribunal de céans afin de voir prononcer la nullité de la marque n°00.3009.533.Elle sollicite, outre des mesures d'interdiction, de confiscation, de publication et l'exécution provisoire sur le tout, l'allocation d'une somme de 500 000 francs à titre de dommages et intérêts en réparation des atteintes portées à ses droits sur les signes distinctifs précités et une indemnité de 50 000 francs en application de l'article 700 du Nouveau Code de Procédure Civile.LA POSTE souligne qu'elle exploite et n'entend exploiter sa marque qu'en association avec la marque notoire LA POSTE, ce qui exclut tout risque de confusion entre les signes en cause. Elle poursuit la nullité partielle de la marque de la demanderesse en ce qu'elle vise "les conseils, les informations ou renseignements d'affaires, la gestion de dossier informatique, les programmations pour ordinateur", en raison de l'imprécision de ces catégories de produits ou services. Elle ajoute que lors du dépôt de sa marque, la société EXACOD a désigné de manière abusive, des services qui ne relèvent pas de son activité principale. Subsidiairement, elle demande qui lui soit donné acte qu'elle s'engage à exploiter sa marque associée à sa marque notoire LA POSTE. Elle réclame le remboursement de ses frais irrépétibles à hauteur de 50 000 francs.La société EXACOD réfute l'argumentation de la défenderesse et maintient l'intégralité de ses demandes.DECISIONI - SUR LA VALIDITE DE LA MARQUE EXACOD :Attendu que LA POSTE conclut que la désignation des services suivants :"conseils, informations ou renseignements d'affaires. Gestion de fichiers informatiques. Programmations pour ordinateur" est insuffisamment précise ;Attendu que, aux termes de l'article L712.2 du Code de la Propriété Intellectuelle le dépôt comprend "l'énumération des produits et services auxquelles elle s'applique" ; que l'arrêté du 31 janvier 1992 précise en son article 2 que "cette énumération peut résulter soit de la désignation individuelle de chacun de ces produits ou services soit de l'énumération de la catégorie à laquelle ils appartiennent. Dans ce dernier cas, les termes employés doivent permettre à toute personne d'en délimiter le contenu de façon immédiate, certaine et constante."Qu'il s'en suit qu'il n'est pas nécessaire qu'une demande d'enregistrement comporte de manière exhaustive tous les produits et services spécifiques que le déposant entend protéger, mais qu'il est possible de ne désigner que les catégories auxquelles appartiennent ses produits ou services ;Attendu que la société EXACOD a visé lors du dépôt de sa marque EXACOD non les intitulés des classes 35 et 36 mais : - deux catégories de services en classe 35 - conseils, informations ou renseignement d'affaires. Gestion de fichiers informatiques- qui recouvrent pour l'une le conseil et l'information des entreprises dans un domaine particulier et pour l'autre la gestion de fichiers informatique ; un service en classe 36, la programmation pour ordinateur, que le contenu de ces services peut être délimité de façon immédiate ;Attendu qu'en second lieu, LA POSTE excipe du caractère abusif de la désignation de services qui ne relèvent pas de l'activité principale du déposant ;Mais Attendu que la fraude ne saurait donc se déduire du fait que la société EXACOD aurait déposé sa marque dans des domaines ...

