Extraction automatique de stéréotypes à partir de données symboliques et lacunaires

Julien Velcin

Université Pierre et Marie Curie
ufr informatique

THESE DE DOCTORAT

Spécialité : Informatique

Présentée par :
Julien VELCIN

Pour l'obtention du grade de
DOCTEUR de l'UNIVERSITE PARIS 6

Extraction automatique de stéréotypes à partir de données symboliques et lacunaires

JURY

M. Jean-Gabriel	GANASCIA	Directeur de thèse
M. Laurent	CHAUDRON	Rapporteur
M. Antoine	CORNUEJOLS	Rapporteur
M. Edwin	DIDAY	Examinateur
Mme Danièle	DUBOIS	Examinatrice
M. Jean-François	PERROT	Examinateur
M. Francis	CHATEAURAYNAUD	Invité

Soutenue publiquement le 30 Novembre 2005

Extraction automatique de stéréotypes à partir de données symboliques et lacunaires

Cette thèse porte sur la construction automatique de stéréotypes à partir d'informations lacunaires. Nous avons choisi comme application les articles de presse car ils présentent justement ce caractère lacunaire. Or, tant en analyse de données qu'en apprentissage artificiel, les valeurs manquantes sont généralement considérées comme des anomalies qui sont traitées avec des techniques spécifiques, comme les k-plus-proches voisins ou l'algorithme EM, avant d'appliquer des méthodes usuelles d'analyse. La formation automatique de stéréotypes à partir de données symboliques partiellement décrites fait appel à un algorithme original d'apprentissage non supervisé fondé sur la notion de subsomption par défaut et à des techniques d'optimisation par recherche locale. La validation s'effectue à la fois à partir de données artificielles dégradées et de données réelles tirées d'articles de presse. Un lien est ainsi établi entre les techniques d'IA et le domaine de l'analyse du contenu de la presse.

Stereotype Extraction from Symbolic and Sparse Data

This thesis deals with the extraction of stereotypes from sparse data using Machine Learning techniques. Newspaper articles were chosen as an application because they have this characteristic of sparseness. In Data Analysis and Machine Learning, missing data are often considered as anomalies that must be corrected with specific techniques, such as the k-nearest neighbors or the EM algorithm, before applying usual analysis methods. Stereotype extraction from symbolic and partially described data uses an original non-supervised learning algorithm based on the default subsomption notion and addresses the problem with a local search strategy. The validation is done with both artificially degraded datasets and real datasets extracted from newspaper articles. Hence, a link is made between AI techniques and press content analysis.

Remerciements

Je tiens tout d'abord à exprimer toute ma reconnaissance à mon directeur de thèse, Jean-Gabriel Ganascia, professeur à l'Université de Paris 6. Le travail que j'ai effectué durant ces trois années au LIP6 doit beaucoup au soutien qu'il m'a toujours témoigné, aussi bien sur le plan scientifique que sur le plan moral. Son enthousiasme communicatif et sa vision de la recherche m'ont permis, non seulement d'aborder les problèmes qui m'étaient posés avec rigueur et passion, mais également de me former au métier de chercheur en informatique.

Je remercie Laurent Chaudron, ingénieur de recherches à l'ONERA de Toulouse, d'avoir accepté de rapporter ma thèse. Je lui suis très reconnaissant pour ses nombreux commentaires éclairants sur mon travail. Ils m'ont été très précieux.

Je remercie Antoine Cornuéjols, maître de conférences au LRI de l'Université d'Orsay, de s'être intéressé à mes travaux et d'avoir accepté de rapporter ma thèse.

Je remercie Edwin Diday, professeur à l'Université Paris Dauphine, Danièle Dubois, directrice de recherches à l'Université Paris 6, Jean-François Perrot, professeur émérite à l'Université de Paris 6, et Francis Chateauraynaud, directeur d'études à l'EHESS, qui ont bien voulu faire partie de mon jury de thèse. Je remercie plus spécialement Francis Chateauraynaud, qui a manifesté, depuis maintenant presque deux ans, un intérêt toujours renouvelé vis-à-vis de mon travail, ainsi que Danièle Dubois, avec qui nous avons eu de très intéressantes conversations concernant la difficile question de la catégorisation.

Ce fut un réel plaisir de faire partie de l'équipe ACASA et je remercie chaleureusement toutes les personnes qui m'ont supporté (dans tous les sens du terme) durant ces années et sont devenus des amis : Isabelle (rare mais précieuse présence féminine), Charles, Aydano, Giordano, Julien... et le "petit" dernier David. Merci pour cette ambiance de travail ô combien agréable, au rythme de la musique brésilienne bien sûr !

Je remercie les nombreuses personnes que j'ai pu côtoyer au LIP6 et qui n'ont pas hésité à prendre de leur temps pour m'apporter leur aide, sans oublier bien sûr les membres du personnel administratif. Un grand merci à Ghislaine qui s'est toujours montrée d'une gentillesse et d'une compétence qui m'a souvent permis de ne pas me perdre dans les méandres de l'administration.

Bien sûr je remercie tous mes amis qui ont toujours répondu présents dans les moments difficiles. Une mention spéciale à Greg et surtout à Agnès qui ont effectué un ingrat mais pourtant indispensable travail de relecture.

Pour finir, je remercie tous les membres de ma famille qui m'ont toujours témoigné une indéfectible confiance et un soutien sans lequel rien n'aurait été possible. Toutes mes pensées vont à ma mère chérie, partie trop tôt, à la mémoire de laquelle cette thèse est dédiée.

À ma mère

Table des matières

Notations

Notations générales

$\mathcal{D}$	L'espace de descriptions
$\mathcal{E}$	L'espace des exemples
$\mathcal{R}$	L'espace des ensembles de stéréotypes couvrants
$\mathcal{R}'$	L'espace des ensembles de stéréotypes couvrants respectant les trois contraintes fixées (poids minimum, non redondance et cohésion cognitive)
E	Un ensemble d'exemples à classer, $E \subset \mathcal{E}$
n_E	Le nombre d'exemples à classer
n_d	Le nombre total de descripteurs dans E
m	La proportions de descripteurs manquants dans E
$d_1, \dots d_n$	Des descriptions, $d_i \in \mathcal{D}$
$s_1, \dots s_n$	Des stéréotypes, $s_i \in \mathcal{D}$
s_T	Le stéréotype vide, $\forall s \in \mathcal{D}, s_T \leq s$
S	Un ensemble de stéréotypes couvrant, $S \in \mathcal{R}$
S^*	L'ensemble de stéréotypes couvrant S sans le stéréotype vide s_T
$\leq$	La relation de subsomption entre deux descriptions
$\leq_D$	La relation de subsomption par défaut entre deux descriptions
δ	La fonction qui associe à chaque exemple e sa description $\delta(e) \in \mathcal{D}$
ρ	La fonction qui associe à chaque exemple e son poids $\rho(e) \in \mathbb{R}_*^+$
ρ_E	La fonction qui associe à un ensemble de stéréotypes couvrant S son poids relativement à E, $\rho_E(S) \in \mathbb{R}^+$
$\rho_{S,E}$	La fonction qui associe à un stéréotype $s \in S$ son poids relativement à S et E, $\rho_E(S) \in \mathbb{R}^+$
ρ_{pds}	Le poids minimum des clusters associés aux stéréotypes
M	La mesure de comparaison utilisée pour associer les exemples aux stéréotypes
I	Les descriptions initiales ayant généré E dans le cas des jeux de données artificiels
n_I	Le nombre de descriptions initiales
dup	Le nombre de duplication de chaque description initiale de I

Le formalisme attribut-valeur

$\mathcal{A}$	L'ensemble des attributs
$V(X)$	L'ensemble des valeurs pouvant être prises par l'attribut X
$(x_i \leq X \leq x_j)$	Un descripteur associé à l'attribut X
$x_?$	Le descripteur indéfini
$c_\emptyset$	Le descripteur absurde
$\mathcal{C}$	L'ensemble des descripteurs
$d_{\|X}$	La projection de la description $d \in \mathcal{D}$ sur l'attribut X
$\|d\|$	La richesse de la description $d \in \mathcal{D}$
D_X	La fonction qui associe à une ou plusieurs descriptions la valeur vraie si celles-ci décrivent l'attribut X
τ_X	La fonction qui vérifie si deux descriptions sont cohérentes suivant l'attribut X
τ	La fonction qui vérifie si deux descriptions sont cohérentes
M_c	La mesure de comparaison (non normalisée) entre deux descriptions
M_N	Les quatre mesures de comparaison normalisées (N = 1, 2, 3 ou 4)
sim, dis	La mesure de similarité (resp. dissimilarité) qui ne prend pas en compte la subsomption par défaut
q_N	Les quatre fonctions d'évaluation (N = 1, 2, 3 ou 4)

Les indices d'évaluation

D	Les descriptions proposées par l'algorithme de classification
n_D	Le nombre de descriptions proposées par l'algorithme de classification
sep_1 et sep_2	Les scores de séparation de D
$couv$	Le taux de couverture des exemples
cmp_1 et cmp_2	Les scores de compacité de D sur E
$adeq_E$, $cont_E$, et $perte_E$	La proportion des descripteurs de E en adéquation, en contradiction, ou non présents dans D
$pred$	Les capacités prédictives de D
$cont_s$ et $cont_p$	Le support et la proportion des descripteurs contradictoires
$adeq_I$, $cont_I$, et $perte_I$	La proportion des descripteurs de I retrouvés, en contradiction, ou n'apparaissant pas dans D
err_C	La proportion des exemples dont la classe est mal prédite

Introduction

L'intuition à la base de cette thèse est la suivante : la description des objets (événements, personnes, objets matériels) que nous percevons à travers les médias est nécessairement partielle et nous pousse à interpréter l'information manquante. Malgré cela, la partie connue des informations qui proviennent de ce type de source sont utilisées pour construire une représentation des ces objets. La partie qui nous est cachée est naturellement sujette à une "pression à l'inférence" [Mos76][Jod89] qui nous fait raisonner sur l'objet de notre attention malgré l'incomplétude de notre connaissance sur un sujet donné.

La plupart des outils utilisés dans le domaine informatique ne sont pas adéquats pour traiter ce type de données. Ceci est principalement dû au fait que ces outils ont été élaborés, dans une large majorité, pour manipuler des données numériques à l'aide de techniques issues des statistiques, alors que de nombreuses informations sont par nature symbolique (couleur et forme d'un objet, nationalité et origine sociale d'un individu). Le caractère lacunaire des données est souvent considéré comme une anomalie qu'il faut traiter à l'aide de techniques spécifiques avant d'appliquer des méthodes standards d'analyse. De surcroît, la proportion des valeurs manquantes considérées dans les expérimentations réalisées jusqu'à présent est souvent faible. Ainsi, elle ne dépasse pas 10% de l'information totale dans [HL04], 30% dans [IL01], voire 50% à de très rares occasions [FKP04]. Nous verrons que le taux de valeurs manquantes dans les données que nous manipulons dépasse aisément ces valeurs.

L'objectif de cette thèse est de proposer un outil capable de construire automatiquement une représentation de données lacunaires par une approche symbolique. Cette représentation, qui ne se targue pas d'être l'unique représentation possible des données sujettes à l'étude, repose sur une caractérisation à base de stéréotypes. Les stéréotypes sont une forme de description des données imparfaitement connues qui s'inspire de la logique des défauts de R. Reiter [Rei80]. Ils sont construits en même temps qu'un ensemble de classes (*clusters*) regroupant chacune des objets jugés similaires au cours d'une tâche de classification non-supervisée (*clustering*). Ils vérifient certains critères, comme celui de couverture par défaut et de cohésion cognitive, qui sont présentés dans le modèle général.

Nous insistons sur le fait que la *découverte* des stéréotypes est au cœur de notre travail et a plus d'importance que la manière dont les objets sont placés dans les catégories. Ces stéréotypes prennent la forme de descriptions riches, facilement interprétables, cohérentes avec les données et permettant de construire des catégories avec une bonne cohésion interne.

Analyse de données et intelligence artificielle

M. Volle définit l'analyse de données comme une *"statistique descriptive perfectionnée"* qui *"permet de décrire plus rapidement et plus sûrement de grands gisements de données. Elle est un outil précieux pour le chercheur qui veut extraire le maximum des données qu'il a collectées."* [Vol81]. Cette discipline s'est, depuis de nombreuses années, attaché au problème de la mise au point de techniques permettant de trouver des caractéristiques intéressantes dans un ensemble de données (corrélations, motifs récurrents, etc.). L'analyse de données [Ba73][DLPT84][Sap90] se divise en deux grandes familles de méthodes : les techniques de classification automatique et les techniques d'analyse factorielle. La plupart des travaux effectués dans ce domaine sont adaptés au traitement des données numériques et reposent sur des modèles statistiques. Les données symboliques, quant à elles, sont souvent traitées sous la forme de variables binaires (tableaux logiques de 0 et de 1). L'analyse de données symboliques [Did87][BD00] propose cependant de manipuler des structures plus complexes, comme des distributions de probabilité ou des intervalles.

Cependant, notre travail ne se résume pas à une analyse descriptive des données et se place plus précisément dans le domaine de l'intelligence artificielle. Il se rapproche de la classification conceptuelle [Mic80], voire de la formation de concepts [GLF92], tous deux généralement associés au champ de l'IA appelé apprentissage artificiel (ou automatique). Bien entendu, le champ qui nous semble le plus adapté est celui de la classification conceptuelle, dans le sens où nous sommes plus intéressés dans l'extraction d'une caractérisation d'un ensemble d'objets que dans la construction effective des catégories. Dans ce contexte, les stéréotypes que nous cherchons à extraire correspondent à un ensemble de concepts décrivant des classes d'objets. Le lien avec la formation de concepts vient du fait que notre modèle peut être étendu afin de considérer les exemples de manière incrémentale, construisant de ce fait les stéréotypes au fur et à mesure de l'arrivée des données. Notons enfin que notre manière de traiter le cas des données lacunaires s'éloigne des méthodes habituellement utilisée en analyse de données et s'inscrit dans une approche beaucoup plus orientée vers l'apprentissage artificiel.

Traitement des valeurs manquantes

Rares sont les bases de données dans lesquelles les valeurs des variables sont toutes parfaitement connues. Or, les techniques d'analyse utilisent usuellement des tableaux de données complets et la qualité des résultats fournis décroît très rapidement avec la proportion de valeurs manquantes. C'est pourquoi de nombreuses techniques ont été mises au point qui permettent de combler automatiquement ces "trous" [LR02]. Parmi celles-ci se trouvent les techniques les plus triviales, comme la suppression pure et simple des données présentant le moindre caractère lacunaire (*listwise deletion*), des techniques d'imputation plus sophistiquées (*hot-desk imputation*, *maximum likelihood imputation*), voire franchement complexes (*multiple imputation*). Quoiqu'il en soit, trouver la technique la plus adaptée à un type de

données lacunaires reste une tâche délicate. Notons le travail original effectué par J.B. Weinberg, G. Biswas et G.R. Koller [WBK92] qui proposent de modifier l'algorithme ITERATE [WBF98] pour traiter le cas des valeurs manquantes d'origine systématique.

L'approche que nous avons choisie, dans notre tâche de classification des données lacunaires, consiste à ne pas considérer la lacunarité des données comme une anomalie mais plutôt comme un caratère propre à un certain type de données qui demande un traitement adapté. Ainsi, il nous semble préférable, lorsque le taux de valeurs manquantes est très important, de travailler directement les données comportant des valeurs manquantes et d'utiliser la plus grande partie de l'information disponible. En d'autres termes, l'analyse est effectuée en prenant en compte le caractère lacunaire des données, au lieu de compléter, dans un premier temps, ces données par une techique d'imputation et d'effectuer, ensuite, une analyse plus classique. Avec cette hypothèse, nous nous plaçons dans un cadre similaire à celui de l'algorithme *Expectation-Maximization* (EM) [DLD77] capable de manipuler des variables aux valeurs inconnues dans un cadre probabiliste et cherchant à maximiser un critère de maximum de vraisemblance.

Cependant, le modèle que nous proposons dans cette thèse diffère de l'algorithme de Dempster suivant au moins trois points fondamentaux. Le premier est que nous ne travaillons pas avec des distributions de probabilité et qu'en conséquence les concepts formés ne sont pas des concepts probabilistes. Il s'agit de descriptions "certaines" comme celles qui sont manipulées dans des algorithmes plus classiques comme celui des c-moyennes. De telles descriptions sont en effet indispensables lorsqu'il s'agit de compléter les données manquantes. Le deuxième point est que nous nous attachons surtout à construire une caractérisation de l'ensemble des données sous la forme de stéréotypes, représentants pertinents considérés plus importants que la catégorisation elle-même. Par représentants pertinents, nous entendons ici des descriptions suffisamment distinctes pour pouvoir discriminer rapidement de nouveaux exemples, mais également suffisamment riches pour effectuer une bonne prédiction des valeurs inconnues. Le dernier point qui différencie notre approche est la volonté de considérer des contraintes d'ordre cognitif, comme la cohérence, assurée par la relation de subsomption par défaut, ou la cohésion cognitive. Il inscrit clairement notre travail dans une perspective résolument tournée vers l'intelligence artificielle.

Données symboliques et comparaison

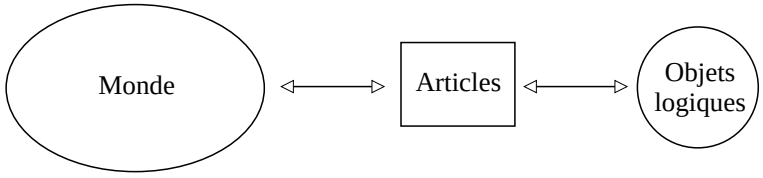
Le cadre des données symboliques (catégorielles et ordonnées) que nous avons choisi pose des problèmes spécifiques. Ainsi, la notion de distance, qui se trouve à la base du calcul des catégories, s'en trouve nécessairement altérée. Il n'est plus possible de recourir à des propriétés géométriques, comme la densité ou la convexité des données, qui perdent une grande part de leur sens. Certaines réponses à ce type de problème sont proposées dans la littérature. Ainsi, D. Gibson [GKR98] utilise une technique itérative de propagation des poids par co-occurrences, basée sur une structure de graphe spectral. S. Guha [GRS00], quant à lui, utilise dans son algorithme ROCK la notion de "lien" à la place de celle de distance dans

un algorithme de clustering hiérarchique.

Le problème se complexifie dès lors que l'on ajoute la contrainte de lacunarité des données. Ainsi, la solution généralement employée lorsque l'on compare deux objets comportant des traits inconnus est, tout simplement, de les ignorer. C'est ce que l'on appelle parfois la technique de *pairwise deletion*. Dans le cadre attribut-valeur, cela consiste à ne prendre en compte que les attributs possédant une valeur dans les deux vecteurs décrivant les objets comparés. Or, cette solution naïve peut amener à de très mauvais résultats. Il nous semble qu'une étude plus fine de la comparaison entre des objets symboliques et lacunaires soit nécessaire. C'est pourquoi nous proposons de comparer l'effet de plusieurs mesures sur la construction des catégories. Ces mesures sont inspirées de mesures de ressemblance classiques, dérivées de la théorie de la mesure [Spä80], comme la mesure de Russel et Rao ou la mesure de Jaccard.

Application à l'analyse du contenu de la presse

L'application privilégiée que nous proposons dans cette thèse concerne l'analyse du discours d'articles tirés de la presse écrite. Les articles, qui constituent déjà une certaine représentation du monde donnée par les médias, sont traduits dans un formalisme logique comme illustré par la figure suivante :



Dans le cadre de notre travail, la traduction des textes dans un tel formalisme a été réalisée manuellement car il n'existe pas, à l'heure actuelle, de programme permettant d'exécuter une tâche aussi complexe automatiquement[1]. C'est une fois cette traduction effectuée que nous nous sommes aperçus du caractère extrêmement lacunaire des descriptions produites. En effet, l'information délivrée par la presse est souvent très faible au regard de la totalité du langage nécessaire pour réaliser le codage. Ce phénomène nous semble parfaitement naturel[2] et demande un traitement adapté. C'est justement cette constatation qui nous a amené à proposer le modèle à base de stéréotypes présenté dans cette thèse.

De nombreux travaux en analyse des discours de presse sont entièrement réalisés à la main [Mos76][Sou92]. Des logiciels d'analyse automatique permettent aujourd'hui de soulager les sociologues d'une partie de ce travail long et fastidieux, même s'ils ne remplacent pas l'analyse réalisée par un ou plusieurs experts. La plupart de ces logiciels, comme le logiciel Alceste [Rei86], repose sur la fréquence de mots-clefs et sur des techniques classiques en analyse de données. Plus récemment, certains logiciels, comme le logiciel Prospero [Cha03] utilisé

[1] Nous parlons bien entendu ici d'un programme qui parviendrait à traduire fidèlement le contenu des discours relatés dans des articles de journaux.

[2] En effet, comment imaginer pouvoir coder absolument toutes les informations relatives à un événement ?

plus spécifiquement en sociologie ou le logiciel NOMINO [PDP98] basé sur des réflexions linguistiques, adoptent une approche qui cherche à représenter le sens des textes à analyser à l'aide de techniques parfois très proches de celles utilisées en intelligence artificielle.

Notre modèle utilise un raisonnement par défaut utilisant une représentation logique des discours relatés dans les articles de presse. Loin des représentations habituellement employées en analyse textuelle, comme les vecteurs basés sur la présence ou la fréquence de mots-clefs, nous essayons de coder le sens des discours à l'aide d'un langage proche de la logique des propositions, souvent employé en IA : le formalisme attribut-valeur. Il donne lieu à des expérimentations réalisées et commentées dans le chapitre 5. Afin de montrer que l'idée qui se trouve derrière notre modèle n'est pas uniquement adaptée à ce langage, nous apportons quelques éléments afin de l'étendre au formalisme beaucoup plus riche des graphes conceptuels élaboré par J.F. Sowa [Sow84].

Enfin, remarquons que notre approche peut être adaptée afin de traiter deux stratégies d'étude de la presse. La première possibilité est de considérer une période de temps dans sa globalité afin d'en extraire une représentation qui pourra être comparée à celle obtenue à une autre période, ou à partir d'autres sources. C'est cette approche, que nous qualifions de globale, que nous avons choisie d'adopter dans cette thèse. La seconde possibilité consiste à effectuer une analyse plus détaillée d'une même période en considérant les informations données jour après jour (approche incrémentale). Cette dernière manière d'aborder la construction des représentations est celle qui se rapproche du champ de la formation de concepts en intelligence artificielle.. Une adaptation mineure de notre modèle devrait être capable de rendre compte de cette approche.

Extraction de stéréotypes

Dans les études sur la catégorisation, la théorie du prototype joue un rôle central. Développée par E. Rosch et ses associées [Ros73], cette théorie cognitive place au centre des représentations des objets typiques appelés prototypes. Les autres objets observés se situent au regard de ces objets spécifiques, réellement existants ou purement imaginaires, en fonction d'un gradient de typicalité. De nouvelles recherches menées dans les années 80 dans le domaine de la linguistique [Lak87] étendent la théorie en réfutant la centralité du prototype et la structure de gradient, retournant à un système basé en priorité sur la notion de ressemblance de famille initiée par le philosophe L. Wittgenstein [Wit53]. Cette théorie étendue de la sémantique du prototype est jugée par certains comme totalement nouvelle, et non comme une simple extension de la théorie classique [Kle90]. L'approche que nous avons choisie se situe dans le cadre de la théorie du prototype au sens général d'organisation formelle des catégories, mais partage plusieurs similitudes avec ces derniers travaux. Nous mettons en effet l'accent sur la notion d'"air de famille" et réfutons l'existence d'un degré d'appartenance à la catégorie basée sur la similarité avec un ou plusieurs exemples typiques.

L'hypothèse principale de notre thèse est qu'un ensemble d'objets dont les observations sont très lacunaires vont former des catégories que nous pouvons décrire à l'aide de sté-

réotypes. Le stéréotype est une sorte de prototype construit, faute d'une information plus complète ou plus savante, à l'aide des rares informations disponibles. Il correspond donc à une certaine vision du monde, souvent déformée, décrite à travers des médias comme la presse écrite. La ressemblance de famille est utilisée comme une contrainte que les catégories doivent respecter en vérifiant un critère de cohésion cognitive, critère basé sur la co-occurrence entre les différentes caractéristiques du stéréotype.

Outre la cohésion interne des catégories, nous imposons une contrainte de non-redondance entre l'information délivrée par les stéréotypes. Cette contrainte peut sembler forte à première vue mais elle correspond bien au concept de stéréotypes et au contexte lacunaire. Elle facilite notamment l'interprétation de ces stéréotypes et leurs capacités discriminantes. De plus, nous proposons de relâcher cette contrainte en prenant en compte à la fois l'homogénéité interne à chaque catégorie, mais également la séparation entre les différents stéréotypes.

Lien avec les travaux antérieurs

Dans les années 30, la logique de la découverte scientique a été étudiée par le philosophe K. Popper [Pop73]. Il y présente notamment le critère de "falsifiabilité" qui permet de déterminer si une théorie peut être ou non considérée comme scientifique. Cette première approche se concentrait surtout sur la structure des découvertes et il faudra attendre les années 60 pour que l'on s'intéresse plus spécifiquement aux processus à l'origine de ces découvertes. Ainsi, la découverte scientifique est devenue un champ de recherche de l'intelligence artificielle [LJ89][Gan02] grâce aux travaux de l'économiste H. Simon, champ où l'on trouve des noms comme P. Langley, J. Zytkow et D. Lenat. L'objectif de ce champ de recherche est la reconstruction rationnelle d'anciennes découvertes scientifiques et la génération de nouvelles théories scientifiques à l'aide de techniques informatiques. Le postulat fondamental est que la découverte scientifique et la créativité peuvent être abordées comme deux cas de résolution de problème et traitées à l'aide des machines.

L'intérêt de ce courant de recherche est double. Tout d'abord, l'intérêt est d'ordre épistémologique puisqu'il s'agit là d'un moyen privilégié pour modéliser l'activité scientifique, a contrario du "eureka" génial mais ô combien hermétique d'Archimède. D'un autre côté, l'intérêt est pratique car l'ordinateur peut être alors utilisé comme un outil secondant l'homme dans la recherche de nouveaux savoirs. Parmi les travaux les plus saillants de ce domaine, on peut citer les systèmes DENDRAL [FBJ71] et BACON [LBS83].

I. Lakatos, inspiré par les théories de Hegel, Marx, Popper et par les travaux du mathématicien G. Polya, pense qu'il n'existe pas de théories scientifiques parfaites, irréfutables [Lak76]. Bien au contraire, parmi les activités scientifiques qui peuvent être simulées, beaucoup reposent sur un raisonnement qui se révèle erronné [Gan04]. La source de ces erreurs est en partie due au manque d'information, mais également à l'intervention de théories implicites sur le domaine. Dans ce cadre, J.-G. Ganascia et V. Corruble ont travaillé sur la découverte des causes du scorbut en utilisant des techniques d'apprentissage symbolique [CG97]. Les méthodes inductives employées, utilisant des règles d'association [Gan90] basées

sur une structure de treillis de gallois [Gan][Gan93a][Gan93b], ont permis d'aboutir à des résultats très prometteurs qui corroboraient certaines découvertes réalisées en médecine au XIX^{e} siècle. Elles ont ainsi démontré l'influence prépondérante des théories implicites qui, s'appuyant sur des données lacunaires, biaisent le processus d'induction.

Ces premiers travaux ont donné lieu à de nouvelles recherches concernant la découverte des causes de la lèpre, recherches qui ont abouti à la thèse de V. Corruble soutenue en 1996 [Cor96]. Cependant, ces recherches restaient circonscrites au champ de la médecine et l'idée fut proposée par J.G. Ganascia d'appliquer les mêmes méthodes d'induction au champ des sciences sociales et au raisonnement de sens commun. Ces réflexions ont conduit à des recherches, réalisées dans le cadre du DEA IARFA, sur la découverte d'explications relatives à des phénomène sociaux [Vel02]. Partant d'articles de journaux, daté de la fin du XIX^{e} siècle et précédant d'un an le début de la célèbre Affaire Dreyfus, l'idée était de construire automatiquement et de manière inductive des explications concernant le désordre de la scène politique française à cette époque. Trois quotidiens aux orientations politiques différentes ont été considérés, ainsi que quatre théories ayant pu avoir cours à cette époque. Ces recherches ont permis de constater, là aussi, l'importance des théories implicites dans la construction des représentations. Confirmant l'intuition première, elles ont permis de montrer l'intérêt des méthodes d'induction dans l'étude du sens commun. Ces travaux ont servi de point de départ à l'élaboration de la présente thèse.

Objectif général

Dans le cadre de cette thèse, notre principal objectif est d'extraire une représentation adaptée aux données lacunaires.

Pour cela, nous proposons, dans un premier temps, un modèle basé sur des objets typiques que nous avons appelé des stéréotypes. Ces stéréotypes peuvent être utilisés pour interpréter plus facilement les données, pour compléter les données manquantes et pour prédire des informations concernant un objet nouvellement observé. Ce modèle est défini sur la base de la structure mathématique de treillis afin de pouvoir l'adapter à plusieurs formalismes, comme celui attribut-valeur ou celui des graphes conceptuels.

Dans un deuxième temps, nous élaborons un algorithme de classification par défaut qui repose sur ce modèle. Cet algorithme a pour objectif d'extraire effectivement les stéréotypes à partir des données. Pour ce faire, nous abordons le problème de la classification automatique comme un problème d'optimisation et utilisons des techniques spécifiques de recherche locale.

L'application qui a guidé notre travail est la construction de représentations à partir d'articles tirés de la presse. En effet, ce type de donnée, qui fait référence à de nombreuses connaissances implicites, présente justement un fort caractère lacunaire. Notre algorithme doit donc être capable d'extraire automatiquement des stéréotypes à partir de ces données.

Plan de la thèse

Le manuscrit est divisé en cinq chapitres.

Le premier chapitre propose un état de l'art divisé en trois partie. La première et la plus importante de ces parties expose, dans le contexte de l'apprentissage non-supervisé, le problème particulier des données lacunaires, puis des données symboliques. Nous donnons ensuite quelques techniques permettant d'évaluer les catégories construites à l'aide d'algorithmes de clustering, avant de parler brièvement de la génération des données artificielles. La deuxième partie de ce premier chapitre aborde la question de la catégorisation d'un point de vue psychologique et social. Elle nous permet de discuter des différences entre les concepts de prototype et de stéréotype. Nous faisons à la fin de cette partie le lien avec l'analyse du contenu des discours de presse. La dernière et plus modeste des parties de l'état de l'art est relative aux méthodes d'optimisation pouvant être utilisée pour résoudre une tâche de clustering. Elle se termine en présentant la méta-heuristique de recherche taboue qui est utilisée dans notre algorithme.

Le deuxième chapitre présente notre modèle de représentation à base de stéréotypes. L'approche repose sur la structure mathématique de treillis et tâche d'être la plus générale possible. Après avoir précisé les motivations qui ont conduit à ce modèle, nous établissons une nouvelle relation entre descriptions appelée subsomption par défaut. Nous définissons alors formellement les notions de stéréotype, de couverture relative, de complétion, et discutons de manière générale sur des propriétés demandées aux mesures de comparaison utilisées. Trois contraintes, enfin, sont proposées afin d'éliminer certains ensembles de stéréotypes qui ne correspondent pas au type de représentation souhaité.

Le troisième chapitre donne une implémentation de notre modèle dans deux langages logiques : d'une part une version évoluée du langage attribut-valeur permettant de manipuler des intervalles de valeurs ordonnées ; d'autre part les graphes conceptuels de Sowa. La première partie détaille ce qui est nécessaire pour parvenir à une implémentation effective de l'algorithme de classification par défaut. Il se termine par une étude de la complexité (en temps machine) de l'algorithme. La seconde partie reste très théorique et n'a pour vocation que de donner quelques pistes vers une adaptation de notre modèle dans le formalisme, très riche mais beaucoup plus difficile à implémenter, des graphes conceptuels.

Le quatrième chapitre décrit la méthodologie qui a été employée pour évaluer le résultat des processus de clustering et détaille les jeux de données employés dans nos expérimentations. Nous donnons un ensemble d'indices pouvant être calculés à partir d'une partition, parmi lesquels se trouvent des indices originaux relatifs à notre problématique. Ces indices sont utilisés dans le chapitre suivant pour interpréter les résultats obtenus. Nous abordons également le problème de l'extraction des descriptions étiquetant les clusters donnés par les algorithmes de clustering classiques. Enfin, nous décrivons la méthode employée pour générer les jeux de données artificiels, avant de détailler les données réelles tirées de la presse du XIX^{e} siècle.

Le cinquième et dernier chapitre est réservé aux expérimentations réalisées à l'aide du

programme implémentant l'algorithme de classification par défaut que nous avons appelé PRESS. Après une rapide description des choix qui ont été fait, le chapitre s'articule autour de deux parties. La première de ces partie donne les résultats obtenus à partir des jeux de données artificiels. Ils permettent tout d'abord d'éliminer des fonctions d'évaluation et des techniques qui conduisent invariablement à de mauvais résultats dans notre problématique. Ils permettent ensuite de se faire une idée de la qualité et de la robustesse des stéréotypes proposés par notre algorithme, notamment en les comparant avec les descriptions associées à trois autres algorithmes de clustering. La deuxième partie permet de tester notre algorithme sur des données réelles et de montrer que les stéréotypes constituent une forme intéressante de représentation du contenu des discours de presse.

Chapitre 1

Etat de l'art

1.1 Apprentissage à partir de données lacunaires

1.1.1 Rappels sur la classification automatique

La classification automatique peut être définie comme regroupant *"l'ensemble des méthodes et algorithmes consistant à découper une population d'objets en plusieurs classes, en tenant compte des variables qui les caractérisent ; ces classes peuvent être plus ou moins liées entre elles, par exemple sous forme d'un arbre dont chaque sommet représente une classe."* [Did79]. Son champ d'application est très large, répondant tout d'abord aux besoins exprimés par les naturalistes dans le classement des plantes et des espèces [SS63] puis s'attachant à des domaines aussi divers que la géologie, la reconnaissance des formes, les sciences sociales, etc. Outre les ouvrages déjà cités, nous nous basons sur plusieurs articles et ouvrage de référence [Spä80][JMF99][Buh02].

1.1.1.1 La classification automatique en apprentissage artificiel

En intelligence artificielle, l'apprentissage se décline suivant deux approches :

L'apprentissage supervisé, aussi appelé apprentissage *par les exemples*, repose sur l'intervention d'un professeur (aussi appelé oracle). A partir, d'une part, d'un ensemble d'observations servant de base à l'apprentissage et, d'autre part, d'un professeur chargé de guider cet apprentissage, l'objectif est d'apprendre un modèle capable de calculer la sortie la plus vraisemblable d'un système quelle que soit l'entrée observée. Ainsi, un modèle de classement (*classification*) est chargé d'estimer la classe d'un nouvel objet à partir d'un ensemble d'objets dont la classe est donnée.

D'un autre côté, l'apprentissage non-supervisé, ou apprentissage *par observations*, permet d'induire de nouvelles informations sans l'intervention d'un tel professeur. Plus spécifiquement, la classification automatique[1] (*clustering*) s'intéresse au processus de formation des catégories, c'est-à-dire à la manière de classer les données observées au sein de structures

[1] Attention à ne pas confondre la classification automatique avec la classification qui est une tâche d'apprentissage supervisée. C'est pourquoi nous préférons employer le terme de "classement" comme dans [CM02].

dans lesquelles celles-ci se regroupent "naturellement". L'objectif du clustering est de placer ensemble dans des classes (ou *clusters*) des observations jugées similaires (similarité intra-classe), toujours au regard d'une certaine métrique, et de placer les observations jugées dissimilaires dans des classes distinctes (dissimilarité inter-classes). La recherche de la "meilleure" classification de ces observations peut être abordée comme un problème d'optimisation d'une fonction d'évaluation au sein d'un espace d'hypothèses [Did79][CM02]. Chaque hypothèse de cet espace est alors une solution potentielle au problème de clustering. Le rôle de la fonction d'évaluation est de comparer ces solutions, selon un ou plusieurs critères fixés à l'avance, afin de trouver l'hypothèse la plus satisfaisante.

Le clustering conceptuel est une amélioration des techniques de clustering habituelles qui a été proposée par R. Michalski, R. Stepp et E. Diday au début des années 80 [Mic80][MSD81]. Les applications sont nombreuses, dans des domaines aussi divers que la chimie, la biologie, la génétique, la sociologie. Cette technique enrichit le clustering en fournissant une caractérisation des clusters à l'aide de concepts et se place traditionnellement dans le domaine de la fouille de données (*data mining*). Cette caractérisation est très importante si l'on souhaite raisonner à partir des clusters qui ont été construits et peut être utilisée afin de classer de nouvelles observations ou prédire leurs caractéristiques. Nous verrons que c'est justement le cas dans notre travail.

1.1.1.2 Deux approches du clustering

Il existe deux manières d'aborder les problèmes de clustering.

Le premier repose sur l'utilisation d'une hiérarchie de concepts afin de caractériser la structure des catégories. Les algorithmes les plus classiques associés à cette approche sont des variantes de l'algorithme hiérarchique ascendant (aussi appelé *bottom-up*) [SS73][Kin67][War63]. Parmi l'ensemble des techniques de clustering hiérarchique se singularise l'algorithme COBWEB de D. Fisher [Fis87] qui repose sur des conclusions du travail de la psychologue E. Rosch, et plus particulièrement sur la notion de *cue validity* (voir la partie 1.2.1.3). L'idée générale est de maximiser la quantité d'information pouvant être prédite dans les classes que construit l'algorithme. Ce dernier est décrit plus précisément dans la section 1.1.1.4. Il est utilisé à titre de comparaison dans les expérimentations réalisées au chapitre 5.

La deuxième approche en classification automatique consiste à construire un recouvrement de l'ensemble des exemples. On restreint la plupart du temps le recouvrement à une partition de l'ensemble des objets de départ, c'est-à-dire à un ensemble de classes disjointes dont l'union est égale à l'ensemble initial. C'est ce que l'on appelle aussi le *hard clustering* (ou *crisp clustering*) par opposition au *fuzzy clustering*[2]. C'est cette structure, la plus classique, que nous avons choisie de manipuler dans ces présents travaux. La figure 1.1 illustre dans ce cas le processus général de classification automatique.

Nous avons choisi d'utiliser trois algorithmes classiques de clustering afin de comparer les

[2]Nous conseillons aux lecteurs intéressés par la structure de recouvrement de se référer aux travaux sur le clustering flou entrepris par L. Zadeh et ses successeurs [Zad65].

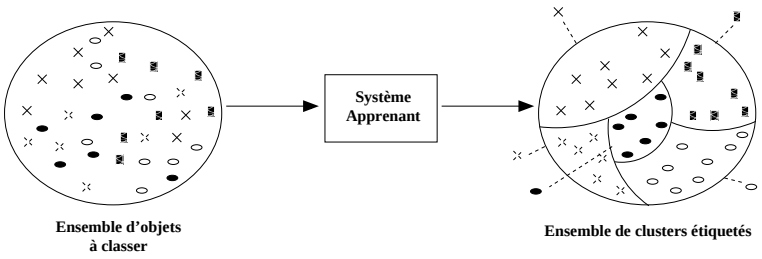


FIG. 1.1 – Apprentissage non-supervisé dans le cas du *crisp clustering*.

performances de l'algorithme d'apprentissage que nous proposons dans le second chapitre. Ces algorithmes sont décrits dans une partie ultérieure. Mais avant, nous nous intéressons aux différents biais qui surviennent lors de l'utilisation de telles techniques.

1.1.1.3 Les biais d'apprentissage

Ces procédés informatiques impliquent l'existence de biais qu'il faut prendre en compte lorsque l'on interprète les résultats obtenus. Ce biais exprime la différence entre l'hypothèse choisie par l'apprenant et celle qu'il aurait dû choisir, en supposant qu'il existe une solution idéale au problème posé. Il a d'ailleurs été prouvé que l'apprentissage était impossible sans l'existence d'un tel biais [Mit80]. Nous présentons ici quatre biais, dont trois sont tirés de [CM02], que nous tâchons de mettre en relation avec notre travail.

Le premier de ces biais est le *biais inductif* (ou *biais*) qui peut être à l'origine d'une erreur dite d'*approximation*. Ce biais est lié à la définition de l'espace des hypothèses, espace qui contient toutes les solutions accessibles à l'apprenant. Or, l'espace des hypothèses dépend nécessairement du langage de représentation utilisé pour coder à la fois les exemples mais aussi les concepts recherchés. Il semble difficile d'être certain que ce langage est le plus adéquat pour la tâche que l'on s'est fixée. Aussi la découverte d'une hypothèse optimisant la fonction d'évaluation ne garantit-elle pas l'obtention d'une solution optimale à notre problème. Cependant, il est possible de réduire ce biais en effectuant des recherches approfondies concernant le thème d'étude choisie (historique, politique, sociologique, etc.) ou en automatisant, du moins en partie, la phase de construction du langage de représentation.

Le second biais est la *variance* qui exprime la dépendance du résultat à l'échantillon des exemples de départ, exemples qui ne constituent souvent qu'un fragment de tous les exemples en rapport avec la problématique choisie. Ce biais provoque une erreur dite d'*estimation* qui peut être réduite dans certains cas. Si l'on considère par exemple les observations tirées de la presse écrite, il peut suffire de définir clairement le thème de l'étude et le mode de sélection des articles. C'est ce que nous avons justement essayé de faire dans notre application sur des données réelles.

Ce second biais nous semble étroitement lié au *biais de restriction*, tiré de [RG90], relatif à la stratégie d'exploration de l'espace des hypothèses. Il s'agit de l'ensemble des choix qui vont sélectionner une partie de l'espace de recherche plutôt qu'une autre, ou bien guider

l'exploration en priorité vers telle ou telle partie. Nous utilisons une stratégie de recherche taboue[3] afin de réduire ce biais et surtout éviter de rester piéger dans des optima locaux. Bien entendu, il n'y a aucun moyen de savoir avec une certitude absolue, dans le cas des données réelles, que la meilleure solution[4] a été obtenue. C'est à la lumière de ces réflexions que l'utilisation de jeux de données artificiels nous semble tout à fait pertinente, dans le sens où la solution optimale (celle qui a permis de générer les données) existe et qu'elle nous est connue.

Le dernier biais, enfin, concerne l'erreur *intrinsèque* provoquée par le "bruit" au sein des données de l'échantillon. En effet, l'information contenue dans les observations peut avoir été modifiée (sciemment ou non), oubliée, ajoutée sans justification, etc. Nous nous intéressons particulièrement au cas où la description des exemples d'apprentissage est très lacunaire. Ainsi, un rédacteur raciste d'une chronique de fait-divers dans un journal pourrait avoir pris l'habitude de préciser systématiquement le pays d'origine d'un agresseur lorsque celui-ci est l'Algérie ou la Tunisie, alors qu'il le passerait sous silence pour une personne à la peau blanche. Qu'ils soient volontaires (choix moral, idéologique, malversation) ou non (négligence, difficultés pour récupérer l'information), ces "trous " constituent justement la base de la création des stéréotypes tels que nous les définissons dans notre modèle. C'est pourquoi nous pensons que ce biais se trouve au cœur même du travail présenté dans cette thèse.

1.1.1.4 Algorithmes classiques de clustering

Nous décrivons à présent trois algorithmes classiques de clustering. Ceux-ci sont utilisés dans nos expérimentations afin de comparer leurs résultats avec ceux obtenus en sortie de notre algorithme.