...qui excéderait celui de son activité sociale, que ce comportement serait éventuellement sanctionné par la déchéance de la marque pour les services inexploités à l'issue d'un délai de cinq ans ; que surabondamment, il peut être noté, que les services visés par la demanderesse dans le dépôt incriminé relèvent de son activité telle que définie lors de son immatriculation au registre du commerce ;Attendu que la demande en nullité partielle de la marque EXACOD ne peut donc prospérer ;II - SUR LA VALIDITE DE LA MARQUE HEXACODES :Attendu qu'aux termes de l'article 711.4 du Code de la Propriété Intellectuelle ne peut être adopté comme marque un signe portant atteinte à des droits antérieurs et notamment :à une marque antérieurement enregistrée,à une dénomination sociale ou raison sociale, s'il existe un risque de confusion dans l'esprit du public,à un nom commercial ...connu sur l'ensemble du territoire national s'il existe un risque de confusion dans l'esprit du public,Attendu que la marque semi-figurative n° 99.796.368 présente un "code barre" gris, avec en dessous les syllabes EXA et COD en couleurs ;Que la présence d'un code barre, ne fait pas perdre au seul élément dénominatif de la marque son caractère distinctif ;Que la marque HEXACODES déposée par LA POSTE reprend dans le même ordre, 6 des 9 lettres de la marque EXACOD ; que la présence de trois lettres supplémentaires n'en modifie pas la prononciation ; qu'il peut donc être relevé une ressemblance visuelle ainsi qu'une similitude phonétique parfaite des signes en présence ;Attendu que la marque n° 99.796.368 désigne parmi les produits et services des classes 9, 16, 35, 36 et 41 les appareils pour le traitement de l'information et les ordinateurs, l'information et le renseignement d'affaires, la gestion de fichiers informatiques, la programmation pour ordinateurs ;Que la marque 00.3009.533 vise dans les classes 9, 16 et 35 les logiciels de gestions de fichiers d'adresses destinés aux routeurs et aux grands émetteurs de courriers, les fiches, la constitution, la centralisation, la tenue et la mise à jour de fichiers d'adresse pour validation du contenu de bases de données (voies, boites postales, localités et Cedex) ;Que ces services entrent dans les catégories visées par la société EXACOD lors de l'enregistrement de sa marque, ce qui peut amener le public, en raison de la quasi-identité des signes à se méprendre sur l'origine des services en cause ;Qu'enfin, les éléments invoqués par LA POSTE concernant les conditions d'exploitation de son signe sont inopérants, la comparaison devant porter exclusivement sur les signes tels qu'enregistrés et le risque de confusion apprécié compte tenu des énonciations du dépôt et non de l'utilisation faite des marques ;Attendu que la nullité de la marque HEXACODES sera donc prononcée par application de l'article L714-3 du Code de la Propriété Intellectuelle, sans qu'il y ait lieu d'examiner l'opposabilité comme antériorité de la dénomination sociale et d'un nom commercial de la société EXACOD ;III - SUR L'ATTEINTE A LA DENOMINATION SOCIALE ET AU NOM COMMERCIAL DE LA SOCIETE EXACOD :Attendu que lors de son immatriculation au registre du commerce la société EXACOD a déclaré comme dénomination sociale EXACOD ; qu'elle a pour activité la réalisation d'inventaires, la gestion de stocks, la rédaction de documents comptables et commerciaux, la commercialisation de banques de données informatiques et de toutes statistiques liées à la structure de stocks et aux achats de la clientèle.Que les documents versés aux débats -une feuille de papier commercial vierge, une carte de visite et une brochure de présentation de son activité- sont insuffisants pour justifier qu'elle ferait un usage effectif du nom commercial EXACOD pour se présenter au public ;Attendu que la marque HEXACODES désigne les logiciels de gestions de fichiers d'adresses destinés aux routeurs et aux grands émetteurs de courriers, les fiches, la constitution, la centralisation, la tenue et la mise à jour de fichiers d'adresses pour validation du contenu de bases de données (voies, boites postales, localités et Cedex) ;Qu'il s'agit d'une activité parfaitement distincte de celle développée et revendiquée par la société EXACOD qui se limite à la réalisation d'inventaire et aux activités annexes en résultant ; que le public ne peut être amené à penser que les services proposés sous la dénomination HEXACODES, qui sont limités à des fichiers d'adresses et les prestations de la société EXACOD qui concernent la réalisation d'inventaires et éventuellement la gestion de stocks émaneraient d'un même opérateur économique ;Que la société EXACOD sera donc déboutée de ses demandes au titre des atteintes à sa dénomination sociale et son nom commercial ;IV - SUR LES MESURES REPARATRICES :Attendu qu'il convient de faire droit aux mesures d'interdiction et de publication sollicitée dans les termes du dispositif ci-dessous ; qu'il n'y a pas lieu d'y ajouter la mesure redondante de confiscation des documents portant la dénomination ...

... HEXACODES ; Attendu que LA POSTE a fait usage de sa marque et a diffusé des brochures publicitaires auprès de la clientèle professionnelle concernée ; qu'elle n'a nullement, comme elle le prétend, associé cet usage à celui de sa marque LA POSTE ; que cet usage a affadi le caractère attractif de la marque de la société EXACOD ; qu'il convient d'allouer à cette dernière la somme de 100 000 francs à titre de dommages et intérêts ;Attendu que l'exécution provisoire sera limitée à la seule mesure d'interdiction ;Attendu que LA POSTE qui succombe sera condamnée aux dépens ; qu'il paraît équitable de fixer à la somme de 15 000 francs en application de l'article 700 du Nouveau Code de Procédure Civile ;PAR CES MOTIFSLE TRIBUNAL,statuant publiquement par jugement contradictoire en premier ressort,Annule la marque n° 00.3009.533 déposée le 23 février 2000 par LA POSTE pour désigner en classes 9, 16 et 35 les logiciels de gestions de fichiers d'adresses destinés aux routeurs et aux grands émetteurs de courriers, les fiches, la constitution, la centralisation, la tenue et la mise à jour de fichiers d'adresses pour validation du contenu de bases de données (voies, boites postales, localités et Cedex) ;Dit que la décision devenue définitive sera transmise sur réquisition du greffier à l'INPI pour inscription au registre national des marques ;Interdit à LA POSTE de faire usage de la dénomination HEXACODES sous astreinte de 1000 francs par infraction constatée à compter de la signification du présent jugement ;Ordonne l'exécution provisoire de cette mesure ;Condamne LA POSTE à payer à la société EXACOD la somme de 100000 francs à titre de dommages et intérêts et une indemnité de 15 000 francs en application de l'article 700 du Nouveau Code de Procédure Civile ;Autorise la société EXACOD a faire publier le présent dispositif dans deux journaux de son choix, aux frais de LA POSTE, le coût total de ces insertions ne pouvant excéder à la charge de cette dernière la somme hors taxes de 40 000 francs ;Déboute les parties, pour le surplus ;Condamne LA POSTE aux entiers dépens ;Accorde à Maître MENDRAS, avocat le droit de recouvrer les dépens dans les conditions prévues par l'article 699 du nouveau Code de procédure civile.

TAB. C.1 – Extrait du texte complet d'une jurisprudence issue de la base documentaire `JURINPI`.

Bibliographie

[Agrawal et al., 1993] Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large database. *In proceedinf of the ACM SIGMOD International Conference of Data*, (pp. 207–216).

[Agrawal & Srikant, 1994] Agrawal, R. & Srikant, R. (1994). : (pp. 487–499). : Morgan Kauffman.

[Ankerst et al., 1996] Ankerst, M., Breunig, M., Kriegel, H.-P., & Sander, J. (1996). Optics : Ordering points to identify the clustering structure. In *ACM SIGMOD* : ACM.

[Auray et al., 1991] Auray, J.-P., Duru, G., & Zighed, A. (1991). *Analyse des données multidimensionnelles.* Lyon : Lacassagne.

[Azé & Roche, 2003] Azé, J. & Roche, M. (2003). Une application de la fouille de textes : l'extraction de règles d'association à partir d'un corpus spécialisé. *EGC*, (pp. 283–294).

[Bauer & Kohavi, 1999] Bauer, E. & Kohavi, R. (1999). An empirical comparison of voting classification algorithms : Bagging, boosting an variants. *Machine Learning*, 36, 105–139.

[Berkin, 2002] Berkin, P. (2002). Survey of clustering data mining techniques. *Technical Report.*

[Biswas et al., 1981] Biswas, G., Jain, A., & Dubes, R. (1981). Evaluation of projection algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3, 701–708.

[Blum & Langley, 1997] Blum, A. & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2), 245–271.

[Borges et al., 2003] Borges, F., Borges, R., & Bourcier, D. (2003). Artificial neural networks and legal categorization. *Legal Knowledge and information systems. Jurix 2003*, (pp. 11–20).

[Brachman et al., 1993] Brachman, R., Halper, F., Selfridge, P., Kirk, T., Terveen, L., Lazar, A., Altman, B., Mcguiness, D., A.Borgida, & Resnick, L. (1993). Integrated support for data archaeology. *IJICIS.*

[Bradley et al., 1998] Bradley, P., Fayyad, U., & Reina, C. (1998). Scaling clustering algorithms to large databases. In *Knowledge Discovery and Data Mining* (pp. 9–15).