• **K-modes :** Cet algorithme est l'un des nombreux dérivés de l'algorithme classique des *c-moyennes* (ou *k-means*) [Jan66][Mac67]. L'idée initiale de cette famille d'algorithmes est d'utiliser un processus itératif afin de minimiser la somme des carrés (*minimum sum-of-squares*) des distances entre les exemples et les centres de gravité (*centroïds*) des clusters. Le critère à minimiser s'écrit grâce à l'équation suivante :

$$\sum_{C_i \in \mathcal{P}} \sum_{x \in C_i} ||x - c_i||^2 \tag{1.1}$$

où les C_i sont les clusters de la partition $\mathcal{P}$, $||.||$ la norme euclienne et c_i le centroïde (vecteur-moyenne) de C_i.

Pour minimiser ce critère, la majorité des techniques repose sur l'alternance de deux étapes : une étape d'*allocation* et une étape de *recentrage*. Tout d'abord, le centroïde caractérisant chacune des classes est calculé (étape d'allocation). Puis, chaque exemple est réparti au sein des différentes classes en fonction de la distance qui le sépare des différents centres

[3] Au sujet de la méta-heuristique de recherche taboue, voir la section 1.3.2 page 49.

[4] "La meilleure solution" est pris au sens d'un optimum global pour la fonction d'évaluation.

calculés à l'étape précédente (étape de recentrage). Il suffit alors de répéter ces deux étapes jusqu'à assister à la convergence des centres de gravité, et donc des classes.

Précisons qu'il s'agit du cas particulier des *nuées dynamiques*. Cette méthode générale, proposée par E. Diday [Did79], repose, d'une part, sur une vision de l'apprentissage non-supervisé comme un cas de recherche local à travers un espace d'hypothèses, et, d'autre part, sur la notion de *noyaux*. Le noyau est une structure plus complexe (une droite, un groupe de points, un centre de gravité, etc.) qui généralise la notion de centroïde utilisée dans l'algorithme des c-moyennes.

De nombreux algorithmes ont vu le jour basé sur les c-moyennes [How66][HM01][Hua97a]. Plus particulièrement, Z. Huang a proposé l'algorithme des *k-modes* [Hua97b] afin de traiter le cas des données catégorielles. Les deux apports principaux consistent, d'une part, à adapter la mesure de comparaison (en l'occurrence une dissimilarité), et, d'autre part, à considérer les valeurs les plus fréquentes (ou *modes*) à la place des valeurs moyennes généralement employées pour les données numériques. Soient deux objets catégoriels X_1 et X_2 définis sur un espace de M dimensions. La mesure de dissimilarité entre ces deux objets peut être définie par le nombre total de catégories d'attributs qui diffèrent entre les deux objets :

$$d(x_1, x_2) = \sum_{i=1}^{M} \delta(x_{1i}, x_{2i}) \tag{1.2}$$

avec

$$\delta(x_{1i}, x_{2i}) = \begin{cases} 0 & \text{si } x_{1i} = x_{2i} \\ 1 & \text{si } x_{1i} \neq x_{2i} \end{cases}$$

La simplicité de la formule à calculer et sa rapidité d'exécution liée à une stratégie de type *descente du gradient* font que l'algorithme des c-moyennes est le plus largement utilisé pour résoudre les tâches de clustering. Ainsi, sa complexité se calcule généralement en $O(nkl)$ où n représente le nombre d'exemples à classer, k le nombre de clusters et l le nombre d'itérations nécessaire à la convergence. Il faut souligner cependant deux inconvénients majeurs :

1. Il est nécessaire de préciser en entrée de l'algorithme le nombre k de clusters que l'on souhaite obtenir. Or, il est souvent difficile d'évaluer ce nombre *a priori*.
2. Une telle stratégie de recherche permet certes de converger rapidement, mais elle conduit au risque de tomber tout aussi vite dans un optimum local. Cela peut amener à obtenir des solutions réellement sous-optimales.

• **EM :** Cette méthode générale proposée par A. Dempster et al. [DLD77] cherche à déterminer les paramètres d'un mélange (*mixture*) de modèles statistiques à partir d'un ensemble de données. Pour ce faire, l'idée est de parvenir à maximiser un critère de vraisemblance (*maximum likelihood*) en se basant sur l'alternance entre deux étapes fondamentales qui sont à l'origine du nom de cette famille d'algorithmes : l'étape d'*Expectation* et l'étape de *Maximization*. Remarquons que cette stratégie est fortement similaire à celle utilisée dans l'algorithme des c-moyennes (étapes d'allocation et de recentrage).

Supposons que l'ensemble des données $\mathcal{X} = \{x_1, \ldots, x_N\}$ est généré par le mélange de densité Θ grâce à l'équation suivante :

$$p(x_i/\Theta) = \sum_{j=1}^{M} p(x_i/w_j; \theta_j) p(w_j) \tag{1.3}$$

où chaque composant du mélange est noté w_j de paramètres θ_j. On suppose que ces données ont été tirées indépendamment et qu'elles sont identiquement distribuées. Il est possible, à partir de cette équation et de ces hypothèses, de définir le log de vraisemblance (*log likelihood*) liée à ces paramètres :

$$l(\Theta/\mathcal{X}) = \sum_{i=1}^{N} \log \sum_{j=1}^{M} p(x_i/w_j; \theta_j) p(w_j) \tag{1.4}$$

En considérant le critère du maximum de vraisemblance, on estime que le meilleur modèle vis-à-vis des données possède les paramètres maximisant l'équation 1.4.

Bien entendu, toute la difficulté réside dans la manière de déterminer les paramètres du modèle qui a le plus de chance d'avoir généré nos données. Pour cela, l'intuition est de considérer une variable cachée z qui indique quel composant du mélange a généré une donnée x de $\mathcal{X}$. Utilisant cette variable, l'équation précédente se réécrit comme suit :

$$l_c(\Theta/\mathcal{X}, \mathcal{Z}) = \sum_{i=1}^{N} \sum_{j=1}^{M} z_{ij} \log p(x_i/z_i; \Theta) p(z_i; \Theta) \tag{1.5}$$

Puisque z est inconnu, on ne peut utiliser l_c directement. C'est pourquoi on utilise à la place son *Expectation* $Q(\Theta/\Theta_k)$ que l'on cherche à estimer. Comme cela est montré par Dempster [DLD77], $l(\Theta/\mathcal{X})$ peut être maximisé en alternant les deux étapes : E, calculant la valeur estimée de la vraisemblance à partir des paramètres de Θ_k, et M, calculant les paramètres d'un nouveau modèle plus fidèle aux données Θ_{k+1}.

Un modèle statistique très largement utilisé à la base du mélange dont on cherche à estimer les paramètres est la distribution normale (ou gaussienne) [RW84]. Dans ce cas, l'étape E se simplifie au calcul de $h_{ij} \equiv E[z_{ij}/x_i, \Theta_k]$, c'est-à-dire la probabilité que la gaussienne j, telle que définie par les paramètres estimée à l'étape k, puisse générer la donnée x_i. L'étape M quant à elle consiste à ré-estimer la moyenne et la variance pour chaque gaussienne en fonction des données pondérées par les h_{ij}.

Ce modèle étant plutôt adapté aux données numériques, on utilise à la place la densité de Bernouilli lorsque l'on souhaite manipuler des variables discrètes binaires. Plus généralement, les données catégorielles peuvent être générées par un mélange de densités multinominales et l'algorithme d'apprentissage est très similaire à ce que nous venons de voir.

Précisons enfin que les algorithmes de la famille EM sont d'une complexité en temps linéaire par rapport au nombre d'exemples à classer. Cela n'a rien d'étonnant si l'on remarque que l'algorithme des c-moyennes est justement un cas particulier de cette famille. Ils se heurtent d'ailleurs à la même difficulté concernant la découverte du bon nombre de clusters

et de nombreuses techniques, comme le BIC (Bayesian Information Criterion) [KR95] ou la validation croisée, peuvent être utilisées afin de résoudre ce problème.

• **COBWEB :** Il s'agit d'un algorithme de clustering conceptuel, incrémental et hiérarchique proposé par D. Fisher en 1987 [Fis87]. Il s'inspire notamment des systèmes UNIMEM [Leb87] et CYRUS [Kol83], ainsi que des travaux réalisés par Michalski et Stepp dans le domaine du clustering conceptuel [Mic80][MS83]. Il effectue une recherche de type *hill-climbing* à travers un espace contenant des clustering hiérarchiques en utilisant un ensemble d'opérateurs. Ces opérateurs ont été choisis afin d'avoir une plus grande souplesse lors de la recherche. L'objectif est d'obtenir une hiérarchie de classes étiquetées par des concepts probabilistes. Un concept probabiliste [SM81] représente chaque classe d'objets à l'aide d'un ensemble de probabilités conditionnelles sur les valeurs d'attribut.

La recherche est guidée par une mesure heuristique appelée *Category Utility* (CU) proposée par M.A. Gluck et J.E. Corter [GC85] et inspirée par la notion de *cue validity* de E. Rosch et C.B. Mervis [Ros75]. Cette mesure peut être vue comme un compromis (*tradeoff*) entre la similarité intra-classe et la dissimilarité inter-classes, thème classique en apprentissage non supervisé. La similarité est représentée par des probabilités conditionnelles de la forme $p(A_i = V_{ij}/C_k)$, où $A_i = V_{ij}$ est une paire attribut-valeur (ou descripteur) et C_k une classe. Plus cette probabilité est grande, plus les membres de classe partagent cette valeur en commun, et donc plus la valeur a de chances d'être prédite à l'intérieur de la classe. A l'inverse, la dissimilarité est représentée par $p(C_k/A_i = V_{ij})$. Plus cette probabilité est grande, moins des objets appartenant à des classes distinctes partagent cette valeur, et plus la valeur est capable de prédire la classe C_k. La mesure générale évaluant la qualité d'une partition est donnée par la formule suivante :

$$\sum_{k=1}^{n} \sum_{i} \sum_{j} p(A_i = V_{ij}) p(C_k/A_i = V_{ij}) p(A_i = V_{ij}/C_k) \tag{1.6}$$

où la probabilité $p(A_i = V_{ij})$ permet de pondérer les valeurs en fonction de leur fréquence au sein des données. En utilisant la loi de Bayes, l'équation ci-dessus se simplifie comme suit :

$$\sum_{k=1}^{n} p(C_k) \sum_{i} \sum_{j} p(A_i = V_{ij}/C_k)^2 \tag{1.7}$$

En d'autres termes, $\sum_i \sum_j p(A_i = V_{ij}/C_k)^2$ représente le nombre espéré des valeurs qui peuvent être correctement devinées pour n'importe quel élément de C_k. Finalement, la CU est définie comme l'apport du clustering en terme de valeurs devinées par rapport à la partition triviale (c'est-à-dire ne comportant qu'une seule classe), le tout divisé par le nombre de classes dans la partition :

$$\frac{\sum_{k=1}^{n} p(C_k)[\sum_i \sum_j p(A_i = V_{ij}/C_k)^2 - \sum_i \sum_j p(A_i = V_{ij})^2]}{n} \tag{1.8}$$

L'algorithme COBWEB considère chaque objet de manière incrémentale et l'insère dans un arbre de classification, chaque nœud de l'arbre étant caractérisé par un concept probabiliste. Quatre opérateurs sont utilisés pour placer un nouvel objet au sein de la hiérarchie,

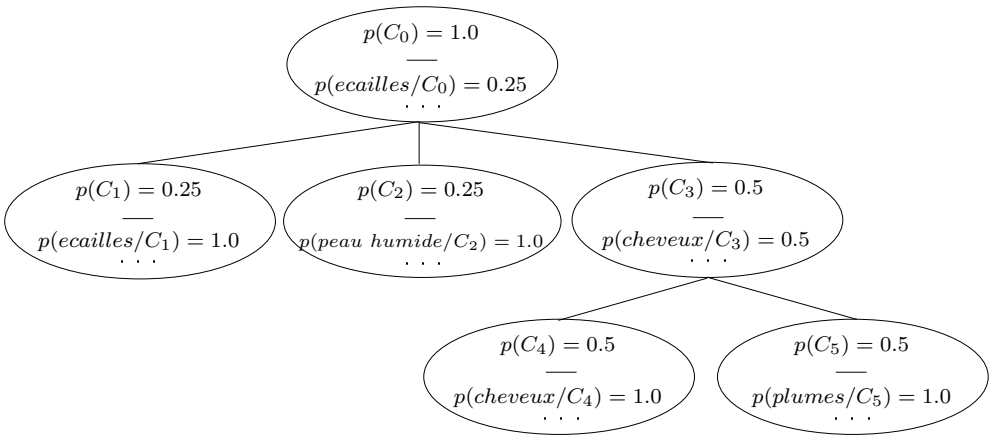


FIG. 1.2 – Exemple de hiérachie obtenue avec COBWEB.

permettant d'ajouter un objet à l'un des nœuds, de fusionner des nœuds ou de les éclater. La mesure de CU est alors utilisée afin de choisir la meilleure hiérarchie à partir de laquelle le processus pourra être itéré à l'aide d'un nouvel objet. La figure 1.2 présente un exemple de hiérarchie obtenue après avoir ajouté les objets 'mammifère' et 'oiseau' à une hiérarchie déjà existante.

Les principaux problèmes rencontrés par COBWEB sont, d'une part, sa grande sensibilité à l'ordre d'entrée des données (due au choix de l'approche incrémentale), et, d'autre part, la stratégie de recherche locale utilisée qui ne donne aucune garantie sur la qualité du résultat en terme d'optimalité. Fisher est bien conscient de ces écueils et propose justement certains opérateurs, comme celui permettant de refusionner des nœuds précédemment éclatés, afin de rendre l'algorithme moins sensible à l'ordre. De plus, il propose dans ses perspectives d'utiliser des algorithmes de recherche plus développés afin de s'extraire des optima locaux. L'algorithme COBWEB a donné lieu à plusieurs améliorations, comme les systèmes CLASSIT [GLF92] et ARACHNE [IL01].

Pour conclure, signalons qu'il est possible d'obtenir une partition des objets à partir de la hiérarchie proposée en considérant les classes situées aux nœuds de l'arbre. L'utilisation d'un paramètre (*cutoff*) permet de fixer à partir de quel moment le gain calculé par l'équation 1.8 devient négligeable et donc de restreindre le phénomène d'éclatement des nœuds de l'arbre. Il est ainsi possible de contrôler, d'une certaine manière, le nombre de classes produites par l'algorithme.

1.1.2 Traitement des données lacunaires

La plupart des méthodes d'analyses de données s'appliquent à des bases d'observations où tous les champs sont parfaitement renseignés. Or, rares sont les données tirées de la vie réelle où ces conditions sont bien remplies. Ainsi, une personne interrogée lors d'une enquête peut refuser de révéler les revenus de son foyer. Des données industrielles rendues incomplètes

à cause d'un matériel défectueux sont un autre exemple de données lacunaires[5]. Précisons qu'il ne faut pas confondre, si l'on reprend l'exemple du sondage, les réponses incertaines ou imprécises avec l'absence de réponse. Les premières sont associées à des descripteurs particuliers ou à une manière particulière de représenter l'information (en utilisant, par exemple, des probabilités ou la logique possibiliste), alors que les secondes impliquent uniquement des champs vides qui peuvent faire l'objet de prédiction.

Aussi, quelles sont les stratégies adoptées généralement face à un tel type de données ? La version la plus naïve consiste à ne traiter que les cas complets, soit en associant des codes spécifiques aux données manquantes, soit en ignorant tout simplement la moindre observation incomplète. On comprend bien que cette stratégie n'est pas adaptée dans la plupart des cas, spécialement si le taux de valeurs manquantes est important. Mais avant de faire un survol des techniques plus fines proposées dans le domaine de l'analyse de données, nous allons regarder quelles sont les différentes familles de données manquantes auxquelles on peut être confronté.

1.1.2.1 Topologie des données lacunaires

Nous nous intéressons ici aux différents mécanismes qui provoquent l'absence de certaines valeurs pour les variables décrivant l'ensemble de données à analyser. Le rôle de ces mécanismes en analyse de données est crucial et a pourtant été ignoré jusqu'aux études menées par Rubin [Rub76] qui a considéré les données manquantes comme des variables aléatoires à part entière. Supposant qu'il existe un mécanisme à l'origine de cette lacunarité, il distingue trois grandes familles :

- **Missing Completely At Random (MCAR) :** Le mécanisme générant les données manquantes ne dépend pas du tout des valeurs (connues ou non) desdites données. Un appareil défectueux empêchant de mesurer l'une des variables une fois sur deux est un exemple de données MCAR, tout comme le responsable d'un sondage égarant (par hasard !) une partie des réponses données par certaines personnes interrogées. Il faut noter que cette hypothèse ne signifie pas pour autant que les "trous" sont générés de manière aléatoire, mais juste qu'ils sont indépendants des descripteurs présents ou des autres valeurs manquantes. L'absence des données peut également être due à des raisons extérieures au champ des variables considérées : ancienneté de la machine, humeur de la personne interrogée, etc.

- **Missing At Random (MAR) :** Le caractère lacunaire des données dépend uniquement des valeurs observées associées aux données. Par contre, il ne dépend pas des autres valeurs manquantes. Ainsi, dans une enquête sur les conditions de vie des Français, il y a plus de probabilités que l'information concernant le poids d'une personne interrogée manque lorsque celle-ci est une femme que s'il s'agit d'un homme. Un autre exemple de MAR est celui de la compréhension écrite : un test de compréhension générale réalisé au début de l'épreuve influera très probablement sur le taux de questions restées sans réponse (car jugées trop

[5] Ces deux exemples sont tirés de [LR02].

difficiles) dans la suite de l'épreuve.

• **Non Missing At Random (NMAR) :** Le choix des valeurs manquantes dépend ici à la fois des valeurs observées au sein des données considérées, mais également des autres valeurs manquantes. C'est le cas le plus général et le plus difficile à appréhender. Le hasard n'intervient plus réellement et les valeurs manquent en suivant une certaine logique qui peut s'avérer complexe et donc difficile à modéliser. Les expérimentations que nous réalisons dans le cas des jeux de données réels se placent précisément dans ce type de cas.

Les mécanismes de type MAR ou NMAR reflètent une logique de répartition des données manquantes qui dépend, pour le premier, des données qui ont été observées, et pour le second, des données observées *et* des données manquantes. C'est bien évidemment dans ce cadre que sont situées les données réelles que nous manipulons. Il n'est en effet pas très réaliste de croire que les données manquant au sein des articles de journaux que nous avons utilisés sont le "fruit du hasard" (en tout cas, au regard des autres données), ce qui correspond à l'erreur intrinsèque décrite dans la section 1.1.1.3. Par contre, les jeux de données artificiels, comme dans la plus grande majorité de (si ce n'est toute) la littérature, se placent dans le cadre des mécanismes MCAR tel que le définit Rubin.

1.1.2.2 Traitement des données lacunaires en analyse de données

Nous donnons un bref aperçu des principales méthodes employées en analyse de données afin de traiter les données lacunaires. Il existe trois grandes familles de méthodes :

• **Les procédures basées sur des données complètes (*listwise deletion*) :** Il s'agit ici d'ignorer les observations comportant la moindre valeur manquante. Si cela peut convenir dans les cas où le taux de données lacunaires est très restreint, on imagine sans peine le biais (de variance) qui en résulte lorsque celui-ci devient important. Dans l'hypothèse où toutes les données comportent des valeurs manquantes (ce qui nous intéresse dans cette thèse), cette solution apparaît parfaitement inadaptée. Une autre alternative, appelée *pairwise deletion*, consiste à ignorer, non pas les exemples comportant des valeurs manquantes, mais les variables dont certaines valeurs manquent au sein des observations à chaque nouveau calcul où elles sont impliquées. Il s'agit de la stratégie adoptée par l'algorithmes des c-moyennes.

• **Les procédures d'imputation :** Avec cette famille de techniques, les valeurs manquantes sont automatiquement complétées dans un premier temps avant d'utiliser les techniques classiques de l'analyse de données (cf. figure 1.3).

La valeurs inconnues sont estimées à l'aide des valeurs prises par les autres variables. Il s'agit de la procédure de *hot-deck*[6], de l'imputation par la moyenne, de l'utilisation de méthodes basées sur la régression ou des techniques plus évoluées de *multiple imputation*. L'imputation peut atteindre des degrés de complexité élevés avec l'estimation des paramètres caractérisant des modèles *a priori* et révisés régulièrement. Parmi toutes ces techniques se

[6]Cette technique correspond à l'approche par les k-plus-proches voisins dans le cas où k=1.

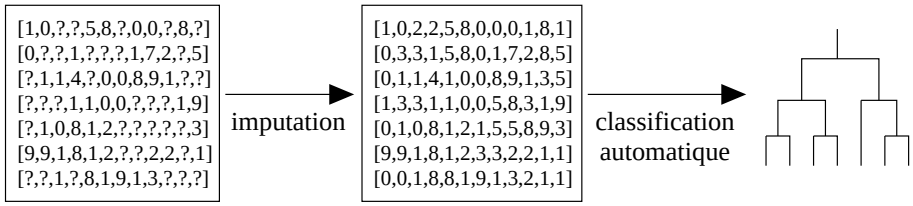


FIG. 1.3 – Application d'une technique d'imputation avant l'analyse.

singularise la famille des algorithmes EM dont nous avons déjà parlé qui effectue l'estimation en maximisant la vraisemblance des données en fonction d'un modèle probabiliste dont les paramètres restent à définir. Celui-ci a d'ailleurs fait l'objet d'études concernant le traitement des données manquantes [LR02][GJ94] auquel il semble particulièrement adapté. C'est pourquoi une implémentation de EM est utilisée dans nos expérimentations par souci de comparaison.

- **Les procédures utilisant des connaissances implicites :** Cette famille regroupe les techniques consistant à utiliser une connaissance extérieure à l'ensemble des données pour compléter automatiquement les valeurs manquantes. La technique la plus simple consiste ainsi à remplir les "trous" avec une valeur par défaut parfaitement arbitraire ou qui dépend du contexte (le 'sans opinion' dans un sondage). Une autre manière de procéder est d'utiliser une logique par défaut qui utilise des règles afin de combler les valeurs manquantes. La prochaine section présente justement l'une de ces logiques introduite dans les années 80.

Nous faisons remarquer que cette classification des procédures utilisées pour gérer les données lacunaires n'est pas universelle et que nous n'avons pas trouvé de parfait concensus à son sujet. Nous conseillerons le lecteur intéressé de se référer à l'ouvrage de Little et Rubin [LR02] pour un aperçu général et à l'article de A. Farhangfar et al. [FKP04] pour la prise en compte de l'apprentissage supervisé et la comparaison effective de différentes méthodes d'imputation.

1.1.2.3 Données lacunaires et raisonnement par défaut

Il semble utopique de croire que les raisonnements que nous formons dans la vie de tous les jours sont parfaitement fondés et irréfutables. Au contraire, ils reposent dans la plupart des situations sur des données trop peu représentatives du phénomène étudié, possédant des descriptions incomplètes et associées à des théories fausses sur le domaine d'apprentissage.

Nous pensons que ce type de raisonnements peut être étudié dans un cadre similaire à celui de la découverte scientifique et aux travaux effectués sur le sens commun [ML79][Jod89]. Comme nous nous trouvons dans un cas où l'information est imparfaite car très lacunaire, nous nous sommes intéressés aux logiques non monotones qui nous y semblaient adaptées. De nombreux travaux ont été effectués dans le cadre de ces logiques, tels ceux de D. McDermott et J. Doyle [MD80] ou ceux de J. McCarthy avec la logique de circonscriptions [McC80]. Nous

avons finalement choisi de nous inspirer de la logique des défauts qui envisage un formalisme de règles permettant de compléter l'information en l'absence d'informations contradictoires.

Ce formalisme a été développé dans les années 80 par R. Reiter [Rei80]. Il s'agit d'une logique pour le raisonnement par défaut basée sur la notion de règles de défaut qui permettent d'inférer de nouvelles formules lorsque les hypothèses ne sont pas incohérentes avec l'état du monde connu. Une règle de défaut possède la forme suivante :

$$\frac{A : B_1, B_2, ...B_n}{C}$$

A est appelé le pré-requis, B_i les justifications et C la conclusion. Cette règle de défaut peut être interprétée de la façon suivante : si A est connu comme étant vrai et qu'il est cohérent de croire que $B_1, B_2, ...B_n$ le sont aussi, alors il est légitime de conclure que C est vraie. Voici un exemple de règle qui pourrait avoir cours dans la France de la fin du XIX^{e} siècle et construite dans le formalisme présenté en annexe page 175 :

$$\frac{\begin{array}{c}(Patriotisme = antipatriote), (MeleAffaireDArgent = oui) : \\ \neg(RelationEtranger = non), \neg(SensMoral = oui)\end{array}}{(Traitre = oui)}$$

Cette règle se traduit de la manière suivante : tout homme politique antipatriote compromis dans des scandales d'argent peut être considéré comme un traître en puissance au service d'un pays étranger, à condition bien sûr qu'il ne soit pas prouvé qu'il n'entretient aucune relation à l'étranger ou qu'il possède un sens moral[7]. Comme nous le verrons, la construction des stéréotypes se fonde également sur cette idée de complétion en l'absence d'information contradictoire, mais en suivant cette fois une approche inductive et non déductive.

1.1.3 Manipulation des données symboliques et mesures de comparaison

Nous donnons dans cette partie un aperçu des différents types de données généralement manipulées en analyse de données et des mesures pouvant être utilisées pour comparer de tes objets deux à deux.

1.1.3.1 Données numériques et données symboliques

Les variables sont usuellement classées en trois grandes familles [Spä80][CM02] :

• **nominal :** Les valeurs associées à une *variable nominale* sont en nombre fini, de telle sorte qu'il est possible d'associer à chaque valeur un code numérique (comme les n premiers entiers). Par exemple, une variable représentant le statut marital d'une personne interrogée accepte les entiers de 1 à 4 pour les valeurs 'célibataire', 'marié', 'divorcé' et 'veuf'. Les notions d'ordre entre les valeurs ou de distance n'ont ici aucun sens. Si la variable ne peut prendre que deux valeurs distinctes (correspondant aux codes 0 et 1), celle-ci est aussi appelée *variable binaire*.

[7] La notion de "sens moral" possédait un sens bien spécifique à cette époque.

• **ordinal :** Les valeurs associées à une *variable ordinale* peuvent être arrangées dans un ordre qui possède un sens. Il peut s'agir, par exemple, de réponses du type 'très mauvaises', 'mauvaises', 'moyennes', 'bonnes', 'très bonnes' pour la variable indiquant les conditions de travail d'un employé utilisée dans un sondage. Par contre, la différence entre deux valeurs successives n'est pas nécessairement toujours la même, voire difficilement quantifiable ou dénuée de sens. C'est pourquoi la notion de distance, au sens géométrique, est généralement absente ici aussi.

• **métrique :** Les valeurs associées à une *variable métrique* se trouvent sur une échelle à intervalles réguliers et possèdent, la plupart du temps, une base absolue (le 0 des entiers par exemple). Des variables comme la température, la taille, le poids sont de bons exemples de variables de type métrique. Les opérations arithmétiques, et particulièrement les mesures de distance comme la distance Euclidienne, peuvent être utilisées avec ce type de variables.

En apprentissage, les variables des deux premières familles (nominale et ordinale) sont regroupées sous le nom de *variables symboliques* (on parle aussi de *variables qualitatives* ou *catégorielles*) alors que celles de la troisième famille sont plus volontiers appelées *variables numériques* ou *quantitatives*. Il arrive souvent que des données de ces différents types se retrouvent mélangées dans un tableau de données, ce qui a donné naissance à une volonté de traiter ces deux problèmes à la fois [DK91][Hua97a][WBF98][CFC+01]. D'un autre côté, des recherches actuelles tâchent de dépasser ces types usuels et de développer la manipulation de structures plus complexes comme des distributions de probabilité ou des intervalles [BD00][Did00]. Précisons enfin que l'on peut s'intéresser à des structures plus riches que des ensembles (ordonné ou non) de valeurs, comme des hiérarchies [Sow84] ou des treillis [CMB03].

Dans cette thèse, nous nous intéressons plus particulièrement à des données de type symbolique, qu'elles soient nominales ou ordinales. Cela aura des incidences sur le choix des mesures de comparaison utilisées, du critère d'optimisation et d'autres points détaillés dans la suite de ce document.

1.1.3.2 Mesures de ressemblance et similarité

La notion de ressemblance, utilisée pour comparer les objets deux à deux, est centrale dans la cognition humaine, et plus particulièrement dans les problèmes de formation des catégories. Les mesures utilisées pour calculer la similitude entre deux objets sont pléthore et il apparaît qu'il n'existe pas de règle générale pour choisir (dans l'absolu) telle mesure plutôt que telle autre. Au contraire, la mesure doit être choisie en fonction de l'application que l'on souhaite en faire [Rif96]. Ainsi, dans notre travail nous nous intéressons aux données de type symbolique et plus précisément aux données comportant de nombreuses valeurs manquantes. Par conséquent, nous avons besoin de considérer une mesure adaptée à ce type de données [HK01]. Nous rappelons à présent quelques définitions générales qui serviront de base aux mesures proposées dans la partie 3.1. Nous nous servons du travail de synthèse de Batagelj

et Bren [BB95], lui-même basé sur les travaux de Sneath et Sokal [SS73], Anderberg [And73], Späth [Spä80].

Définition 1.1.1 *Soit un ensemble d'objets $\mathcal{E}$, une* mesure de ressemblance *sur $\mathcal{E}$ est une application r de $\mathcal{E} \times \mathcal{E}$ dans $\mathbb{R}^+$ telle que r vérifie à la fois la propriété P1 et soit la propriété P2.a, soit la propriété P2.b :*

P1. - r est symétrique : $\forall (x, y) \in \mathcal{E} \times \mathcal{E}$, $r(x, y) = r(y, x)$.

*P2.a - r est dite "vers l'avant" (*forward*) : $\forall (x, y) \in \mathcal{E} \times \mathcal{E}$, $r(x, x) \leq r(x, y)$.*

*P2.b - r est dite "vers l'arrière" (*backward*) : $\forall (x, y) \in \mathcal{E} \times \mathcal{E}$, $r(x, x) \geq r(x, y)$.*

Une mesure de ressemblance satisfaisant la propriété P2.a est notée d, alors qu'une mesure satisfaisant P2.b est notée s. Pour plus de clarté, nous appelons s une *mesure de similitude*. Les mesures de ressemblance r et s peuvent être raffinées selon deux nouveaux types de mesure si l'on considère la nouvelle contrainte P3 :

P3. $\exists r^* \in \mathbb{R} / \forall x \in \mathcal{E}$, $r(x, x) = r^*$.

Définition 1.1.2 *Une* mesure de similarité *est une application de $\mathcal{E} \times \mathcal{E}$ dans $\mathbb{R}^+$ satisfaisant les trois propriétés $P1$, $P2.b$ et $P3$.*

La mesure de similarité est également appelée *indice de ressemblance* [Did79]. En posant d tel que $d(x, y) = r(x, y) - r^*$, nous définissons la mesure symétrique suivante :

Définition 1.1.3 *Une* mesure de dissimilarité *est une application de $\mathcal{E} \times \mathcal{E}$ dans $\mathbb{R}^+$ satisfaisant les trois propriétés suivantes :*

R1. $\forall (x, y) \in \mathcal{E} \times \mathcal{E}$, $d(x, y) \geq 0$.

R2. $\forall x \in \mathcal{E}$, $d(x, x) = 0$.

R3. $\forall (x, y) \in \mathcal{E} \times \mathcal{E}$, $d(x, y) = d(y, x)$.

La mesure de dissimilarité est également appelée *indice de dissemblance* [Did79]. Comme nous pouvons le remarquer, nous avons défini la similarité et la dissimilarité comme deux mesures symétriques. Ceci est la raison pour laquelle seules des mesures de ressemblances seront utilisées dans notre travail, nous évitant de multiplier des expérimentations inutiles.

1.1.3.3 Mesures adaptées aux données catégorielles

Passées ces quelques définitions très générales, nous donnons quelques mesures de ressemblances classiques utilisées pour traiter des données de type symbolique. Les variables considérées sont binaires, c'est-à-dire qu'elles adoptent uniquement les valeurs 0 ou 1. Elles expriment généralement le fait qu'un attribut est présent (1) ou absent (0) dans la définition d'un objet. On considère souvent que, pour traiter le cas multi-modal, il suffit de procéder à une "dichotomisation" des variables qui permet de transformer une variable nominale à m modalités en $m - 1$ variables binaires (ou *dummies*) de présence/absence.

Bien que ces mesures puissent être dérivées de la théorie de la mesure, nous avons choisi une manière plus intuitive de les présenter. Considérant deux vecteurs X et Y dans un espace à p variables, $X = (x_1, \ldots, x_p)$ et $Y = (y_1, \ldots, y_p)$, nous posons :

- a = le nombre de variables prenant la valeur 1 à la fois dans X et Y.
- b = le nombre de variables prenant la valeur 0 dans X et 1 dans Y.
- c = le nombre de variables prenant la valeur 1 dans X et 0 dans Y.
- d = le nombre de variables prenant la valeur 0 à la fois dans X et Y.

On note que $a+b+c+d = p$. Nous donnons ci-dessous trois mesures classiques exprimées dans ce formalisme qui permettent de calculer une similarité entre deux objets :

1. $s_1(X,Y) = \frac{a}{a+b+c+d}$ (Russel et Rao)
2. $s_2(X,Y) = \frac{a+d}{a+b+c+d}$ (Kendall ; Sokal et Michener)
3. $s_3(X,Y) = \frac{a}{a+b+c}$ (Jaccard)

Notons la propriété suivante :

Propriété 1.1.1 $\forall X, Y, s_1(X,Y) \leq s_3(X,Y) \leq s_2(X,Y)$

Ces trois mesures de forment la base de trois des quatre mesures de comparaison que nous utilisons pour construire les stéréotypes dans la partie 3.1. La différence entre les mesures s_1 et s_3 réside dans le coefficient de normalisation. Ainsi, la mesure de Jaccard n'utilise pas d, ce qui implique que cette mesure respecte la propriété P3 et qu'il s'agit donc d'une mesure de similarité, contrairement à la mesure de Russel et Rao qui est une mesure de similitude. Comparativement, s_2 utilise d à la fois au numérateur et au dénominateur, ce qui en fait également une mesure de similarité. C'est ce dernier type de mesure qu'utilise un algorithme comme les k-modes.

Les mesures proposées jusqu'à présent respectent la propriété de symétrie qui en font donc des mesures de similarités. La nécessité d'une telle propriété est mise en cause par A. Tversky [Tve77] qui rejette l'hypothèse de représentation métrique des objets. Ainsi, il affirme qu'un sujet A peut être davantage similaire à un référent B que B n'est similaire à A car l'établissement d'une similarité est directionnelle. On dira plus facilement, par exemple, que "les turcs combattent comme des tigres" que "les tigres combattent comme des turcs". Les tigres, connus pour leur esprit de combat, sont utilisés comme référence plus que comme sujet.

Ces dernières réflexions donneront lieu à la prise en compte d'une quatrième mesure de comparaison non symétrique prenant le stéréotype comme référent. Le lecteur intéressé par les questions spécifiques concernant les mesures de comparaison pourra se reporter à la thèse de M. Rifqi [Rif96].

1.1.4 Evaluation des catégories issues du clustering

La classification automatique est un champ de recherche qui concerne de très nombreux domaines. Ceci explique le nombre sans cesse grandissant d'algorithmes existants pour résoudre ce problème, algorithmes adaptés aux différents contextes et hypothèses considérés.

Il est important de parvenir à évaluer le résultat des systèmes de clustering, c'est-à-dire à la fois la qualité des classes effectivement construites, mais également des représentants de ces classes. Un tel retour à partir du résultat obtenu permet de déterminer les mesures de distance qu'il convient d'utiliser, de paramétrer correctement les algorithmes employés, voire de choisir quels algorithmes sont les plus adaptés à la tâche considérée. Il est souvent difficile de juger objectivement de la qualité d'un clustering car cela dépend grandement de l'application que l'on souhaite en faire. La procédure d'évaluation du résultat d'un algorithme de clustering est connue sous le nom de *cluster validity* [HBV02a].

La validation manuelle, c'est-à-dire réalisée par des humains élevés au rang d'experts du domaine, est la méthode la plus intuitive car elle permet de juger directement du résultat obtenu en fonction de l'application qui a motivé l'étude. Elle a été largement utilisée pour juger de la sortie de clustering sur des données à deux, voire trois, dimensions. Mais ce type de validation n'est pas forcément toujours faisable ou souhaitable, et elle devient très difficile avec des données décrites dans de grandes dimensions.

C'est pourquoi il convient de prendre en considération des méthodes automatiques quantitatives afin de donner une idée sur la qualité des clustering obtenus. Nous allons présenter ici quelques-unes des techniques couramment employées et qui permettent de comparer les résultats obtenus avec différents algorithmes de clustering. Nous nous basons sur la distinction entre *class conformation* et *cluster distribution* effectuée par He et al. [HTTS02]. Cette distinction se retrouve dans [HBV02a] et peut être mise en relation avec les deux approches nommées respectivement *external criteria* et *relative criteria*. Les deux différences fondamentales sont que l'approche à "critère externe" repose sur des fondements statistiques (elle utilise notamment la technique de test Monte-Carlo) et qu'elle possède un coût de calcul beaucoup plus élevé car les exemples sont comparés deux à deux.

1.1.4.1 Conformation de classe

La première manière employée pour évaluer le résultat des algorithmes de clustering est celle dite de *class conformation*, termes que nous traduisons par "conformation de classe". Elle correspond au "critère externe" d'Halkidi et al. [HBV02a]. Celle-ci suppose l'existence d'une distribution idéale des données avec laquelle il est possible de comparer le résultat donné par l'algorithme de clustering. Grâce à cette distribution, une classe peut être assignée à chacun des objets de l'ensemble à classer. L'objectif du clustering est alors de retrouver automatiquement le même étiquetage des objets. Des objets ayant la même classe seront placés dans le même cluster par le processus de classification, alors que des objets de différentes classes apparaîtront dans des clusters différents.

- **Données réelles et données artificielles :**

La conformation de classe peut être utilisée avec des jeux d'essais réels et des jeux d'essais artificiellement créés.

La première technique consiste à considérer une base réelle déjà étiquetée, en général par des experts humains, puis à tâcher de retrouver automatiquement ces classes. Il suffit de

comparer les classes découvertes (clusters) aux classes réelles (qui ont été cachées durant le processus de clustering). Pour réaliser cette comparaison, il existe plusieurs mesures dont les principales sont données dans le prochain paragraphe. Nous insistons sur le fait que cette technique considère l'étiquetage qui est donné comme étant la catégorisation idéale. Cette catégorisation idéale sert de modèle pour affirmer qu'un clustering est plus ou moins juste.

La seconde technique utilise une base d'exemples créée artificiellement. Son utilisation est très largement répandue dans les recherches concernant le clustering. Ainsi, il est très courant d'effectuer des premiers tests à partir de tels jeux de données, plus facilement contrôlables, avant de s'attaquer à des problèmes réels. La première partie des expérimentations réalisées dans le chapitre 5 manipulent ce type de données. Nous discutons de la génération des jeux d'essai artificiels dans la section 1.1.4.3.

- **Validation du clustering :**

La mesure la plus largement répandue est aussi la plus simple. Elle consiste à calculer l'erreur de classification, c'est-à-dire le pourcentage des instances de la base qui se retrouvent "mal étiquetées" après le processus de classification. Cela suppose d'associer chaque cluster qui vient d'être construit avec l'une des classes de départ. Il est ainsi possible de prédire le cluster dans lequel devrait se retrouver chacune des instances de l'ensemble à catégoriser. Chaque instance qui se retrouve dans un cluster différent est jugée mal classée et compte pour l'erreur globale de classification. Remarquons que plusieurs clusters peuvent être associés à la même classe d'origine. Cela ne pose pas de problème particulier tant que la classe prédite pour chaque exemple est la bonne. Cette mesure permettant de juger de la qualité d'un clustering peut se décomposer en deux mesures complémentaires plus fines.

Boley [Bol98] introduit deux mesures dont l'approche est basée sur le principe d'entropie de l'information. La première mesure est l'entropie de cluster et reflète la qualité des clusters de manière individuelle, c'est-à-dire en terme d'homogénéité des instances à l'intérieur de chacun des clusters. L'entropie du cluster c_i (*cluster entropy*) est calculée de la manière suivante :

$$E_{c_i} = -\sum_j \frac{n(l_j, c_i)}{n(c_i)} \log \frac{n(l_j, c_i)}{n(c_i)} \tag{1.9}$$

où $n(l_j, c_i)$ est le nombre d'instances de c_i étiquetées par le label l_j et $n(c_i) = \sum_j n(l_j, c_i)$ est le nombre total d'instances dans c_i. L'entropie de cluster globale E_c (*overall cluster entropy*) est donnée par la somme pondérée des entropies individuelles de chaque cluster :

$$E_c = \frac{1}{\sum_i n(c_i)} \sum_i n(c_i) E_{c_i} \tag{1.10}$$

où $n(c_i)$ correspond au nombre d'instances appartenant au cluster c_i et E_{c_i} est l'entropie (individuelle) du cluster c_i.

L'entropie de cluster ne prend pas en compte le nombre de clusters produits et aura même tendance à baisser si ce nombre augmente. C'est pourquoi nous avons besoin d'une mesure complémentaire d'entropie appelée entropie de classe. Celle-ci mesure la manière dont une

même classe est représentée au sein des différents clusters. L'entropie de la classe l_j (*class entropy*) est calculée de la manière suivante :

$$E_{l_j} = -\sum_i \frac{n(l_j, c_i)}{n(l_j)} \log \frac{n(l_j, c_i)}{n(l_j)} \tag{1.11}$$

où $n(l_j, c_i)$ est le nombre d'instances de c_i étiquetées par le label l_j et $n(l_j) = \sum_i n(l_j, c_i)$ est le nombre total d'instances étiquetées par l_j. L'entropie de classe globale E_l (*overall class entropy*) est donnée par la somme pondérée des entropies individuelles de classe :

$$E_l = \frac{1}{\sum_j n(l_j)} \sum_j n(l_j) E_{l_j} \tag{1.12}$$

où $n(l_j)$ correspond au nombre d'instances étiquetées par la classe l_j et E_{l_j} est l'entropie (individuelle) de la classe l_j.

On remarque que l'entropie de cluster diminue alors que l'entropie de classe augmente, et réciproquement. He propose de considérer une combinaison linéaire de ces deux mesures afin d'obtenir une mesure générale d'entropie prenant en compte ces deux phénomènes complémentaires. Ainsi, l'entropie globale (*overall entropy*) est définie par la formule suivante :

$$E_{cl}(\beta) = \beta.E_c + (1 - \beta).E_l \tag{1.13}$$

où $beta \in [0, 1]$ est la valeur de compromis (*trade-off*) entre l'entropie de cluster globale E_c et l'entropie de classe globale E_l.

Il est important de souligner que ces mesures liées à la conformation de classe sont indépendantes du choix de l'espace des représentations et des mesures de distance employées pour effectuer le clustering. En ce sens, elles sont capables d'évaluer à peu près n'importe quel système générant un clustering à partir du moment où les données d'entrées peuvent être fiablement étiquetées. C'est justement là où la plus grande difficulté réside : le processus d'étiquetage peut parfois être très complexe, coûteux, voire impossible à réaliser. C'est pourquoi nous allons aborder dans la section suivante les techniques d'évaluation ne nécessitant aucun étiquetage préalable.

1.1.4.2 Distribution des clusters

La deuxième manière envisagée pour évaluer le résultat d'un algorithme de clustering est celle dite de *cluster distribution* que nous traduisons par "distribution des clusters". Elle correspond au "critère relatif" d'Halkidi [HBV02a] en ne considérant que des structures "simples" (partitions). Contrairement à l'approche précédente, aucun étiquetage préalable des données ne permet de comparer le résultat du clustering à un quelconque modèle idéal. De nombreux indices de validité ont été proposés et des travaux récents attestent de la vitalité de cette perspective de recherche [HTTS02][HBV02a][HBV02b][CSL03]. Ils se basent sur la recherche d'un compromis entre les principes de similarité (ou homogénéité) intra-classe et de dissimilarité (ou séparation) inter-classes. Remarquons que cette approche peut être utilisée en complément de celle de conformation de classe lorsqu'un étiquetage est possible.

Par analogie aux différentes familles d'algorithmes de clustering, deux approches générales se dégagent afin de définir un indice de validité.

La première est centrée sur la notion d'*adéquation* entre le recouvrement des objets obtenu grâce au clustering et la représentation des classes de ce recouvrement. E. Diday [Did79] propose ainsi de définir le critère à optimiser comme une application à valeurs positives qui sert à mesurer cette adéquation. Autrement dit, on cherche à maximiser l'adéquation entre les exemples de l'ensemble d'apprentissage et les représentants (moyenne, centroïde, prototype, stéréotype) des clusters qui les couvrent. C'est principalement cette approche que nous développons avec la mesure de qualité générale de [HTTS02] car elle est naturellement adaptée à la thématique du clustering conceptuel.

Cette première approche s'oppose à celle plus exhaustive comparant les exemples deux à deux, c'est-à-dire prenant en considération la structure du clustering mais ignorant les concepts étiquetant les différents clusters. Un indice caractéristique de cette approche est celui de Dunn et de ses suivants [Dun74] que nous allons détailler par souci de comparaison. Nous ne prétendons pas bien entendu être exhaustifs car les indices existants sont nombreux et conseillons au lecteur de se référer à l'aperçu général donné par Halkidi [HBV02b].

• **Indice de Dunn :** Cet index, proposé dans [Dun74], tâche d'identifier des clusters compacts et bien séparés. Soit une fonction d mesurant la dissimilarité entre les deux clusters c_i et c_j telle que $d(c_i, c_j) = \min_{x \in c_i, y \in c_j} d(x, y)$ et une fonction $diam$ calculant le diamètre d'un cluster, l'indice de Dunn est défini grâce à l'équation suivante :

$$D_{n_c} = \min_{i=1,\ldots,n_c} \left\{ \min_{j=i+1,\ldots,n_c} \left\{ \frac{d(c_i, c_j)}{\max_{k=1,\ldots,n_c} diam(c_k)} \right\} \right\} \tag{1.14}$$

Le diamètre d'un cluster c, reflétant la dispersion au sein de ce cluster, peut être défini de la manière suivante :

$$diam(c) = \max_{x,y \in c} d(x, y) \tag{1.15}$$

Si les données sont organisées en clusters compacts et bien séparés, la distance entre les clusters sera grande et les clusters seront de faibles diamètres, ce qui se traduira par une valeur de D_{n_c} d'autant plus grande. Le nombre de clusters ne rentrant pas en compte dans l'équation 1.14, une valeur maximum pour cet indice peut donner une indication sur le nombre adéquat de clusters correspondant aux données.

Les deux problèmes principaux relevés par Halkidi sont : (i) la complexité importante en temps de calcul (tous les exemples sont comparés deux à deux) ; (ii) la sensibilité à la présence de bruit (notamment les *outliers*) dans les données, phénomène lié au calcul du diamètre des clusters. D'autres indices basés sur le même principe [PB97] sont davantage résistants au bruit et sont connus comme les indices *Dunn-like*.

A ces deux inconvénients, nous ajoutons le fait que seule l'organisation des exemples au sein des clusters est prise en compte en ignorant la représentation de ces clusters. Or, notre travail cherche précisément à découvrir en priorité une bonne représentation des données. C'est pourquoi nous présentons dans la suite des mesures dans lesquelles les exemples ne sont

plus comparés deux à deux mais directement aux représentants des clusters dans lesquels ils ont été placés.

• **Mesures de compacité-séparation :** He et al. [HTTS02] proposent une mesure basée sur la généralisation des travaux de Halkidi et al. [HVB00]. Ces mêmes travaux étaient déjà une adaptation aux partitions "rigides" des indices définis dans [RLR98] pour l'algorithme des c-moyennes floues. Cette mesure repose sur les notions duales de compacité (*compactness*) et de séparation (*separation*). Celles-ci traduisent quantitativement les notions centrales du clustering que sont, pour la première, l'homogénéité interne des exemples à l'intérieur d'une même classe, et pour la seconde, la distinction suffisante entre des exemples de classes différentes. La notion de compacité repose sur une définition généralisée de la variance donnée dans [HTTS02] et que nous rappelons ici :

$$v(X) = \sqrt{\frac{1}{|X|} \sum_{x_i \in X} d^2(x_i, \overline{x})} \tag{1.16}$$

où $d(x_i, x_j)$ est une distance métrique entre les deux vecteurs x_i et x_j, et $\overline{x} = \frac{1}{|X|} \sum_{x_i \in X} x_i$ est le vecteur-moyenne (ou centroïde) de X.

Il est clair qu'une valeur de variance plus petite pour un ensemble X est le signe d'une plus grande homogénéité entre les vecteurs de X, au sens bien entendu de la distance d. Comme nous le verrons lorsqu'il s'agira d'évaluer le résultat de notre système, et à la lumière des réflexions déjà réalisées sur les mesures de comparaison, il existe ainsi plusieurs manières de calculer la variance suivant la distance qui a été choisie. Remarquons également que si l'espace X est décrit dans un espace à une seule dimension et si d est la distance euclidienne, $v(X)$ est la variance statistique usuelle $\sigma(X)$ de l'ensemble X.

A partir de cette définition de la variance, il est possible de définir la mesure de compacité (*cluster compactness*) d'un ensemble $\mathcal{C}$ de clusters recouvrant un ensemble $\mathcal{X}$[8] :

$$Cmp(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{c_i \in \mathcal{C}} \frac{v(c_i)}{v(\mathcal{X})} \tag{1.17}$$

Cette mesure évalue l'apport du clustering, en terme de diminution de la variance, en comparaison du clustering trivial formé par l'unique cluster X. Lorsque d est la distance euclidienne, on retrouve l'index d'*average cluster scattering* entrant dans le calcul de l'indice SD [HVB00]. Une valeur inférieure de compacité est le signe d'une plus grande homogénéité moyenne à l'intérieur des clusters.

On s'aperçoit tout de suite qu'une compacité optimale est obtenue dans le cas extrême où chaque instance constitue un cluster à la sortie du système. Cela signifie que la mesure de compacité ne constitue pas à elle seule une manière correcte d'évaluer la qualité d'un clustering, mais qu'elle doit être contrebalancée par la mesure duale de séparation. La séparation

[8] La notation a été légèrement changée mais la formule reste la même.

(*cluster separation*) d'un ensemble $\mathcal{C}$ de clusters est définie de la manière suivante[9] :

$$Sep(\mathcal{C}) = \frac{1}{|\mathcal{C}|(|\mathcal{C}|-1)} \sum_{c_i \in \mathcal{C}} \sum_{c_j \in \mathcal{C}, c_j \neq c_i} \exp - \frac{d^2(\overline{x}_{c_i}, \overline{x}_{c_j})}{2\sigma^2} \tag{1.18}$$

où σ est une constante gaussienne, $\overline{x}_{c_k}$ est le vecteur-moyenne (ou centroïde) du cluster $c_k \in \mathcal{C}$, et $d(\overline{x}_{c_i}, \overline{x}_{c_j})$ la distance entre les vecteurs-moyenne $\overline{x}_{c_i}$ et $\overline{x}_{c_j}$.

Il faut souligner que la distance entre les vecteurs-moyenne des clusters pris deux à deux (*pairwise distance*) est la clef de la mesure de séparation. La fonction gaussienne et la fraction de gauche ont pour fonction de normaliser la mesure entre 0 et 1. Une valeur plus petite de la mesure de séparation indique une dissimilarité plus importante entre les clusters. Cependant, tout comme pour la mesure de compacité, cette nouvelle mesure n'est pas capable d'évaluer à elle-seule la qualité d'un clustering. En effet, il suffit de considérer deux représentants arbitraires aussi éloignés que possible l'un de l'autre pour obtenir une très bonne (sinon parfaite) séparation. C'est pourquoi He [HTTS02] propose une mesure calculant la qualité globale d'un cluster (*overall cluster quality*) basée sur une combinaison linéaire, tout comme dans le cas des mesures d'entropie citées plus haut, des deux mesures de compacité et de séparation :

$$Ocq(\beta, \mathcal{C}) = \beta.Cmp(\mathcal{C}) + (1-\beta).Sep(\mathcal{C}) \tag{1.19}$$

où $beta \in [0,1]$ est la valeur de compromis (*trade-off*) entre la similarité intra-classe et la dissimilarité inter-classes.

Bien que cette mesure globale puisse faciliter le travail de comparaison entre deux partitions, résultats d'un processus de classification, il nous paraît clair que le choix d'utiliser une combinaison linéaire n'est pas la solution idéale pour rendre compte à la fois d'une bonne homogénéité et d'une bonne séparation des clusters. Il peut ainsi être difficile d'interpréter la valeur finale obtenue, d'autant que les scores Cmp et Sep ne sont pas mesurés selon la même échelle. Nous nous limitons cependant dans le cadre de cette thèse, conscients de ces limitations, à cette solution proposée dans la littérature. Nous tâchons toutefois, à chaque fois que cela est possible, de considérer ces indices séparemment afin de réduire l'erreur d'approximation introduite.

1.1.4.3 Génération des données artificielles

La très grande majorité des algorithmes de clustering passe par l'étape obligée de validation à l'aide de données artificielles. En effet, ce type de données est beaucoup plus facile à trouver, à traiter et à manipuler. Il permet surtout de simplifier l'étape de validation des résultats même s'il ne remplace pas les expérimentations sur des jeux réels.

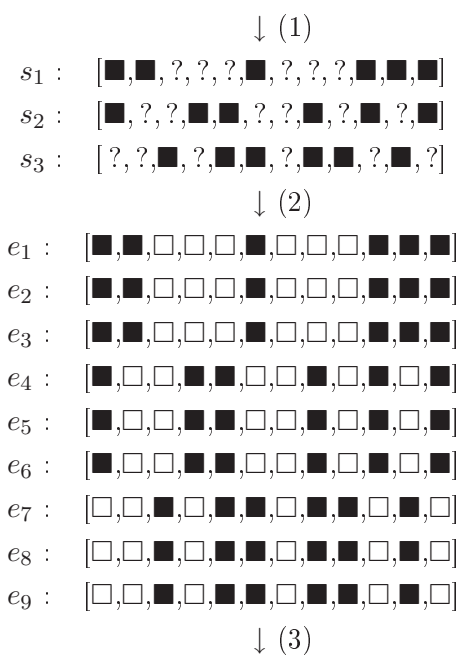
Le point de départ, dans la plupart des cas, consiste à choisir un nombre k de classes à chacune desquelles est associée une distribution gaussienne spécifique suivant chacune des variables. Les données sont ensuite générées à l'aide de ces distributions et étiquetées par les

[9] Même remarque que précédemment.

numéros des classes correspondantes. Un poids peut être assigné à chacune de ces distributions afin de signifier la probabilité qu'elles ont de générer un exemple [FL00]. Ce que l'on obtient est un mélange (ou "mixture") de plusieurs distributions normales probabilistes. Les étiquettes sont alors cachées à l'algorithme de clustering qui tente de regrouper les exemples générés par une même distribution. Cette méthode de validation est classiquement utilisée dans le contexte du traitement d'image dans $\mathbb{R}^2$ ou $\mathbb{R}^3$ [HVB00][CSL03].

Si maintenant nous laissons de côté ce modèle probabiliste, largement utilisé, les expériences réalisées se font plus rares. W. Iba [IL01] cite le travail de J. Gennari qui, étudiant justement le cas des données lacunaires, utilise des jeux de données synthétiques. Il y est dit tout d'abord que chaque instance est, comme auparavant, étiquetée suivant la classe à laquelle elle appartient. Elle est en sus décrite par douze attributs, six numériques et six nominaux, avec à chaque fois trois attributs de chaque type considérés comme non pertinents pour la classe. En clair, cela signifie que les classes sont générées à partir d'un ensemble d'objets de départ[10] comportant six attributs pertinents, et que les trous sont comblés avec des valeurs aléatoires. Ces expérimentations considèrent donc la moitié des descripteurs utilisés comme étant du bruit qu'il faut savoir ignorer pour découvrir les classes qui y sont dissimulées.

Le schéma suivant résume la stratégie adoptée par Gennari pour générer des données artificielles :



(1) génération de trois "seeds" comportant chacun 6 variables possédant des valeurs pertinentes ■.

(2) recopie des descriptions initiales et complétion à l'aide des valeurs aléatoires □. Les exemples e_1, e_2 et e_3 correspondent à la description initiale s_1 ; e_4, e_5 et e_6 à s_2 ; e_7, e_8 et e_9 à s_3.

[10] Le terme *seed* est employé pour désigner ces objets de départ et correspond bien à l'idée de "graine".

$\downarrow$ (3)

e'_1 : [■,■, ?,□,□, ?, ?,□,□,■,■, ?]
e'_2 : [■,■,□, ?,□, ?,□,□, ?, ?,■,■]
e'_3 : [■, ?,□,□, ?,■,□,□,□,■,■, ?]
e'_4 : [?,□, ?,■,■, ?,□,■,□,■,□, ?]
e'_5 : [?,□,□,■,■,□, ?,■, ?, ?, ?,■]
e'_6 : [■,□, ?,■,■,□,□, ?,□, ?, ?,■]
e'_7 : [?,□,■,□,■,■,□,■, ?,□, ?, ?]
e'_8 : [?,□, ?,□,■,■,□,■, ?,□,■, ?]
e'_9 : [□, ?,■,□,■, ?,□,■,■,□, ?, ?]

(3) suppression aléatoire d'un pourcentage p de valeurs au sein des exemples.

L'objectif de l'algorithme de classification est alors de retrouver les classes initiales à partir de l'ensemble des exemples $e'_1, \ldots, e'_9$ produits par ce processus.

Une autre stratégie, cette fois totalement adaptée au cas des données symboliques, est proposée par S. Guha et al. et traite le cas d'un panier à provisions virtuel [GRS00]. L'ensemble est constitué de 114 586 transactions dont 5 456 (5%) sont des *outliers*, c'est-à-dire des données aléatoires qui n'appartiennent à aucune classe précise (bruit). Les autres transactions appartiennent à l'une des 10 classes dont la taille varie de 5 000 à 15 000 individus. Chaque cluster est défini par un ensemble d'items. Environ 40% des items sont communs avec d'autres clusters alors que 60% sont caractéristiques du cluster. Au sein de chaque cluster, une transaction est générée en tirant aléatoirement des items au sein de cet ensemble d'items. Les outliers sont générés totalement aléatoirement. Le paramètre de taille de la transaction possède une distribution normale avec une valeur moyenne de 15. La figure suivante résume la stratégie adoptée en se limitant, pour une question de lisibilité, à 12 items et 3 classes :

$\downarrow$ (1)

it_1 : [■,■,■,□,□,□,□,□,□,■,■,□]
it_2 : [□,□,□,■,■,■,□,□,□,□,■,■]
it_3 : [□,□,□,□,□,□,■,■,■,■,□,■]

$\downarrow$ (2)

e_1 : [■,□,□,□,□,□,□,□,□,■,■,□]
e_2 : [□,□,■,□,□,□,□,□,□,■,□,□]
e_3 : [□,■,■,□,□,□,□,□,□,■,□,□]
e_4 : [□,□,□,□,□,■,□,□,□,□,□,■]
e_5 : [□,□,□,□,■,■,□,□,□,□,□,□]
e_6 : [□,□,□,■,■,□,□,□,□,□,□,■]
e_7 : [□,□,□,□,□,□,□,□,■,■,□,■]
e_8 : [□,□,□,□,□,□,■,■,■,□,□,□]
e_9 : [□,□,□,□,□,□,■,□,■,□,□,□]

$\bigoplus$ (3)

e_{10} : [□,■,□,□,■,■,□,□,□,■,□,□]
e_{11} : [■,□,□,□,□,□,■,□,■,□,□,□]

(1) génération de trois ensembles d'items it_1, it_2 et it_3. Le carré noir ■ signifie que l'item est présent, alors que le carré blanc □ signifie qu'il est absent.

(2) génération du jeu de données par tirage aléatoire au sein de ces ensembles d'items. Les exemples e_1, e_2 et e_3 correspondent à l'ensemble it_1 ; e_4, e_5 et e_6 à it_2 ; e_7, e_8 et e_9 à it_3.

(3) ajout des outliers e_{10} et e_{11} (les proportions ne sont pas respectées).

L'objectif est de retrouver les trois classes initiales à partir de l'ensemble des exemples $e_1, \ldots, e_{11}$. Même si on ne manipule pas directement des données manquantes (un item appartient ou n'appartient pas à une transaction), la problématique développée est similaire. Il est bien évident que chaque transaction ne peut comporter tous les items du panier. Cependant, les éléments absents de cette transaction peuvent être considérés, d'une certaine manière, comme manquant au panier "idéal" représenté par l'ensemble des items associés.

La méthode que nous employons pour générer les ensembles de données artificielles est présenté dans la partie 4.2.1 et possède de nombreux points communs avec les deux approches que nous venons de décrire.

1.2 Catégorisation et analyse de la presse

Dans cette partie, nous donnons tout d'abord les principales approches utilisées pour traiter le problème de la catégorisation, approches dont les principes sont une importante source d'inspiration pour nos propres travaux. Nous présentons ensuite le concept de stéréotype et tâchons de détailler les relations qu'il entretient avec celui de prototype défini dans la partie précédente. Enfin, nous faisons le lien avec le domaine des représentations sociales et l'analyse du contenu de la presse.

1.2.1 Principes de catégorisation

La catégorisation est un thème très ancien en philosophie qui se trouve aujourd'hui au carrefour de nombreux domaines comme la psychologie cognitive, les sciences sociales, la linguistique et l'intelligence artificielle. C'est un problème central car, sans cette capacité à classer les objets similaires dans les mêmes groupes, nous serions plongés dans un chaos de cas particuliers. Ce problème est fortement lié à celui de l'induction qui propose de dépasser les entités individuelles pour découvrir les propriétés générales du monde qui nous entoure.

Nous donnons dans cette partie un aperçu des principaux courants traitant de la catégorisation et tâchons d'y situer nos travaux. Précisons immédiatement que nous faisons parfois un amalgame volontaire entre les théories psychologiques, sociologiques et linguistiques de la catégorisation. En effet, notre objectif est moins de discuter de l'apport exact de chacune de ces disciplines que d'y puiser notre inspiration dans l'élaboration d'un modèle cognitif. Ainsi, le problème lié au langage est évacué dans le sens où nous travaillons uniquement sur les concepts et que nous assumons l'existence d'un langage formel objectif permettant de décrire les objets du monde.

Après avoir abordé l'approche logique classique, dite aristotélicienne, qui a régné en maître durant des siècles, nous donnons les principales critiques qui lui furent opposées. Nous présentons ensuite la théorie cognitiviste de E. Rosch, s'inscrivant dans une étude d'ordre psychologique, qui vint bouleverser le paysage des idées reçues à propos des principes sous-jacent aux catégories. Nous parlerons enfin de la révision de la théorie du prototype proposée dans le domaine de la linguistique et principalement initiée par G. Lakoff.

1.2.1.1 L'approche classique

La théorie classique présente principalement la *thèse logique* (aussi appelée *thèse de définissabilité*) selon laquelle la catégorisation se fait sur la base de propriétés communes. Elle se place dans la tradition *objectiviste* qui assume que la pensée rationnelle manipule des symboles abstraits, symboles qui trouvent leur sens *via* une correspondance avec le monde, indépendamment de tout être vivant. Ainsi, pour décider de l'appartenance de l'objet x à la catégorie associée au concept "chien", il suffit de vérifier si le x en question possède les attributs qui constituent le dénominateur commun de la catégorie, autrement dit s'il est un animal, un mammifère, etc. S'il vérifie ces propriétés, ce sera un chien. Dans l'hypothèse contraire, il ne fait pas partie de la catégorie et ne peut donc pas être considéré comme étant un chien. Les concepts sont alors définis par ces propriétés, qui sont autant de conditions nécessaires dont la conjonction est suffisante pour décider de l'appartenance à l'une ou l'autre des catégories. C'est pourquoi ce modèle s'appelle le modèle CNS (comme Conditions Nécessaires et Suffisantes), également nommé par R.W. Langacker en linguistique modèle des attributs critériaux [Lan87].

Sans entrer plus avant dans les détails de cette approche, nous soulignons les points essentiels qui permettront d'effectuer une comparaison avec les approches ultérieures. Tout d'abord, l'appartenance d'un objet à une catégorie se calcule simplement en vérifiant le faisceau des CNS. Ces CNS constituent des traits qui déterminent totalement l'extension, c'est-à-dire tous les objets qui tombent sous le concept associé à la catégorie. Ensuite, tous les membres de la catégorie sont considérés comme étant équivalents car vérifiant, tous de la même façon, les CNS. Enfin, les frontières de la catégorie sont parfaitement définies, rigides : on appartient ou on n'appartient pas à cette catégorie. Nous simplifions sciemment les réflexions qui sont faites à ce sujet dans de nombreux ouvrages [Lak87][Kle88][Kle90].

1.2.1.2 Critique de l'approche par CNS

Cette théorie logique, qui ne prend pas en compte les agents cognitifs que sont les êtres humains ni leurs interactions avec le monde[11], est fortement critiquée dans les années 70 par Rosch et ses collaborateurs dans ce que M. Posner appelle la "révolution roschienne" [Pos86]. Le reproche principal adressé à cette approche est celui d'impliquer des catégories arbitraires qui dépendent fortement d'un relativisme culturel. Rosch lui oppose une vision "écologique" relative à l'existence de catégories naturelles [Ros73]. Ainsi, les résultats empiriques ont prouvé l'existence d'un niveau conceptuel privilégié, le niveau de base, autour duquel s'organise la plus grande partie de nos connaissances, et ce quelle que soit notre culture ou notre éducation.

Un deuxième point en défaveur du modèle CNS est sa vision très figée des catégories. Il ne semble pas très naturel de considérer des frontières rigides, notamment lorsqu'il s'agit de traiter les cas marginaux. De plus, tous les objets étant équivalents au sein de la catégorie, ils sont tous d'aussi bons représentants de cette catégorie. Là encore, l'interprétation des

[11] G. Lakoff emploie le terme *disembodied* à son sujet que l'on pourrait traduire par "désincarné".

résultats expérimentaux a permis de démentir les fondements psychologiques de l'approche classique.

Enfin, le principe même de conditions nécessaires et suffisantes semble ne pas être adapté à de nombreux exemples. Ainsi, L. Coleman et P. Kay [CK81] remettent en question le caractère nécessaire des trois conditions associées au verbe *to lie*. L'exemple lui-même du *bachelor* comme un "homme adulte non marié", emblématique du modèle CNS, est mis en cause avec des cas comme celui du pape ou des vieux couples homosexuels. En bref, il manque à ce modèle la flexibilité nécessaire à une véritable applicabilité dans notre quotidien.

1.2.1.3 La "révolution roschienne"

La psychologue E. Rosch pose la catégorisation comme l'un des piliers des sciences cognitives. Elle oppose aux catégories dites savantes, construites artificiellement, comme par exemple en formation de concepts[12], des catégories sémantiques naturelles. Ces catégories ont une structure basée sur le concept de *prototype*, exemplaire typique autour duquel s'organisent les autres objets de la catégorie.

Deux principes généraux sous-tendent à la formation des catégories.

Le premier est le principe d'*économie cognitive* qui propose d'obtenir le maximum d'informations pour une dépense minimum de ressources cognitives. Ce principe impose que les catégories tendent à être vues comme les plus distinctes les unes des autres et selon des coupures les plus nettes possibles. Le second est le principe selon lequel le monde perçu est *intrinsèquement structuré*. En effet, le monde qui nous entoure n'est pas un monde "gris" dans lequel toutes les propriétés sont équiprobables, mais au contraire un monde dans lequel celles-ci sont fortement corrélées. Ainsi, les propriétés "avoir des plumes" et "voler" ont beaucoup plus de probabilités de se retrouver ensemble que, par exemple, "donner du lait" et "pondre des œufs".

En combinant ces deux principes de base, on voit que l'on obtient le maximum d'informations pour le moindre coût cognitif si les catégories sont en correspondance aussi étroite que possible avec la structure du monde tel que nous le percevons. Le point de vue de cette approche est dit réaliste car celle-ci se base sur le monde perçu et non sur un monde métaphysique indépendant du sujet. Il est intéressant de noter que la vision de Rosch change sensiblement tout au long des années 70 [Ros73][Ros75][Ros78]. Le prototype "meilleur exemplaire" (en extension) devient progression une image abstraite (en intension). Elle remet finalement en cause plusieurs de ses hypothèses et ses conclusions se rapprochent de ce que G. Kleiber appelle, en linguistique, la version *étendue* de la sémantique du prototype [Kle90]. Nous reviendrons sur cette nouvelle approche de la catégorisation plus loin dans ce document parce qu'il nous faut à présent détailler davantage les principes sur lesquels reposent aujourd'hui la plupart des travaux sur la catégorisation.

La théorie du prototype peut être abordée selon deux dimensions.

[12] Il s'agit ici d'études effectuées en philosophie à cette époque, et non du champ rattaché plus récemment à l'IA.

La première est la dimension *verticale* et concerne les différents niveaux d'inclusion des catégories. Ces dernières sont représentées par un arbre allant des catégories les plus générales (en haut de l'arbre) aux catégories les plus spécifiques (vers les feuilles de l'arbre). Cette manière d'appréhender la catégorisation est celle des techniques de clustering hiérarchique, comme c'est le cas par exemple avec l'algorithme COBWEB de D. Fisher (voir section 1.1.1.4). Nous ne détaillons pas cet aspect de la théorie car nous nous intéressons principalement à l'aspect horizontal lié aux problèmes de clustering non hiérarchique, c'est-à-dire situés à un même niveau d'abstraction.

La seconde est la dimension *horizontale* qui reflète l'organisation interne des catégories. Rosch met en évidence l'existence d'un niveau de base qui est le niveau d'abstraction où les catégories correspondent le mieux à la structure des attributs tels que nous les percevons. Ce niveau peut être formalisé à l'aide du concept probabiliste de *cue validity* [RMG+76]. La validité, en tant que prédicteur de la classe y, du *cue* x (c'est-à-dire la probabilité conditionnelle $p(y/x)$) augmente avec la fréquence avec laquelle x est associé à y et diminue avec la fréquence avec laquelle x est associé à une autre catégorie que y [Bea64][Ree72]. La valeur de *cue validity* d'une catégorie est obtenue par la somme sur chaque attribut des *cue validity* individuelles. Une catégorie avec une forte valeur de *cue validity* est par définition mieux différenciée d'une autre catégorie qu'une catégorie possédant une faible valeur.

1.2.1.4 Le principe de typicalité

La structure des catégories repose sur le principe de typicalité (aussi appelé typicité) qui met en scène des objets appelés prototypes. Voilà comment E. Rosch explique en quelques mots le principe de typicalité :

"La perception de différences de typicalité est, d'abord, un fait empirique relatif aux jugements que les gens émettent sur l'appartenance à une catégorie. C'est maintenant un fait bien établi que les sujets s'accordent massivement dans leurs jugements relatifs au fait qu'un exemplaire est un exemple clair ou représentatif d'une catégorie, même pour des catégories dont ils discutent les frontières" [Ros78].

A la différence de la théorie classique, tous les exemples d'une même catégorie ne sont pas jugés également représentatifs de cette catégorie. Les éléments se distribuent ainsi selon un *gradient de typicalité* calculé suivant leur distance à un *prototype*. Le prototype résume l'ensemble des propriétés (corrélées) de la plupart des exemplaires en fonction du principe d'économie cognitive.

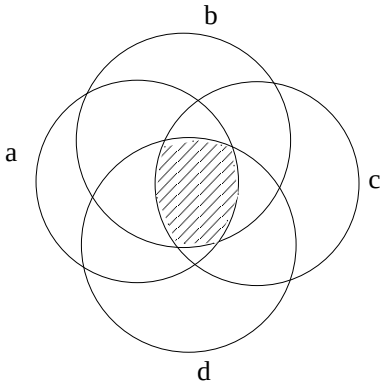
Dans les premiers travaux de Rosch [Ros73], le prototype est un membre caractéristique de la catégorie, c'est-à-dire le meilleur exemplaire communément associé à celle-ci. Ainsi, le *moineau* est le prototype de la catégorie *oiseaux*, tout comme le *chien* celui des *animaux domestiques*. Il est ensuite considéré comme un point de référence cognitif [Ros75], pour finalement devenir une construction mentale issue d'opérations cognitives (telles que les activités

de discrimination) [Ros78]. Ainsi, D. Dubois écrit :

"Le prototype peut être formé d'une combinaison jamais rencontrée de valeurs, même si chacune de ces valeurs se trouve le plus fréquemment rencontrée" [Dub86].

Le degré d'appartenance à la catégorie est alors considéré comme isomorphe au degré de similarité existant entre l'objet à classer et le prototype de cette catégorie. L'objet n'est donc pas obligé de posséder toutes les caractéristiques associées à la catégorie. La base de cette opération de catégorisation est appelé principe d'appariement (*matching principle*).

T. Givon [Giv86] illustre la théorie standard du prototype avec le schéma suivant :



Les objets situés dans la partie hachurée possèdent les quatre propriétés a, b, c et d. Ils sont donc les membres prototypiques de la catégorie. Ceux qui n'en possèdent que trois sont moins typiques et s'éloignent des instances centrales, mais sont toutefois moins marginaux que ceux qui n'en possèdent que deux ou une, ces derniers se situant à la périphérie de la catégorie. Il est important de souligner que la partie hachurée peut être vide, aucun exemplaire ne vérifiant les quatre propriétés à la fois. Cependant, n'oublions pas que ces quatre propriétés devront être fréquemment observées chez les membres de la catégorie.

Contrairement au modèle CNS, les membres d'une même catégorie, comme nous venons de la voir, ne partagent pas tous les mêmes traits en commun. Quelle est alors la relation qui les unit ? La réponse provient des réflexions du philosophe allemand L. Wittgenstein et a trait à la notion de *ressemblance de famille* (ou *air de famille*) [Wit53][Sha04][Nar01] . Comme l'expliquent E. Rosch et C.B. Mervis :

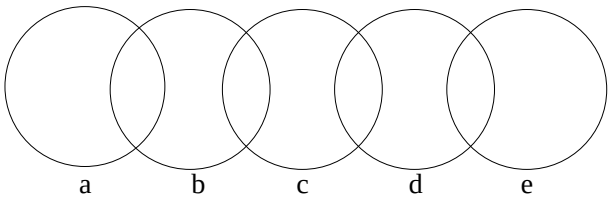
"Le but de la présente recherche était d'explorer un des principes structuraux majeurs qui, croyons-nous, peut régler la formation de la structure prototypique des catégories sémantiques. Ce principe a été suggéré en premier en philosophie ; Wittgenstein (1953) a postulé que les référents d'un mot n'ont pas besoin d'avoir d'éléments en commun pour être compris et employés dans le fonctionnement normal du langage. Il a suggéré qu'il s'agissait plutôt d'une ressemblance de famille qui reliait les différents référents d'un mot. Une structure de ressemblance de famille prend la forme AB, BC, CD, DE. C'est-à-dire que chaque item a au

moins un et probablement plusieurs éléments en commun avec un ou plusieurs autres items, mais aucun ou peu d'élément sont communs à tous les items" [RM75].

Dans la version standard de la théorie du prototype, la structure en ressemblance de famille a pour vertu essentielle d'expliquer que les membres d'une catégorie ne sont rassemblés ni au hasard ni sur la base d'une identité en termes de CNS. Ce qui les réunit, ce sont des similarités, des ressemblances qui s'entrecroisent, se recouvrent partiellement. Cependant, comme le souligne G. Kleiber [Kle90], si la structuration prototypique de la catégorie peut correspondre à une structuration en ressemblance de famille, la réciproque n'est pas forcément vraie.

1.2.1.5 L'approche par ressemblance de famille

La version étendue de la sémantique du prototype proposée dans le domaine de la linguistique prend les leçons de ces dernières réflexions. Sa caractéristique fondamentale est de redonner ses "lettres de noblesse" à la ressemblance de famille en la replaçant à sa juste place, c'est-à-dire à la base de la structure des catégories. Voici le nouveau schéma proposé par T. Givon [Giv86] afin d'illustrer le principe de la ressemblance de famille dans la théorie étendue :



On voit bien que, pour qu'il y ait ressemblance de famille, il faut et il suffit que chaque membre de la catégorie partage au moins une propriété avec un autre membre de la catégorie. Le schéma donné p.38 dans le cadre de la théorie standard du prototype est un cas particulier de celui-ci. La différence est que les exemples ne sont plus obligés de partager au moins une propriété commune avec le prototype. Les différentes instances d'une même catégorie ne convergent plus vers une même entité centrale qui, dans la version standard, constitue le ciment de la catégorie et qui fait de cette version prototypique un *compromis* entre le modèle classique et la théorie de la ressemblance de famille [Giv86].

Le concept de prototype central autour duquel s'organise la catégorie s'efface pour laisser la place à une structure plus générale qui est une véritable rupture avec la théorie standard pour certains, et non une simple évolution. Bien sûr, le concept de prototype est conservé mais l'interprétation des résultats des expériences réalisées dans les années 70 est reconsidérée [Ros78][Lak87]. Le prototype est à présent vu comme un *effet* de la structure des catégories, et non comme une entité fondatrice, et l'on préfèrera parler de degrés de prototypicalité. Rosch renonce à l'idée que le prototype puisse être un concept servant à représenter la catégorie :

"Les prototypes ne constituent pas une théorie de la représentation des catégories" [Ros78].

Puisque le prototype n'est plus cette entité organisatrice de la catégorie, on lui enlève aussi tout pouvoir pour expliquer l'appartenance d'une entité à celle-ci. Voici le deuxième dogme de la théorie standard qui s'écroule :

"Les prototypes ne constituent pas un modèle de calcul particulier pour les catégories" [Ros78].

La thèse des frontières floues et celle de l'assimilation du degré de représentativité au degré d'appartenance à la catégorie sont aussi abandonnées. Ainsi, même si *poussin* est un moins bon exemplaire de la classe *oiseau*, il n'en est pas moins un oiseau : ce n'est pas *plus ou moins* un oiseau, même si cet exemplaire est atypique de la classe.

Pour résumer, de toutes les propositions qui étaient à la base de la théorie standard, il ne reste que les deux suivantes, dont la première a en plus subi une modification cruciale :

- Il n'y a plus que des effets prototypiques : le prototype comme représentant des concepts des catégories et comme structuration de la catégorie a disparu ;
- La relation qui unit les différents membres d'une même catégorie, quelle que soit cette catégorie, est celle de ressemblance de famille.

Pour G. Lakoff, les effets prototypiques ne sont qu'une conséquence, un *by-product* des structures des catégories. Celui-ci propose une structure nouvelle sous la forme de *modèles cognitifs idéalisés* (*Idealized Cognitive Models* ou ICM en anglais) [Lak87], que nous ne détaillons pas ici. Rosch suppose, elle, qu'une description intensionnelle unique et rigide de l'*essence* d'une catégorie est impossible [Gee].

Nous ne détaillerons pas plus avant cette théorie récente qui impose un changement radical de point de vue, mais nous verrons que le modèle présenté dans cette thèse possède des similitudes avec cette version étendue car il donne également une grande importance à la notion d'air de famille. Nous allons à présent aborder le concept de stéréotype qui est très lié à celui de prototype et à la théorie de la catégorisation.

1.2.2 Le concept de stéréotype

Le concept de stéréotype est bien plus ancien que celui de prototype, au sens de la psychologie cognitive, et les deux sont souvent confondus. A tel point que de nombreuses publications mettent ces deux concepts en parallèle et s'attachent à éclaircir un peu le paysage de la typicité [Kle88][DRR93][Pla99]. Nous allons à présent brosser un aperçu des différentes manières d'aborder le concept de stéréotype.

Le stéréotype, au sens de représentation ou d'image stéréotypée, est apparu dans les années 20 avec le publiciste américain W. Lippman dans un ouvrage sur l'opinion publique [Lip22]. Rapidement les travaux sur ce sujet se multiplièrent, notamment aux U.S.A. à propos des problèmes raciaux, ainsi qu'autour de la seconde guerre mondiale à propos des stéréotypes nationaux. Mais ces travaux portaient surtout sur les contenus représentationnels des stéréotypes, plutôt que sur le processus cognitif de stéréotypisation. Il faudra attendre le

mouvement cognitiviste des années 70 et les recherches de Rosch pour s'interroger sur ces processus et effectuer de réelles validations empiriques.

Un travail fondamental est réalisé par H. Putnam en 1975 [Put75] dans le domaine de la psychologie sociale où le stéréotype est vu comme une représentation mentale idéalisée. Cette représentation peut être erronée mais possède la caractéristique d'être partagée par un grand nombre de personnes, ce qui la définit comme un *stéréotype social*, concept actuellement étudié au sein du champ de la psychologie sociale. L'exemple couramment utilisé pour illustrer cette notion est celui de l'or, dont la valeur par défaut associée au stéréotype est jaune, bien que l'on puisse trouver de l'or blanc. Un autre stéréotype donné comme exemple est celui du tigre dont le pelage porte des rayures, bien qu'on ne sache pas s'il s'agit réellement du trait le plus couramment observé chez cet animal. Ainsi, la fréquence correspondant à la caractéristique observée n'est pas déterminante et peut s'effacer devant la convention sociale.

Dans la lignée des travaux de M. Minsky sur les *frames* [Min75] ou de R.C. Schank et R.P. Abelson sur les *scripts* [SA77], E. Rich propose en 1979 [Ric79] d'utiliser des réseaux de stéréotypes, organisés en hiérarchies, afin d'élaborer un modèle de lecteur au sein d'une bibliothèque. Son système, Grundy, construit au fur et à mesure de la conversation un *User Synopsis* (USS) en combinant l'information donnée directement par le lecteur, des inférences déduites des actions de celui-ci et des prédictions effectuées à l'aide des stéréotypes qui ont été déclenchés avec des activateurs (*triggers*). Ce modèle est ensuite utilisé afin de conseiller l'utilisateur sur les ouvrages les plus adaptés à sa personnalité. Un mécanisme d'apprentissage par renforcement est utilisé afin d'adapter les stéréotypes (et les activateurs) en modifiant la valeur des attributs associés et leurs indices de confiance, en fonction des utilisateurs rencontrés. La création de nouveaux stéréotypes est envisagée dans les perspectives.

En 1985, C. Schwarze [Sch85], cité par G. Kleiber et D. Dubois, distingue ces deux réalités en nommant la première, c'est-à-dire l'objet qui est le meilleur exemplaire de la catégorie, prototype, et la seconde, c'est-à-dire le concept qui la décrit, stéréotype. Le stéréotype est la représentation mentale, présentée *en intension*, d'une représentation *en extension* qui est en fait le prototype lui-même pris dans sa première version de 1973 [Ros73]. De son côté, G. Lakoff rapproche en 1987 [Lak87] les stéréotypes de la notion de représentation stabilisée et fixée en mémoire artificielle telles que les frames ou les scripts manipulées en IA à cette époque. Ce point de vue le rapproche des expériences réalisées par E. Rich huit ans plus tôt [Ric79].

D. Dubois parle d'un "glissement" du concept de prototype vers celui de stéréotype :

"Ce glissement du domaine psychologique individuel à la norme sociale imposée par la langue, à travers le passage par une mémoire artificielle impersonnelle, non spécifique d'une langue et donc universelle, conduit à introduire dans sa définition même l'idée que 'l'analyse associative s'appuie sur une relation a priori, qui fait partie du savoir présumé communément partagé sur les choses' (Kleiber), c'est-à-dire une définition d'un savoir normé par un contexte social et culturel, donc un stéréotype". [DRR93]

Ce glissement s'explique par un phénomène de généricité, d'acceptation d'un certain ordre du monde par un groupe social. Le prototype devient un stéréotype dans la mesure où il est une représentation stabilisée d'un ensemble physique universel : "le monde réel". D. Geeraerts souligne d'ailleurs le rapprochement entre la notion de prototype et le stéréotype social de Putnam :

"Le prototype porte en lui-même les traits les plus communément *associés à la catégorie. Ce passage de l'individuel à l'ensemble des locuteurs, au 'conventionnel' en somme, ouvre en même temps sur une dimension collective qui rapproche la sémantique du prototype de la théorie du stéréotype de H. Putnam, d'origine 'sociale"'* [Gee].

Il précise aussi que les perspectives sont différentes dans le sens où le prototype décrit plutôt une réalité psychologique d'économie conceptuelle alors que le stéréotype décrit une convention sociale. Cette opinion est assez largement répandue, aussi bien dans le domaine de la psychologie cognitive que dans celui de la psychologie sociale [Sem89].

Cependant, la frontière entre ces deux notions souvent amalgamées n'est pas aussi claire qu'une distinction entre réalité psychologique et réalité sociologique. L'assimilation est d'ailleurs largement favorisée par le fait que le terme "prototype" a plus le sens de stéréotype que celui de prototype en tant que premier exemplaire [Kle90]. On serait même poussé à croire que les deux notions se rejoignent dans le cas de la théorie standard, dans la mesure où les données sémantiques les plus importantes d'un point de vue social sont aussi celles qui sont les plus importantes dans l'organisation cognitive des catégories.

Enfin, pour J.L. Plane [Pla99], les travaux qui se réfèrent à la notion de prototype le font généralement pour mettre l'accent d'une façon ou d'une autre sur les différenciations et la structuration internes aux catégories, alors que ceux qui recourent plus volontiers à la notion de stéréotype tendent à négliger cet aspect pour mettre plutôt l'accent sur les contrastes ou différenciations entre catégories. Nous allons voir à présent comment situer nos propres travaux dans un contexte aussi complexe.

1.2.3 Les représentations sociales

Pour maîtriser l'infinie complexité du monde qui nous entoure, nous avons besoin d'utiliser des principes de catégorisation pour classer les objets, les événements, les personnes. *"Ces procédures se rapportent à la manière dont nous décomposons, disséquons, divisons et ordonnons notre réalité physique et sociale".* Cette citation, que nous empruntons à G.R. Semin [Sem89], montre bien l'intérêt porté aux théories de la catégorisation par les personnes travaillant en représentations sociales, domaine affilié à la psychologie sociale et plus généralement aux sciences sociales. Semin étudie les similitudes d'approche entre les représentantions sociales et la psychologie cognitive.