[Breiman, 1996a] Breiman, L. (1996a). Bagging predictors. *Machine Learning*, 24(2), 123–140.

[Breiman, 1996b] Breiman, L. (1996b). Some properties of splitting criteria. *Machine Learning*, 21, 41–47.

[Caruana & Freitag, 1994] Caruana, R. & Freitag, D. (1994). Greedy attribute selection. *Proceeding of the Eleventh International Conference on Machine Learning*, (pp. 28–36).

[Chan & Stolfo, 1995] Chan, P. & Stolfo, S. (1995). A comparative evaluation of voting and meta learning on partitioned data. *ICML*.

[Chan & Stolfo, 1996] Chan, P. & Stolfo, S. (1996). Scaling learning by meta-learning over disjoint and partially replicated data. *Proc. Ninth Florida Artificial Intelligence Research Symposium*, (pp. 151–155).

[Cheeseman & Stutz, 1996] Cheeseman, P. & Stutz, J. (1996). Bayesian classification (autoclass) : Theory and results. *Advances in Knowledge Discovery and Data Mining*.

[Cross & Sudcamp, 1991] Cross, V. & Sudcamp, T. (1991). An empirical study of fuzzy compatibility measures and aggregation operators. 1607, 415–428.

[Crémilleux, 2000] Crémilleux, B. (2000). Classification interactive. *Apprentissage par l'interaction*, (pp. 207–240).

[Deer & Eklund, 1996] Deer, J. & Eklund, P. (1996). On the fusion of image features. *First International Discourse on Fuzzy Logic and the Management of Complexity*, 1, 45–50.

[Dong & Li, 1998] Dong, G. & Li, J. (1998). Interstingness of discovered association rules in terms of neighborhood-based unexpectedness. *Proceedings of the Second Pacific-Asia Conference on Knowledge Discovery and Data Mining*, (pp. 72–86).

[Dougherty et al., 1995] Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. *International Conference on Machine Learning*, (pp. 194–202).

[Dreyfus, 2002] Dreyfus, G. (2002). *Réseaux de neurones : méthodologie et applications*. Eyrolles.

[Durand, 2004] Durand, N. (2004). Extraction de clusters à partir du treillis de concepts : Application à la découverte de communautés d'intérêt pôur améliorer l'accés à l'information.

[Ester et al., 1996] Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density based algorithm for discovering clusters in large spatial databases with noise. In *2nd ACM SIGKDD* : ACM.

[Famili, 1995] Famili, A. (1995). The role of data pre-processing in intelligent data analysis. *Proceedings of the IDA-95 Symposium, Baden-Baden, Germany(1995) P. 54-58.*

[Fayad et al., 1996] Fayad, U., Piatetsky-Shapiro, G., & P.Smyth (1996). The kdd process for extracting useful knowldege from volumes of data. *Communication of the ACM.*

[Fayyad & Irani, 1993] Fayyad, U. & Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. *International Joint Conferences on Artificial Intelligence.*

[Fayyad et al., 1996a] Fayyad, U., Piatetski-Shapiro, G., & Smyth, P. (1996a). From data mining to knowledge discovery : An overview. *Advances in Knowledge Discovery and Datta Mining*, (pp. 1–34).

[Fayyad et al., 1996b] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). Knowledge discovery and data mining : towards a unifying framework. *Second International Conference on knowledge Discovery an Data Mining.*

[Feldman et al., 1998] Feldman, R., Fresko, M., Hirsh, H., Aumann, Y., Liphstat, O., Schler, Y., & Rajman, M. (1998). Knowledge management : A text mining approach. *In Proceedings of the 2nd International Conference on Practical Aspects of Knowledge Management (PAKM98).*

[Fournier, 2001] Fournier, D. (2001). *Etude de la qualité de données à partir de l'apprentissage automatique. Application aux arbres d'induction.* PhD thesis, Université de Caen - Campus II.

[Fraley & Raferty, 1999] Fraley, C. & Raferty, A. (1999). Mclust : Software for model-based cluster and discriminant analysis. *Technical Report 342.*

[Freitas, 1998] Freitas, A. (1998). On objective measures of rule surprisingness. *Proceedings of the Second Eurpoean Conference on the Principles of Data Mining and Knowledge Discovery*, (pp. 1–9).

[Freund & Schapire, 1996] Freund, Y. & Schapire, R. (1996). Experiments with a new boosting algorithm. In *International Conference on Machine Learning* (pp. 148–156).

[Gago & Bentos, 1998] Gago, P. & Bentos, P. (1998). A metric for selection of the most promising rules. *Proceedings of the Second Eurpoean Conference on the Principles of Data Mining and Knowledge Discovery*, (pp. 19–27).

[Gordonn & des Jardins, 1995] Gordonn, D. & des Jardins, M. (1995). Evaluation and selection of biases in machine learning. *MLJ 1995.*

[Gray & Orlowska, 1998] Gray, B. & Orlowska, M. (1998). Ccaiia : clustering categorical attributes into interesting association rules. *Proceedings of the Second Pacific-Asia Conference on Knowledge Discovery and Data Mining*, (pp. 132–143).

[Guha et al., 1998] Guha, S., Rastogi, R., & Shim, K. (1998). Cure : an efficient clustering algorithm for large databases. *ACM SIGMOD*, (pp. 73–84).

[Hamilton & Fudger, 1995] Hamilton, H. & Fudger, D. (1995). Estimating dblearn's potential for knowledge discovery in databases. *Computational Intelligence*, 11(2), 280–296.

[Hamilton et al., 1997] Hamilton, H., Shan, N., & Ziarko, W. (1997). Machine learning of credible classifications. *Proceedings of the Tnth Australian Conference on Artificial Intelligence*, (pp. 330–339).

[Hartigan, 1975] Hartigan, J. (1975). Clustering algorithms.

[Hartigan & Wong, 1979] Hartigan, J. & Wong, M. (1979). Algorithme as136 : A k-means clustering algorithm. *Applied statistics*, 28.

[Hilario, 2002] Hilario, M. (2002). Model complexity and algorithm selection in classification. *METAL*.

[Hilderman & Hamilton, 1999] Hilderman, R. & Hamilton, H. (1999). *Knowledge discovery and interestingness measures : A survey*. Technical report.

[Hinneburg & Keim, 1998] Hinneburg, A. & Keim, D. (1998). An efficient approach to clustering large multimedia databases with noise. In *4th ACM SIGKDD* : ACM.

[Holte, 1993] Holte, R. (1993). Very simple classification rules perform well on most commonnely used datasets. *Machine Learning*, 11, 63–90.

[Institute, 1989] Institute, S. (1989). *SAS/STAT User's Guide, Version 6 (4th Ed.)*.

[John et al., 1994] John, G., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. *International Conference on Machine Learning*, (pp. 121–129). Journal version in AIJ.

[J.Petrak, 2000] J.Petrak (2000). Fast subsampling performance estimates for classification algorithm selection. *Workshop at ECML 2000*.

[Kamber & Shinghal, 1996] Kamber, M. & Shinghal, R. (1996). Evaluating the interestingness of characteristic rules. *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, (pp. 263–266).

[Kaufman & Rousseeuw, 1990] Kaufman, L. & Rousseeuw, P. (1990). *Finding groups in data : An introduction toi Cluster Analysis*. John Willey and Sons, NY.

[Kaynak, 1997] Kaynak, C. (1997). Multistage classification by cascaded classifiers.

[Kaynak & Alpaydin, 2000] Kaynak, C. & Alpaydin, E. (2000). Multistage cascading of multiple classifiers : One man's noise is another man's data. *In proceeding of the 17th International Conference on Machine Learning (ICML)*, (pp. 455–462).

[Kerber, 1992] Kerber, R. (1992). Chimerge : discretization of numeric attributes. *Proceedings of the Tenth NAtional Conference on Artificial Intelligence*, (pp. 123–128).

[Klemettinen et al., 1994] Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., & Verkamo, A. (1994). Finding interestinng rules from large sets of discovered association rules. *Proceedings of the Third International Conference on Information and Knowledge Mamagement*, (pp. 401–407).

[Kohavi & John, 1997] Kohavi, R. & John, G. (1997). Wrappers for feature subset selection. *Artificial Intelligence archive*, 97(1-2), 273–324.