S. Moscovici pose les bases d'un nouveau domaine de recherche, celui des représentations sociales, dans son ouvrage intitulé *La psychanalyse : son image et son public* paru pour la

première fois en 1961 [Mos76]. En quelques mots, une représentation sociale est *"un type idéal constitué à partir des traditions, des œuvres politiques ou philosophiques"*, qu'il distingue de la notion d'idéologie. Voici quelques précisions qu'il apporte à cette notion dans la réédition de son ouvrage :

"Les représentations sociales sont des entités presque tangibles. Elles circulent, se croisent et se cristallisent sans cesse à travers une parole, un geste, une rencontre, dans notre univers quotidien. La plupart des rapports sociaux noués, des objets produits ou consommés, des communications échangées, en sont imprégnées. Nous le savons, elles correspondent d'une part à la substance symbolique qui entre dans l'élaboration et d'autre part à la pratique qui produit ladite substance, tout comme la science ou les mythes correspondent à une pratique scientifique et mythique" [Mos76].

A cette lecture, il apparaît que le concept des représentations sociales ne possède pas, par nature, une définition parfaitement claire. Enfin, il pâtit d'un contenu trop large et mal défini [HL83]. D. Jodelet donne une définition qui nous semble beaucoup plus accessible [Jod89] : les représentations sociales nous guident dans la façon de nommer et de définir ensemble les différents aspects de notre réalité de tous les jours, dans la façon de les interpréter, de statuer sur eux et, le cas échéant, de prendre une position à leur égard et de la défendre. C'est un savoir naïf élaboré à partir des éléments épars de la vie de tous les jours. De nombreux travaux réalisés dans ce domaine ont mis en évidence la possibilité d'étudier la structure interne et la dynamique de ce type de représentations [Abr03].

1.2.4 L'analyse du contenu de la presse

Ce domaine de recherche étudie la formation des représentations à travers le discours de presse. Parmi les premiers travaux de ce domaine, citons celui entrepris par S. Moscovici et ses collaborateurs [Mos76]. Tâche titanesque que celle d'étudier "à la main" les représentations de la psychanalyse comme un modèle de communication et d'expression à travers 1640 coupures parues dans 230 journaux entre janvier 1952 et juillet 1956. Utilisant une multitude de tableaux et de "schémas de message", ce travail remarquable pour l'époque nous conforte aujourd'hui dans l'idée qu'une automatisation au moins partielle des processus d'analyse est indispensable.

De nombreux logiciels d'analyse automatique des données textuelles permettent aujourd'hui de soulager les sociologues de ce travail long et fastidieux. Parmi ceux-ci se distingue Alceste [Rei86] qui, après avoir extrait les mots les plus importants du texte étudié (on parle de *mots-clefs*), construit grâce à une classification descendante hiérarchique les structures signifiantes les plus fortes. Il utilise également des techniques classiques de l'analyse des données comme l'analyse factorielle des correspondances. Plus atypique est le logiciel Prospero élaboré par F. Chateauraynaud [Cha03] qui se propose de suivre et d'étudier des dossiers complexes, tels celui de l'Amiante ou de la crise de la vache folle. Il mêle des techniques

habituelles de statistiques à des techniques de linguistique, voire d'intelligence artificielle, afin d'effectuer une analyse du discours tenu dans ce type d'affaires. Citons enfin le logiciel Nomino [PDP98] développé au département de linguistique de l'université du Québec à Montréal. Son objectif est de construire des bases de connaissances en utilisant le langage naturel comme support d'information. Pour cela, il se base sur la construction et la catégorisation d'une structure de *fiches*, représentant l'information de manière plus formalisée, pour effectuer une analyse sémantique et conceptuelle des textes.

Notre thèse propose une approche nouvelle concernant l'étude des représentations sociales à travers les discours de presse. Elle suppose qu'une première analyse sémantique a été effectuée et extrait à partir de ces données formalisées une forme de représentations que sont les stéréotypes. Ainsi, les articles de journaux constituent l'application principale de notre modèle car l'information qu'ils fournissent est nécessairement lacunaire. De plus, cette information reflète la plupart du temps la nature de la source d'où elle est extraite du fait des opérations de filtrage qui y sont effectuées. Que l'opération soit effectuée sciemment ou non, les informations disponibles dans les médias concernant un objet social (qui est souvent un enjeu social) proposent une vision biaisée, décalée du monde qu'elles sont pourtant sensées décrire. Nous pensons qu'en cherchant à extraire automatiquement une représentation, à partir des médias, par un mécanisme d'induction, nous rentrons dans le cadre de l'analyse du contenu de la presse. L'intérêt principal est bien entendu l'automatisation d'une partie du processus d'analyse qui demande énormément de temps lorsqu'il est effectué à la main, et une meilleure objectivisation avec la suppression de certaines des (sinon toutes les) étapes du processus.

1.2.5 Du choix des termes

L'un des aspects des représentations sociales passe, et cela semble assez naturel, par l'étude des stéréotypes sociaux tels qu'ils ont pu être définis par Putnam [Put75]. Le modèle cognitif proposé dans cette thèse utilise précisément des ensembles de stéréotypes afin de représenter un corpus d'articles de journaux relatifs à un phénomène que l'on souhaite étudier. Le choix du terme de "stéréotype" ne s'est pas fait sans mal, considérant d'autres termes candidats comme celui de "lieu commun", "cliché", "caricature" et surtout "prototype". Nous allons passer rapidement en revue chacun de ces termes.

Alors que le *lieu commun* est une idée banale en soi, sans intérêt, éculée, le *cliché* représente en linguistique une expression toute faite devenue banale à force d'être répétée. Ces deux termes ne nous ont pas semblé satisfaisants, de par la nature linguistique de leur contexte d'utilisation courante et surtout de par l'absence de "déformation" des représentations. Remarquons cependant que le cliché est parfois vu comme un stéréotype de l'expression linguistique [DIC].

La *caricature* est une image non conforme à la réalité qu'elle représente ou suggère, et dont elle constitue une altération déplaisante ou ridicule. Le moindre défaut est gonflé jusqu'à la démesure, les traits les plus saillants sont mis en évidence et exagérés [Kot03]. La caricature

se démarque de notre travail selon deux points : d'une part, elle est tournée vers l'exagération de traits saillants alors que nous nous sentons plus proches de l'idée d'une déformation liée à l'ignorance d'une partie de l'information ; d'autre part, la caricature a souvent un objectif (comique, politique) que nous ne revendiquons pas. Il est cependant clair que le terme de "caricature" possède des liens avec ce que nous cherchons à réaliser dans notre travail.

1.2.6 Des stéréotypes pour l'analyse du contenu de la presse

Nous avons choisi d'employer le terme de stéréotype pour qualifier les représentants de données lacunaires. Nos travaux s'inscrivent naturellement dans les théories sur la catégorisation de Rosch et cherchent à construire des éléments représentatifs centraux qui constituent une forme de représentation des données. Plusieurs raisons ont motivé ce choix par rapport au terme plus usuel de prototype.

La première raison est que nous ne nous situons pas exactement dans la théorie standard du prototype, telle que nous l'avons présentée dans la partie 1.2.1.4. En effet, nous ne prétendons pas que le stéréotype est le meilleur représentant de la catégorie, mais plutôt une combinaison de traits qui permet à la catégorie de conserver sa cohésion. D'ailleurs, les exemples les plus proches en terme de similarité avec le stéréotype qui les couvre ne sont pas considérés comme des *meilleurs* membres de la catégorie. Certaines observations spécifiques semblent ainsi jouer un rôle important dans la cohésion des catégories malgré un faible degré de similarité avec leur stéréotype. De plus, le principe d'appariement n'est pas respecté car nous utilisation une fonction qui affecte à chaque objet un, et un seul, stéréotype. Enfin, nous donnons une importance majeure à la ressemblance de famille comme c'est le cas dans la version étendue de la théorie du prototype. Ceci est à l'origine du développement de la contrainte de *cohésion cognitive* que nous présentons dans les sections 2.2.7 et 3.1.5.2.

La seconde raison est que notre application considère des articles de journaux qui sont diffusés par un groupe social pour un autre groupe social. Ceux-ci influencent inévitablement les représentations élaborées par le groupe de personnes au contact de ces informations. L'intuition nous pousse à croire que ces représentations seront d'autant plus similaires d'une personne à l'autre que le nombre de sources concernant le même phénomène est faible. Précisons que notre modèle s'inscrit clairement dans la vision cognitiviste de la théorie du prototype même si l'application qu'il traite concerne des objets aux enjeux d'ordre social. C'est pourquoi nous avons choisi de ne pas laisser de côté cet aspect qui rapproche notre travail du domaine des représentations sociales.

La troisième raison provient du cadre des données à valeurs manquantes. Même si les aspects psychologique et social sont intimement liés, nous pensons en effet que l'étude des prototypes concerne plus particulièrement les observations complètes, dans lesquelles la majorité de l'information "objective" est disponible. Plus le nombre de valeurs absentes est important, plus la représentation élaborée s'éloigne de la réalité telle que nous la percevons. Ainsi, bien que cette représentation nous semble "coller" à toute l'information dont nous disposons, celle-ci en constitue dans une large majorité des cas une vision déformée, non vérifiée.

En ce sens, elle mérite selon nous davantage l'étiquette de stéréotype que celle de prototype. Dans la même veine que le stéréotype social de Putnam, la représentation stéréotypée que nous souhaitons construire automatiquement est très probablement erronée, mais susceptible d'être partagée par un groupe important de personnes. Le concept de stéréotype est vu ici comme une représentation déformée (nous pourrions presque dire dégénérée) du concept de prototype causée par l'absence d'une majeure partie de l'information. Là où le prototype est associé à des calculs de moyenne ou de fréquence, comme c'est majoritairement le cas en apprentissage artificiel, le stéréotype est plutôt lié à une idée de combinaison de propriétés pertinentes.

Nous ne cherchons pas à décrire explicitement les représentations mentales des personnes placées au contact d'une ou plusieurs sources d'information. Les représentations que nous construisons à l'aide de stéréotypes ne sont que le reflet de ce que les médias peuvent diffuser, volontairement (comme dans le cas de la propagande) ou non. Nous ne spéculons pas sur l'usage exact qu'un lecteur peut faire des médias dans l'élaboration de sa propre construction mentale.

La dernière raison invoquée concerne l'utilisation effective des représentations que nous sommes amenés à construire à l'aide de notre modèle. En effet, celles-ci présentent des caractéristiques importantes de discrimination (grâce notamment à la notion de séparation) et de prédiction (grâce à la richesse des descriptions obtenues) qui correspondent à l'idée qu'en donne Lippman [Lip22].

Toutes ces raisons nous amènent à préférer le terme de "stéréotype", en précisant toutefois que nous nous trouvons bien dans un cadre d'ordre psychologique en nous situant au niveau d'un individu, ce qui n'est pas du tout incompatible. Nous partageons également plusieurs idées avec celles développées dans la version étendue de la sémantique du prototype. Par contre, nous continuons à chercher une image centrale à la catégorie (notre stéréotype), alors que la version étendue rejette cette idée.

1.3 Méthodes d'optimisation pour la tâche de clustering

Nous traitons dans cette partie des méthodes pouvant être utilisées afin d'optimiser une fonction dans le cadre d'une tâche d'apprentissage non-supervisé. Après avoir discuté du problème général posé lorsque l'on souhaite procéder à une telle optimisation, nous détaillons la méta-heuristique de recherche taboue et discutons brièvement du compromis entre intensification et diversification. Cette stratégie de recherche est adaptée et mise en œuvre dans l'algorithme que nous utilisons pour effectuer les expérimentations du chapitre 5.

1.3.1 Problèmes d'optimisation

1.3.1.1 Optimisation et $\mathcal{NP}$-complétude

On considère souvent deux grandes classes lorsque l'on s'attaque à un problème. La première de ces classes regroupe tous les problèmes dans lesquels rien ne prouve qu'une solution

existe. Le fameux "problème de l'arrêt d'un programme" de Turing en est une illustration. La seconde classe regroupe les problèmes dont on sait qu'ils possèdent une solution. Trouver la (ou les) valeur(s) qui optimise(nt) une fonction, comme nous le faisons dans le cadre de cette thèse, est un problème de ce type. Seulement, ce n'est pas parce qu'il existe une solution que celle-ci peut être si aisément découverte.

Beaucoup de problèmes d'algorithmique peuvent être résolus en temps polynomial. Il est d'usage d'utiliser le symbole $\mathcal{P}$ pour représenter l'ensemble de ces problèmes qui possèdent une complexité en $\mathcal{O}(n^k)$, où k est une constante. Malheureusement, cela n'est pas toujours le cas et certains problèmes ne peuvent être résolus en temps polynomial. Ainsi, la classe des problèmes dits $\mathcal{NP}$-complets regroupe justement une famille de problèmes dont la solution optimale prendra (très probablement du moins) un temps exponentiel pour être découverte. Outre les problèmes d'optimisation, un exemple célèbre de problèmes de ce type est celui du *voyageur de commerce* qui cherche à trouver le trajet le plus court qui passe (une seule fois) par un ensemble de villes et revient finalement au point de départ.

Comme nous l'avons vu précédemment, les algorithmes de clustering se basent, dans une large majorité des cas, sur l'optimisation d'une fonction traduisant la qualité de la partition recherchée (somme des carrés, vraisemblance, etc.). Or, découvrir la meilleure valeur de cette fonction, c'est-à-dire minimisant la variance intra-classe et maximisant la variance inter-classes, est un problème $\mathcal{NP}$-complet [Bru78]. Cela signifie qu'il n'existe très probablement pas d'algorithme permettant de le résoudre de manière exacte en un temps polynomial $\mathcal{O}(n^k)$.

Deux approches peuvent être utilisées dans le cas d'un problème $\mathcal{NP}$-complet [CLR92]. D'une part, si les entrées du problème sont petites, un algorithme ayant un temps d'exécution exponentiel (comme la technique du *Branch and Bound*) peut parfaitement convenir. Il s'agit ici d'effectuer une énumération "intelligente" de toutes les situations possibles afin de découvrir la solution du problème. D'autre part, il est possible de trouver des solutions *presque optimales* à l'aide d'heuristiques s'effectuant en temps polynomial. Ce type de résultat approché peut convenir dans une grande majorité des applications. C'est pourquoi notre problème d'*optimisation* peut être résolu par l'utilisation d'une heuristique qui donnera une bonne approximation de la solution optimale.

1.3.1.2 Heuristiques pour l'optimisation

De nombreuses heuristiques ont été proposées dans le domaine de l'IA pour résoudre ce type de problèmes. Parmi celles-ci se distinguent les méta-heuristiques dite de *recherche locale* qui sont des méthodes générales pour trouver une solution approchée à un problème d'optimisation. La descente du gradient, les algorithmes génétiques ou la recherche taboue sont des exemples de telles méthodes. Ces techniques ont déjà été utilisées pour effectuer une tâche de clustering [ASK96][BK99][Ben03].

Dans un premier temps, nous présentons la méthode la plus élémentaire cherchant à trouver la solution d'une fonction à optimiser. Cette technique de recherche locale, basée sur le calcul du voisinage de la solution courante, est appelée soit *descente du gradient*, soit

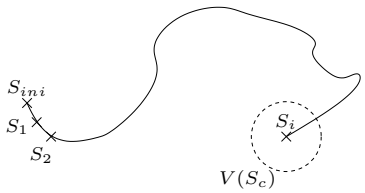


FIG. 1.4 – Recherche locale à travers l'espace des solutions.

hill-climbing, suivant que l'on cherche à minimiser ou maximiser la fonction. Cette stratégie de base est ensuite enrichie grâce à l'utilisation de la méta-heuristique de recherche taboue décrite dans la section suivante.

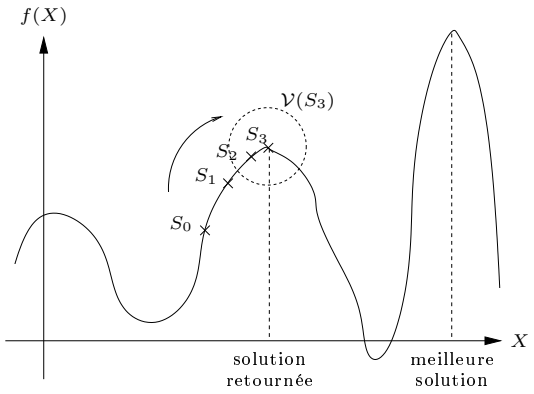
La recherche locale est une heuristique qui vise à trouver l'optimum d'une fonction f de manière itérative. On considère à chaque étape un voisinage $\mathcal{V}$ de la solution courante S_c, voisinage au sein duquel on choisit la meilleure solution $\arg\max_{S\in\mathcal{V}} f(S)$ qui devient la nouvelle solution courante. Nous résumons l'algorithme implémentant une stratégie de type hill-climbing :

1. Une solution initiale $S_{ini} \in \mathcal{S}$ est choisie (de manière aléatoire ou non). Celle-ci devient la solution courante $S_c = S_{ini}$ dont la valeur pour la fonction est $f(S_c)$.
2. Un voisinage $\mathcal{V}$ est calculé à partir de la solution courante S_c. Chaque nouvelle solution potentielle de $\mathcal{V}(S_c)$ est obtenue à partir de S_c et d'un ensemble de *mouvements* autorisés. Le choix des mouvements possibles est très important car il détermine l'efficacité et le temps d'exécution de la recherche. On calcule la valeur de f pour chacune des solutions du voisinage.
3. Le mouvement qui mène à la meilleure solution, c'est-à-dire la solution S_{max} de $\mathcal{V}(S_c)$ qui obtient le meilleur score $\max_{S\in\mathcal{V}(S_c)} f(S)$ pour la fonction f, est choisi. S_{max} devient la nouvelle solution courante S_c.
4. Le processus reprend à l'étape 2 un nombre *iter* de fois. Le meilleure solution découverte jusqu'à présent S_b est enregistrée.
5. Lorsque *iter* itérations ont été effectuées, la meilleure solution S_b est retournée.

Cette méthode très générale pour explorer un espace d'hypothèses, illustrée par la figure 1.4, possède deux défauts majeurs bien connus.

Le premier est le temps d'exécution qui peut croître très rapidement avec l'augmentation de la taille du voisinage. En effet, il est indispensable de calculer la qualité de toutes les solutions potentielles se trouvant aux alentours de la solution courante. Suivant la fonction employée, cela peut prendre du temps. C'est pourquoi il ne faut pas envisager trop de mouvements possibles.

Le deuxième est le danger d'être rapidement coincé dans un optimum local et de ne parvenir à en sortir. La figure ci-dessous illustre ce problème lorsque l'on cherche à maximiser la fonction f à partir de la solution initiale S_{ini} :



Nous allons voir que la recherche taboue permet justement de pallier ce dernier problème.

1.3.2 La Recherche Taboue

1.3.2.1 Principes généraux

La recherche taboue est une technique, proposée par F. Glover en 1986 [Glo86], qui se base sur l'utilisation de structures de mémoire à court et à long terme. Avant toute chose, précisons que cette technique reste basée sur les principes de la recherche locale tels que nous venons de les voir. C'est-à-dire que l'on garde la notion essentielle d'exploration itérative du voisinage d'une solution courante. Cependant, la recherche taboue propose d'inclure des méthodes reposant sur des principes de l'intelligence humaine.

Cette méta-heuristique ajoute certaines restrictions dans le but de guider la recherche vers des régions difficiles à atteindre avec des méthodes classiques. Ces restrictions prennent plusieurs formes comme l'exclusion pure et simple de certaines alternatives, la modification de l'évaluation des solutions ou de la probabilité de leur sélection. Elles font toujours référence à une structure de mémoire adaptée à cet effet.

Les deux principales caractéristiques de la recherche taboue sont, d'une part, l'ajout d'une mémoire adaptative qui consigne (notamment) les dernières solutions rencontrées, et, d'autre part, l'utilisation d'une exploration dite sensible. Nous détaillons ces deux caractéristiques dans les deux prochaines sections avant de discuter d'une technique spécifique de mémoire à long terme.

1.3.2.2 Utilisation de la mémoire

La recherche taboue utilise une mémoire adaptative illustrée par l'exemple de l'ascension d'une montagne par un grimpeur présenté dans [GL97]. Ce dernier, pour trouver le chemin le plus adéquat, doit se souvenir des différents indices rencontrés lors de ses précédentes ascensions. Ainsi, il possède dans sa tête des schémas (*samples*) de situations passées qui lui serviront à trouver un chemin plus efficace. Les structures de mémoire utilisées font référence à quatre dimensions (primeur, fréquence, qualité et influence) que nous ne détaillons pas ici.

Les deux éléments à retenir sont les suivants. Tout d'abord, il est possible d'enregistrer les solutions soit de manière *explicite* (toutes les solutions rencontrées ou juste les meilleures), soit de manière *attributive* (on enregistre les attributs des solutions rencontrées). Alors que la première stratégie est coûteuse en place mémoire et peut être utilisée pour lancer de nouvelles recherches dans des régions encore inexplorées de l'espace, la seconde est plus économique et permet de guider les mouvements de la recherche. Ensuite, deux types de mémoire peuvent être considérés : la mémoire à court terme (*recency-based memory*) et la mémoire à long terme (*frequency-based memory*).

1.3.2.3 Mémoire à court terme

La mémoire à court terme est certainement la forme la plus connue et la plus communément utilisée de recherche taboue. A tel point que cette dernière se résume dans l'écrasante majorité des cas à ce type de mémoire. Elle permet de réduire le voisinage $\mathcal{V}(S)$ de la solution courante S, guidant ainsi le choix du mouvement à effectuer lors de la prochaine itération. Nous noterons par $\mathcal{V}^*(S)$ le nouveau voisinage obtenu à l'aide de cette méthode. Ainsi, la recherche taboue peut être vue comme une recherche où le voisinage de la solution courante S est dynamique et dépend de l'historique de la recherche.

La manière la plus simple de prendre en compte une mémoire à court terme est d'utiliser une *liste taboue* T contenant explicitement t différentes solutions (mémoire explicite). Le nouveau voisinage $\mathcal{V}^*(S)$ consiste alors à retirer de $\mathcal{V}$ les solutions contenues dans T. Cette mémoire permet notamment d'éviter de tomber dans des cycles de longueur inférieure ou égale à $card(T)$ sur la trajectoire de la recherche. Cependant, la mémoire à court terme est le plus souvent basée sur l'enregistrement, non pas des solutions complètes, mais des mouvements qui ont été effectués dans un passé récent. C'est pourquoi cette mémoire est qualifiée de *recency-based*. Ces mouvements correspondent souvent à des attributs constitutifs de la solution (comme un arc, un nœud).

En général, la prochaine solution choisie au sein du voisinage $\mathcal{V}^*(S)$ est celle ayant une meilleure valeur pour la fonction f que l'on cherche à optimiser. Cependant, il peut arriver qu'aucune de ces solutions ne convienne et que l'on se retrouve bloqué. L'un des moyens de passer outre ce problème d'optima local est d'utiliser un *critère d'aspiration*. Il permet de modifier exceptionnellement le statut d'un élément de T afin que celui-ci puisse être à nouveau choisi. Il existe plusieurs critères de ce type : aspiration par défaut, aspiration objective, etc. Il arrive cependant que l'on se retrouve parfois bloqué, malgré l'utilisation d'une liste taboue, dans un optima local. L'intégration d'une mémoire à long terme devient alors indispensable.

1.3.2.4 Intensification et Diversification

La mémoire à long terme dans la recherche taboue est encore très peu usitée au sein de la communauté scientifique. Elle repose sur l'enregistrement des meilleures solutions découvertes jusqu'à présent (en général, les optima locaux), qui seront utilisées comme base de départ pour de futures recherches, et sur la prise en compte d'informations de fréquence complémentaires

à celles de la mémoire à court terme.

Ces informations basées sur la fréquence sont typiquement des ratios dont le numérateur correspond à deux mesures : une mesure de *transition* correspondant au nombre d'itérations où un attribut de la solution a été changé (c'est-à-dire qu'il est entré ou sorti de la liste taboue) ; et une mesure de *résidence* correspondant, par exemple, au nombre d'itérations où un attribut a appartenu à la solution visitée. Ces ratios représentent des fréquences de transition et des fréquences de résidence qui gardent une trace de l'utilisation des attributs durant la recherche. Ils peuvent permettre de lancer une phase d'intensification ou de diversification.

L'*intensification* est une stratégie basée sur la modification des règles de choix parmi les solutions du voisinage, encourageant des combinaisons de mouvements et la réutilisation de caractéristiques (attributs) de solutions déjà rencontrées et jugées intéressantes. Cela peut notamment permettre de retourner dans des régions considérées comme attractives pour les exploiter davantage. L'intensification peut revêtir plusieurs aspects comme l'utilisation de fonctions d'évaluation modifiées ou la combinaison de plusieurs solutions. Des techniques comme le *path-relinking* ou l'oscillation s'inscrivent dans une telle stratégie.

La *diversification* est une stratégie très importante qui rappelle, d'une certaine manière, la technique de "restart" utilisée par exemple en recherche adaptative. Le principe fondamental encourage, au contraire de l'intensification, les processus de recherche permettant d'examiner des régions restées jusqu'alors inexplorées. Les solutions trouvées peuvent être très différentes de celles découvertes depuis le début de la recherche. Il existe de nombreuses méthodes permettant de réaliser cette stratégie, que ce soit en modifiant de façon significative la fonction à optimiser ou même en générant de façon aléatoire des plans entiers de la solution courante.

Comme c'est souvent le cas lorsque deux critères aussi opposés sont présentés, il est important de trouver un bon compromis entre les stratégies d'intensification et de diversification. En effet, une trop grande intensification conduit à un enfermement au sein de certaines régions de l'espace de recherche, alors qu'une trop grande diversification réduit considérablement les chances d'augmenter la valeur des solutions découvertes.

Chapitre 2

Modèle de représentation à base de stéréotypes

2.1 Motivations du modèle

2.1.1 Un modèle général

Le modèle théorique présenté dans ce chapitre a bien sûr pour objectif de permettre la classification de données symboliques lacunaires, mais surtout d'extraire des stéréotypes caractérisant ces données. Il s'inspire de la logique des défauts de R. Reiter [Rei80] et utilise l'information disponible dans les différents clusters afin de compléter les valeurs manquantes des objets partiellement observés. Surtout, il s'agit d'un modèle général basé sur la structure de treillis et pouvant ainsi être adapté à différents formalismes. Nous proposons, dans le cadre de cette thèse, deux adaptations de notre modèle. La première, dans le formalisme attribut-valeur, est la plus complète et donne lieu à une validation expérimentale. La seconde propose des pistes afin de pouvoir utiliser notre modèle dans le formalisme plus riche des graphes conceptuels.

Nous pensons que ce modèle à base de stéréotypes peut constituer un nouvel outil pour manipuler les grandes bases de textes rendues de plus en plus disponibles grâce à internet. Les outils actuels, si l'on s'attache plus particulièrement aux tâches de classification, sont essentiellement basés sur des méthodes statistiques et ne prennent pas assez en compte le sens exprimé dans les textes. L'approche par mots-clefs et l'algorithme TF-IDF [SAB94] constituent une bonne illustration de ces méthodes. Or, de nouveaux travaux voient actuellement le jour [PDP98][XP05][SM05] qui prennent davantage en considération le contenu des textes que leur forme. Il nous paraît important d'élaborer des outils prévus pour manipuler directement des objets logiques traduisant le sens des textes écrits en langage naturel.

L'un des besoins cruciaux de notre modèle est de pouvoir juger du caractère cohérent ou contradictoire de certaines informations. Pour donner un exemple simple, considérons deux articles de journaux : le premier relate un vol à la tire ayant eu lieu dans le métro à Paris, tandis que le second relate le braquage d'une banque à Amiens. Ces deux articles ne pourront

pas être mis dans un même groupe qui prendrait en compte le critère de localité, car l'action ne peut avoir lieu *à la fois* à Paris et à Amiens. On comprend bien que le système utilisant la fréquence des mots ne peut pas, à lui seul, faire la distinction ; le "sac de mots" (*bag of words*) étiquetant le groupe où se trouvent ces deux articles contient les deux mots Paris et Amiens sans que cela ne pose le moindre de problème. Par contre, ces articles peuvent être classés sous la dénomination "vol ayant eu lieu en France" dès lors que l'on prend la peine de considérer la relation géographique entre les lieux Paris, Amiens et France.

Pour conclure sur les motivations générales du modèle à base de stéréotypes, nous pensons que les années à venir vont voir le développement d'une analyse prenant davantage en compte le sens des textes écrits en langage naturel, probablement associée à des analyses statistiques plus traditionnelles. Les données n'auront plus à être extraites à la main à partir des corpus d'articles et pourront être directement présentés en entrée de notre modèle. De cette manière, le processus entier, partant des textes et terminant avec l'extraction des stéréotypes, pourra être entièrement automatisé, ce qui réduira, dans une certaine mesure, la dimension subjective de la traduction manuelle.

Nous développons ces motivations suivant deux points plus spécifiques : le premier propose une réflexion sur la stratégie générale des méthodes de clustering, alors que le second effectue un rapprochement avec les travaux effectués sur l'étude du sens commun.

2.1.2 La validité sous contrainte

Notre travail se propose de traiter le problème de la classification automatique avec une approche d'optimisation combinatoire où les stéréotypes sont recherchés *en même temps* que la catégorisation des données, tout en prenant en compte un ensemble de contraintes sur les solutions que l'on souhaite atteindre. Partant d'une solution initiale, nous utilisons une méthode de recherche locale afin de contruire, de proche en proche, une partition de notre ensemble d'exemples. Pour ce faire, un voisinage est calculé à chaque étape et la meilleure solution est choisie dans ce voisinage. Les contraintes permettent d'éliminer les solutions jugées mauvaises : clusters de trop petite taille, mauvaise cohésion interne des classes, description des clusters insuffisamment distincte. De plus, la recherche locale est améliorée grâce à l'utilisation de la recherche taboue qui permet principalement de s'échapper des optima locaux. Cette approche est reconnue comme étant plus coûteuse en temps de calcul mais donne généralement de meilleurs résultats qualitatifs. Bien que l'utilisation de la recherche taboue pour le clustering ait déjà fait l'objet de recherches ces dernières années, spécialement dans le domaine de la reconnaissance de formes [AS95][SJ00][NW02], nous développons cette approche afin de l'adapter au cas des données symboliques et lacunaires. Cela se traduit par la définition d'une nouvelle relation de subsomption, inspirée de la logique des défauts, et par une réflexion sur la mesure de similarité à employer. De plus, des contraintes spécifiques sont ajoutées, comme celle de cohésion cognitive.

La prise en compte de contraintes dans le domaine du clustering fait déjà l'objet de nombreuses recherches [Gor96][TNLH01][HM01]. Ainsi, K. Wagstaff [WCRS01] propose d'ajou-

ter des connaissances implicites données par des experts afin de guider la recherche. Ces contraintes correspondent souvent au désir de regrouper certaines instances (*must-link constraint*) ou de les séparer (*cannot-link constraint*). On parle alors de contraintes "au niveau des instances". Des contraintes de plus haut niveau sont proposées par F. Rossi et F. Vautrain [RV00] qui parlent de contraintes "au niveau symbolique". D'autres types de contraintes sont bien sûr étudiés, comme celui de la taille minimale des clusters [BBD00].

Sur le plan technique, l'originalité de notre travail consiste ainsi à combiner une approche du clustering par des méthodes d'optimisation avec une approche prenant en compte des contraintes, le tout inscrit dans la problématique spécifique des données symboliques et lacunaires.

2.1.3 Une approche du sens commun

Pour finir, nous pensons que le modèle à base de stéréotypes peut être vu comme un nouvel outil pour étudier le sens commun. Cette problématique n'est pas nouvelle en IA et débute véritablement dans les années 80 avec les travaux de J. McCarthy, V. Lifschitz, R. Reiter [ML79][McC80][Rei80]. Un symposium sur la formalisation logique du raisonnement de sens commun est organisé dans les années 90, événement auquel participaient les personnes que nous venons de citer. E. Davis donne un bon aperçu de ce qui a été fait dans ce domaine en résumant les travaux présentés dans la cinquième édition de ce symposium tenu à New-York en mai 2001 [DM04]. Au regard de cette communauté, nos propres travaux se situent dans une approche résolument tournée vers la construction automatique du sens commun, et non, comme c'est le cas par exemple avec le projet CYC [LG90][Len95], dans une approche constructiviste systématique et encyclopédique du sens commun.

2.2 Représentation à l'aide de stéréotypes

Cette première partie décrit le modèle de représentation des données symboliques et lacunaires à l'aide d'ensembles de stéréotypes. Celui-ci peut être adapté à plusieurs types de formalismes du moment qu'ils sont basés sur la structure de treillis. Nous précisons que les termes observations, instances et exemples sont indifféremment utilisés pour désigner les objets que l'on souhaite classer.

2.2.1 Espace des descriptions

Commençons par considérer un espace de descriptions que l'on notera $\mathcal{D}$. Une description d de $\mathcal{D}$ est un objet représentant de manière logique une personne, un événement, une observation, etc. Elle peut être manipulée, par exemple, sous la forme d'un vecteur ou d'un graphe conceptuel. Nous supposons que $\mathcal{D}$ peut être structuré sous la forme d'un treillis grâce aux deux opérateurs de généralisation/spécialisation $\wedge$ et $\vee$. Nous rappelons la définition formelle de treillis :

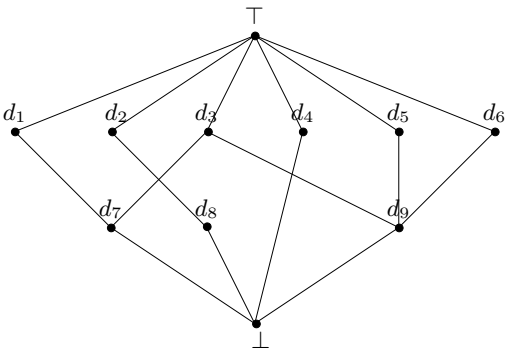


FIG. 2.1 – Un treillis de descriptions.

Définition 2.2.1 *Une algèbre* $\langle \mathcal{D}; \wedge, \vee \rangle$ *est appelée* treillis *si* $\mathcal{D}$ *est un ensemble non vide et si* $\wedge$ *et* $\vee$ *sont deux opérateurs binaires vérifiant les propriétés d'idempotence, de commutativité, d'associativité, ainsi que la loi d'absorption.*

Un ensemble partiellement ordonné (ou *poset* pour *partially ordered set*) peut être déduit de ce treillis en utilisant la relation de subsomption définie comme suit :

Définition 2.2.2 *Etant données deux descriptions* $(d_1, d_2) \in \mathcal{D}^2$, d_1 subsume d_2 *(noté* $d_1 \leq d_2$*) si et seulement si* $d_1 \wedge d_2 = d_1$.

L'expression $d_1 \leq d_2$ signifie que toutes les observations qui vérifient les caractéristiques de la description d_2 (nous dirons qu'elles sont *couvertes* par d_2) sont également couvertes par d_1. En d'autres termes, d_1 est une description, plus générale que d_2, qui peut couvrir en conséquence un nombre plus important d'observations. Si d_1 et d_2 sont deux descriptions de $\mathcal{D}$, $d_1 \wedge d_2$ correspond à la moindre des généralisations communes à d_1 et d_2, c'est-à-dire $\forall d/\ d \leq d_1$ et $d \leq d_2 \Rightarrow d \leq d_1 \wedge d_2$. De la même façon, $d_1 \vee d_2$ correspond à la plus grande des spécialisations communes à d_1 et d_2, c'est-à-dire $\forall d/\ d_1 \leq d$ et $d_2 \leq d \Rightarrow d_1 \vee d_2 \leq d$. $\wedge$ et $\vee$ peuvent également être vues comme étant les bornes respectivement inférieure et supérieure de d_1 et d_2 considérant la relation de subsomption $\leq$.

Les descriptions remarquables $\top$ et $\bot$ appartiennent nécessairement au treillis que nous venons de définir. $\top$ est la description la plus générale qui subsume toutes les autres descriptions, c'est-à-dire $\forall d \in \mathcal{D}, \top \leq d$. Nous utilisons aussi l'expression "description vide" car il s'agit de la description qui apporte le moins d'information. Elle vérifie les propriétés : $\forall d \in \mathcal{D}, d \wedge \top = \top$ et $d \vee \top = d$. D'un autre côté, $\bot$ est la "description absurde" qui est subsumée par toutes les autres descriptions, c'est-à-dire $\forall d \in \mathcal{D}, d \leq \bot$. Elle vérifie les propriétés : $\forall d \in \mathcal{D}, d \vee \bot = \bot$ et $d \wedge \bot = d$.

La figure 2.1 donne un exemple de treillis basé sur les onze descriptions d_1, $d_2 \ldots d_9$, $\top$, $\bot$. Nous pouvons remarquer que $d_1 \vee d_3 = d_7$, mais $d_2 \vee d_4 = \bot$. Si nous utilisons le second opérateur, nous obtenons par exemple $d_1 \wedge d_4 = \top$ et $d_7 \wedge d_9 = d_3$. Ce modèle peut être adapté à tous les formalismes qui permettent le calcul d'une description unique comme constituant le résultat des opérateurs de généralisation et spécialisation. Le formalisme attribut-valeur,

qui est abordé dans la section 3.1, est un exemple d'un tel espace de descriptions dans lequel $\vee$ et $\wedge$ sont les opérateurs ensemblistes $\cup$ et $\cap$. Les graphes conceptuels de J. Sowa [Sow84], du moins une restriction de cette théorie[1], constituent un second exemple d'implémentation possible de notre modèle. Une adaptation à d'autres structures, comme celles utilisées dans le modèle ER (entité-relation) des bases de données, semble parfaitement envisageable.

Avant d'aborder la prochaine partie, nous définissons la relation de complétion entre descriptions afin de faciliter la compréhension de ce qui suit :

Définition 2.2.3 *Etant données deux descriptions* $(d_1, d_2) \in \mathcal{D}^2$*, on dit que* d_2 complète *la description* d_1 *si et seulement si* $d_1 \leq d_2$.

La description d_2 est considérée plus *riche* que d_1 dans le sens où elle apporte des informations plus spécifiques sur les objets qu'elle couvre. Ainsi, les objets vérifiant la description d_2 vérifient également la description plus générale d_1. Une définition plus concrète de cette notion de richesse d'une description est donnée dans le formalisme attribut-valeur.

2.2.2 Subsomption par défaut

A partir du treillis de subsomption, une description d de $\mathcal{D}$ peut être complétée de manière à obtenir une description plus riche. On dit alors que d_1 subsume d_2 *par défaut* si d_2 peut être complétée en une description d_c qui est subsumée (au sens classique du terme) par d_1. Cette opération est bien entendu possible s'il n'existe pas de contradiction entre les informations fournies par d_1 et celles fournies par d_2. En effet, les objets vérifiant la description d_c vérifient également la description plus générale d_2 complétée par d_c. Si d_1 et d_2 sont contradictoires, ces objets ne peuvent pas vérifier la description d_1 et donc en constituer une spécialisation. Voici une première définition de la relation de subsomption par défaut :

Définition 2.2.4 *Pour tout couple* $(d_1, d_2) \in \mathcal{D}^2$, d_1 subsume d_2 par défaut *(noté* $d_1 \leq_D d_2$*) si et seulement si* $\exists\ d_c \in \mathcal{D}$, $d_c \neq \perp$, *telle que* d_c *complète* d_2 *et* $d_1 \leq d_c$.

Remarquons que d_c est un majorant commun à d_1 et d_2 dans le treillis de subsomption tel que nous l'avons défini dans la section précédente. Cela nous amène à introduire la définition suivante dans la théorie des treillis :

Définition 2.2.5 *Pour tout couple* $(d_1, d_2) \in \mathcal{D}^2$, $d_1 \leq_D d_2 \Leftrightarrow d_1 \vee d_2 \neq \perp$.

La relation de subsomption par défaut répond aux trois propriétés suivantes :

Propriété 2.2.1 *La subsomption par défaut est une relation réflexive.*

Preuve 2.2.1 *Trivial en considérant la loi d'absorption de* $\vee$.

Propriété 2.2.2 *La subsomption par défaut est plus générale que la subsomption classique : si* d_1 *subsume* d_2*, alors* d_1 *subsume* d_2 *par défaut. La réciproque n'est pas vraie.*

[1] En effet, il faut considérer qu'il n'existe qu'une seule spécialisation et qu'une seule généralisation pour tout couple de graphes conceptuels.

Preuve 2.2.2 *Comme $d_1 \leq d_2$, alors on pose $d_c = d_2$. Puisque $d_2 = d_c \leq d_c$ et $d_1 \leq d_c = d_2$, on obtient (définition 2.2.4) $d_1 \leq_D d_2$.*

Propriété 2.2.3 *La subsomption par défaut est une relation symétrique : $\forall (d_1, d_2) \in \mathcal{D}^2$, si $d_1 \leq_D d_2$ alors $d_2 \leq_D d_1$.*

Preuve 2.2.3 *Comme $d_1 \leq_D d_2$ alors $d_1 \vee d_2 \neq \perp$ (définition 2.2.5). L'opérateur $\vee$ étant commutatif, on a aussi $d_2 \vee d_1 \neq \perp$. Cela nous permet de conclure que $d_2 \leq_D d_1$.*

Cette dernière propriété peut paraître étrange aux personnes habituées à la relation de subsomption classique. Ainsi, elle ne définit par de relation d'ordre, même partiel, sur l'espace de descriptions $\mathcal{D}$. On peut se demander alors ce qui a motivé l'utilisation du symbole $\leq_D$ pour désigner une relation symétrique. Nous avons fait ce choix pour refléter l'idée "peut être considéré comme plus général", même s'il s'avère qu'une description subsumant (par défaut) une autre définition peut également être subsumée (par défaut) par celle-ci.

2.2.3 Stéréotypes et comparaison

Etant donné l'ensemble de descriptions $\mathcal{D}$, nous définissons formellement la notion d'exemple :

Définition 2.2.6 *Un* exemple *est un objet logique e défini par une description $\delta(e) \in \mathcal{D}$ et un poids $\rho(e) \in \mathbb{R}_*^+$.*

où δ est la fonction qui associe à l'exemple e sa description dans $\mathcal{D}$ et ρ un nombre entier strictement positif représentant le poids de e. En effet, nous supposons que certains exemples peuvent prendre plus d'importance que d'autres, ce qui se traduit par un poids plus grand. Le poids par défaut d'un exemple est fixé à 1. On calcule le poids d'un ensemble d'exemples E en effectuant la somme sur tous les exemples de E :

$$\rho(E) = \sum_{e \in E} \rho(e) \tag{2.1}$$

Nous utilisons le symbole $\mathcal{E}$ pour représenter l'ensemble de tous les exemples possibles et E pour représenter l'ensemble des exemples que nous cherchons à classer ($E \subset \mathcal{E}$). A présent, nous définissons un stéréotype comme suit :

Définition 2.2.7 *Un* stéréotype *est une description particulière $s \in \mathcal{D}$.*

L'ensemble des stéréotypes est donc le même que l'ensemble des descriptions, c'est-à-dire $\mathcal{D}$. Ainsi, un stéréotype est une description qui n'est pas, à l'origine, assujettie à un ensemble d'exemples en particulier. Par contre, il est susceptible de couvrir une partie des exemples de E et un poids peut lui être associé suivant son contexte d'utilisation. Nous verrons par la suite comment calculer le poids de s.

L'objectif de notre travail est de trouver l'ensemble des stéréotypes qui couvrent les données de la meilleure façon possible. Pour traduire cette idée d'*adéquation* entre un stéréotype

et les exemples qu'il couvre dans le cadre d'une tâche de classification non-supervisée, nous devons être capables de calculer le degré de ressemblance entre deux descriptions de $\mathcal{D}$.

Nous avons fait le choix de nous intéresser aux mesures basées sur la notion de similarité pour plusieurs raisons. La première est liée au fait que nous traitons des données de type symbolique et que les notions de distance, de convexité, de barycentre, et la plupart des propriétés géométriques perdent de leur pertinence. La seconde raison vient du fait que la théorie du prototype utilise la notion de degré de similarité entre le prototype et les objets de la catégorie. La troisième raison est l'aspect dual entre distance et similarité, c'est-à-dire qu'il est toujours possible de calculer une distance entre deux objets à partir de leur score de similarité. C'est pourquoi nous restreignons notre étude au cas des mesures de ressemblance.

2.2.4 Mesures de comparaison choisies

Dans ce cadre encore très général, nous ne pouvons donner une définition précise du type de mesures pouvant être employé pour calculer la similarité entre deux descriptions. Tout du moins est-il possible d'en donner une définition générale :

Définition 2.2.8 *M est une fonction de $\mathcal{D} \times \mathcal{D}$ dans $\mathbb{R}^+$ qui traduit le degré de similarité existant entre deux descriptions d_1 et d_2. Plus d_1 et d_2 se ressemblent, plus le score obtenu à l'aide de M est grand.*

Nous proposons à présent deux propriétés fondamentales qui doivent être respectées par la mesure M :

Propriété 2.2.4 $\forall (d_1, d_2) \in \mathcal{D}^2$, *si* $d_1 \not\leq_D d_2$ *alors* $M(d_1, d_2) = 0$.

Nous fixons ainsi, par convention, la similarité de deux descriptions contradictoires à 0. Cette contrainte peut sembler forte car elle propose d'ignorer les caractéristiques communes des descriptions si celles-ci se contredisent ne serait-ce que sur un point. Elle repose surtout sur le fait que nous travaillons dans un cadre lacunaire au sein duquel chaque unité d'information est importante. Soulignons également que cette mesure de similarité n'est utilisée que pour comparer un stéréotype avec les exemples qu'il couvre, et non pour comparer des exemples entre eux.

Propriété 2.2.5 $\forall d_1 \in \mathcal{D}, \forall d_2 \in \mathcal{D}, M(d_1, d_2) \leq M(d_1, d_1)$.

Cette propriété exige que la maximum de la fonction M pour une description d soit obtenu avec cette même description[2].

Des propriétés citées par Batagelj et Bren dans leur travail sur les mesures de ressemblance[3], nous ne conservons donc que la propriété P2.b. La relation de symétrie P1 est en effet sujet à controverse dans le domaine de la catégorisation. Nous laissons ainsi la possibilité de faire la distinction entre sujet et référent (exemple et stéréotype dans notre travail). Cette

[2] Il serait absurde en effet d'envisager une description d' ressemblant davantage à d que d elle-même !

[3] Au sujet des mesures de ressemblance, voir la section 1.1.3.2 page 23.

distinction est notamment illustrée par les travaux de A. Tversky[4]. Nous relâchons également la contrainte P3 car celle-ci nous semble trop restrictive dans le cadre des données lacunaires. Les expérimentations menées dans le chapitre 5 confirment cette première intuition. M n'est donc pas une mesure de similarité, ni même une mesure de similitude. Nous employons à la place l'expression générique de mesure de comparaison, même si les termes de similarité ou ressemblance pourront apparaître dans ce document. Des mesures plus spécifiques sont proposées dans la partie relative au formalisme attribut-valeur.

2.2.5 Couverture relative des exemples

Considérons à présent un stéréotype $s \in \mathcal{D}$ et un exemple $e \in \mathcal{E}$. Nou établissons que s *couvre par défaut* e si s subsume par défaut la description de e. En d'autres termes :

Définition 2.2.9 *Soient $s \in \mathcal{D}$ et $e \in \mathcal{E}$, s* couvre e par défaut *si et seulement si $s \leq_D \delta(e)$.*

Cette notion de couverture par défaut peut être étendue à un ensemble S de stéréotypes et un ensemble E d'exemples en vérifiant que chaque exemple de E est couvert par au moins un stéréotype de S :

Définition 2.2.10 *Soient $S = \{s_1, s_2 \dots s_n\} \subset \mathcal{D}$ et un ensemble d'exemples $E \subset \mathcal{E}$, S* couvre *E si et seulement si tout élément $e \in E$ est couvert par défaut par* au moins *un élément $s \in S$.*

Dans notre tâche de classification, nous avons besoin de trouver un ensemble de stéréotypes qui couvre la totalité des exemples de E. Pour cela, nous définissons la notion d'ensemble de stéréotypes couvrant :

Définition 2.2.11 *Un ensemble de stéréotypes S est dit* couvrant *s'il vérifie les conditions suivantes :*

1. *S est de la forme $S = \{s_1, s_2 \dots s_n, s_\top\} \subset \mathcal{D}$,*
2. *$\forall s \in S, s \neq \bot$,*
3. *$s_\top$ est le stéréotype par défaut, équivalent à la description $\top$, qui regroupe tous les exemples non couverts par les autres stéréotypes.*

L'ensemble regroupant tous les ensembles couvrant est noté $\mathcal{R}$. La différence avec $\mathcal{P}(\mathcal{D})$ consiste en réalité à prévoir une catégorie qui regroupe les exemples non couverts par les autres stéréotypes. La principale propriété des éléments de $\mathcal{R}$ est la suivante :

Propriété 2.2.6 *Quel que soit un ensemble d'exemples $E \subset \mathcal{E}$, un ensemble de stéréotypes $S \in \mathcal{R}$ couvre nécessairement tous les exemples de E.*

Preuve 2.2.4 *Soit E l'ensemble des exemples considérés et e l'un des exemples de E. On peut affirmer que e est couvert par défaut par au moins une description de S : le stéréoptype*

[4] Le sujet est discuté dans la thèse de M. Rifqi [Rif96].

par défaut $s_\top$, équivalent à la description vide $\top$. En effet, d'après les propriétés de treillis que nous avons rappelées précédemment : $\forall d \in \mathcal{D}\top \leq d$. Grâce à la propriété 2.2.2, on sait que : $s_\top \leq_D d = \delta(e)$. Comme tous les exemples de E peuvent au moins être couverts de cette manière par $s_\top \in S$, on en déduit que S couvre bel et bien E suivant la définition 2.2.10.

Nous notons par S^* l'élément S de $\mathcal{R}$ duquel a été retiré le stéréotype-vide $s_\top$, c'est-à-dire :

$$S^* = \{s \in S / \ s \neq s_\top\}$$

Dans le cadre de cette thèse, nous nous restreignons au cas du *crisp clustering*, c'est-à-dire aux partitions de E. Chaque exemple est alors associé à un seul stéréotype de S à la fois. Bien sûr, nous n'affectons pas arbitrairement les exemples aux stéréotypes de S, mais à celui qui en est le plus proche, c'est-à-dire celui qui maximise la mesure M. En clair, le stéréotype choisi pour couvrir un exemple e est le stéréotype s_i de S dont la description, à la fois, subsume $\delta(e)$ par défaut et maximise $M(\delta(e), s_i)$. S'il n'y a aucun candidat respectant ces deux conditions, l'exemple est associé au stéréotype vide $s_\top$. Nous utilisons le terme de couverture relative pour désigner cette application dont voici la définition formelle :

Définition 2.2.12 *L'exemple $e \in E$ est* couvert *par s* relativement *à l'ensemble de stéréotypes $S = \{s_1, s_2 \dots s_n, s_\top\} \in \mathcal{R}$, ce qui est noté $s = C_S(e)$, si les conditions suivantes sont respectées :*

1. $s \in S$,
2. $s \leq_D \delta(e)$,
3. $\forall s' \in S, s' \neq s, M(\delta(e), s) \geq M(\delta(e), s')$.

En cas d'égalité du score obtenu par M, le stéréotype couvrant est choisi arbitrairement, par exemple en conservant celui de plus petit indice. S'il n'existe pas de stéréotype permettant d'obtenir une similarité supérieure à 0, l'exemple est placé par défaut sous le stéréotype vide $s_\top$. Cette décision, qui peut surprendre, est pourtant celle qui est utilisée quasi-systématiquement dans tous les algorithmes de clustering.

Il est possible d'ajouter une contrainte de couverture stricte pour bannir tout arbitraire de la tâche de classification. Pour cela, il suffit de remplacer la relation $\geq$ par la relation stricte $>$ et de placer les exemples qui obtiennent un même (meilleur) score sous le stéréotype vide. Bien entendu, on s'expose ici à obtenir un nombre d'objets non classés plus important. Cette option "prudente" doit être adoptée suivant les circonstances en répondant à la question suivante : est-il préférable de laisser de côté les exemples ambigus plutôt que de les mal classer ? Nous ne détaillons pas plus avant cette question et utilisons uniquement la fonction C_S dans nos expérimentations.

Notons qu'il s'agit ici d'une formulation différente de celle utilisée usuellement dans le domaine du clustering pour affecter les exemples à leur classe respective (aussi appelée étape d'*allocation*) [Did79]. La figure 2.2 illustre le mécanisme de couverture relative de l'exemple e par s_3. Les scores obtenus avec M sont indiqués en étiquette sur les arêtes reliant le stéréotype aux exemples.

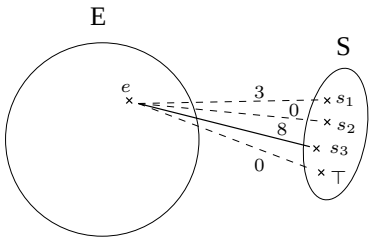


FIG. 2.2 – L'exemple e est couvert par s_3 relativement à S.

Le poids $\rho_{S,E}(s)$ de chacun des stéréotypes s de S est calculé à partir de l'ensemble des exemples E de la manière suivante :

$$\rho_{S,E}(s) = \sum_{e \in E/\ C_S(e)=s} \rho(e) \tag{2.2}$$

Il représente l'importance du stéréotype en terme de couverture. Nous verrons par la suite qu'obtenir des stéréotypes de poids trop faible n'est pas toujours souhaitable. Le poids total de couverture $\rho_E(S)$ des stéréotypes S est calculé en additionnant le poids des stéréotypes, sans prendre bien sûr en compte le stéréotype vide $s_\top$:

$$\rho_E(S) = \sum_{s \in S^*} \rho_{S,E}(s) \tag{2.3}$$

2.2.6 Catégorisation et complétion des exemples

A partir d'un ensemble de stéréotypes $S \in \mathcal{R}$, il est possible de construire une partition de tout ensemble d'exemples E. Pour cela, il suffit de regrouper dans un même cluster tous les exemples couverts par le même stéréotype relativement à S (calculé grâce à la définition 2.2.12). Le stéréotype vide $s_\top$ est associé à une classe $E_\top$, que nous appelons la classe "reste", qui regroupe tous les exemples n'ayant pas trouvé de stéréotype plus adéquat. Nous rappelons que l'ensemble $\{E_1, E_2, ...E_n\}$ constitue une partition de E si et seulement si :

1. $\forall i \in [1,n], E_i \subset E$
2. $\bigcup_{i \in [1,n]} E_i = E$
3. $\forall (i,j) \in [1,n]^2, E_i \cap E_j = \emptyset$

La catégorisation calculée à partir de S prend donc la forme d'une partition. Chaque classe E_i de cette partition regroupe des exemples similaires couverts par un même stéréotype s_i tel que : $E_i = \{e \in E/\ C_S(e) = s_i\}$.

Un stéréotype peut être utilisé afin de compléter la description des exemples qu'il couvre. La description complétée de e relativement à un ensemble S de stéréotypes est calculée comme suit :

$$\delta_S(e) = \delta(e) \vee C_S(e) \tag{2.4}$$

Puisque $C_S(e) \leq_D \delta(e)$, on est bien assuré que $\delta_S(e) \neq \perp$. Il est possible de généraliser cette formule pour calculer un nouvel ensemble d'exemples E' à partir de E et d'un ensemble de

stéréotypes S :

$$E' = \{e' \in \mathcal{E} / \ \delta(e') = \delta_S(e) \text{ pour } e \in E\} \tag{2.5}$$

Dans la partie suivante, nous tâchons d'apporter quelques réflexions sur la nature des ensembles de stéréotypes que nous cherchons à extraire.

2.2.7 Contraintes sur les stéréotypes

Tout d'abord, il apparaît rapidement évident que l'ensemble des solutions potentielles peut être réduit en un ensemble plus restreint $\mathcal{R}' \subset \mathcal{R}$ qui constituera notre espace d'hypothèses. Nous allons proposer dans cette partie un certain nombre de contraintes qui permettent de retirer certaines solutions jugées mauvaises. Le lecteur est averti que cette partie reste encore très théorique et que des illustrations concrètes se trouvent au chapitre suivant.

• **Poids minimum :** La première contrainte, primordiale, exige qu'aucune classe vide ne soit construite à partir d'un ensemble de stéréotypes (exceptée bien sûr la classe $E_\top$ associée à $s_\top$). Cela signifie qu'il doit exister, pour chaque stéréotype, au moins un exemple couvert par celle-ci relativement à l'ensemble S. Il est possible d'étendre cette contrainte à la nécessité de couvrir un poids minimum $\rho_{pds} \in \mathbb{R}^+$ d'exemples de E. Etant donnés S et E, cette contrainte s'écrit alors :

$$\forall s \in S^*, \rho_{S,E}(s) \geq \rho_{pds}. \tag{2.6}$$

• **Non-redondance :** Cette deuxième contrainte, qui ne figurait pas dans le modèle initialement développé, a été ajoutée afin de garantir une parfaite séparation des différentes catégories construites à partir d'un ensemble de stéréotypes. Elle propose d'ignorer les ensembles de stéréotypes dans lesquels l'information fournie est redondante. Nous la détaillons dans la partie relative au formalisme attribut-valeur.

Cette contrainte nous semble tout à fait cohérente avec le cadre des données lacunaires. Elle correspond à notre interprétation du concept de stéréotype[5]. Ainsi, elle permet de construire des ensembles contrastés très discriminants puisqu'un minimum d'information permet de classer très vite un objet (anciennement ou nouvellement observé) sous l'un des stéréotypes. De plus, les stéréotypes sont ainsi beaucoup plus lisibles, faciles à interpréter, que lorsque les traits sont redondants. L'aspect typique prend le pas sur le savoir encyclopédique, au détriment bien sûr de la richesse de l'information délivrée.

• **Cohésion cognitive :** Cette contrainte repose sur la notion de ressemblance de famille initiée par le philosophe L. Wittgenstein puis largement utilisée dans la théorie de la catégorisation[6]. Dans notre travail, nous proposons l'idée d'un stéréotype reflétant les relations qui unissent les objets couverts. Cette nouvelle contrainte est chargée de vérifier la cohésion interne à chaque catégorie en utilisant le stéréotype comme "lien" entre les exemples.

[5] Au sujet du concept de stéréotype, voir la section 1.2.2 page 40.
[6] Pour un aperçu concernant la ressemblance de famille, voir les section 1.2.1.4 et 1.2.1.5.

Pour ce faire, il faut s'assurer qu'il existe toujours un chemin reliant deux informations appartenant au stéréotype. Ces deux informations atomiques peuvent être deux descripteurs dans le formalisme attribut-valeur ou deux concepts dans celui des graphes conceptuels. Le chemin emprunté pour aller de l'une à l'autre consiste en une succession d'exemples qui justifient, de proche en proche, la raison pour laquelle ces deux informations se trouvent ensemble au sein du stéréotype. C'est à cause de cette idée sous-jacente de justification de l'information, nous pourrions presque dire d'explication, que l'épithète "cognitive" a été ajouté. Une définition plus formelle de la contrainte de cohésion cognitive est donnée dans le langage attribut-valeur[7].

[7]La définition de la contrainte de cohésion cognitive dans le formalisme attribut-valeur est donnée dans la section 3.1.5.2 page 75.

Chapitre 3

Implémentation dans deux formalismes

3.1 Le formalisme attribut-valeur

Le principal langage de descriptions que nous avons considéré dans nos travaux est dérivé du formalisme attribut-valeur bien connu en intelligence artificielle. Il s'agit d'une version étendue de la logique des propositions dans laquelle les variables sont appelées des attributs et acceptent chacune un ensemble fini de valeurs. Dans notre travail, un effort a été réalisé afin de traiter le cas des attributs ordinaux, c'est-à-dire acceptant des valeurs sur lesquelles un ordre est fixé. Pour ce faire, nous avons choisi de considérer les attributs nominaux comme un cas particuliers des attributs ordinaux. Précisons cependant que la prise en compte des attributs ordinaux ne fait pas l'objet d'une validation spécifique. Le lecteur intéressé par une présentation plus succinte qui ne considère que des attributs de type nominal pourra consulter [VG05a][VG05b].

3.1.1 Précisions sur le langage de descriptions

Nous nous situons dans un cadre symbolique tel qu'il a été discuté dans la partie 1.1.3. Nous notons par $\mathcal{A}$ l'ensemble des attributs servant à décrire les observations et posons la variable $n_{\mathcal{A}}$ comme le nombre d'éléments de $\mathcal{A}$, c'est-à-dire $card(\mathcal{A})$. Chaque attribut $X \in \mathcal{A}$ peut prendre une ou plusieurs valeur(s) symbolique(s) dans l'ensemble $V(X) = \{x_0, x_1, \dots x_n\}$. Ces valeurs sont ordonnées de la manière suivante : $x_0 \leq x_1 \dots \leq x_n$. Si l'attribut est nominal, cet ordonnancement peut être fixé de manière arbitraire. Nous définissons la notion de descripteur de la manière suivante :

Définition 3.1.1 *Un* descripteur *est l'association d'un attribut $X \in \mathcal{A}$ avec un intervalle de valeurs $[x_i, x_j]$ tel que $(x_i, x_j) \in V(X)^2$ et $x_i \leq x_j$. Il est noté : $(x_i \leq X \leq x_j)$ si $x_i \neq x_j$ et $(X = x_i)$ (voire directement x_i) si $x_i = x_j$.*

Un descripteur associé à un attribut nominal possède deux valeurs aux bornes confondues et est systématiquement noté $(X = x_i)$ ou x_i s'il n'y a pas d'ambiguïté. Un descripteur peut

par exemple représenter la nationalité d'un sujet (*Nationalité = belge*) ou sa taille (*1m72* $\leq$ *Taille* $\leq$ *1m82*). Nous notons par $\mathcal{C}$ l'ensemble de tous les descripteurs possibles et fixons par convention $c_\emptyset$ comme étant le descripteur absurde. Un attribut qui accepte toutes les valeurs possibles de $V(X)$ est dit *indéfini* ou *manquant*. Cela signifie que sa valeur existe bien mais qu'elle n'est pas connue. Ce type de descripteur ($x_0 \leq X \leq x_n$) est noté plus simplement ($X =?$) ou $x_?$, et on le qualifie d'*indéfini* ou *manquant*.

Nous pouvons à présent donner une définition plus précise de la notion de description dans le cadre attribut-valeur :

Définition 3.1.2 *Une* description *d est un ensemble de descripteurs $C = \{c_1, c_2 \dots c_n\} \in \mathcal{C}^n$ dans lequel chaque élément de C est associé à l'un des attributs X de $\mathcal{A}$.*

La seule exception est la description absurde $\perp$ qui possède comme unique descripteur le descripteur absurde $c_\emptyset$. Au contraire, tous les descripteurs de la description vide $\top$ sont indéfinis et donc équivalents à $x_?$ quel que soit l'attribut X considéré.

Cette définition ensembliste remplace celle plus usuelle reposant sur des vecteurs. Ce choix s'explique, comme nous le verrons pas la suite, par le lien plus naturel qui s'effectue avec le modèle théorique basé sur la structure de treillis. Pour revenir justement à notre modèle, rappelons que l'ensemble des descriptions possibles est noté $\mathcal{D}$. Suit un exemple de description comportant trois descripteurs, dont deux sont de type nominal et un de type ordinal :

$$d = \{(Nationalité = belge), (Religion = protestant), (1m72 \leq Taille \leq 1m82)\}$$

Nous rappelons que δ définit une fonction associant à chaque exemple e de la base d'exemples E sa description $\delta(e) \in \mathcal{D}$. Les descriptions étant traitées sous la forme d'ensembles de descripteurs, nous utilisons la relation ensembliste $\in$ pour indiquer si un descripteur appartient ou non à une description.

Pour simplifier la taille des exemples, nous allons à présent considérer un langage artificiel dans lequel les attributs sont désignés par les lettres de l'alphabet en majuscules ($A, B \dots Z$), alors que les valeurs qui leur sont associées sont désignées par la lettre de l'attribut en minuscule suivi d'un indice indiquant son rang dans la liste des valeurs possibles ($a_0, a_1, \dots, a_n$ pour l'attribut A). Ces notations sont utilisées à plusieurs reprises dans la suite de ce mémoire.

Considérant ce langage artificiel, voici la description $\delta(e)$ qui peut être associée à un exemple e :

$$\delta(e) = \{(A = a_0), (B = b_0), (C =?), (D = d_3), (e_0 \leq E \leq e_1)\}$$

si A accepte les valeurs a_0, a_1, a_2 ; B les valeurs b_0, b_1 ; C, c_0, c_1 ; D, d_0, d_1, d_2, d_3 et E est un attribut ordonné tel que $e_0 < e_1 < e_2$. Comme nous l'avons vu tout à l'heure, la description ci-dessus peut s'écrire de manière équivalente en gardant uniquement les valeurs et en ignorant les attributs indéfinis :

$$\delta(e) = \{a_0, b_0, d_3, e_0, \leq E \leq e_1\}$$

Cependant, il faut bien garder à l'esprit que les descripteurs correspondant à des attributs indéfinis font toujours partie de la description, même s'ils n'apparaissent pas dans son écriture pour une question évidente de lisibilité.

Pour retrouver le descripteur associé à un attribut en particulier au sein d'une description, nous définissons la fonction de projection suivante :

Définition 3.1.3 *La* projection *d'une description $d \in \mathcal{D}$ sur un attribut $X \in \mathcal{A}$, notée $d_{|X}$, est définie comme suit :*

$$
\begin{array}{rccl}
_{|X} : & \mathcal{D} & \longrightarrow & \mathcal{C} \\
 & d & \longmapsto & \left\{ \begin{array}{l} c = (x_i \leq X \leq x_j) / \ c \in d \text{ si } d \neq \bot \ , \\ c_\emptyset \text{ sinon.} \end{array} \right.
\end{array}
$$

Un tel descripteur existe nécessairement au vu de la définition 3.1.2. De plus, si un descripteur, différent du descripteur absurde ou du descripteur indéfini, est associé à un attribut X dans une description, alors nous disons que celle-ci *décrit* X :

Définition 3.1.4 *Une description $d \in \mathcal{D}$* décrit *l'attribut $X \in \mathcal{A}$, ce qui est noté $D_X(d)$, si $d_{|X} \neq x_?$ et $d_{|X} \neq c_\emptyset$.*

Nous généralisons cette définition pour n descriptions :

Définition 3.1.5 *Les descriptions $(d_1, \ldots, d_n) \in \mathcal{D}^n$* décrivent *l'attribut $X \in \mathcal{A}$, noté $D_X(d_1, \ldots, d_n)$, si $\forall i, 1 \leq i \leq n, D_X(d_i)$.*

A partir de là, nous définissons un moyen de calculer la richesse de l'information délivrée par une description :

Définition 3.1.6 *La* richesse *d'une description $d \in \mathcal{D}$, notée $|d|$, est le nombre d'attributs décrits par d, c'est-à-dire :*

$$|d| = card(\{X \in \mathcal{A} / \ D_X(d)\}) \tag{3.1}$$

A partir de cette définition, on calcule les scores suivant : $|\{e_1\}| = 1$, $|\{a_0, d_1, f_1 \leq F \leq f_4\}| = 3$ et $|\bot| = |\top| = 0$.

Intéressons-nous à présent au cas des attributs ordinaux. Pour parvenir à les traiter, nous avons besoin d'établir une relation d'ordre partiel (emboîté) $\subseteq$ entre les descripteurs associés à un même attribut :

Définition 3.1.7 *Soit $X \in \mathcal{A}$ un attribut associé à l'ensemble des valeurs $V(X)$ telles que $x_i < x_j$ si $(x_i, x_j) \in V(X)^2$ et $i < j$. Nous établissons alors la relation d'*emboîtement $\subseteq$ *entre les descripteurs : $(x_{i1} \leq X \leq x_{j1}) \subseteq (x_{i2} \leq X \leq x_{j2})$ si et seulement si $x_{i2} \leq x_{i1}$ et $x_{j1} \leq x_{j2}$. De plus, $\forall (x_i, x_j) \in V(X)^2, c_\emptyset \subseteq (x_i \leq X \leq x_j) \subseteq x_?$.*

Nous obtenons, par exemple, que $(a_1 \leq A \leq a_2) \subseteq (a_1 \leq A \leq a_5)$ et $a_0 = (A = a_0) = (a_0 \leq A \leq a_0) \subseteq (a_0 \leq A \leq a_3) = (A =?) = a_?$ si A admet 4 valeurs possibles.

Nous définissons à présent la longueur d'un descripteur :

Définition 3.1.8 *La* longueur *d'un descripteur $c \in \mathcal{C}$, notée $|c|$, est le nombre de valeurs pouvant être prises par l'attribut concerné au sein de ce descripteur, c'est-à-dire si $c = (x_i \leq X \leq x_j)$:*

$$|c| = |(x_i \leq X \leq x_j)| = card(\{v \in V(X)/\ v \geq x_i \text{ et } v \leq x_j\}) \tag{3.2}$$

On fixe par convention $|c_\emptyset| = 0$.

De cette manière, $|a_0| = |(A = a_0)| = 1$; $|(B =?)| = |(b_0 \leq B \leq b_1)| = 2$ si B est un attribut binaire admettant les valeurs b_0 et b_1 ; et $|(c_1 \leq C \leq c_3)| = 3$. Nous avons également besoin de définir deux opérateurs sur les descripteurs : l'intersection $\cap$ et l'union $\cup$.

Définition 3.1.9 *L'*intersection *de deux descripteurs c_1 et c_2 associés au même attribut X admettant les valeurs ordonnées $x_0 < ... < x_n$ est définie comme suit :*

$$\begin{array}{rccl} \cap : & \mathcal{C} \times \mathcal{C} & \longrightarrow & \mathcal{C} \\ & (c_1, c_2) & \longmapsto & \left\{ \begin{array}{l} c_\emptyset \text{ si } \nexists\ v \in V(X)/\ v \in c_1 \text{ et } v \in c_2, \text{ sinon} \\ (x_{k0} \leq X \leq x_{k1})/ \\ \quad \forall v \in V(X)/\ v \geq x_{k0} \text{ et } v \leq x_{k1}, \text{ alors } v \in c_1 \text{ et } v \in c_2. \end{array} \right. \end{array}$$

où $v \in c$ signifie que la valeur v fait partie de l'intervalle de valeurs possibles de l'attribut X définies par le descripteur c. Remarquons que l'on a toujours $v \notin c_\emptyset$.

Par contre, l'union doit être définie différemment suivant qu'il s'agit d'un attribut ordinal ou d'un attribut nominal :

Définition 3.1.10 *L'*union *de deux descripteurs $c_1 = (X = x_1)$ et $c_2 = (X = x_2)$ associés au même attribut* nominal *$X \in \mathcal{A}$ admettant les valeurs $x_0, ...x_n$ est définie comme suit :*

$$\begin{array}{rccl} \cup : & \mathcal{C} \times \mathcal{C} & \longrightarrow & \mathcal{C} \\ & (c_1, c_2) & \longmapsto & \left\{ \begin{array}{l} (X = x_1) \text{ si } x_1 = x_2, \\ x_? \text{ si } x_1 \neq x_2. \end{array} \right. \end{array}$$

Définition 3.1.11 *L'*union *de deux descripteurs $c_1 = (x_{10} \leq X \leq x_{11})$ et $c_2 = (x_{20} \leq X \leq x_{21})$ associés au même attribut* ordinal *$X \in \mathcal{A}$ admettant les valeurs ordonnées $x_0 < ... < x_n$ est définie comme suit :*

$$\begin{array}{rccl} \cup : & \mathcal{C} \times \mathcal{C} & \longrightarrow & \mathcal{C} \\ & (c_1, c_2) & \longmapsto & \left\{ \begin{array}{l} (x_{30} \leq X \leq x_{31})/\ x_{30} = \min(x_{10}, x_{20}) \text{ et } x_{31} = \max(x_{11}, x_{21}) \\ \quad \text{si } c_1 \neq c_\emptyset \text{ et } c_2 \neq c_\emptyset, \\ c_\emptyset \text{ si } c_1 = c_\emptyset \text{ ou si } c_2 = c_\emptyset. \end{array} \right. \end{array}$$

De plus, et ce quelle que soit la nature de l'attribut, on pose : $\forall c \in \mathcal{C}, c \cup c_\emptyset = c_\emptyset \cup c = c$. De cette définition de l'union, il est possible de déduire la propriété suivante sur la relation $\subseteq$:

Propriété 3.1.1 $\forall (c, c') \in \mathcal{C}^2 / c \neq c_\emptyset \text{ et } c' \neq c_\emptyset, c \subseteq c \cup c'$

Preuve 3.1.1 *Si l'attribut considéré X est nominal, alors $c = x_?$ ou $c = (X = x_i)$. Avec $c = x_?$, $c \cup c' = x_? = c$, d'où l'on déduit $c = c \cup c'$ donc $c \subseteq c \cup c'$. Avec $c = (X = x_i)$, $c \cup c'$ peut prendre les valeurs $(X = x_i)$ ou $x_?$. Dans les deux cas, on a bien $(X = x_i) \subseteq c \cup c'$*

Si maintenant l'attribut est ordinal, alors $c = (x_i \leq X \leq x_j)$. *Si* $c' = (x_k \leq X \leq x_l)$ *alors* $c' \cup c = (x_m \leq X \leq x_n)/\ x_m = min(x_i, x_k)$ *et* $x_n = max(x_j, x_l)$. *Puisque* $min(x_i, x_k) \leq x_i$ *et* $max(x_j, x_l) \geq x_j$, *on a bien* $c = (x_i \leq X \leq x_j) \subseteq (x_m \leq X \leq x_n) = c' \cup c$ *(cf. définition 3.1.7).*

Concernant maintenant les descriptions de $\mathcal{D}$, les relations du treillis $\vee$ et $\wedge$ sont traduites dans ce formalisme à l'aide des opérateurs booléens $\cup$ et $\cap$. Nous définissons l'intersection puis l'union de deux descriptions de la manière suivante :

Définition 3.1.12 *L'*intersection *de deux descriptions* d_1 *et* d_2 *de* $\mathcal{D}$ *est définie comme suit :*

$$\begin{array}{rccl} \cap : & \mathcal{D} \times \mathcal{D} & \longrightarrow & \mathcal{D} \\ & (d_1, d_2) & \longmapsto & d/\ \forall X \in \mathcal{A}, d_{|X} = d_{1_{|X}} \cup d_{2_{|X}}. \end{array}$$

Définition 3.1.13 *L'*union *de deux descriptions* d_1 *et* d_2 *de* $\mathcal{D}$ *est définie comme suit :*

$$\begin{array}{rccl} \cup : & \mathcal{D} \times \mathcal{D} & \longrightarrow & \mathcal{D} \\ & (d_1, d_2) & \longmapsto & \left\{ \begin{array}{l} \bot \ \textit{si} \ \exists\ X \in \mathcal{A}/\ d_{1_{|X}} \cap d_{2_{|X}} = c_\emptyset, \ \textit{sinon} \\ d/\ \forall X \in \mathcal{A}, d_{|X} = d_{1_{|X}} \cap d_{2_{|X}}. \end{array} \right. \end{array}$$

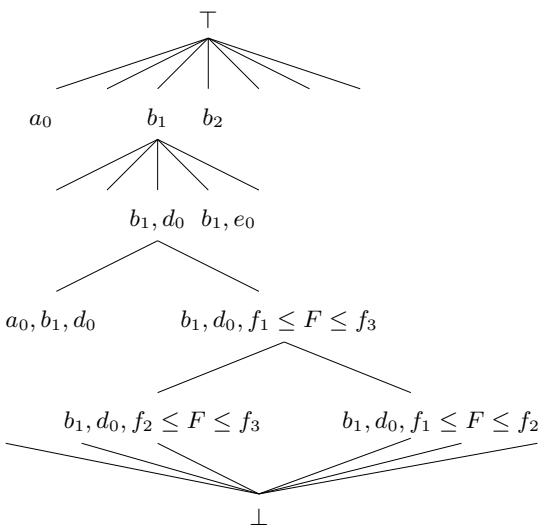
Rappelons que la description vide $\top$ est une description où tous les attributs sont indéfinis, ce qui signifient qu'ils peuvent prendre toutes les valeurs possibles. A l'inverse, la description absurde $\bot$ est celle qui ne contient que le descripteur absurde $c_\emptyset$. Il faut être vigilant sur le fait que l'intersection de deux descriptions amène à construire une description plus générale et qu'en conséquence il faut bien utiliser l'opérateur union $\cup$ dans le cas des attributs ordinaux. Ainsi, $(a_0 \leq A \leq a_3) = (a_0 \leq A \leq a_2) \cup (a_2 \leq A \leq a_3)$ est bien plus général que chacun des descripteurs $(a_0 \leq A \leq a_2)$ et $(a_2 \leq A \leq a_3)$.

3.1.2 Subsomption par défaut

Nous définissons à présent la relation de subsomption $\leq$ entre deux descriptions dans le formalisme attribut-valeur :

Définition 3.1.14 *Soient* d_1 *et* d_2 *deux descriptions de* $\mathcal{D}$, $d_2 \neq \bot$, *alors* d_1 subsume d_2, *noté* $d_1 \leq d_2$ *si* $\forall X \in \mathcal{A}/D_X(d_1)$, *on a* $D_X(d_2)$ *et* $d_{2_{|X}} \subseteq d_{1_{|X}}$. *On fixe par convention :* $\forall d \in \mathcal{D}, d \leq \bot$.

Il s'agit bien d'une relation d'ordre partiel pouvant être représentée sous la forme d'un treillis de descriptions dont nous donnons une illustration ci-dessous :



On vérifie aisément les relations de subsomption suivantes : $\{b_1\} \leq \{a_0, b_1, d_0\}$ et $\{b_1, d_0, (f_1 \leq F \leq f_3)\} \leq \{b_1, d_0, (f_2 \leq F \leq f_3)\}$. Par contre, on obtient également : $\{b_1, e_0\} \not\leq \{a_0, b_1, d_0\}$, $\{b_1, d_0, (f_2 \leq F \leq f_3)\} \not\leq \{b_1, d_0, (f_1 \leq F \leq f_2)\}$ et $\{b_1, d_1\} \not\leq \{b_1, d_0\}$.

Voici la définition de la relation de subsomption par défaut, dérivée de la définition 2.2.5, dans le formalisme attribut-valeur :

Définition 3.1.15 $\forall (d_1, d_2) \in \mathcal{D}^2, d_1 \leq_D d_2 \Leftrightarrow d_1 \cup d_2 \neq \bot$.

Nous établissons maintenant la notion de cohérence entre deux descriptions selon un attribut X, et par voie de conséquence entre deux descriptions de manière générale :

Définition 3.1.16 *Les descriptions* $(d_1, d_2) \in \mathcal{D}^2$ *sont* cohérentes selon l'attribut $X \in \mathcal{A}$, *ce qui est noté* $\tau_X(d_1, d_2)$, *si* $d_{1_{|X}} \cap d_{2_{|X}} \neq c_\emptyset$.

Définition 3.1.17 *Les descriptions* $(d_1, d_2) \in \mathcal{D}^2$ *sont* cohérentes, *ce qui est noté* $\tau(d_1, d_2)$, *si,* $\forall X \in \mathcal{A}, \tau_X(d_1, d_2)$.

Deux descriptions incohérentes exhibent donc au moins un caractère contradictoire. Dans ce cas, il ne peut y avoir de relation de subsomption par défaut entre celles-ci. Cela se traduit par la propriété suivante :

Propriété 3.1.2 $\forall (d_1, d_2) \in \mathcal{D}^2 /\ d_1 \leq_D d_2 \Rightarrow \forall X \in \mathcal{A}, \tau(d_1, d_2)$.

Preuve 3.1.2 *Soient deux descriptions* d_1 *et* d_2 *telles que* $d_1 \leq_D d_2$. *Si l'on suppose que* $\neg\tau(d_1, d_2)$ *alors* $\exists X \in \mathcal{A} / \neg\tau_{d_2}(d_{1_{|X}})$ *(définition 3.1.17), ce qui entraîne :* $\exists X \in \mathcal{A} / d_{1_{|X}} \cap d_{2_{|X}} = c_\emptyset$ *(définition 3.1.16). Or, cela signifie que* $d_1 \cup d_2 = \bot$ *(définition 3.1.13) et* $d_1 \not\leq_D d_2$, *ce qui est absurde au vu de notre hypothèse.*

En utilisant l'exemple précédent, on obtient $\{b_1, e_0\} \leq_D \{a_0, b_1, d_0\}$ car $\{b_1, e_0\} \cup \{a_0, b_1, d_0\} = \{a_0, b_1, d_0, e_0\}$. De la même façon, $\{b_1, d_0, (f_2 \leq F \leq f_3)\} \leq_D \{b_1, d_0, (f_1 \leq F \leq f_2)\}$ car $\{b_1, d_0, (f_2 \leq F \leq f_3)\} \cup \{b_1, d_0, (f_1 \leq F \leq f_2)\} = \{b_1, d_0, f_2\}$. Par contre, $\{b_1, d_1\} \not\leq_D$

$\{b_1, d_0\}$ car $\{b_1, d_1\} \cup \{b_1, d_0\} = \perp$ du fait que $d_0 \cap d_1 = c_\emptyset$. La contradiction entre les descripteurs d_0 et d_1 empêche ces deux descriptions d'être impliquées dans une relation de subsomption par défaut.

3.1.3 Complétion des exemples

Considérons à présent un ensemble E contenant 5 exemples couverts par un stéréotype s[1]. Nous prenons comme langage de descriptions 8 attributs représentés par les 8 premières lettres de l'alphabet. Ainsi, A peut prendre les valeurs a_0 et a_1, B les valeurs b_0, b_1 et b_2. C est un attribut ordinal acceptant les valeurs $c_0 < c_1 < c_2 < c_3$, etc. Voici les six descriptions associées au couple (s, E) :

		A	B	C	D	E	F	G	H
	s :	?	b_2	$(c_0 \leq C \leq c_1)$	?	?	f_0	$(g_1 \leq G \leq g_4)$	h_1
	e_1 :	a_0	?	c_0	?	?	?	$(g_2 \leq G \leq g_4)$	?
	e_2 :	?	b_2	?	d_1	e_2	f_0	?	?
E	e_3 :	?	b_2	?	?	?	f_0	g_3	h_1
	e_4 :	a_1	?	c_1	d_2	e_1	?	?	h_1
	e_5 :	?	b_2	$(c_1 \leq C \leq c_2)$	d_0	?	?	g_4	?

Les descripteurs indéfinis sont simplement notés par le symbole ?. Restreignons à présent les descriptions aux attributs X dont les valeurs sont décrites par s, c'est-à-dire tels que $D_X(s)$:

		B	C	F	G	H
	s :	b_2	$(c_0 \leq C \leq c_1)$	f_0	$(g_1 \leq G \leq g_4)$	h_1
	e_1 :	?	c_0	?	$(g_2 \leq G \leq g_4)$	?
	e_2 :	b_2	?	f_0	?	?
E	e_3 :	b_2	?	f_0	g_3	h_1
	e_4 :	?	c_1	?	?	h_1
	e_5 :	b_2	$(c_1 \leq C \leq c_3)$	?	g_4	?

En utilisant la description de s, nous pouvons à présent compléter la description des exemples couverts par s sur les attributs concernés[2] :

		B	C	F	G	H
	s :	b_2	$(c_0 \leq C \leq c_1)$	f_0	$(g_1 \leq G \leq g_4)$	h_1
	e_1 :	b_2	c_0	f_0	$(g_2 \leq G \leq g_4)$	h_1
	e_2 :	b_2	$(c_0 \leq C \leq c_1)$	f_0	$(g_1 \leq G \leq g_4)$	h_1
E	e_3 :	b_2	$(c_0 \leq C \leq c_1)$	f_0	g_3	h_1
	e_4 :	b_2	c_1	f_0	$(g_1 \leq G \leq g_4)$	h_1
	e_5 :	b_2	c_1	f_0	g_4	h_1

Notons que, dans le cas des attributs ordinaux, la valeur complétée est égale à l'intersection des valeurs prises dans le stéréotype et dans l'exemple traité. Cette complétion de la description des exemples est rendue possible grâce à la relation du subsomption par défaut qui assure la cohérence entre les exemples de E et le stéréotype couvrant s.

[1] Les exemples sont bien sûr couverts relativement à un ensemble S dont s fait partie, mais nous n'avons pas besoin de la totalité de S pour cette illustration.

[2] Voir à propos de la complétion des exemples la section 2.2.6 page 62.

3.1.4 Mesures de comparaison

Les mesures que nous présentons dans le formalisme attribut-valeur respectent les propriétés énoncées dans la section 2.2.4. Elles sont fortement inspirées des mesures classiques utilisées dans le cadre des données catégorielles qui ont été évoquées dans la partie 1.1.3.3. La différence provient du fait que nous travaillons avec des variables pouvant accepter plus de deux valeurs et surtout de l'adaptation au cas des données lacunaires. Là où l'on vérifie habituellement si une variable binaire prend la valeur 1 dans les deux descriptions comparées, nous regardons dans notre cas si les deux descripteurs associés sont identiques (ou très proches dans le cas des attributs ordonnés) ou non. La valeur 0 prend chez nous le sens de la valeur indéfinie $x_?$. Le tableau ci-dessous résume le parallélisme que nous faisons pour un attribut X, selon qu'il s'agisse d'un attribut nominal ou d'un attribut binaire :

Classification de données binaires		Classification par défaut	
$d1_{\|X}$	$d2_{\|X}$	$d1_{\|X}$	$d2_{\|X}$
$(X=1)$	$(X=1)$	x_i	x_i
$(X=1)$	$(X=0)$	x_i	$x_?$
$(X=0)$	$(X=1)$	$x_?$	x_i
$(X=0)$	$(X=0)$	$x_?$	$x_?$

où $dk_{|X}$ correspond à la projection de la description d_k selon l'attribut X.

La première mesure de comparaison proposée est définie de la manière suivante :

$$M_c(d_1,d_2) = \begin{cases} \sum_{X/\ D_X(d_1 \cap d_2)} \frac{|d_{1_{|X}} \cap d_{2_{|X}}|}{|d_{1_{|X}} \cup d_{2_{|X}}|} & \text{si } d_1 \leq_D d_2, \\ 0 & \text{si } d_1 \not\leq_D d_2 \end{cases} \tag{3.3}$$

Remarque 3.1.1 *Si l'attribut considéré X est nominal, alors* $\frac{|d_{1_{|X}} \cap d_{2_{|X}}|}{|d_{1_{|X}} \cup d_{2_{|X}}|} = 1$.