[Kohavi & John, 1998] Kohavi, R. & John, G. (1998). The wrapper approach.

[Kopanas et al., 2002] Kopanas, I., Avouris, N., & Daskalaki, S. (2002). The role of doamine knowledge in large scale data mining project. *LNAI*, (2308), 288–299.

[Lavrac & Dzeroski, 1994] Lavrac, N. & Dzeroski, S. (1994). *Inductive Logi Programming : Techniques and Applications*. Ellis Horwood, New York.

[Little & Rubin, 1986] Little, R. & Rubin, D. (1986). *statistical analysis with missing data*. John Wiley and Sons, Inc.

[Littlestone & Warmuth, 1989] Littlestone, N. & Warmuth, M. (1989). The weighted majority algorithm. *Technical Report*.

[Liu et al., 1997] Liu, B., Hsu, W., & Chen, S. (1997). Using general impressions to analyze discovered classification rules. *In Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, (pp. 31–36).

[Liu et al., 1999] Liu, H., Lu, H., Feng, L., & Hussain, F. (1999). Efficient search of reliable exceptions. *In Proceedings of the Third Pacific-Asia Conference on Knowledge Discovery and Data Mining*, (pp. 194–203).

[Major & Mangano, 1993] Major, J. & Mangano, J. (1993). Selecting among rules induced from hurricane database. *In Knowledge Discovery in Databases*, Workshop, 28–41.

[Mannila, 1996] Mannila, H. (1996). Data mining : Machine learning, statistics, and databases. In *Statistical and Scientific Database Management* (pp. 2–9).

[McLachlan & Krishnan, 1997] McLachlan, G. & Krishnan, T. (1997). The em algorithm and extensions.

[Mitra et al., 2002] Mitra, S., Pal, S., & Mitra, P. (2002). Data mining in soft computing framework : A survey. *Ieee Trans. On Neural Networks*, 13.

[Morel & Lacheret-Dujour, 1998] Morel, M. & Lacheret-Dujour, A. (1998). Utilisation d'une structure arborescente pour une hiérarchisation fine des règles de transcription graphème-phonème. *XXIIèmes Journées d'Etudes sur la Parole*, (pp. 151–154).

[Morel & Lacheret-Dujour, 2001] Morel, M. & Lacheret-Dujour, A. (2001). Le logiciel de synthèse vocale kali : de la conception à la mise en œuvre. *TAL*, 42, 193–221.

[Nédellec, 2000] Nédellec, C. (2000). Knowledge extraction from text, a machine learning approach. *Learning's WWW*, (pp. 107–117).

[Piatesky-Shapiro, 1991] Piatesky-Shapiro, G. (1991). Discovery, analysis and presentation of strong rules. *In Knowledge Discovery in Databases*, (pp. 229–248).

[Portinale & Saitta, 2002] Portinale, L. & Saitta, L. (2002). Feature selection. *Technical report.*

[Quinlan, 1993] Quinlan, J. (1993). *C4.5 : programs for machine learning.* Morgan Kaufmann Publishers Inc.

[Renaux, 2003] Renaux, P. (2003). Catmine : Computer assisted trade-mark infringement evaluation. *Legal Knowledge and Information Systems - JURIX03*, (pp. 61–70).

[Renaux, 2005a] Renaux, P. (2005a). Classification de documents électroniques juridiques. *International Workshop on Legal Electronic Documents - Beyrouth - Liban.*

[Renaux, 2005b] Renaux, P. (2005b). Interactive classification of legal electronic documents. *Human System Learning Who is in control ? CAPS05*, (pp. 303–314).

[Renaux, 2006] Renaux, P. (2006). Catmine : A virtual office approach for magistrate. *International Workshop on New Trends in Information Technology - NTIT2006.*

[Renaux & Zreik, 2004] Renaux, P. & Zreik, K. (2004). Trade mark registration : the multilanguage aspect. *International Conference on Technology : from Theory to Application*, (pp. summary : 439–440).

[Rioult, 2002] Rioult, F. (2002). Représentation condensée de bases de données et valeurs manquantes.

[Rissanen, 1986] Rissanen, J. (1986). Stochastic complexity and modeling. *Ann. Statist*, 14, 1080–1100.