Preuve 3.1.3 *Si l'attribut X est nominal et si $d_1 \leq_D d_2$, alors $D_X(d_1,d_2) \Leftrightarrow d_{1_{|X}} = d_{2_{|X}}$. En effet, si $d_{1_{|X}} = c_1 \neq c_2 = d_{2_{|X}}$ alors $c_1 = (X = x_1) \neq (X = x_2) = c_2$, d'où $c_1 \cap c_2 = c_\emptyset$ qui implique $d_1 \cup d_2 = \perp$ et donc $d_1 \not\leq_D d_2$ (contradiction). Puisque $d_{1_{|X}} = d_{2_{|X}}$, $d_{1_{|X}} \cup d_{2_{|X}} = d_{1_{|X}} \cap d_{2_{|X}} = c$ et l'on conclut* $\frac{|d_{1_{|X}} \cap d_{2_{|X}}|}{|d_{1_{|X}} \cup d_{2_{|X}}|} = \frac{c}{c} = 1$.

En d'autres termes, cette mesure de comparaison consiste à compter les descripteurs communs lorsque l'on ne considère que des attributs de type nominal. Une valeur située entre 0 et 1 est utilisée dans le cas des attributs ordinaux. M_c est une mesure de similitude qui vérifie les propriétés précédemment définies 2.2.4 et 2.2.5 :

Preuve 3.1.4 *La propriété 2.2.4 est vérifiée trivialement par la définition de la mesure M_c.*

Preuve 3.1.5 *Soit deux descriptions d_1 et d_2 appartenant à $\mathcal{D}$. Si $d_1 \not\leq_D d_2$, alors $M_c(d_1,d_2) = 0 \leq M_c(d_1,d_1)$ car $M_c(d_1,d_1) \geq 0$. Considérons que $d_1 \leq_D d_2$. Le cas où $d_1 = \perp$ ou*

$d_2 = \perp$ est trivial. Dans le cas général, on a $\forall X \in \mathcal{A}, (d_1 \cap d_1)_{|X} = d_{1_{|X}} \subseteq d_{1_{|X}} \cup d_{2_{|X}}$ (cf. propriété 3.1.1). Ainsi, $D_X(d_1 \cap d_2) \Rightarrow D_X(d_1)$. En effet, si on considère que $d_{1_{|X}} = c_?$ alors on tombe sur la contradiction $(d_1 \cap d_2)_{|X} = d_{1_{|X}} \cup d_{2_{|X}} = c_?$; et si $d_{1_{|X}} = c_\emptyset$ alors $d_1 = \perp \Rightarrow d_1 \cup d_2 = \perp \Rightarrow d_1 \not\leq_D d_2$ (contradiction). Soit l'ensemble H tel que $H = \{X \in \mathcal{A} / D_X(d_1 \cap d_2)\}$. Considérons à présent l'un des $X \in H$. Posons $A(X) = \frac{|d_{1_{|X}} \cap d_{2_{|X}}|}{|d_{1_{|X}} \cup d_{2_{|X}}|}$ et $B(X) = \frac{|d_{1_{|X}} \cap d_{1_{|X}}|}{|d_{1_{|X}} \cup d_{1_{|X}}|}$. On peut être sûr que $A(X) \leq 1$ car $|d_{1_{|X}} \cap d_{2_{|X}}| \leq |d_{1_{|X}} \cup d_{2_{|X}}|$. $B(X) = 1 \Rightarrow \forall X \in H, A(X) \leq B(X)$, puis $\sum_{X \in H} A(X) \leq \sum_{X \in H} B(X)$ et donc $M_c(d_1, d_2) \leq M_c(d_1, d_1)$. La propriété 2.2.5 de maximalité est vérifiée.

Nous proposons à présent quatre mesures M_{N_1}, M_{N_2}, M_{N_3} et M_{N_4} basées sur quatre normalisations différentes de M_c dont les trois premières sont inspirées des mesures sur les données catégorielles données à la section 1.1.3.3. La première mesure M_{N_1} correspond à la mesure de Russel et Rao appliquée au cas des données lacunaires :

$$M_{N_1}(d_1, d_2) = \frac{M_c(d_1, d_2)}{n_{\mathcal{A}}} \tag{3.4}$$

M_{N_1} est une mesure de similitude (elle ne vérifie pas la propriété P_3). Celle-ci est bornée par 1 car le score maximum pouvant être obtenu par la mesure M_c est égale à $n_{\mathcal{A}}$, c'est-à-dire au nombre d'attributs appartenant à $\mathcal{A}$.

Nous définissons de manière similaire la mesure M_{N_2}, correspondant à la mesure de Kendall, en prenant cette fois en compte les attributs non décrits par l'une et l'autre des descriptions. Nous avons besoin ici de différencier explicitement le cas où d_1 subsume par défaut d_2 de celui où il ne la subsume pas.

$$M_{N_2}(d_1, d_2) = \begin{cases} \frac{1}{n_{\mathcal{A}}} \times [M_c(d_1, d_2) + card(\{X \in \mathcal{A} / d_{1_{|X}} = d_{2_{|X}} = x_?\})] & \text{si } d_1 \leq_D d_2 \\ 0 & \text{si } d_1 \not\leq_D d_2 \end{cases} \tag{3.5}$$

La mesure M_{N_2} est une mesure de similarité car $\forall d \in \mathcal{D}, M_{N_2}(d, d) = 1$.

La mesure M_{N_3} correspond à la mesure classique de Jaccard :

$$M_{N_3}(d_1, d_2) = \frac{M_c(d_1, d_2)}{|d_1 \cup d_2|} \tag{3.6}$$

M_{N_3} est une mesure de similarité car $\forall d \in \mathcal{D}, M_{N_3}(d, d) = 1$. En effet, le dénominateur ne prend pas en compte les attributs non décrits dans l'une et l'autre des descriptions (le cas $0/0$ des attributs binaires).

Enfin, voici la définition de la mesure M_{N_4} :

$$M_{N_4}(d_1, d_2) = \frac{M_c(d_1, d_2)}{|d_1|} \tag{3.7}$$

Cette mesure non symétrique se rapproche des mesures de satisfiabilité inspirée par l'approche de Tversky [Tve77]. Nous cherchons avec M_{N_4} à calculer la proportion des attributs décrits par l'exemple de description d_1 qui se retrouve dans le stéréotype d_2 jouant ici le rôle de référent. Ce n'est ni une mesure de similarité, ni même une mesure de ressemblance.

Nous définissons également deux mesures qui ne prennent pas en compte la relation de subsomption par défaut. Ces mesures sont utilisées dans le calcul de certains indices pour l'évaluation du clustering (comme le score de séparation). La première calcule la similarité de deux descriptions :

$$sim(d_1, d_2) = \sum_{X \in \mathcal{A}/\ D_X(d_1 \cap d_2)} \frac{|d_{1_{|X}} \cap d_{2_{|X}}|}{|d_{1_{|X}} \cup d_{2_{|X}}|} \tag{3.8}$$

La seconde calcule la dissimilarité de deux descriptions et se déduit de la précédente :

$$dis(d_1, d_2) = 1 - sim(d_1, d_2) \tag{3.9}$$

Remarquons que ces deux mesures ne sont pas normalisées. En effet, le choix du dénominateur, comme nous l'avons vu précédemment, n'est pas anodin et dépend de l'utilisation de la mesure. C'est pourquoi la normalisation est effectuée suivant leur contexte d'utilisation.

3.1.5 Contraintes sur les stéréotypes

Parmi les contraintes sur les ensembles de stéréotypes discutées dans la section 2.2.7, nous ne revenons que sur celles de non-redondance et de cohésion cognitive. Un algorithme permettant de vérifier cette dernière contrainte est proposé et sa complexité est analysée.

3.1.5.1 Contrainte de non-redondance

Tout d'abord, intéressons-nous à la contrainte interdisant la redondance de descripteurs entre les stéréotypes d'un même ensemble de $\mathcal{R}$. La contrainte de non-redondance peut s'écrire comme suit :

Définition 3.1.18 *Etant donné un ensemble de stéréotypes $S \in \mathcal{R}$, la contrainte de non-redondance est vérifiée si $\forall (s_1, s_2) \in S^2, \forall\ X \in \mathcal{A}/\ D_X(s_1, s_2), s_{1_{|X}} \cap s_{2_{|X}} = c_\emptyset$.*

Autrement dit, il ne doit exister aucun attribut décrit par deux stéréotypes de S par des descripteurs identiques ou qui possèdent une intersection non vide dans le cas d'un attribut ordinal.

Le tableau ci-dessous présente deux ensembles de stéréotypes, le premier S_1 vérifiant la contrainte de non-redondance, le second S_2 ne la respectant pas.

		A	B	C	D	E	F	G	H
	s_1 :	a_0	?	c_0	?	?	$(f_0 \leq F \leq f_1)$	?	h_2
S_1	s_2 :	?	b_2	c_1	d_1	e_2	f_2	g_0	?
	s_3 :	?	b_3	c_2	$(d_2 \leq D \leq d_3)$	e_0	f_3	g_3	h_1
	s_1 :	a_0	?	c_0	$(d_0 \leq D \leq \mathbf{d_1})$	e_0	f_2	g_2	?
S_2	s_2 :	a_1	$\mathbf{b_2}$	?	$\mathbf{d_1}$	e_2	f_0	g_0	$\mathbf{h_1}$
	s_3 :	?	$\mathbf{b_2}$	c_1	?	?	f_1	g_1	$\mathbf{h_1}$

3.1.5.2 Contrainte de cohésion cognitive

Nous proposons à présent une définition de la contrainte de cohésion cognitive adaptée au formalisme attribut-valeur ainsi qu'une illustration inspirée des expérimentations réalisées dans le chapitre 5. Cette contrainte est vérifiée si, étant donnés deux attributs X_1 et X_2 décrits par s, il est toujours possible de trouver une suite d'exemples, choisis parmi les exemples couverts par s, tels que l'on peut passer par corrélations successives de X_1 à X_2. En voici une définition plus formelle :

Définition 3.1.19 *Etant donnés un stéréotype $s \in S$ et un ensemble $E/\ \forall e \in E, C_S(e) = s$, on dit que s et E vérifient la contrainte de* cohésion cognitive *si $\forall (X_1, X_2) \in \mathcal{A}^2/\ D_{X_1}(s)$ et $D_{X_2}(s), \exists\ u_n = e_{i(1)}, e_{i(2)} \ldots e_{i(m)}/\ e_{i(k)} \in E, D_{X_1}(\delta(e_{i(1)})), D_{X_2}(\delta(e_{i(m)}))$ avec $\forall j \in [1, m-1], \exists\ X/\ D_X(s)$ et $D_X(\delta(e_{i(j)}), \delta(e_{i(j+1)}))$.*

Bien sûr, la contrainte se généralise à un ensemble de stéréotypes :

Définition 3.1.20 *Etant donnés un ensemble de stéréotypes couvrant $S \in \mathcal{R}$ et un ensemble d'exemples E, on dit que S vérifie la contrainte de* cohésion cognitive *sur E si et seulement si $\forall s \in S, s$ vérifie la contrainte de cohésion cognitive sur $E' = \{e \in E/\ C_S(e) = s\}$.*

Afin d'illustrer cette nouvelle contrainte inspirée de la notion de ressemblance de famille, nous reprenons l'exemple de l'homme politique de la fin du XIX^{e} siècle. Nous ne manipulons dans un premier temps, par souci de simplicité, que des attributs nominaux. Prenons le cas du stéréotype d'un "opportuniste juif, malhonnête, lié aux franc-maçons et impliqué dans une action de diffamation", qui pourrait être tiré d'un journal extrémiste de l'époque comme La Libre Parole. Dans cet exemple, nous nous restreignons aux attributs décrits par le stéréotype, c'est-à-dire : *Parti*, *Religion*, *Honnêteté*, *Lié-FM* et *Evénement*. Dans ce langange, l'attribut *Parti* accepte les valeurs *radical*, *opportuniste*, *monarchiste*, etc. Un descriptif détaillé du langage de description est donné en annexe page 175.

La figure 3.1 présente deux cas de figure pouvant survenir : un cas où les exemples couverts vérifient la contrainte de cohésion cognitive et un second cas où ils ne la vérifient pas. Nous en détaillons ci-dessous les raisons.

Le stéréotype s vérifie la contrainte avec l'ensemble d'exemples E_1. Ainsi, il est toujours possible de passer par corrélation de l'un des attributs *Parti*, *Religion*, *Honnêteté*, *Lié-FM*, *Evénement* définis par les descripteurs *(Parti = opportuniste)*, *(Religion=juif)*, *(Honnêteté = non)*, *(Lié-FM = oui)*, *(Evénement = diffamation)* à un autre attribut décrit par s. Par exemple, pour passer de l'attribut *Evénement* à l'attribut *Lié-FM*, il suffit de suivre les exemples e_1, e_{42} et e_8, comme illustré par le schéma suivant :

		Parti	Religion	Honnêteté	Lié-FM	Evénement
	s :	opportuniste	juif	non	oui	diffamation
	e_1 :	opportuniste	?	?	?	diffamation
	e_2 :	opportuniste	juif	?	?	?
E_1	e_6 :	?	?	non	?	?
	e_8 :	?	juif	non	oui	?
	e_{42} :	opportuniste	?	non	?	?

		Parti	Religion	Honnêteté	Lié-FM	Evénement
	s :	opportuniste	juif	non	oui	diffamation
	e_1 :	opportuniste	?	?	?	diffamation
	e_2 :	opportuniste	juif	?	?	?
E_1	e_6 :	?	?	non	?	?
	e_8 :	?	juif	non	oui	?
	e_{42} :	opportuniste	?	non	?	?
	e_0 :	opportuniste	?	?	?	?
	e_8 :	?	?	?	oui	?
E_2	e_9 :	opportuniste	juif	?	?	?
	e_{51} :	?	?	non	?	diffamation
	e_{101} :	?	?	non	?	diffamation

FIG. 3.1 – Deux ensembles E_1 et E_2 pouvant être couverts par s.

Ce chemin explique, d'une certaine façon, la présence de ces deux descripteurs appartenant au même stéréotype. Cette "explication" est constituée par une succession d'observations similaires deux à deux.

Par contre, le stéréotype s ne vérifie pas la contrainte avec l'ensemble d'exemples E_2. Ainsi, il n'existe aucun chemin reliant les attributs *Parti* et *Honnêteté*. Si on regarde la description de ces exemples de plus près, on remarque bien que cet ensemble est constitué de trois descriptions bien distinctes : $d_1 = \{$ *(Parti = opportuniste)*, *(Religion = juif)*$\}$, $d_2 = \{$ *(Lié-FM = oui)*$\}$ et $d_3 = \{$ *(Honnêteté = non)*, *(Evénement = diffamation)*$\}$:

	e_0 :	opportuniste	?	?	?	?
	e_8 :	?	?	?	oui	?
E_2	e_9 :	opportuniste	juif	?	?	?
	e_{51} :	?	?	non	?	diffamation
	e_{101} :	?	?	non	?	diffamation

Les attributs ordinaux demandent une attention particulière. Considérons la situation suivante où A et C sont des attributs nominaux et B est un attribut ordinal :

		A	B	C
	s :	a_0	$(b_0 \leq B \leq b_1)$	c_1
E	e_1 :	a_0	b_0	?
	e_2 :	?	b_1	c_1

Contrairement au cas des attributs nominaux, la dichotomie que l'on aurait tendance à exhiber, c'est-à-dire $\{a_0, b_0\}/\{b_1, c_1\}$, n'est pas aussi évidente. Ce sentiment est renforcé par le fait que $\{a_0, b_0\} \cup \{b_1, c_1\} = c_\emptyset$. Cependant, la relation d'ordre sur B implique que les valeurs b_0 et b_1 ne sont pas égales ou contradictoires, mais qu'elles sont *plus ou moins* égales et *plus ou moins* contradictoires. Ainsi, une femme mesurant 1m75 est très similaire à une femme d'1m80 à l'échelle de tous les mammifères, mais elle lui est éloignée si l'on considère uniquement des mannequins devant mesurer entre 1m70 et 1m82. C'est pourquoi la solution choisie consiste à considérer les descripteurs associés à un tel attribut comme similaires et, dans l'exemple précédent, accepter le lien entre les attributs A et C. Nous considérons cette

opération légitime car nous manipulons des exemples couverts par un même stéréotype. De cette manière, la contrainte de cohésion cognitive n'écarte pas ce type de solutions et nous laissons le soin à la fonction d'évaluation de décider si la solution $S_2 = \{\delta(e_1), \delta(e_2), s_\top\}$ est ou n'est pas une meilleure solution que $S_1 = \{s, s_\top\}$.

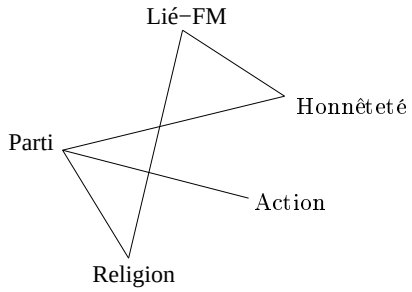
3.1.5.3 Généralisation de la cohésion cognitive

Nous allons aborder maintenant le problème de la cohésion cognitive du point de vue de la théorie des graphes. Considérons le graphe d'adjacence non orienté G formé par l'ensemble des nœuds N et l'ensemble des arêtes A. Soient s un stéréotype et E un ensemble d'exemples couverts par s, N est constitué des attributs décrits par s. L'arête $(n_1, n_2) \in N^2$ fait partie de A s'il existe un nombre suffisant d'exemples dont la description décrit à la fois n_1 et n_2. Cela peut se traduire ainsi :

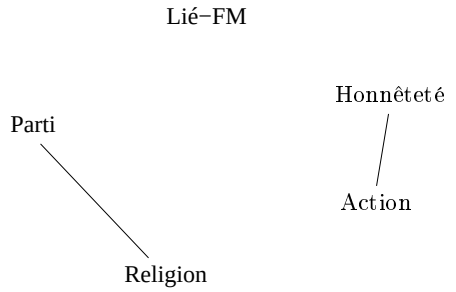
$$A = \{(n_1, n_2) \in N^2 / \sum_{e \in E/\ D_{n_1}(\delta(e)) \text{ et } D_{n_2}(\delta(e))} \rho(e) \geq \rho_{sup}\} \tag{3.10}$$

où $\rho_{sup} \in \mathbb{R}_*^+$ représente le poids minimum des exemples exhibant une corrélation entre les attributs n_1 et n_2. Ces exemples forment le *support* de la corrélation entre les deux attributs. Nous fixons par défaut la valeur de ρ_{sup} à 1. La variation de ce paramètre peut faire l'objet de futures expériences.

Si l'on reprend l'exemple présenté à la figure 3.1, le graphe relatif à s et E_1 a la forme suivante :



A contrario, le graphe relatif à s et E_2 a la forme suivante :



La première remarque que l'on peut faire en observant ces deux graphes, c'est que le premier ne possède qu'une composante connexe alors que le second en possède trois. Nous rappelons quelques définitions de la théorie des graphes concernant la connexité :

Définition 3.1.21 *Etant donné un graphe non orienté G, une* chaîne *est une suite d'arêtes consécutives. La* longueur *d'une chaîne est le nombre d'arêtes qui la composent. Son* poids *est la somme du poids des arêtes.*

Définition 3.1.22 *Etant donné un graphe non orienté G, G est dit* connexe *si pour tout couple de nœuds appartenant à G, il existe une chaîne qui les relie.*

Définition 3.1.23 *Etant donné un graphe non orienté G, le plus grand sous-graphe connexe G' de G contenant un nœud n s'appelle la* composante connexe *de n. Le graphe G s'écrit alors comme une réunion disjointe de ses composantes connexes.*

Si nous prenons le graphe $G = (N, A)$ tel qu'il a été précédemment défini à partir du stéréotype s et de l'ensemble des exemples couverts E, nous pouvons redéfinir la contrainte de cohésion cognitive grâce à la théorie des graphes :

Définition 3.1.24 *Etant donnés un stéréotype $s \in S$, un ensemble $E/\ \forall e \in E, C_S(e) = s$ et le graphe d'adjacence G défini à partir de s et E, on dit que s et E vérifient la contrainte de* cohésion cognitive *si et seulement si G est connexe.*

En faisant varier le poids minimum ρ_{sup}, on exige une plus ou moins forte corrélation entre les attributs. Le cas extrême où l'on fixe $\rho_{sup} = \sum_{e \in E} \rho(e) = \rho(E)$ revient à exiger que tous les attributs soient décrits dans la description de tous les exemples.

3.1.5.4 Implémentation de la cohésion cognitive

La figure 3.2 présente l'algorithme, écrit en pseudo-code, qui vérifie la contrainte de cohésion cognitive dans le formalisme attribut-valeur. A partir de la définition donnée dans la théorie des graphes, on se rend compte qu'il se résume à un parcours (en largeur ou en profondeur) de graphe. Pour traiter le cas d'un ensemble de stéréotypes, il suffit d'appliquer cet algorithme successivement à chacun des stéréotypes de S. Suite à la remarque du paragraphe précédent, nous traitons de la même manière les attributs nominaux et ordinaux.

Pour implémenter ce problème, trois structures sont principalement utilisées. Une liste L d'attributs tout d'abord, gérée comme une file, qui permet de savoir quels attributs décrits par s restent à explorer. C'est pourquoi on a besoin d'une fonction *tete* qui extrait le premier attribut en tête de file et le supprime de la liste, ainsi que d'une fonction + qui concatène un nouvel attribut à la fin de la file. La deuxième structure, sous-jacente, est utilisée afin de connaître, sans avoir à les recalculer à chaque fois, tous les attributs v' voisins de v, c'est-à-dire pour lesquels il se trouve un minimum d'exemples décrivant ces deux attributs. La troisième, enfin, est un tableau *marqueur* qui permet de savoir si un descripteur de s a déjà été visité ou non. Il suffit de tester ce tableau à la fin de l'algorithme pour savoir si tous les attributs décrits par le stéréotype ont bien été visités.

Algorithme de vérification de la cohésion cognitive

entrée : s = un stéréotype de S
entrée : E = l'ensemble des exemples couverts par s relativement à S

sortie : vrai si la contrainte de cohésion est vérifiée, *faux* sinon

Soit L une file initialisée avec un attribut X arbitrairement choisi / $D_X(s)$.
Pour tout $X \in \mathcal{A}$ faire $marqueur(X) \leftarrow$ faux
Tant que (L non vide) faire
{
  $X \leftarrow tete(L)$
  si ($marqueur(X) = faux$) alors
  {
    $marqueur(X) \leftarrow vrai$
    $V(X) \leftarrow \{X' \in \mathcal{A}/\ D_{X'}(s), marqueur(X) = faux$ et
      $\sum_{e \in E/\ D_X(\delta(e)), D_{X'}(\delta(e))} \rho(e) \geq \rho_{sup}\}$.
    Pour tout $X' \in V(X)$ faire $L \leftarrow L + X'$
  }
}
S'il existe $X \in \mathcal{A}/\ D_X(s)$ et $marqueur(X) = faux$ retourner $faux$
Sinon retourner $vrai$

FIG. 3.2 – Vérification de la cohésion cognitive.

3.1.5.5 Complexité en temps de vérification de la cohésion cognitive

La cohésion cognitive est globalement vérifiée si chacun des stéréotypes composant S vérifie localement la contrainte. La complexité est donc en $\mathcal{O}(|S| \times isCoh)$, où $isCoh$ correspond à la complexité associée au calcul de la contrainte pour un stéréotype et l'ensemble des exemples que celui-ci couvre. Pour chaque stéréotype $s \in S$, le graphe des attributs, restreint aux exemples couverts par s et aux attributs décrits par s, est parcouru afin d'en vérifier la connexité[3]. Cela revient donc à effectuer un parcours (en largeur ou en profondeur) de graphe.

Le nombre de nœuds visités est nécessairement borné par le nombre d'attributs décrits par le stéréotype considéré[4], c'est-à-dire $|s|$. A chaque nœud visité, on considère l'ensemble des attributs voisins possibles, c'est-à-dire une nouvelle fois $|s|$. La complexité $isCoh$ est donc évaluée à $\mathcal{O}(|s|^2)$, ce qui nous amène à une complexité totale en :

$$\mathcal{O}(|S| \times M^2) \tag{3.11}$$

où $M = \max_{s \in S}\{|s|\}$ est la richesse maximum des stéréotypes de S.

Le nombre de descripteurs d'un stéréotype s étant borné par le nombre d'attributs $n_{\mathcal{A}}$, nous obtenons la formule suivante :

$$\mathcal{O}(|S| \times n_{\mathcal{A}}^2) \tag{3.12}$$

Nous en aurons besoin pour calculer la complexité globale de l'algorithme de classification par défaut. Cependant, dans certaines applications, comme en fouille de textes où le nombre de descripteurs peut prendre très vite de grandes proportions, il est probable que la valeur estimée $n_{\mathcal{A}}$ s'éloigne considérablement de M. C'est pourquoi il vaut mieux rester prudent, dans la pratique, concernant la prise en considération de ces calculs de complexité.

3.1.6 Classification par défaut

La recherche du meilleur ensemble couvrant les exemples de E est une tâche de classification automatique non supervisée. Comme le modèle général présenté dans le chapitre 2 repose sur la relation de subsomption par défaut, nous avons choisi d'appeler cette technique *classification par défaut*. Entendons par là que cette classification utilise un raisonnement par défaut afin de trouver une bonne représentation des données, évitant toute connotation péjorative pouvant être associée au terme "défaut". Nous avons évoqué précédemment comment calculer une partition à partir d'un ensemble de stéréotypes avec la fonction de couverture relative et une mesure de comparaison. C'est en utilisant d'une part cet ensemble $S \in \mathcal{R}'$ et d'autre part la partition inférée que nous jugeons de la qualité de S. A partir de là, notre objectif est de trouver, parmi tous les ensembles existant, l'ensemble de stéréotypes qui correspond de la meilleure façon possible aux données. Pour ce faire, et devant la taille

[3] Voir la section 3.1.5.3 page 77.

[4] Attention à ne pas confondre ce nombre avec le nombre de stéréotypes $|S|$.

		A	B	C	D	E	F
S	s_1 :	a_0	b_2	c_1	?	e_0	f_2
	s_2 :	a_1	?	c_3	d_0	e_1	?
	$s_\top$:	?	?	?	?	?	?
E_1	e_1 :	a_0	?	?	d_2	?	?
	e_2 :	?	?	c_1	d_1	e_0	f_2
	e_3 :	?	b_2	?	?	e_0	?
	e_4 :	a_0	?	c_1	d_1	?	f_2
E_2	e_5 :	?	b_2	c_3	d_0	e_1	f_2
	e_6 :	?	b_0	?	?	e_1	?
	e_7 :	a_1	?	c_3	?	?	f_0
$E_\top$	e_8 :	a_0	b_2	c_0	?	?	?
	e_9 :	?	b_2	c_3	d_0	e_3	?

FIG. 3.3 – $S = \{s_1, s_2, s_\top\}$ couvre $E = \{e_1, \ldots, e_9\}$.

extrêmement combinatoire de l'espace $\mathcal{R}'$, nous avons choisi d'aborder ce problème comme un problème d'optimisation.

Dans un premier temps, nous revenons sur plusieurs fonctions d'évaluation que nous comptons évaluer avant de les utiliser. Ensuite, nous décrivons brièvement les choix qui ont été faits pour implémenter l'algorithme de recherche local dans le formalisme attribut-valeur. Nous donnons finalement les stratégies employées pour implémenter la méta-heuristique de recherche taboue.

3.1.6.1 Fonctions d'évaluation

Nous proposons dans un premier temps une fonction uniquement basée sur le principe d'homogénéité intra-classe car la séparation entre les stéréotypes est supposée parfaite[5]. La figure 3.3 présente un ensemble de stéréotypes et un ensemble des exemples qu'ils couvrent. Cette illustration est utilisée dans le calcul des fonctions d'évaluation proposées plus loin.

L'homogénéité des exemples à l'intérieur des catégories est basée sur l'adéquation qui existe entre les descriptions des exemples et leur stéréotype couvrant (voir figure 3.4). Ainsi, plus les exemples possèdent des caractéristiques communes avec leur stéréotype, plus ce dernier reflète l'information fournie par les exemples. Ce choix de comparer les exemples avec leur stéréotype a été fait car, d'une part, nous nous intéressons en priorité aux descriptions étiquetant nos catégories, et, d'autre part, nous sommes dans un cadre lacunaire privilégiant la relation de ressemblance de famille. Or, il se peut très bien que deux exemples d'une même catégorie ne partagent que peu (voire pas du tout) de caractéristiques communes, ce qui nous amènerait à leur attribuer un score négligeable. Ce type de comportement n'est pas souhaitable dans le cadre que nous nous sommes fixés. Une réflexion similaire se retrouve dans les travaux de S. Guha et al. [GRS00] avec la notion de liens (*links*) entre les données.

Nous définissons à présent une fonction d'évaluation basée sur les quatre mesures de

[5] Voir à ce sujet la contrainte de non-redondance aux sections 2.2.7 et 3.1.18.

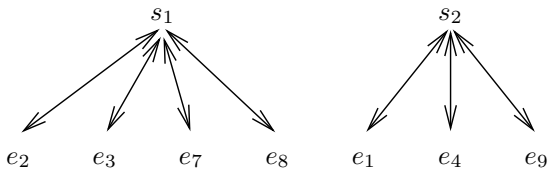


FIG. 3.4 – Calcul de l'homogénéité intra-cluster.

comparaison M_{N_1}, M_{N_2}, M_{N_3} et M_{N_4} :

$$q_N(S,E) = \frac{\sum_{e \in E} (M_N(\delta(e), C_S(e)) \times \rho(e))}{\rho(E)} \tag{3.13}$$

où N peut prendre la valeur N_1, N_2, N_3 ou N_4. Cette fonction générique se décline donc suivant quatre fonctions d'évaluation q_{N_1}, q_{N_2}, q_{N_3} et q_{N_4} prenant une valeur normalisée entre 0 et 1. Si nous utilisons l'exemple présenté dans la figure 3.3, voici les quatre valeurs prises en fonction de la mesure choisie :

$$q_{N_1}(S,E) = \frac{1}{9} \times \frac{(1+3+2+3)+(3+1+2)+(0+0)}{6} = \frac{15}{54} \simeq 0.28$$

$$q_{N_2}(S,E) = \frac{1}{9} \times [(\frac{1}{6}+\frac{3}{6}+\frac{2}{5}+\frac{3}{6})+(\frac{3}{6}+\frac{1}{5}+\frac{2}{5})+(\frac{0}{3}+\frac{0}{4})] \simeq \frac{2.67}{9} \simeq 0.3$$

$$q_{N_3}(S,E) = \frac{1}{9} \times \frac{(1+3+3+3)+(3+2+3)+(3+2)}{6} = \frac{23}{54} \simeq 0.43$$

$$q_{N_4}(S,E) = \frac{1}{9} \times [(\frac{1}{2}+\frac{3}{4}+\frac{2}{2}+\frac{3}{4})+(\frac{3}{5}+\frac{1}{2}+\frac{2}{3})+(\frac{0}{3}+\frac{0}{4})] \simeq \frac{4.77}{9} \simeq 0.53$$

Nous avons donc à notre disposition quatre fonctions différentes permettant de calculer l'adéquation entre les exemples et l'ensemble des stéréotypes couvrant, reflet de l'homogénéité intra-cluster.

Cette première fonction ne prend pas en compte la séparation inter-clusters car elle présuppose la vérification de la contrainte de non-redondance. Imaginons maintenant que cette contrainte ne soit pas vérifiée et que nous souhaitons établir une fonction plus générale afin de refléter la qualité globale d'un ensemble S de stéréotypes. Nous inspirant des travaux réalisés par He et al. [HTTS02], nous proposons la mesure suivante :

$$q_{N,\alpha}(S,E) = \alpha \times q_N(S,E) + (1-\alpha) \times (1 - sep(S)) \tag{3.14}$$

où $sep(S)$ est la mesure de séparation décrite par l'équation 1.18 adaptée à notre formalisme. En fait, $sep(S)$ peut être calculée de deux manières différentes suivant le cœfficient de normalisation choisi pour la mesure de dissimilarité. Ainsi, soit :

$$sep(S) = \frac{1}{|S^*| \times (|S^*|-1)} \sum_{s \in S^*} \sum_{s' \in S^* / \ s' \neq s} \exp - \frac{dis^2(s,s')}{n_{\mathcal{A}} \times 2\sigma^2} \tag{3.15}$$

où $|S^*|$ correspond au nombre de stéréotype de S différents du stéréotype vide $s_\top$, c'est-à-dire $card(S^*)$. Soit :

$$sep(S) = \frac{1}{|S^*| \times (|S^*| - 1)} \sum_{s \in S^*} \sum_{s' \in S^* / \ s' \neq s} \exp - \frac{dis^2(s, s')}{card(\{X \in \mathcal{A} / \ D_X(s) \text{ ou } D_X(s')\}) \times 2\sigma^2} \tag{3.16}$$

Le choix de la fonction calculant la séparation se fait bien entendu de manière cohérente avec la stratégie choisie pour calculer la similarité dans la mesure d'adéquation. Ainsi, l'utilisation des mesures de comparaison M_{N_1} et M_{N_3} implique nécessairement le choix du dénominateur contenant $n_\mathcal{A}$ car la somme est effectuée sur tous les attributs de $\mathcal{A}$. D'un autre côté, M_{N_2} implique le choix du dénominateur contenant $card(\{X \in \mathcal{A} / \ D_X(s) \text{ ou } D_X(s')\})$, c'est-à-dire du nombre d'attributs décrits par l'un ou l'autre des stéréotypes. Quant à M_{N_4}, le choix n'est pas aussi évident et l'on peut donc choisir l'une ou l'autre des fonctions de séparation. Bien entendu, ce choix doit alors être pris en compte dans l'interprétation des résultats obtenus.

3.1.6.2 Choix concernant la recherche

Tout d'abord, nous avons décidé de partir de la solution vide $S_{ini} = \{s_\top\}$ afin de n'imposer aucune contrainte préalable sur les descripteurs. Ce choix, bien sûr, peut être sujet à discussion et il nous semble possible d'améliorer l'efficacité de l'algorithme en partant de "bonnes" solutions initiales. Le calcul de telles solutions, prometteuses *a priori* et pouvant conduire à de bons résultats plus rapidement, peut faire l'objet de futurs développements.

Dans un second temps, nous donnons le principe de construction du voisinage d'une solution courante. Le calcul du voisinage est basé sur l'ajout ou le retrait de descripteurs. Nous illustrons ce passage à l'aide de l'ensemble S_e de stéréotypes suivant :

		A	B	C	D	E
	s_0 :	$(a_0 \leq A \leq a_2)$	?	c_0	$(d_1 \leq D \leq d_2)$	?
S_e	s_1 :	?	b_0	?	?	?
	s_2 :	?	b_1	c_2	d_3	?

Etant donnée une solution courant $S_c = \{s_1, \ldots, s_n, s_\top\}$, les mouvements envisagés sont les suivants :

1. Spécialisation du descripteur associé à l'attribut X de l'un des stéréotypes s_k de S_c. La nouvelle solution potentielle s'écrit alors :

   $$S = \{s_1, \ldots, s_k \cup \{c\}, \ldots, s_n, s_\top\}$$

   à condition que $s_k \cup \{c\} \neq \perp$. Si X est un attribut nominal, cette opération revient à affecter une valeur V_i à X dans s_k.

   L'exemple ci-dessous présente quatre stéréotypes s_a, s_b, s_c et s_d, résultats d'un mouvement possible de spécialisation calculé sur le stéréotype s_0 de S_e (les descripteurs modifiés sont soulignés en gras) :

	A	B	C	D	E
s_0 :	$(a_0 \leq A \leq a_2)$	?	c_0	$(d_1 \leq D \leq d_2)$	?
s_a :	$\mathbf{(a_1 \leq A \leq a_2)}$	?	c_0	$(d_1 \leq D \leq d_2)$	?
s_b :	$(a_0 \leq A \leq a_2)$	$\mathbf{b_1}$	c_0	$(d_1 \leq D \leq d_2)$	?
s_c :	$(a_0 \leq A \leq a_2)$	?	c_0	$\mathbf{d_2}$	?
s_d :	$(a_0 \leq A \leq a_2)$	?	c_0	$(d_1 \leq D \leq d_2)$	$\mathbf{e_2}$

2. Ajout d'un descripteur c comme constituant un nouveau stéréotype de S_c. La nouvelle solution potentielle s'écrit alors :

$$S = \{s_1, \ldots, s_n, \{c\}, s_\top\}$$

Voici un nouvel ensemble S'_e de stéréotypes obtenu à partir de S_e en effectuant ce type de mouvement avec $c = e_1$:

		A	B	C	D	E
	s_0 :	$(a_0 \leq A \leq a_2)$	?	c_0	$(d_1 \leq D \leq d_2)$	?
S'_e	s_1 :	?	b_0	?	?	?
	s_2 :	?	b_1	c_2	d_3	?
	$\mathbf{s_4}$:	?	?	?	?	$\mathbf{e_1}$

3. Généralisation du descripteur associé à l'attribut X de l'un des stéréotypes s_k de S_c. La nouvelle solution potentielle s'écrit alors :

$$S = \{s_1, \ldots, s, \ldots, s_n, s_\top\}$$

telle que $s = [s_k - s_{k|_X}] \cup \{c\}$ avec $s \neq \top$ et $s_{k|_X} \subseteq c$.

L'exemple ci-dessous présente quatre stéréotypes s_a, s_b, s_c et s_d, résultats d'un mouvement possible de généralisation calculé sur le stéréotype s_2 de S_e (les descripteurs modifiés sont soulignés en gras) :

	A	B	C	D	E
s_2 :	?	b_1	c_2	d_3	?
s_a :	?	$\mathbf{?}$	c_2	d_3	?
s_b :	?	b_1	$\mathbf{?}$	d_3	?
s_c :	?	b_1	c_2	$\mathbf{(d_2 \leq D \leq d_3)}$	?
s_d :	?	b_1	c_2	$\mathbf{(d_3 \leq D \leq d_4)}$	?

Notons au passage qu'aucun mouvement de généralisation ne peut être effectué à partir du stéréotype s_4 de S'_e car il mène invariablement à la description vide $\top$.

4. Retrait d'un stéréotype s_k de S_c tel que $|s_k| = 1$. La nouvelle solution potentielle s'écrit alors :

$$S = \{s_1, \ldots, s_{k-1}, s_{k+1}, \ldots, s_n, s_\top\}$$

Cela revient à retirer l'un des stéréotypes de la solution. En réalité, il s'agit d'un cas particulier du mouvement de généralisation.

Voici un nouvel ensemble de stéréotypes S''_e obtenu à partir de S_e en effectuant ce type de mouvement avec le stéréotype s_1 :

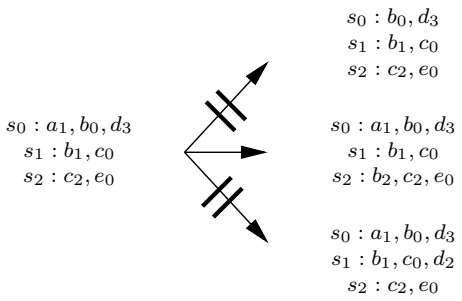


FIG. 3.5 – Exemples de mouvements autorisés et interdits si $T = \{A, D\}$.

S''_e	A	B	C	D	E
s_0 :	$(a_0 \leq A \leq a_2)$	?	c_0	$(d_1 \leq D \leq d_2)$	?
s_2 :	?	b_1	c_2	d_3	?

Le voisinage de S_c, noté $\mathcal{V}(S_c)$, est l'ensemble contenant toutes les solutions potentielles résultant des mouvements effectués à partir de S_c. Plus l'éventail des mouvements possibles est large, plus l'ensemble $\mathcal{V}(S_c)$ est grand. Il est bien sûr possible d'imaginer de nombreux autres types de mouvements. Nous nous sommes cependant limités à cet ensemble de mouvements minimum qui nous garantit l'accès à toutes les solutions potentielles de $\mathcal{R}'$.

Pour finir, nous avons fixé un nombre maximum d'itérations *maxIter* durant lequel la recherche doit se poursuivre. Chaque itération voit une modification de la solution courante à l'aide du type de mouvement que nous venons de définir. Au terme de la recherche, la meilleure solution S_b trouvée jusqu'à présent, c'est-à-dire celle maximisant notre fonction d'évaluation, est donnée comme résultat.

3.1.6.3 Implémentation de la recherche taboue

Nous présentons les choix que nous avons dû faire pour implémenter la recherche taboue. Celle-ci repose principalement sur une liste taboue T ayant une taille constante $k = |T|$.

Le premier choix à faire concerne les éléments de T. Ces éléments sont des attributs de $\mathcal{A}$ qui sont marqués comme tabous s'ils appartiennent à T. Un attribut tabou ne peut être utilisé pour effectuer un mouvement à partir de la solution courante S_c. La figure 3.5 donne quelques exemples de mouvements autorisés et interdits lorsque $T = \{A, D\}$ où $(A, D) \in \mathcal{A}^2$.

Le deuxième choix concerne le critère d'aspiration utilisé. Contrairement à l'algorithme initial de la recherche taboue, nous ne vérifions pas ce critère systématiquement (c'est-à-dire à chaque itération de la recherche) mais à intervalles réguliers. En effet, la liste taboue nous sert non seulement à intégrer une mémoire à court terme afin d'éviter les cycles dans la recherche, mais aussi à réduire la taille du voisinage et ainsi les temps de calcul. Vérifier ce critère, à chaque étape, rendrait caduc ce bénéfice en regard du temps d'exécution de la recherche. Dans notre algorithme, l'aspiration revient à ignorer le statut tabou des attributs toutes les n_{aspi} itérations de la recherche afin de vérifier qu'il n'existe pas de solution meilleure que toutes celles rencontrées jusqu'à présent (et non meilleure que la solution courante).

Enfin, nous avons choisi d'appliquer une heuristique d'intensification à la fin de la recherche. Elle consiste à continuer la recherche et ce, même si le nombre maximum d'itérations $maxIter$ est dépassé, tant que la solution courante obtient un score meilleur en regard de la fonction d'évaluation. Cette opération est effectuée sans prendre en compte la liste taboue. Elle permet de trouver le prochain optimum local qui est susceptible d'améliorer le résultat final.

3.1.6.4 Algorithme de classification par défaut

Dans cette partie, nous décrivons l'algorithme de classification par défaut qui repose sur le modèle que nous venons de définir. Il s'agit d'une version simplifiée qui ne prend pas en compte les attributs ordinaux et n'effectue pas les mécanismes d'intensification (comme le critère d'aspiration) décrits dans la section précédente. L'algorithme est présenté en pseudo-code dans la figure 3.6.

Les paramètres donnés en entrée sont, tout d'abord, les données spécifiques à l'ensemble que l'on souhaite classifier : l'ensemble des exemples E, ainsi que le langage de description défini par l'ensemble des attributs $\mathcal{A}$ et l'ensemble des descripteurs $\mathcal{C}$. Ensuite viennent les deux composantes majeures de la recherche proprement dite : la taille de la liste taboue $t = |T|$ et le nombre maximum d'itérations $maxIter$.

La recherche commence par initialiser la solution avec le descripteur vide $s_\top$. La liste taboue T est vide au départ, ce qui signifie qu'elle ne contient encore aucun attribut. Ensuite vient le corps principal de la recherche, c'est-à-dire une boucle contrôlée par la variable $iter$ et bornée par le nombre maximum d'itérations $maxIter$. Une itération se déroule en trois phases que nous détaillons.