[Sahar, 1999] Sahar, S. (1999). Interestigness via what is not interesting. *KDD99*, (pp. 332–336).

[Sammon, 1969] Sammon, J. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18, 401–409.

[Schafer & Graham, 2002] Schafer, J. & Graham, J. (2002). Missing data : Our view of the state of the art. *Psychological Methods*, 7, 147–177.

[Schapire, 1990] Schapire, R. (1990). The strength of weak learnability. *Machine Learning*, 5, 197–227.

[Sebag & Gallinari, 2002] Sebag, M. & Gallinari, P. (2002). Apprentissage artificiel : acquis, limites et enjeux. *2ème assises nationales du GDR i3*, (pp. 303–333).

[Silberschatz & Tuzhilin, 1995] Silberschatz, A. & Tuzhilin, A. (1995). On subjective measures of interestingness in knowledge discovery. *In Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, (pp. 275–281).

[Silberschatz & Tuzhilin, 1996] Silberschatz, A. & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *Ieee Trans. On Knowledge And Data Engineering*, (pp. 970–974).

[Smyth & Goodman, 1991] Smyth, P. & Goodman, R. (1991). Rule induction using information theory. *In Knowledge Discovery in Databases*, (pp. 159–176).

[Ting, 1994] Ting, K. (1994). Ddiscretization of continuous-valued attributes and instances-based learning. *Technical Report 491.*

[Todorovski & Dzeroski, 2000] Todorovski, L. & Dzeroski, S. (2000). Combining multiple models with meta decision trees. *Proceedings of the Fourth European Conference on Principles of Data Mining and Knowledge Discovery*, (pp. 54–64).

[Toivonen, 1996] Toivonen, H. (1996). Sampling large databases for association rules. In T. M. Vijayaraman, A. P. Buchmann, C. Mohan, & N. L. Sarda (Eds.), *In Proc. 1996 Int. Conf. Very Large Data Bases* (pp. 134–145). : Morgan Kaufman.

[Tsymbal & Puuronen, 2000] Tsymbal, A. & Puuronen, S. (2000). Bagging and boosting with dynamic integration of classifiers. *PKDD*, (pp. 116–125).

[Venturini et al., 1997] Venturini, G., Silmane, M., Morin, F., & de Beauville, J. A. (1997). On using interactive genetic algorithms for knowledge discovery in databases. *In Th. Bäck, éd, Proc. of the 7th Internationl Conference on Genetic Algorithms*, (pp. 696–703).

[Vilalta & Drissi, 2002] Vilalta, R. & Drissi, Y. (2002). A perspective view and survey of meta-learning. *Artificial Intelligence Review*, Volume 18, 77–95.

[Wallace & Dowe, 1994] Wallace, C. & Dowe, D. (1994). Intrinsic classification by mml - the snob program. In *7th Australian Joint Conference on Artificial Intelligence* (pp. 37–44). : World Scientific Publishing C.

[Wettschereck & Aha, 1995] Wettschereck, D. & Aha, D. W. (1995). Weighting features. In M. Veloso & A. Aamodt (Eds.), *Case-Based Reasoning, Research and Development, First International Conference* (pp. pages 347–358). Berlin : Springer Verlag.

[Witten & Eibe, 2005] Witten, I. & Eibe, F. (2005). *Data Mining : Practical machine learning tools and techniques.* Morgan Kaufmann.

[Zenko et al., 2001] Zenko, B., Todorovski, L., & Dzeroski, S. (2001). A comparison of stacking with meta decision trees to other combining methods. *ICML 2001.*

[Zhao & Karypis, 2001] Zhao, Y. & Karypis, G. (2001). Criterion functions for document clustering : Experiments and analysis.

[Zhao & Karypis, 2004] Zhao, Y. & Karypis, G. (2004). Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3), 311–331.

[Zhong et al., 1999] Zhong, N., Yao, Y., & Oshuga, S. (1999). Peculiarity-oriented multi-database mining. *In Proceedings of the Third European Conference on the Principles of Data Mining and Knowledge Discovery*, (pp. 136–146).

[Zighed et al., 1999] Zighed, D., Rabaséda, S., Rakotomalala, R., & Feschet, F. (1999). Discretization methods in supervised learning. *Encyclopedia of Computer Science and Technology*, 40, 35–50.