La première phase consiste à calculer un voisinage $\mathcal{V}(S_c)$ à partir de la solution courante S_c. Pour cela, on parcourt tous les attributs qui ne se trouvent pas dans la liste taboue et on utilise les descripteurs associés pour modifier à chaque fois l'un des stéréotypes de S_c. Trois types de mouvements sont alors considérés : spécialiser l'un des stéréotypes en lui affectant un nouveau descripteur ; généraliser l'un des stéréotypes en lui retirant l'un de ses descripteurs (pouvant aboutir à la suppression du stéréotype) ; ajouter un nouveau stéréotype formé uniquement de ce descripteur. La prise en compte d'attributs ordinaux conduit à une définition plus fine des mouvements qui ne pose pas de problème particulier mais alourdit les notations.

La deuxième phase consiste à retirer du voisinage $\mathcal{V}(S_c)$ toutes les solutions qui ne satisfont pas les contraintes que nous avons fixées, autrement dit les contraintes de poids minimum, de cohésion cognitive et de non-redondance. Bien entendu, une implémentation plus optimale aurait pu être réalisée en utilisant directement cette dernière contrainte dans le choix des mouvements à effectuer. Nous avons préféré opter pour une version générale de l'algorithme qui nous permet de prendre en compte le critère de séparation.

La troisième phase consiste à choisir, dans le voisinage $\mathcal{V}(S_c)$ de la solution courante S_c,

Algorithme de classification par défaut

entrée : E = l'ensemble des exemples à classifier
entrée : $\mathcal{A}$ et $\mathcal{C}$ = le langage de descriptions dans le formalisme attribut-valeur
entrée : t = la taille de la liste taboue T
entrée : $maxIter$ = le nombre maximum d'itérations de la recherche

sortie : S_b = le meilleur ensemble de stéréotypes au sens de la fonction d'évaluation q_N

$S_c \leftarrow \{s_\top\}$; initialisation de la solution courante
$S_b \leftarrow S_c$; initialisation de la meilleure solution découverte
$T \leftarrow \emptyset$; la liste des attributs-tabous est initialement vide
pour $iter \leftarrow 1$ à $maxIter$ faire
{
  soit $S_c = \{s_\top, s_1, s_2 \dots s_k\}$
  $\mathcal{V}(S_c) \leftarrow \emptyset$; on initialise le voisinage
  pour tout $A_i \in \mathcal{A} / \ A_i \notin T$ faire
  {
    pour $m \leftarrow 1$ à k faire
    {
      pour tout descripteur $c \leftarrow (A_i = V_{ij})$ faire
      {
        ; on spécialise le stéréotype s_m
        $s \leftarrow s_m \cup \{c\}$
        si $s \neq \bot$ alors $\mathcal{V}(S_c) \leftarrow \mathcal{V}(S_c) \cup \{s_\top, s_1 \dots s_{m-1}, s, s_{m+1} \dots s_k\}$
      }
      ; on généralise le stéréotype s_m
      $s \leftarrow [s_m - \{s_{m|_{A_i}}\}] \cup \{a_{i?}\}$
      si $s \neq \top$ alors $\mathcal{V}(S_c) \leftarrow \mathcal{V}(S_c) \cup \{s_\top, s_1 \dots s_{m-1}, s, s_{m+1} \dots s_k\}$
      sinon $\mathcal{V}(S_c) \leftarrow \mathcal{V}(S_c) \cup \{s_\top, s_1 \dots s_{m-1}, s_{m+1} \dots s_k\}$
    }
    ; on ajoute un nouveau stéréotype possédant un seul descripteur
    pour tout descripteur $c \leftarrow (A_i = V_{ij})$ faire
      $\mathcal{V}(S_c) \leftarrow \mathcal{V}(S_c) \cup \{s_\top, s_1 \dots s_k, \{c\}\}$
    }
  }
  $\mathcal{V}(S_c) \leftarrow \mathcal{V}(S_c) - \{S / S$ ne vérifie pas les contraintes fixées $\}$.
  $S_n \leftarrow \arg\max_{S \in \mathcal{V}(S_c)}(q_N(S))$
  T est mise à jour en fonction de l'attribut A_i qui a permis de passer de S_c à S_n
  $S_c \leftarrow S_n$
  si $q_N(S_c) > q_N(S_b)$ alors $S_b \leftarrow S_c$
  si $q_N(S_c) = q_N(S_b)$ et $|S_c| < |S_b|$ alors $S_b \leftarrow S_c$; rasoir d'Occam
}
retourner la meilleure solution S_b

FIG. 3.6 – Algorithme de classification par défaut

la meilleure solution S_n selon la fonction d'évaluation q_N[6]. Cette solution devient la nouvelle solution courante à partir de laquelle la recherche peut continuer. La meilleure solution découverte S_b, qui sera retournée par l'algorithme, est mise à jour suivant la valeur obtenue par la nouvelle solution S_c.

3.1.6.5 Complexité en temps de l'algorithme de classification par défaut

L'algorithme de classification par défaut repose sur la répétition des trois phases que nous venons de présenter un nombre $maxIter$ de fois. Il nous suffit donc de décrire la complexité de chacune de ces phases pour calculer sa complexité globale :

• **Calcul du voisinage :** Le voisinage est calculé à partir de la solution courante S_c en passant tous les attributs (qui ne sont pas tabous) en revue et en essayant tous les stéréotypes de S_c. Il se réalise donc en :

$$\mathcal{O}(n_{\mathcal{A}} \times |S|) \tag{3.17}$$

Le calcul de la valeur prise par la fonction d'évaluation q_N et la vérification des contraintes sont effectués dans une phase ultérieure.

• **Vérification des contraintes :** A cette phase, on parcourt le voisinage $\mathcal{V}(S_c)$ pour tester si les solutions S vérifient bien les contraintes qui ont été fixées. Du fait de sa construction, le parcours est proportionnel au nombre d'attributs $n_{\mathcal{A}}$. Pour la contrainte de non-redondance, il suffit de parcourir la description des stéréotypes de la solution analysée et le calcul se fait en $\mathcal{O}(|S| \times n_{\mathcal{A}})$. Par contre, la contrainte de cohésion cognitive, comme nous l'avons vu dans la section 3.1.5.5, demande un temps en $\mathcal{O}(|S| \times n_{\mathcal{A}}^2)$. Enfin, la contrainte de poids minimum demande d'avoir affecté les exemples de E au bon stéréotype de S. Il faut avoir comparé chaque exemple avec chaque stéréotype en utilisant la mesure de comparaison M_N. Par conséquent, la complexité de cette dernière contrainte est en $\mathcal{O}(|S| \times n_E \times n_{\mathcal{A}})$. La complexité totale de cette phase est donc en :

$$\mathcal{O}(|S| \times n_E \times n_{\mathcal{A}}^3) \tag{3.18}$$

• **Choix de la meilleure solution :** Dans cette phase, il suffit juste de calculer le score obtenu par la fonction d'évaluation pour chaque solution du voisinage ayant satisfait les contraintes de la phase précédente. Pour ce faire, il faut effectuer la somme des scores de comparaison obtenus entre les stéréotypes et les exemples qu'ils couvrent, puis de diviser par le poids total de E, ce qui est réalisé en $\mathcal{O}(n_E \times n_{\mathcal{A}})$. Le temps peut être réduit à $\mathcal{O}(n_E)$ si l'on sauvegarde les scores obtenus lors de la phase précédente. Prenant cette dernière remarque en compte, voici la complexité totale de cette dernière phase :

$$\mathcal{O}(n_E \times n_{\mathcal{A}}) \tag{3.19}$$

• **Complexité générale de l'algorithme :**

[6] Il suffit de remplacer q_N par la fonction $q_{N,\alpha}$ pour prendre en compte le critère de séparation.

Nous supposons que le nombre de stéréotypes appartenant à la solution courante est faible, hypothèse que nous justifions par l'utilisation des trois contraintes de poids minimum, de cohésion cognitive et de non-redondance. La complexité de l'algorithme de classification par défaut, ici calculée en considérant uniquement des attributs nominaux, est donc en :

$$\mathcal{O}(maxIter \times n_E \times n_{\mathcal{A}}^3) \tag{3.20}$$

Comparativement à l'algorithme des k-modes, nous constatons que nous restons linéaires par rapport au nombre d'itérations et au nombre d'exemples. Par contre, la stratégie de recherche locale, utilisant un voisinage calculé sur le nombre d'attributs, et la contrainte de cohésion cognitive, parcourant un graphe constitué des attributs décrits par le stéréotype, apportent chacune un facteur $n_{\mathcal{A}}$ supplémentaire dans le calcul de la complexité.

3.2 Le formalisme des Graphes Conceptuels

3.2.1 Généralités sur les graphes conceptuels

Les graphes conceptuels constituent un système de représentation des connaissances basé sur les réseaux sémantiques[7] et sur la logique de C.S. Peirce[8]. Ils expriment la connaissance sous une forme logique et graphique, aisément lisible, dont la manipulation peut être automatisée par l'utilisation d'une machine. D'une certaine manière, les graphes conceptuels peuvent être considérés comme un langage intermédiaire entre un formalisme logique computationnel et le langage naturel. Nous ne développons ici qu'une version très simplifiée du formalisme proposé par Sowa en 1984 [Sow84] dont l'hypothèse fondamentale est la suivante :

Hypothèse 3.2.1 *Un* graphe conceptuel *est un graphe fini, connexe et bi-partite.*

- *Les deux types de nœud du graphe bi-partite sont les* concepts *et les* relations conceptuelles.
- *Chaque relation conceptuelle a un ou plusieurs* arc(s)*, chaque arc devant être relié à l'un des concepts du graphe.*
- *Si une relation a n arcs, elle est dite* n-adique *et ses arcs sont numérotés de* 1 *à n. Le terme* monadique *est synonyme de 1-adique,* dyadique *de 2-adique, et* triadique *de 3-adique.*
- *Un concept peut à lui seul former un graphe conceptuel, mais les arcs de toutes les relations conceptuelles doivent être reliés à l'un ou l'autre des concepts du graphe.*

La figure 3.7 présente un exemple de graphe conceptuel. Celui-ci traduit le repas pris par Jules, composé d'un plat principal, lui-même constitué par une viande dont l'accompagnement n'est pas précisé, de vin rouge et terminé par un café. Ce graphe peut s'écrire également sous la forme linéaire suivante :

[7] Voir [Sow] pour une introduction aux réseaux sémantiques.

[8] Un aperçu des principaux apports de Peirce à la logique a été écrit par H. Putnam [Put90]. Une version commentée par J.F. Sowa du texte majeur sur les graphes existentiel MS 514 peut être trouvé à l'adresse suivante : `http ://www.jfsowa.com/peirce/ms514.htm`.

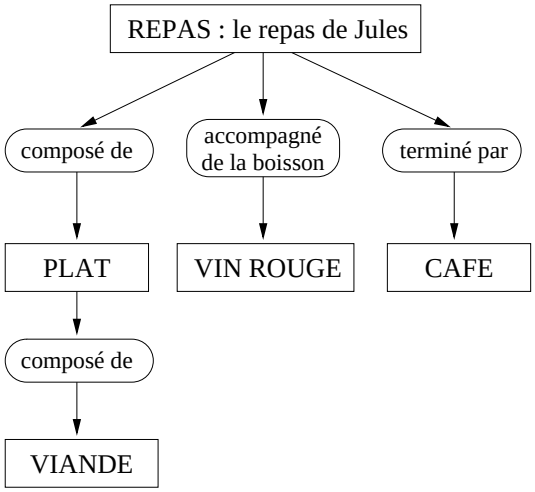


FIG. 3.7 – Exemple de graphe conceptuel.

```
[REPAS : le repas de Jules]-
    (composé de)→[PLAT]-
        (composé de)→[VIANDE]
    (accompagné de la boisson)→[VIN ROUGE]
    (terminé par)→[CAFE].
```

Chaque concept est associé à un *type* de concept (Repas, Plat, Viande, Vin-Rouge, Café) et à un *référent* (le-repas-de-Jules). L'absence de référent signifie que le concept est associé au *marqueur générique* $*$. Ainsi, [Vin Rouge] équivaut à [Vin Rouge : $*$] qui correspond au concept générique du vin rouge, par opposition à [Vin Rouge : #12387] qui fait référence à un vin rouge en particulier auquel on a pu associer l'étiquette #12387. Cette étiquette est un *marqueur individuel* représenté par un unique symbole ou nombre. Si on effectuait une analogie avec la logique classique, le marqueur individuel correspondrait aux constantes alors que le marqueur générique correspondrait aux variables.

Pour finir, rappelons que les graphes conceptuels peuvent être organisés en un treillis de subsomption grâce aux relations duales de généralisation/spécialisation. Dire que g_2 est une *spécialisation* de g_1 (noté $g_2 \leq g_1$) signifie que g_2 est dérivable de g_1 par l'application d'un nombre fini d'opérateurs de spécialisation. Par exemple, l'opérateur de *jointure* consiste à former, à partir de deux graphes conceptuels g_1 et g_2, un nouveau graphe g_3 en fusionnant des concepts et des relations identiques se trouvant dans g_1 et g_2. La figure 3.8 donne une illustration de l'application de l'opérateur de jointure à partir du concept [PLAT]. De la même manière, le graphe g_2 dérivé de tels opérateurs est dit subsumé par le graphe initial g_1 qui est alors une *généralisation* de g_2. Il est possible de calculer la *projection* πg_1 dans g_2 en prenant le sous-graphe de g_2 décrivant les concepts et relations présents dans g_1. Nous avons choisi de ne pas détailler le formalisme, très précis, des graphes conceptuels et enjoignons le lecteur à consulter l'ouvrage de référence écrit par Sowa [Sow84].

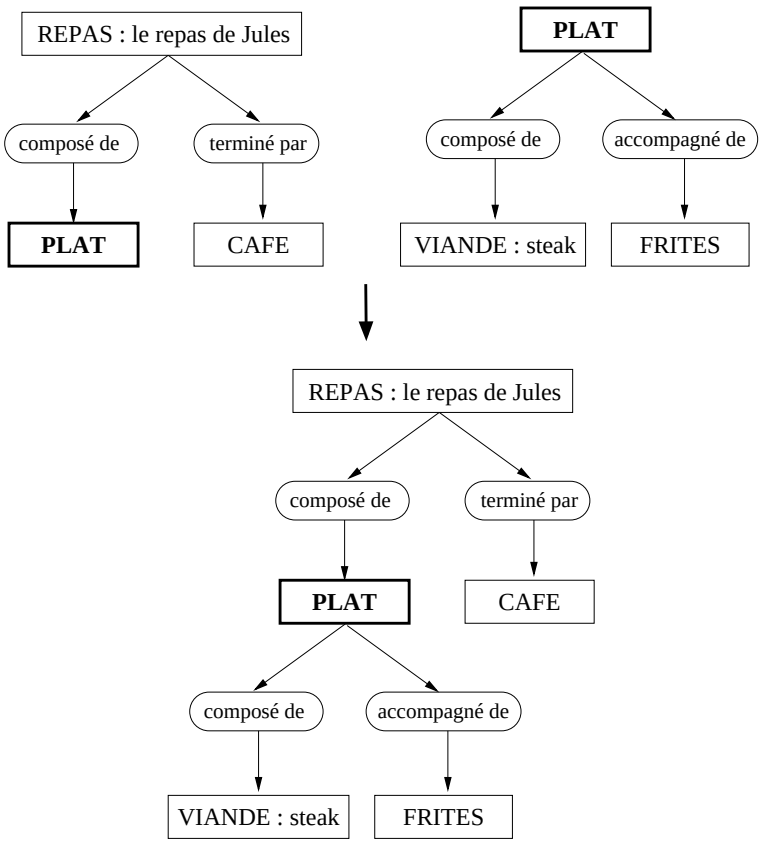


FIG. 3.8 – Exemple de jointure entre deux graphes conceptuels basée sur le concept [PLAT].

Il faut souligner que la notation choisie pour signifier la relation de subsomption est l'inverse de celle qui a été utilisée dans le formalisme attribut-valeur. Cette différence fondamentale est déjà remarquée dans la thèse de C. Faron [Far98] qui, par ailleurs, propose également une présentation plus détaillée du formalisme des graphes conceptuels. Il convient donc de rester vigilant quant aux notations employées.

3.2.2 Lien avec le modèle de représentation à base de stéréotypes

La redéfinition de la complétion dans le formalisme des graphes conceptuels souligne déjà l'importance de ne pas confondre les notations employées :

Définition 3.2.1 *Soient g_1 et g_2 deux graphes conceptuels, alors g_2* complète *g_1 si et seulement si $g_2 \leq g_1$.*

Un stéréotype est un graphe conceptuel spécifique. La définition 2.2.4 de subsomption par défaut se réécrit donc de la manière suivante :

Définition 3.2.2 *Soient g_1 et g_2 deux graphes conceptuels, alors $g_1 \leq_D g_2$ si et seulement s'il existe un graphe g_S tel que g_S complète g_2, $g_S \neq \perp$ et $g_S \leq_D g_1$ (i.e. g_S est une spécialisation de g_1).*

Cette première définition amène, comme dans le cas attribut-valeur, à une définition plus spécifique au formalisme des graphes conceptuels :

Définition 3.2.3 *Soient g_1 et g_2 deux graphes conceptuels, alors $g_1 \leq_D g_2$ si et seulement s'il existe un graphe g_S qui est une spécialisation commune aux graphes g_1 et g_2.*

La figure 3.9 présente le fait "Le repas de Jules est composé d'un steak, accompagné de vin rouge et terminé par un grand café" qui est subsumé par défaut par le stéréotype "un repas composé d'un steak-frites, accompagné par de la baguette et terminé par un café". Le graphe g situé en dessous est le résultat de plusieurs opérations de jointure successives entre s et f : il s'agit donc d'une spécialisation commune à ces deux graphes. Si le stéréotype avait présenté un repas terminé, non par un café, mais par une liqueur, alors cela n'aurait pas correspondu avec le fait et il n'y aurait eu aucune relation de subsomption possible.

Nous souhaitons à présent établir le lien entre la subsomption par défaut et la notion de compatibilité développée par Sowa. Nous devons tout d'abord rappeler la définition présentant la compatibilité [Sow84] :

Définition 3.2.4 *Considérons que les deux graphes conceptuels g_1 et g_2 possèdent une généralisation commune g_c de projections respectives $\pi_1 : g_c \rightarrow g_1$ et $\pi_2 : g_c \rightarrow g_2$. Les deux projections π_1 et π_2 sont dites* compatibles *si, pour chaque concept c de g_c, les conditions suivantes sont vérifies :*

- *$type(\pi_1 c) \cap type(\pi_2 c) > \perp$,*
- *Les référents de $\pi_1 c$ et $\pi_2 c$ sont conformes au type $type(\pi_1 c) \cap type(\pi_2 c)$,*

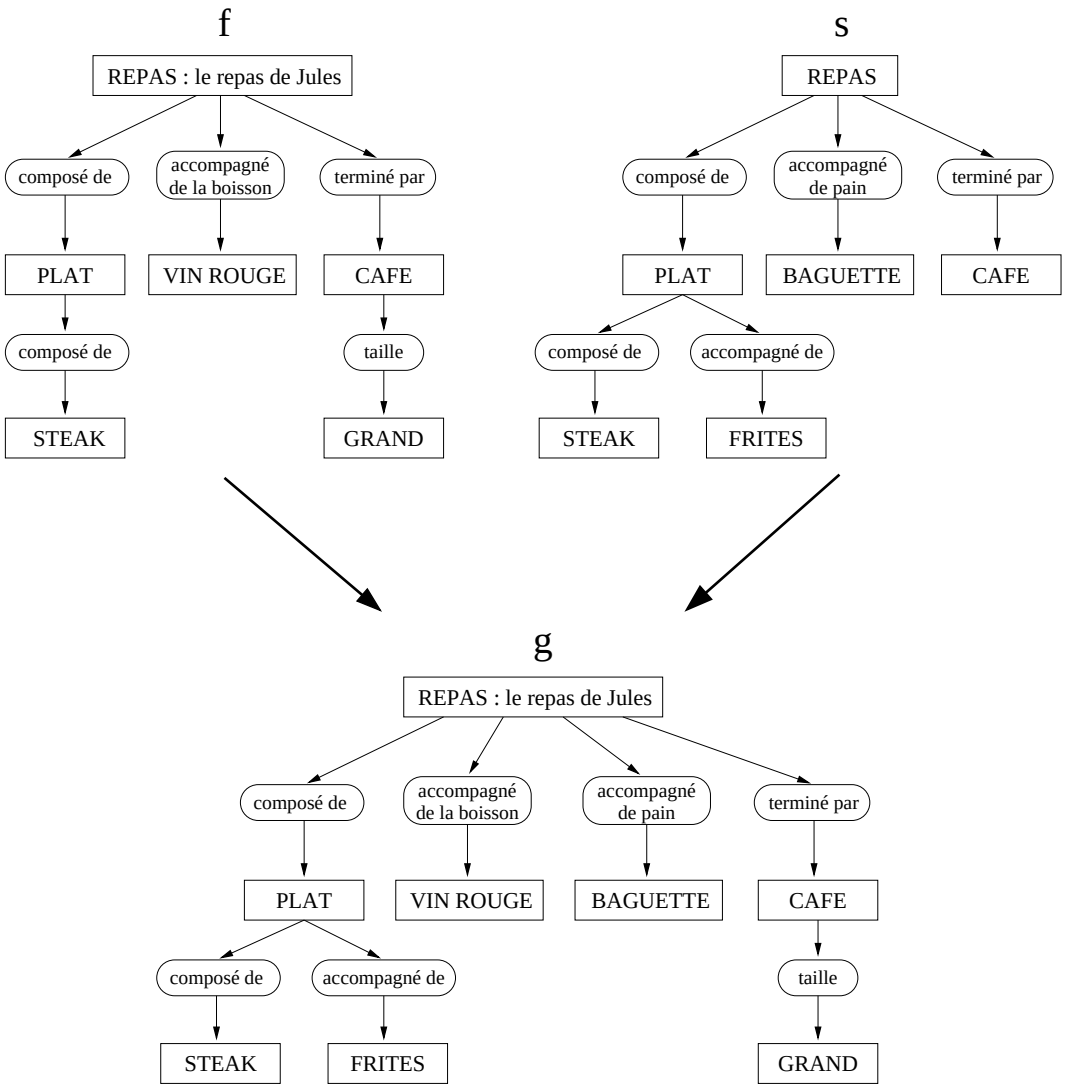


FIG. 3.9 – Le stéréotype s subsume le fait f par défaut. Le graphe g présenté en-dessous est le résultat de plusieurs opérations de jointure entre s et f.

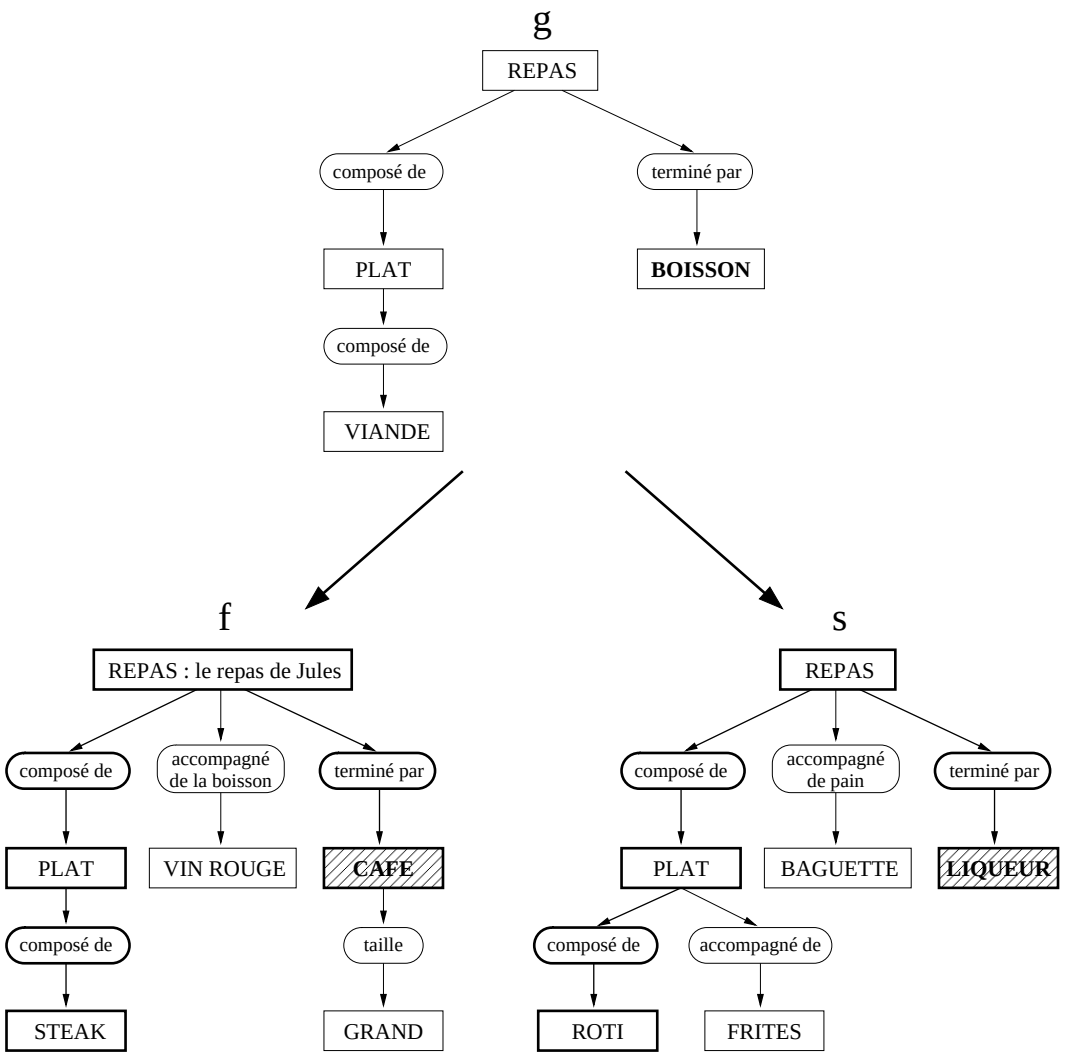


FIG. 3.10 – Le stéréotype s ne subsume pas le fait f par défaut. En effet, les projections (représentées avec des lignes épaisses) ne sont pas compatibles.

- *Si le référent de $\pi_1 c$ est un marqueur individuel i, alors le référent de $\pi_2 c$ est soit i, soit le marqueur générique $*$.*

La figure 3.10 présente deux graphes très similaires à ceux de la figure 3.9 où les deux projections ne sont pas compatibles car $type(CAFE) \cap type(LIQUEUR) = \bot$. Cet exemple nous donne déjà l'intuition que le stéréotype s ne peut pas subsumer le fait f par défaut, ce que nous formalisons grâce au théorème suivant :

Théorème 3.2.1 *Considérons que les deux graphes conceptuels g_1 et g_2 possèdent une généralisation commune g_c de projections respectives $\pi_1 : g_c \rightarrow g_1$ et $\pi_2 : g_c \rightarrow g_2$. Alors π_1 et π_2 sont compatibles si et seulement si g_1 subsume g_2 par défaut.*

Preuve 3.2.1 *Si π_1 et π_2 sont compatibles, alors il existe une spécialisation commune w à g_1 et g_2 (cf. theorème 3.5.7 dans [Sow84]). Suivant la définition 3.2.3, on peut conclure que g_1 subsume g_2 par défaut. Réciproquement, si g_1 subsume g_2 par défaut, il existe une spécialisation w commune aux deux graphes. Supposons que π_1 et π_2 ne soient pas compatibles. Ainsi, il existe au moins un concept c de w tel que $type(\pi_1 c) \cap type(\pi_2 c) = \perp$, ou avec l'un des référents $\pi_1 c$ ou $\pi_2 c$ non conforme à $type(\pi_1 c) \cap type(\pi_2 c)$, ou avec $referent(\pi_1 c) = i$ et $referent(\pi_2 c) = j$ tel que $i \neq j$. Ces trois cas de figure sont absurdes dans le sens où ils contredisent la construction de w. On en conclut que π_1 et π_2 sont nécessairement compatibles.*

Contrairement au formalisme précédent, nous ne proposons pas de mesures de similarité ni d'algorithmes adaptés aux graphes conceptuels. Concernant la similarité entre graphes, il s'agit d'un problème combinatoire beaucoup plus difficile à résoudre qu'un appariement entre vecteurs. J. Zhong et al. [ZZLY02] proposent une méthode utilisant une intéressante heuristique afin de calculer une telle similarité. Le lecteur intéressé par une présentation synthétique de la classification par défaut dans le formalisme des graphes conceptuels pourra consulter [GV04][VG04].

Bien qu'aucune expérimentation n'ait été entreprise utilisant le formalisme des graphes conceptuels, une première étude a été réalisée dans le cadre d'articles de journaux relatifs à l'Affaire Dreyfus[9]. Ainsi, une hiérarchie de types incluant 399 concepts et 174 relations a été construite dans le cadre d'un stage de DEA effectué par A. Remillieux. En plus de cela, un graphe "type" a été proposé afin d'aider à traduire les articles de journaux dans ce formalisme. La figure 3.11 montre un exemple de graphe construit sur ce modèle.

Ce graphe représente un extrait d'article de presse extrait du quotidien L'éclair utilisant la librairie CoGITaNT développée par D. Genest et E. Salvat [GS98]. Il peut être traduit en langage naturel de cette façon : "l'article tiré du quotidien L'éclair affirme explicitement qu'Alfred Dreyfus est coupable parce qu'Esterhazy a été innocenté par la justice". Une fois que tous les articles, ou extraits d'articles, ont été traduits de cette manière, les graphes peuvent devenir l'object d'une classification par défaut afin d'obtenir des stéréotypes d'argumentation concernant cette affaire.

[9] Une présentation de cette période historique est proposée en annexe page 173.

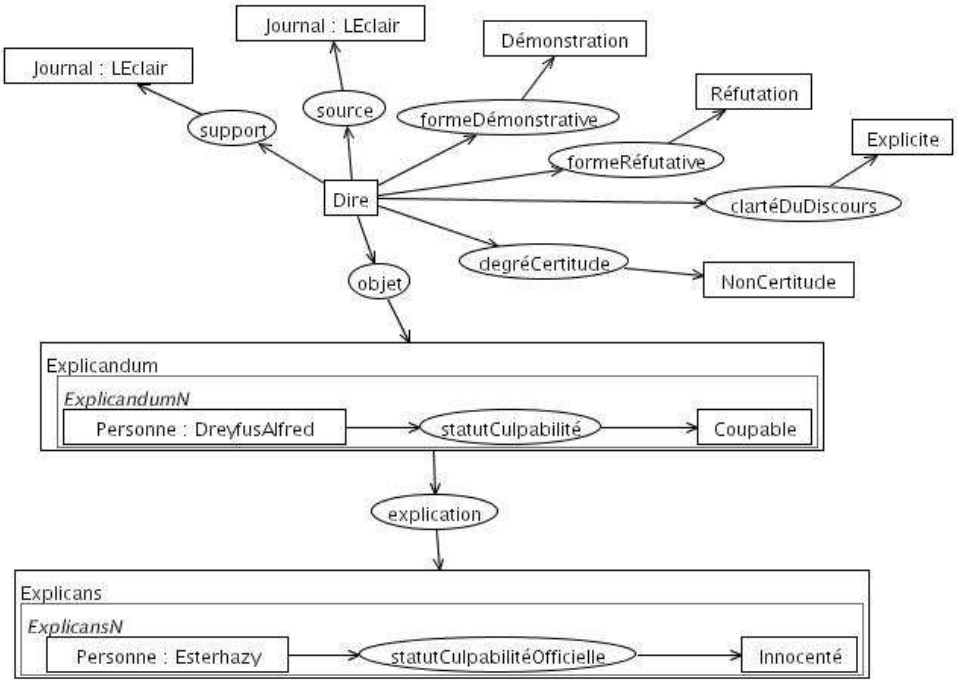


FIG. 3.11 – Un exemple de graphe conceptuel traduisant un extrait d'article relatif à l'Affaire Dreyfus.

Chapitre 4

Méthodologie employée

4.1 Méthodes d'évaluation

Avant d'aborder les expérimentations proprement dites, nous détaillons la méthodologie employée afin de juger de la qualité des partitions obtenues, à la fois grâce à notre algorithme de classification, mais aussi avec les autres algorithmes de clustering utilisés à titre de comparaison. Cette partie est indispensable à une bonne compréhension des expérimentations réalisées dans le chapitre 5.

4.1.1 Généralités

Bien que la validité des clusters (*cluster validity*) soit un thème largement étudié dans la littérature[1], il nous est apparu très vite indispensable d'adapter les indices à notre problématique et surtout d'en proposer de nouveaux. Les indices relatifs à l'homogénéité intra-cluster (*compactness*, adéquation) et à la séparation entre les clusters (*separation*), ainsi que ceux relatifs à l'erreur de classification ayant été déjà abordés[2], nous n'y reviendrons pas dans le détail pour nous intéresser davantage aux nouveaux indices proposés. A ces indices calculés "en sortie" se superposent des paramètres d'entrée liés aux jeux d'essai artificiels, comme le nombre de descriptions initialement générées ou le taux de descripteurs manquants.

Considérons à présent la figure 4.1. Celle-ci donne une illustration de la répartition des indices dans notre problématique de classification automatique. D est l'ensemble des descriptions associées aux clusters construit par l'algorithme de classification[3], c'est-à-dire des stéréotypes pour l'algorithme de classification par défaut ou des centroïdes pour celui des c-moyennes. E est l'ensemble des exemples que nous cherchons à classifier. I est l'ensemble des descriptions initiales dans le cas des jeux de données artificiels.

Dans le cas des jeux d'essai basés sur des données réelles, le processus se résume aux deux objets principaux D et E liés par la relation (A). L'évaluation du clustering est ef-

[1] Au sujet des mesures de validité, voir la section 1.1.4 page 25.

[2] Voir au sujet des mesures de compacité et de séparation les sections 1.1.4 et 3.1.6.1.

[3] La description absurde $\perp$ ne peut donc pas faire partie de D, nous considérons cela comme implicite dans tous les calculs.

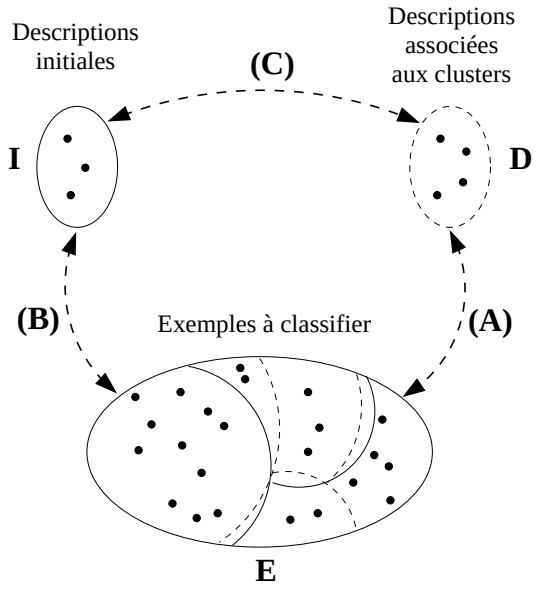


FIG. 4.1 – Schéma de répartition des indices d'évaluation.

fectuée sans connaître de classes a priori, c'est-à-dire en calculant uniquement la qualité de recouvrement des exemples de E par les descriptions de D. Si l'on s'intéresse aux données artificielles, le troisième objet I intègre le schéma accompagné de deux nouvelles relations (B) et (C). Puisque l'on connaît les objets qui ont permis de générer les classes, de nouveaux indices de validation peuvent être pris en compte. Cette distinction est détaillée dans les deux prochaines sections et correspondent aux *cluster distribution* et *class conformation* de He et al. [HTTS02]. Un tableau récapitulant tous les indices donnés dans cette partie est proposé en annexe page 171.

4.1.2 Evaluation sans classes a priori

Nous allons nous intéresser dans cette partie aux indices les plus généraux pouvant être calculés quel que ce soit le type de jeu d'essai employé. Ils peuvent se répartir en trois familles, chacune associée à l'un des items de la figure 4.1.

4.1.2.1 Indices associés aux descriptions D

Cette famille d'indices ne s'intéresse qu'à l'ensemble D des descriptions étiquetant les clusters.

- **Le nombre de descriptions (n_D) :** Il correspond au nombre de descriptions, distinctes de $\top$, qui appartiennent à D. Cet indice n_D donne une première idée sur la complexité de la représentation engendrée. L'interprétation en est difficile parce que l'on ne peut affirmer qu'une représentation à n descriptions soit un meilleur résultat qu'une représentation à $n-1$ ou $n+1$ descriptions. La plupart du temps, comme dans le cas des c-moyennes, ce nombre est fixé avant l'exécution de l'algorithme.

• **La séparation (sep_1 et sep_2) :** L'indice de séparation correspond à la dissimilarité inter-classes et donne une idée du pouvoir de discrimination des descriptions de D. Plus celui-ci est bas, plus les exemples sont rapidement classés et appartiennent clairement à l'une ou l'autre des classes. Nous rappelons que la séparation peut être calculée de deux manières différentes suivant le cœfficient de normalisation choisi pour la mesure de dissimilarité.

Voici les deux équations permettant de calculer la séparation dans le formalisme attribut-valeur :

$$sep_1(D) = \frac{1}{n_D \times (n_D - 1)} \sum_{d \in D} \sum_{d' \in D/\ d' \neq d} \exp - \frac{dis^2(d, d')}{2\ n_{\mathcal{A}}} \tag{4.1}$$

Soit :

$$sep_2(D) = \frac{1}{n_D \times (n_D - 1)} \sum_{d \in D} \sum_{d' \in D/\ d' \neq d} \exp - \frac{dis^2(d, d')}{2\ card(\{X \in \mathcal{A}/\ D_X(d) \text{ ou } D_X(d')\})} \tag{4.2}$$

La constante σ a été fixée à la valeur 1, ce qui ne change rien aux comparaisons effectués avec cette mesure.

Notons que si la contrainte de non-redondance que nous avons préalablement définie est respectée, cela implique nécessairement une séparation parfaite entre les descriptions étiquetant les clusters. Dans le calcul de sep_1, $dis(d, d') = n_{\mathcal{A}}$, et dans celui de sep_2, $dis(d, d') = card(\{X \in \mathcal{A}/\ D_X(d, d')\})$. Le score minimal pouvant être obtenu par sep_1 et sep_2 est de 0.6065. Nous fixons cette valeur par défaut lorsque $n_D = 1$.

4.1.2.2 Indices associés aux données E

Cette famille d'indices s'intéresse uniquement à l'ensemble des données E que l'on cherche à classer. On y trouve :

• **Le nombre d'exemples (n_E) :** Il s'agit du nombre d'exemples que l'on cherche à classer. Cet indice, égal à $card(E)$, est noté plus simplement n_E. Il est imposé dans le cas de jeux de données réels. Dans le cas des données artificielles, il s'agit d'un paramètre que l'on peut faire varier afin d'étudier son influence sur les résultats obtenus. Le poids total des exemples de E est calculé par l'équation suivante (cf. section 2.2.3) :

$$\rho(E) = \sum_{e \in E} \rho(e) \tag{4.3}$$

• **Le nombre total de descripteurs (n_d) :** Cet indice correspond au nombre de descripteurs qui décrivent les exemples e de E en prenant en compte les poids $\rho(e)$. Il est calculé de la manière suivante :

$$n_d(E) = \sum_{e \in E} |\delta(e)| \times \rho(e) \tag{4.4}$$

Dans le cas où tous les attributs sont décrits par tous les exemples, on obtient le cas limite où $n_d(E) = \sum_{e \in E} n_{\mathcal{A}} \times \rho(e) = n_{\mathcal{A}} \times \rho(E)$. L'ensemble E de la figure 4.2 présente un total de 14 descripteurs si l'on considère un poids uniforme sur les exemples. En effet, $n_d(E) =$

		A	B	C	D	E
	e_1 :	?	b_2	$(c_0 \leq C \leq c_4)$	d_2	?
	e_2 :	a_1	?	?	?	e_3
E	e_3 :	a_0	b_2	c_1	d_2	e_0
	e_4 :	?	?	?	d_4	?
	e_5 :	?	b_2	$(c_1 \leq C \leq c_2)$	?	e_3

FIG. 4.2 – Un ensemble E comportant 5 exemples de poids 1.

$|\{b_2, c_0 \leq C \leq c_4, d_2\}| \times 1 + |\{a_1, e_3\} \times 1 + |\{a_0, b_2, c_1, d_2, e_0\}| \times 1 + |\{d_4\}| \times 1 + |\{b_2, c_1 \leq C \leq c_2, e_3\}| \times 1 = 3 + 2 + 5 + 1 + 3 = 14$. Ce score est bien borné par le nombre maximum de descripteurs égal à $n_{\mathcal{A}} \times \rho(E) = 5 \times 5 = 25$.

- **Le taux de valeurs manquantes (m) :** Il s'agit de la proportion de descripteurs manquant au sein de la description des exemples. Il est calculé à partir du nombre de descripteurs n_d :

$$m = 1 - [\frac{n_d(E)}{n_{\mathcal{A}} \times \rho(E)}] \tag{4.5}$$

Comme nous l'avons vu, $n_d(E)$ est nécessairement borné par $n_{\mathcal{A}} \times \rho(E)$. Le score m est donc bien situé entre 0 (tous les attributs sont décrits par tous les exemples) et 1 (tous les exemples possèdent une description vide $\top$). Le taux de valeurs manquantes correspondant à l'exemple de la figure 4.2 est calculé comme suit : $m = 1 - [\frac{14}{5 \times 5}] = 0.44$. Il manque en conséquence 44% des descripteurs au sein de l'ensemble E. Remarquons que, dans le cas des jeux de données artificiels, ce score n'a pas besoin d'être calculé car il est donné en paramètre d'entrée de l'algorithme de génération.

4.1.2.3 Indices associés à la couverture des données (A)

Cette famille d'indices s'intéresse aux relations existant entre l'ensemble des descriptions D et l'ensemble des exemples à couvrir E. Elle permet notamment de traiter l'homogénéité interne aux clusters (cmp_1, cmp_2, $adeq_E$), la richesse prédictive des descriptions de D ($pred$) et les incohérences pouvant survenir entre D et E ($cont_E$, $cont_s$ et $cont_p$).

Avant de décrire ces indices, nous posons la fonction C_D qui étend la fonction de couverture relative[4] aux ensembles quelconques de descriptions obtenues en sortie d'autres algorithmes de clustering. Etant donné un exemple e de E et un ensemble de descriptions D, $C_D(e)$ retourne la description $d \in D$ associée à la classe dans laquelle se trouve e. Si l'exemple e n'a pas de classe associée, $C_D(e)$ est égale à la description vide $\top$ (équivalente au stéréotype vide $s_\top$). Le poids associé à D et à ses éléments peut être calculé de la même façon que dans le cas des ensembles de stéréotypes en remplaçant la fonction C_S par C_D.

- **Le taux de couverture ($couv$) :** Cet indice rend compte de la proportion des exemples

[4] Voir la section 2.2.5 page 60 où la couverture relative est définie pour les ensembles de stéréotypes.

de E qui sont couverts par au moins une description de D. Il se calcule de la manière suivante :

$$couv(D,E) = \frac{1}{\rho(E)} \times \sum_{e \in E / C_D(e) \neq \top} \rho(e) \tag{4.6}$$

Il peut également se calculer à partir du taux de couverture individuel de chaque description de D :

$$couv(D,E) = \frac{1}{\rho(E)} \times \sum_{d \in D} \rho_{D,E}(d) \tag{4.7}$$

Dans le cas des trois algorithmes de clustering considérés, tous les exemples sont nécessairement classés et donc couverts par l'une des descriptions de D. La couverture $couv(D,E)$ est donc égale à $\frac{1}{\rho(E)} \times \sum_{e \in E} \rho(e) = 1$.

Les deux prochains paragraphes regroupent des indices qui donnent une indication sur l'homogénéité des exemples à l'intérieur des clusters. Pour ce faire, nous considérons la similitude existant entre chaque description de D et la description des exemples qu'elle couvre. Une autre approche, comparant entre elles les descriptions des exemples appartenant à un même cluster, n'a pas été retenue car notre objectif est d'extraire des descriptions représentant les classes.

• **La compacité des clusters (cmp_1 et cmp_2) :** Les deux indices cmp_1 et cmp_2 correspondent à une traduction dans le formalisme attribut-valeur de la mesure de compacité[5]. Etant donné un ensemble de descriptions D, il existe en effet deux manières de la calculer suivant la manière dont est opérée la normalisation :

$$cmp_1(D,E) = \frac{1}{\rho(E)} \sum_{d_i \in D} \frac{v_1(E_i,d_i)}{v_1(E,d_G)} \rho(E_i) \tag{4.8}$$

$$cmp_2(D,E) = \frac{1}{\rho(E)} \sum_{d_i \in D} \frac{v_2(E_i,d_i)}{v_2(E,d_G)} \rho(E_i) \tag{4.9}$$

où E_i regroupe les exemples couverts par d_i et $v_k(E,d)$ ($k \in \{1,2\}$) correspond à la variance calculée sur E en utilisant comme vecteur-moyenne d. $v_k(E,d)$ est calculée ainsi :

$$v_1(E,d) = \sqrt{\frac{1}{\rho(E)} \sum_{e \in E} \frac{dis^2(\delta(e),d)}{n_{\mathcal{A}}} \rho(e)} \tag{4.10}$$

$$v_2(E,d) = \sqrt{\frac{1}{\rho(E)} \sum_{e \in E} \frac{dis^2(\delta(e),d)}{card(\{X \in \mathcal{A} / D_X(\delta(e)) \text{ ou } D_X(d)\})} \rho(e)} \tag{4.11}$$

On reconnaît les deux types de normalisation utilisés pour calculer la séparation. Par contre, si D est un ensemble de stéréotypes, nous devons prendre ici en compte le stéréotype vide $s_\top$. Ainsi, le cas contraire nous amène à considérer le cas extrême $D = \{s_\top\}$ comme marquant le meilleur score de compacité, ce qui est absurde.

[5] Au sujet de la mesure de compacité, voir la section 1.1.4.2 page 30.

Il nous reste à définir d_G. Cette description correspond au vecteur-moyenne formé à partir de la totalité de l'ensemble E. Pour ce faire, nous utilisons les descripteurs les plus fréquents pour chaque attribut afin de former d_G tout en interdisant la moindre contradiction entre la description de D et la description des exemples[6]. Ainsi, nous obtenons une description subsumant tous les exemples par défaut et restons dans un contexte comparable à celui à base de stéréotypes.

• **L'adéquation, la contradiction et la perte ($adeq_E$, $cont_E$ et $pert_E$) :** Il s'agit de trois indices complémentaires reflétant l'utilisation des descripteurs de E dans la construction des descriptions de D.

L'indice $adeq_E$ calcule la proportion des descripteurs dont l'attribut X est décrit à la fois par l'exemple e et par sa description associée $d = C_D(e)$, et qui sont *cohérents* avec d[7]. Il donne une idée de la quantité d'information tirée de E qui sert à construire les descriptions de D et se calcule de la sorte :

$$adeq_E(D,E) = \frac{1}{n_d(E)} \times \sum_{e \in E} card(\{X \in \mathcal{A}/\ D_X(\delta(e), C_D(e)) \text{ et } \tau_X(C_D(e), \delta(e))\}) \times \rho(e) \tag{4.12}$$

L'indice $cont_E$ calcule la proportion des descripteurs dont l'attribut X est décrit par l'exemple e et sa description associée $d = C_D(e)$, et qui sont, cette fois, *incohérents* avec d. Il donne une idée du taux de contradiction entre les exemples de E et les descriptions couvrantes de D. Il se calcule comme suit :

$$cont_E(D,E) = \frac{1}{n_d(E)} \times \sum_{e \in E} card(\{X \in \mathcal{A}/\ D_X(\delta(e), C_D(e)) \text{ et } \neg\tau_X(C_D(e), \delta(e))\}) \times \rho(e) \tag{4.13}$$

Enfin, l'indice $perte_E$ calcule la proportion des descripteurs dont l'attribut X est décrit par l'exemple e mais qui n'est pas décrit (donc ni cohérent, ni incohérent) par sa description associée $d = C_D(e)$. Il donne une idée de la quantité d'information non utilisée pour construire les représentants de E. Il se calcule comme suit :

$$perte_E(D,E) = \frac{1}{n_d(E)} \times \sum_{e \in E} card(\{X \in \mathcal{A}/\ D_X(\delta(e)) \text{ et } \neg D_X(C_D(e))\}) \times \rho(e) \tag{4.14}$$

Comme nous l'avons écrit, ces trois indices sont complémentaires et vérifient la propriété suivante :

Propriété 4.1.1 $\forall D, E, adeq_E(D,E) + cont_E(D,E) + perte_E(D,E) = 1$.

Preuve 4.1.1 *La richesse de la description $\delta(e)$ est calculée comme suit (définition 3.1.6) : $|\delta(e)| = card(\{X \in \mathcal{A}/\ D_X(\delta(e))\})$. Etant donnée une description $d \in \mathcal{D}$, elle se décompose ainsi : $|\delta(e)| = card(\{X \in \mathcal{A}/\ D_X(\delta(e), d)$ et $\tau_X(d, \delta(e))\}) + card(\{X \in \mathcal{A}/\ D_X(\delta(e), d)$ et $\neg\tau_X(d, \delta(e))\}) + card(\{X \in \mathcal{A}/\ D_X(\delta(e))$ et $\neg D_X(d)\})$. On en déduit que : $n_d(E) = \sum_{e \in E} |\delta(e)| \times \rho(e) = n_d(E) \times [adeq_E(D,E) + cont_E(D,E) + perte_E(D,E)]$, d'où le résultat.*

[6] Cela correspond à la technique T_2 décrite dans la section 4.1.4 appliquée à l'ensemble des exemples E.

[7] La notion de cohérence est définie dans la section 3.1.2 page 70 où la fonction τ est introduite.

• **La cohérence de couverture ($cont_s$ et $cont_p$) :** Cette partie regroupe deux indices qui affinent le score de contradiction $cont_E$ tel qu'il a été défini dans la partie précédente. Ces indices reflètent la cohérence entre les descripteurs des exemples de E et les descripteurs des descriptions associées dans D. Le tableau ci-dessous présente plusieurs incohérences entre une description couvrante et cinq exemples associés :

	A	B	C	D	E
d :	a_0	$(b_0 \leq B \leq b_1)$	c_0	d_2	e_3
e_1 :	a_0	?	?	$(d_1 \leq D \leq d_2)$	?
e_2 :	$\boldsymbol{a_1}$	b_1	?	?	e_3
e_3 :	a_0	$\boldsymbol{b_2}$	c_0	d_2	?
e_4 :	?	?	c_0	$\boldsymbol{(d_3 \leq D \leq d_4)}$	e_3
e_5 :	?	?	c_0	?	e_3

L'indice $cont_s$ (support de la contradiction) calcule la proportion des exemples de E exhibant *au moins un* descripteur contradictoire avec son représentant dans D :

$$cont_s(D, E) = \frac{1}{\rho(E)} \times \sum_{e \in E / \neg\tau(\delta(e), C_D(e))} \rho(e) \tag{4.15}$$

L'indice $cont_p$ (proportion de la contradiction) considère ces exemples et calcule la proportion des descripteurs effectivement en contradiction avec la description associée :

$$cont_p(D, E) = A \times B \tag{4.16}$$

avec :

$A = cont_s(D, E) \times \rho(E)$

$B = \sum_{e \in E / \neg\tau(\delta(e), C_S(e))} \frac{card(\{X \in \mathcal{A} /\ D_X(\delta(e), C_D(e)) \text{ et } \neg\tau_{C_D(e)}(\delta(e)_{|X}))\})}{|\delta(e)|} \times \rho(e)$

A correspond au poids total des exemples contenant au moins une contradiction avec leur représentant, alors que B calcule la proportion de la description $\delta(e)$ en contradiction avec $C_D(e)$. Comme nous le verrons, ces indices liés à la cohérence des représentants construits au regard des exemples à classer sont très importants pour montrer l'une des caractéristiques majeures de la classification par défaut.

• **La capacité prédictive ($pred$) :** Cet indice prend en compte la richesse des descriptions, au sens du nombre d'attributs décrits[8], afin de donner une idée sur leur capacité prédictive. Il correspond à la proportion des attributs dont la valeur peut être donnée à l'aide de D. Pratiquement, il s'agit du score moyen de richesse de D pondéré par la taille des clusters associés et par le nombre d'attributs :

$$pred(D, E) = \frac{1}{n_{\mathcal{A}} \times \rho(E)} \times \sum_{d \in D} |d| \times \rho_{D,E}(d) \tag{4.17}$$

où $\rho_{D,E}(d)$ correspond au poids des exemples de E couverts par la description $d \in D$.

[8] Concernant la richesse des descriptions, voir la définition qui en est donnée dans la section 3.1.6 page 67.

		A	B	C	D	E
S	s_1 :	?	b_1	c_1	d_2	?
	s_2 :	a_1	?	c_3	?	e_3
	s_3 :	a_0	b_2	c_4	d_1	?
E_1	e_1 :	?	b_1	c_1	?	?
	e_2 :	a_0	b_1	c_1	d_2	?
E_2	e_3 :	a_1	b_0	?	?	?
	e_4 :	?	?	c_3	d_2	?
	e_5 :	?	?	c_3	?	?
	e_6 :	a_1	b_1	?	?	e_3
E_3	e_7 :	a_0	b_2	c_4	d_1	e_4
$E_\top$	e_8	a_0	b_0	?	d_0	?

FIG. 4.3 – Trois stéréotypes couvrant l'ensemble $E_1 \cup E_2 \cup E_3 \cup E_\top$.

Considérons l'ensemble de stéréotypes présenté figure 4.3 et couvrant un ensemble $E = E_1 \cup E_2 \cup E_3 \cup E_\top$ dont les 8 exemples ont un poids uniforme de 1[9]. Le stéréotype s_i couvre les exemples contenus dans E_i relativement à S. Le score de prédiction se calcule alors ainsi : $pred(S, E) = \frac{1}{5\times 8} \times [|\{b_1, c_0 \leq C \leq c_2, d_2\}| \times 2 + |\{a_1, c_3, e_3\}| \times 4 + |\{a_0, b_2, c_4 \leq C \leq c_5, d_1\}| \times 1] = \frac{1}{40} \times [6 + 12 + 4] = 0.55$. Cela signifie que la valeur de 55% des descripteurs peut être donnée à partir des stéréotypes s_1, s_2 et s_3.

4.1.3 Evaluation avec classes a priori

Nous nous intéressons dans cette seconde partie aux indices qui peuvent être calculés lorsque E correspond à un jeu de données artificiels. Le processus de génération de ces données est décrit dans la section 4.2.1 page 110.

Chaque exemple de E étant produit à partir d'une description initiale de I, il est possible de calculer une partition "idéale" des données. Pour cela, il suffit d'étiqueter chaque exemple avec la description qui l'a généré. Posons la fonction μ qui associe à tout exemple de $e \in E$ la description initiale $\mu(e) \in I$ qui lui correspond. Etant donné un ensemble de descriptions D, nous étendons la fonction μ afin qu'elle permette de calculer la classe dominante du cluster associé à l'une des descriptions de D :

$$\mu_{D,I,E}(d) = \arg\max_{i \in I} \sum_{e \in E/\ C_D(e)=d \text{ et } \mu(e)=i} \rho(e) \tag{4.18}$$

que nous notons plus simplement $\mu_I(d)$. La description vide $\top$ correspondant aux exemples non classés est automatiquement affectée à la description absurde $\bot$, c'est-à-dire $\mu_I(\top) = \bot$.

Les indices peuvent se répartir en trois familles, chaque famille étant associée à l'un des items de la figure 4.1.

[9] Nous nous limitons pour des raisons de clarté à des attributs nominaux.

$$I \begin{cases} i_1 : & a_0 & ? & c_1 & d_1 & e_1 & ? \\ i_2 : & a_1 & b_0 & ? & d_0 & e_0 & f_1 \\ i_3 : & ? & b_1 & c_0 & ? & ? & f_0 \end{cases} \quad \longleftrightarrow \quad D \begin{cases} d_1 : & a_0 & b_1 & c_1 & ? & ? & ? \\ d_2 : & a_1 & b_0 & ? & d_0 & e_1 & f_1 \\ d_3 : & ? & ? & ? & ? & ? & f_0 \\ d_4 : & ? & ? & c_0 & d_1 & ? & ? \end{cases}$$

FIG. 4.4 – Correspondance entre les descriptions initiales et les descriptions découvertes.

4.1.3.1 Indices associés aux descriptions initiales I et à la génération des données artificielles

Ces indices sont associés à la fois aux descriptions initiales et à la manière dont les données artificielles ont été générées. Ils sont entrés en tant que paramètres dans nos expérimentations afin de tester la validité de nos hypothèses.

• **Le nombre de descriptions (n_I) :** Cet indice indique le nombre de descriptions initiales générées.

• **Le nombre de duplications (dup) :** Cet indice indique le nombre d'exemples dont la description est tirée de chaque description initiale de I. Le nombre total d'exemples se calcule de la manière suivante : $n_E = n_I \times dup$.

• **Le taux de données manquantes (m) :** Cet indice indique la proportion de descripteurs manquant dans la description des exemples.

4.1.3.2 Indices associés à l'adéquation entre D et I (C)

Ces indices sont très importants lorsque l'on tâche de redécouvrir les descriptions initiales I, qui ont permis de générer les données lacunaires artificielles E, sous la forme d'un ensemble de descriptions D. Il faut veiller à ne pas considérer la description vide $\top$ si celle-ci appartient à D (c'est le cas des ensembles de stéréotypes).

La figure 4.4 propose quatre descriptions découvertes associées grâce à la fonction μ_I à trois descriptions initiales de cette manière : $\mu_I(d_1) = i_1$, $\mu_I(d_2) = i_2$ et $\mu_I(d_3) = \mu_I(d_4) = i_3$. Les poids des quatre clusters associés à D sont : $\rho_{D,E}(d_1) = 8$, $\rho_{D,E}(d_2) = 6$, $\rho_{D,E}(d_3) = 3$ et $\rho_{D,E}(d_4) = 5$. Cet exemple permet d'illustrer le calcul des indices que nous présentons ci-dessous.

• **L'adéquation, la contradiction et la perte ($adeq_I$, $cont_I$ et $perte_I$) :** Il ne faut pas confondre ces trois indices complémentaires avec les scores d'adéquation, de contradiction et de perte calculés entre D et E. Nous cherchons ici à déterminer à quel point les descriptions D données par l'algorithmes de clustering correspondent aux descriptions initiales I qui ont généré les données artificielles.

Le score d'adéquation $adeq_I$ traduit la proportion de descripteurs appartenant aux descriptions initiales I qui sont retrouvées au sein des descriptions découvertes D et peuvent compléter les valeurs manquantes dans les exemples de E. Notons que cet indice est normalisé avec le poids total des descriptions $\rho_E(D) = \sum_{d \in D/d \neq \top} \rho_{D,E}(d)$ et non avec le poids total

des exemples $\rho(E)$. C'est pourquoi le pourcentage de couverture doit être pris en considération lorsque l'on interpréte les résultats en fonction de ces mesures. L'indice d'adéquation est défini comme suit :

$$adeq_I(D,I,E) = \frac{\sum_{d\in D/\ d\neq\top} \frac{sim(d,\mu_I(d))}{|\mu_I(s)|} \times \rho_{D,E}(d)}{\rho_E(D)} \tag{4.19}$$

Nous soulignons le facteur utilisé pour normaliser la mesure de similarité. En effet, nous cherchons à comparer le nombre de descripteurs retrouvés dans le stéréotype *au regard* du nombre de descripteurs total appartenant à la description initiale.

L'adéquation de l'exemple donné à la figure 4.4 est calculé comme suit : $adeq_I(D,I,E) = \frac{1}{8+6+3+5} \times [\frac{2}{4} \times 8 + \frac{4}{5} \times 6 + \frac{1}{3} \times 3 + \frac{1}{3} \times 5] \simeq \frac{4+4.8+1+1.67}{22} \simeq 0.52$. Cela signifie que 52% des descripteurs de D se retrouvent dans les descriptions initiales associées I en prenant en compte le poids respectif des classes sur E.

Le score de contradiction $cont_I$ traduit la proportion des descripteurs de I qui ne se retrouvent pas dans les descripteurs de D. Cet indice est défini comme suit :

$$cont_I(D,I,E) = \frac{\sum_{d\in D/d\neq\top} \frac{cont(d,\mu_I(d))}{|\mu_I(d)|} \times \rho_{D,E}(d)}{\rho_E(D)} \tag{4.20}$$

où *cont* est définie de manière duale par rapport à *sim* :

$$cont(d_1,d_2) = \sum_{X\in\mathcal{A}/D_X(d_1\cap d_2)} 1 - \frac{|d_{1_{|X}} \cap d_{2_{|X}}|}{|d_{1_{|X}} \cup d_{2_{|X}}|} \tag{4.21}$$

De la même façon que précédemment, le facteur de normalisation de $cont(d,\mu_I(d))$ correspond au nombre de descripteurs de la description initiale $|\mu_I(d)|$.

La contradiction de l'exemple donné à la figure 4.4 est calculée comme suit : $cont_I(D,I,E) = \frac{1}{8+6+3+5} \times [0\times 8 + \frac{1}{5}\times 6 + 0\times 3 + 0\times 5] = \frac{0+1.2+0+0}{22} \simeq 0.05$. Cela signifie que 5% des descripteurs de D (en l'occurence e_0 et e_1) entrent en contradiction avec les descriptions initiales associées I.

Le score de perte $perte_I$ traduit la proportion des descripteurs de I dont l'attribut n'est même pas décrit par les descriptions de D correspondantes. Cet indice est défini comme suit :

$$perte_I(D,I,E) = \frac{\sum_{d\in D/d\neq\top} \frac{abst(d,\mu_I(d))}{|\mu_I(s)|} \times \rho_{D,E}(d)}{\rho_E(D)} \tag{4.22}$$

où $abst(d_1,d_2)$ calcule le nombre de descripteurs de d_2 qui n'apparaissent pas dans d_1 :

$$abst(d_1,d_2) = card(\{X \in \mathcal{A}/D_X(d_2) \text{ et } d_{1_{|X}} = x_?\}) \tag{4.23}$$

La perte de l'exemple donné à la figure 4.4 est calculé comme suit : $perte_I(D,I,E) = \frac{1}{8+6+3+5} \times [\frac{2}{4} \times 8 + 0 \times 6 + \frac{2}{3} \times 3 + \frac{2}{3} \times 5] \simeq \frac{4+0+2+3.33}{22} \simeq 0.42$. Cela signifie que 42% des descripteurs de D correspondant à des descripteurs dans I sont indéfinis.

Comme pour les indices d'adéquation par rapport à E, ces trois indices sont complémentaires et vérifient la propriété suivante[10] :

[10] Le score de 0.99 obtenu dans l'exemple présenté est dû à des erreurs d'approximation.

Propriété 4.1.2 $\forall D, E, adeq_I(D,E) + cont_I(D,E) + perte_I(D,E) = 1.$

Preuve 4.1.2 *La fonction cont étant définie de manière duale à sim, on peut tout de suite remarquer que :* $sim(d_1,d_2) + cont(d_1,d_2) = card(\{X \in \mathcal{A}/D_X(d_1 \cap d_2)\})$ *[1]. Or, si l'on considère que la description absurde* $\perp$ *ne fait pas partie de D, on obtient :* $D_X(d_1 \cap d_2) \Rightarrow D_X(d_1)$ *et* $D_X(d_2)$*. En effet, si on a* $\neg D_X(d_1)$ *alors* $d_{1_{|X}} = x_?$*,* $(d_1 \cap d_2)_{|X} = d_{1_{|X}} \cup d_{2_{|X}} = x_?$ *et donc* $\neg D_X(d_1 \cap d_2)$*, ce qui est absurde au vu de nos hypothèses. Le même raisonnement peut être tenu pour* d_2*. [1] se réécrit alors* $sim(d_1,d_2)+cont(d_1,d_2) = card(\{X \in \mathcal{A}/D_X(d_1)$ *et* $D_X(d_2)\})$*.* $\neg D_X(d_1)$ *est ici équivalent à* $d_{1_{|X}} = x_?$*. On a donc :* $sim(d_1,d_2)+cont(d_1,d_2)+abst(d_1,d_2) = card(\{X \in \mathcal{A}/D_X(d_1)$ *et* $D_X(d_2)\}) + card(\{X \in \mathcal{A}/D_X(d_2)$ *et* $\neg D_X(d_1)\}) = card(\{X \in \mathcal{A}/D_X(d_2)\}) = |d_2|$ *(définition 3.1.6). Si l'on remplace* d_1 *par d et* d_2 *par* $\mu_I(d)$*, on obient finalement :* $sim(d,\mu_I(d)) + cont(d,\mu_I(d)) + abst(d,\mu_I(d)) = |\mu_I(d)|$*. A partir de là, la démonstration de la formule 4.1.2 est immédiate.*

Remarquons qu'il ne faut pas confondre les indices d'adéquation, contradiction et perte relatifs à I, que nous venons de décrire, et les indices relatifs à E définis dans la section 4.1.2.3. Ici, l'idée est de calculer le rapport entre les représentants proposés par un algorithme de clustering et les descriptions initiales I qui ont permis de générer les données artificielles. Cette évaluation peut bien sûr être complétée par les indices calculant l'adéquation, la contradiction et la perte entre les représentants D et les données elles-mêmes E, sans faire d'hypothèse sur l'existence de classes a priori. Il s'agit donc de deux manières différentes, mais pas nécessairement contradictoires, d'aborder l'évaluation du résultat du clustering.

4.1.3.3 l'erreur de classification (err_C)

L'erreur de classification err_C est calculée en comptant la proportion des exemples dont l'étiquette $\mu(e)$ ne correspond pas à l'étiquette affectée au cluster où il se trouve, c'est-à-dire $\mu_{D,I,E}(C_D(e))$.

$$err_C(D,I,E) = \frac{1}{\rho(E)} \times \sum_{e \in E/\ \mu(e) \neq \mu_{D,I,E}(C_D(e))} \rho(e) \tag{4.24}$$

Remarquons que les exemples non classés, c'est-à-dire associés à la description absurde $\perp$ par μ, sont naturellement comptés dans l'erreur de classification.

4.1.4 Extraction des descriptions à partir d'un clustering

Une partie de l'étape de validation est de parvenir à évaluer les résultats de notre programme PRESS au regard des résultats obtenus avec d'autres algorithmes de classification automatique. Pour ce faire, nous avons besoin de connaître les descriptions étiquetant les clusters proposés par ces algorithmes parce que c'est sur ce type d'objet, et non sur le clustering lui-même, que porte la comparaison. Plus encore, nous devons avoir à notre disposition des descriptions "certaines" exprimées sous la forme d'un ensemble de descripteurs, et non d'un ensemble de distributions de probabilité comme dans le cas de l'algorithme EM.

Dans cette partie, nous allons détailler quatre stratégies permettant d'extraire des descriptions à partir d'une catégorisation donnée à l'avance. Grâce à ces descriptions et à cette catégorisation, il est alors possible de comparer les résultats obtenus avec les différents algorithmes. Nous n'affirmons pas, bien entendu, qu'il s'agit des seules techniques pouvant être utilisées. Cependant, elles nous ont semblé refléter plutôt fidèlement les résultats proposés par les algorithmes que nous utilisons à titre de comparaison. De plus, seul le cas des attributs de type nominal est pris en compte car les algorithmes choisis ne traitent pas les attributs ordinaux. Précisons enfin que les classes vides générées parfois par les algorithmes de clustering ont été supprimées car, détériorant inutilement les résultats obtenus, ils biaisent la comparaison effectuée avec PRESS.

4.1.4.1 Utilisation du mode

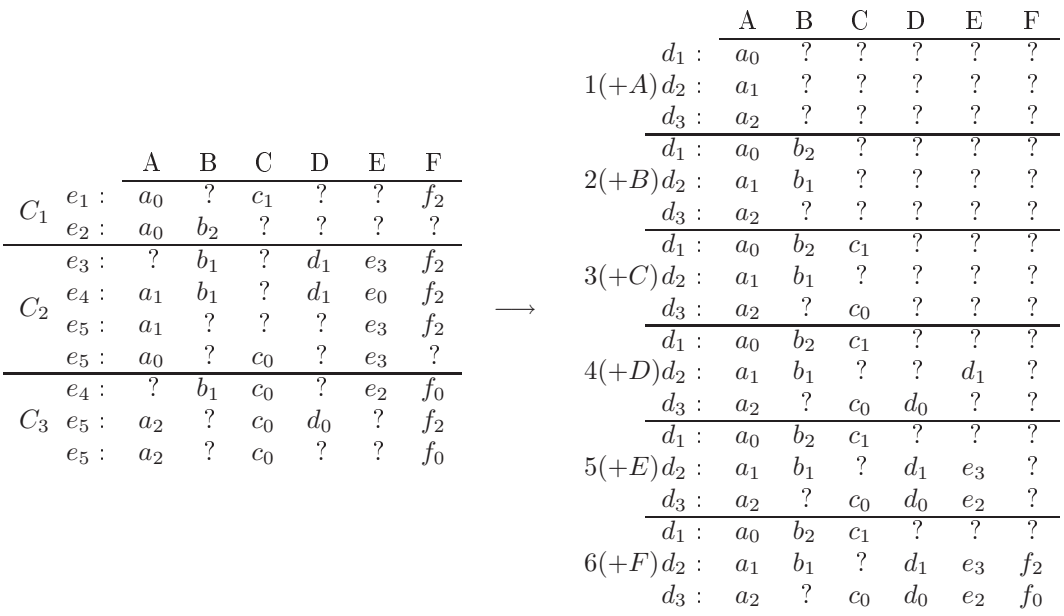
Cette première technique d'extraction, que nous notons T_1, est la plus simple. Elle consiste à conserver, pour chaque attribut, le descripteur le plus fréquent. Si deux descripteurs sont d'égale fréquence, le choix du descripteur est arbitraire. Cette technique correspond à la manière d'extraire les centroïdes à partir des clusters dans l'algorithme des k-modes. Le schéma suivant illustre ce procédé au niveau de l'un des clusters :

		A	B	C	D	E	F
	e_1 :	a_0	?	c_1	?	?	?
	e_2 :	a_0	b_2	?	?	?	?
C	e_3 :	a_3	b_2	c_0	d_1	e_3	?
	e_4 :	?	b_2	c_1	?	e_2	?
	e_5 :	a_1	?	c_1	?	?	?
	d :	a_0	b_2	c_1	d_1	e_3	?

4.1.4.2 Mode avec prise en compte des incohérences

La deuxième technique d'extraction T_2 est identique à la première dans le sens où elle consiste à ne considérer que les descripteurs les plus fréquents. Par contre, si tout autre descripteur apparaît au sein des exemples du cluster, alors l'attribut correspondant est considéré comme indéfini dans la description du représentant. Cette technique permet d'obtenir une cohérence parfaite entre le représentant et les exemples qu'il couvre, dans le sens où il ne peut exister aucune contradiction entre eux. Le schéma suivant illustre ce procédé au niveau de l'un des clusters :

		A	B	C	D	E	F
	e_1 :	a_0	?	c_1	?	e_3	?
	e_2 :	a_0	b_2	?	?	?	?
C	e_3 :	?	?	?	?	e_3	f_2
	e_4 :	?	b_1	?	?	e_2	?
	e_5 :	a_0	?	?	?	?	?
	d :	a_0	?	c_1	?	?	f_2



		A	B	C	D	E	F
C_1	e_1 :	a_0	?	c_1	?	?	f_2
	e_2 :	a_0	b_2	?	?	?	?
C_2	e_3 :	?	b_1	?	d_1	e_3	f_2
	e_4 :	a_1	b_1	?	d_1	e_0	f_2
	e_5 :	a_1	?	?	?	e_3	f_2
	e_5 :	a_0	?	c_0	?	e_3	?
C_3	e_4 :	?	b_1	c_0	?	e_2	f_0
	e_5 :	a_2	?	c_0	d_0	?	f_2
	e_5 :	a_2	?	c_0	?	?	f_0

		A	B	C	D	E	F
	d_1 :	a_0	?	?	?	?	?
$1(+A)$	d_2 :	a_1	?	?	?	?	?
	d_3 :	a_2	?	?	?	?	?
	d_1 :	a_0	b_2	?	?	?	?
$2(+B)$	d_2 :	a_1	b_1	?	?	?	?
	d_3 :	a_2	?	?	?	?	?
	d_1 :	a_0	b_2	c_1	?	?	?
$3(+C)$	d_2 :	a_1	b_1	?	?	?	?
	d_3 :	a_2	?	c_0	?	?	?
	d_1 :	a_0	b_2	c_1	?	?	?
$4(+D)$	d_2 :	a_1	b_1	?	?	d_1	?
	d_3 :	a_2	?	c_0	d_0	?	?
	d_1 :	a_0	b_2	c_1	?	?	?
$5(+E)$	d_2 :	a_1	b_1	?	d_1	e_3	?
	d_3 :	a_2	?	c_0	d_0	e_2	?
	d_1 :	a_0	b_2	c_1	?	?	?
$6(+F)$	d_2 :	a_1	b_1	?	d_1	e_3	f_2
	d_3 :	a_2	?	c_0	d_0	e_2	f_0

FIG. 4.5 – Construction progressive des descriptions représentant les clusters.

4.1.4.3 Répartition des descripteurs

La troisième technique d'extraction T_3 consiste à répartir les descripteurs entre les différents représentants étiquetant les clusters. Chaque attribut est analysé l'un après l'autre. Le représentant du cluster exhibant le descripteur le plus fréquent associé à cet attribut se voit attribuer ce descripteur. Ce processus est itéré tant qu'il subsiste un descripteur non attribué appartenant à un cluter étiqueté par un représentant où cet attribut n'est pas encore défini. Puis on passe à l'attribut suivant jusqu'à ce que ceux-ci aient tous été traités. Ce processus est illustré par la figure 4.5.

Remarquons que la séparation entre les clusters est nécessairement parfaite, au sens des mesures que nous avons définies préalablement, puisqu'il n'y a aucune redondance entre les descripteurs.

4.1.4.4 Répartition avec prise en compte des incohérences

Cette dernière technique, notée T_4, repose sur la même idée que la précédente, mais en considérant (comme pour la technique T_2) les incohérences pouvant survenir entre la description des exemples couverts et les représentants associés. Comme pour T_3, on remarque que l'on obtient nécessairement une séparation parfaite entre les clusters, au sens des mesures que nous avons définies préalablement. Ce processus est illustré par la figure 4.6.

4.2 Jeux de données utilisés

Cette partie donne une description des deux types de jeux de données employés dans nos expérimentations : d'une part des jeux artificiellement générés, d'autre part des jeux tirés

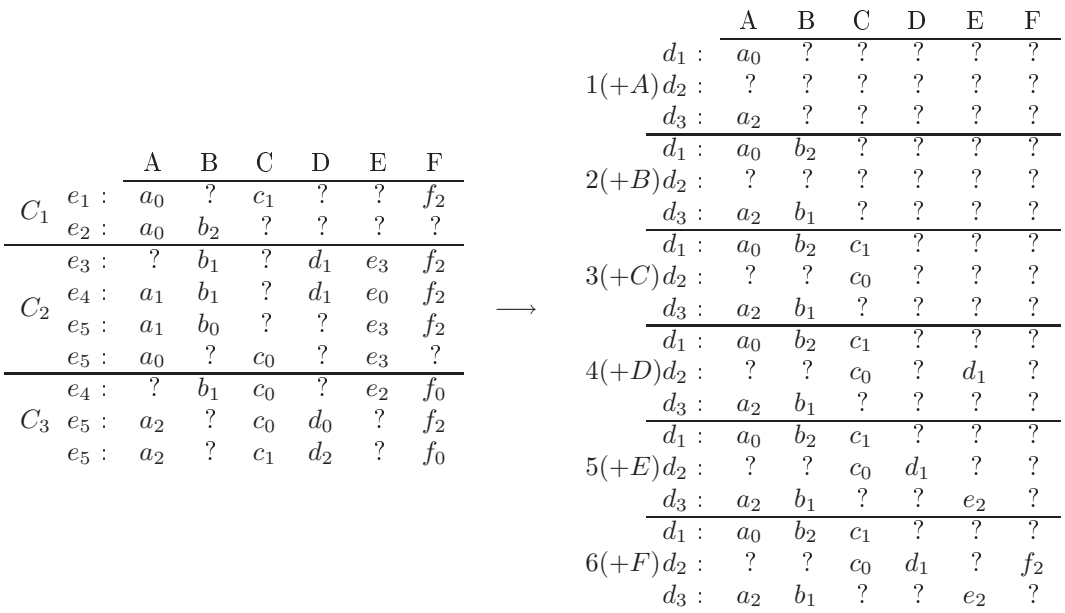


FIG. 4.6 – Répartition prenant en compte la cohérence.

d'articles de presse.

4.2.1 Les jeux de données artificiels

Il s'agit de jeux de données générés artificiellement afin d'évaluer de manière plus précise notre algorithme d'apprentissage. Le processus de génération de ces données est décrit ci-dessous :

- Générer un langage de description composé de $n_{\mathcal{A}}$ attributs nominaux acceptant chacun *mod* valeurs différentes. Si $mod = 2$, alors il s'agit d'attributs binaires classiques. C'est le cas de figure que nous avons suivi dans nos expérimentations.
- Générer aléatoirement n_I descriptions initiales qui respectent la contrainte de non-redondance, c'est-à-dire qui ne partagent aucun descripteur en commun. Cet ensemble intitulé I contient les "graines de départ" (*seed*) à l'origine du jeu de données. Pour s'assurer de la non-redondance, on attribue les *mod* descripteurs possibles pour chaque attribut à l'une des descriptions de I. Si $mod < n_I$ alors les attributs ne sont pas décrits par toutes les descriptions.
- Dupliquer à l'identique chaque description initale de I un nombre *dup* de fois. Nous obtenons alors $n_I \times dup$ descriptions que l'on associe à autant d'exemples de poids 1. Ces $n_E = n_I \times dup$ exemples forment l'ensemble des données artificielles.
- Remplir les attributs encore indéfinis avec des descripteurs aléatoires. Cela fait exactement $(n_I - mod) \times n_{\mathcal{A}} \times dup$ descripteurs à ajouter (si bien sûr $n_I \geq mod$). Ces descripteurs correspondent à du "bruit" inséré dans les données qui est d'autant plus important que la différence $n_I - mod$ est élevée.
- Retirer aléatoirement un pourcentage m de descripteurs aux $n_I \times dup \times n_{\mathcal{A}}$ descripteurs

qui décrivent les exemples artificiels.

Ce processus est illustré par la figure 4.7 qui présente la même décomposition que celle utilisée à la section 1.1.4.3. La technique que nous utilisons est d'ailleurs très similaire à celle choisie par Gennari lorsqu'il traite justement le cas des données lacunaires. Précisons une nouvelle fois que chaque descripteur de chaque attribut se trouve au plus une fois dans les descriptions de I. Dans l'exemple présenté, les valeurs suivantes ont été fixées : $n_{\mathcal{A}} = 10$, $n_I = 3$, $mod = 2$, $dup = 4$ et $m = 0.6$. Le tableau de la figure 4.8 résume les différents paramètres pouvant être ajustés dans la génération d'un tel jeu d'essai.

4.2.2 Les jeux de données tirés de la presse

Nous présentons dans cette partie quatre jeux de données tirés de la presse. Chacun d'eux est associé à un quotidien à grand tirage de la fin du XIX^{e} siècle en France, respectivement : Le Matin, La Libre Parole, La Croix et Le Petit Journal. L'information extraite de ces journaux a la particularité, une fois traduite dans notre formalisme attribut-valeur, de présenter un taux très élevé de valeurs manquantes et d'entrer parfaitement dans notre problématique. Après avoir détaillé le cadre historique duquel ces données sont extraites, nous proposons un langage de description permettant d'effectuer cette traduction ainsi qu'un schéma général d'extraction des stéréotypes à partir de celles-ci.

4.2.2.1 Cadre historique

La problématique utilisée dans ces expérimentations est la même que celle présentée dans [Vel02] et nous la détaillons à nouveau ici. Elle s'appuie sur plusieurs lectures relatives à la période très riche en événements de la fin du XIX^{e} siècle en France, comme *La Couleur et le Sang* du sociologue P.-A. Taguieff [Tag02] dont un chapitre traite de l'antisémitisme et du nationalisme, *L'Affaire* de J.-D. Bredin [Bre83] qui résume admirablement l'Affaire Dreyfus, ou *La France Juive* écrite par E. Drumont [Dru86], ouvrage emblématique de l'antisémitisme de cette fin de siècle.

Le thème retenu est celui du désordre de la scène politique française dont la presse, média privilégié à cette époque, faisait ses choux gras. Dans le contexte des années 1890, les gouvernements se succèdent à une cadence effrénée et la troisième République passe par deux crises conséquentes dont la seconde manque de finir en coup d'Etat : le scandale de Panama, scandale financier impliquant des politiciens connus comme Clémenceau ou Reinach, et surtout l'Affaire Dreyfus, histoire d'une erreur judiciaire commise à l'encontre d'un officier juif et remettant en cause la raison d'Etat et la toute puissance de l'armée comme défenseur de la patrie[11].

La période que nous avons plus particulièrement choisie d'étudier est la première décade de septembre 1893, intéressante à deux titres. D'une part, elle précède d'un peu plus d'un an l'éclatement au grand jour de l'Affaire Dreyfus et préfigure l'état d'esprit dans lequel l'opinion publique percevra l'affaire dans les premiers temps. D'autre part, elle est le cadre

[11] Des détails historiques sur ces deux affaires sont donnés en annexe page 173.

		A	B	C	D	E	F	G	H	I	J
		a_0	b_0	c_0	d_0	e_0	f_0	g_0	h_0	i_0	j_0
		a_1	b_1	c_1	d_1	e_1	f_1	g_1	h_1	i_1	j_1
						↓ (1)					
	i_1 :	a_0	b_1	c_0	d_0	e_1	f_0	g_1	?	?	j_1
I	i_2 :	a_1	?	?	d_1	?	?	g_0	h_0	i_1	?
	i_3 :	?	b_0	c_1	?	e_0	f_1	?	h_1	i_0	j_0
						↓ (2)					
	e_1 :	a_0	b_1	c_0	d_0	e_1	f_0	g_1	?	?	j_1
	e_2 :	a_0	b_1	c_0	d_0	e_1	f_0	g_1	?	?	j_1
	e_3 :	a_0	b_1	c_0	d_0	e_1	f_0	g_1	?	?	j_1
	e_4 :	a_0	b_1	c_0	d_0	e_1	f_0	g_1	?	?	j_1
	e_5 :	a_1	?	?	d_1	?	?	g_0	h_0	i_1	?
	e_6 :	a_1	?	?	d_1	?	?	g_0	h_0	i_1	?
	e_7 :	a_1	?	?	d_1	?	?	g_0	h_0	i_1	?
	e_8 :	a_1	?	?	d_1	?	?	g_0	h_0	i_1	?
	e_9 :	?	b_0	c_1	?	e_0	f_1	?	h_1	i_0	j_0
	e_{10} :	?	b_0	c_1	?	e_0	f_1	?	h_1	i_0	j_0
	e_{11} :	?	b_0	c_1	?	e_0	f_1	?	h_1	i_0	j_0
	e_{12} :	?	b_0	c_1	?	e_0	f_1	?	h_1	i_0	j_0
						↓ (3)					
	e_1 :	a_0	b_1	c_0	d_0	e_1	f_0	g_1	h_0	i_1	j_1
	e_2 :	a_0	b_1	c_0	d_0	e_1	f_0	g_1	h_1	i_1	j_1
	e_3 :	a_0	b_1	c_0	d_0	e_1	f_0	g_1	h_1	i_1	j_1
	e_4 :	a_0	b_1	c_0	d_0	e_1	f_0	g_1	h_0	i_0	j_1
	e_5 :	a_1	b_0	c_1	d_1	e_1	f_0	g_0	h_0	i_1	j_1
	e_6 :	a_1	b_1	c_0	d_1	e_1	f_0	g_0	h_0	i_1	j_0
	e_7 :	a_1	b_0	c_0	d_1	e_0	f_0	g_0	h_0	i_1	j_0
	e_8 :	a_1	b_1	c_1	d_1	e_0	f_1	g_0	h_0	i_1	j_1
	e_9 :	a_1	b_0	c_1	d_0	e_0	f_1	g_1	h_1	i_0	j_0
	e_{10} :	a_0	b_0	c_1	d_1	e_0	f_1	g_0	h_1	i_0	j_0
	e_{11} :	a_0	b_0	c_1	d_0	e_0	f_1	g_1	h_1	i_0	j_0
	e_{12} :	a_0	b_0	c_1	d_1	e_0	f_1	g_0	h_1	i_0	j_0
						↓ (4)					
	e_1 :	?	b_1	c_0	?	?	f_0	g_1	h_0	?	j_1
	e_2 :	a_0	?	?	?	?	f_0	?	h_1	?	j_1
	e_3 :	a_0	b_1	?	?	?	f_0	?	h_0	?	?
	e_4 :	?	?	c_0	d_0	e_1	?	?	?	i_0	?
	e_5 :	a_1	c_1	?	d_1	?	?	g_0	?	?	?
E	e_6 :	?	?	c_0	d_1	?	f_0	?	?	i_1	?
	e_7 :	?	?	?	?	?	?	g_0	h_0	?	j_0
	e_8 :	?	?	?	d_1	e_0	f_1	?	?	?	j_1
	e_9 :	?	?	c_1	d_0	e_0	?	?	?	i_0	?
	e_{10} :	?	b_0	c_1	d_1	?	?	g_0	?	?	j_0
	e_{11} :	?	b_0	?	d_0	?	?	g_1	?	?	?
	e_{12} :	?	?	?	?	e_0	f_1	?	h_1	?	?

(1) génération de trois descriptions initiales par répartition des descripteurs de chaque attribut binaire.

(2) chaque description initiale est dupliquée 4 fois. Les exemples e_1, e_2, e_3 et e_4 correspondent à la description d_1 ; e_5, e_6, e_7 et e_8 à d_2 ; e_9, e_{10}, e_{11} et e_{12} à d_3.

(3) complétion des descriptions avec des descripteurs aléatoires.

(4) suppression aléatoire de 60% des valeurs au sein de la description des exemples.

FIG. 4.7 – Génération d'un jeu de données artificiel.

Paramètre	Description
$n_{\mathcal{A}}$	Nombre d'attributs du langage de description.
mod	Modalité des attributs du langage de description.
n_I	Nombre de descriptions dans l'ensemble initial I.
dup	Nombre de duplicaton de chaque description de I.
m	Pourcentage de descripteurs manquants.

FIG. 4.8 – Paramètres des jeux de données artificiels.

du second tour des élections législatives de l'époque, ce qui nous promet une source riche en information concernant la scène politique française : duels électoraux, diffamations, trafics de voix, accusations, etc. Il s'agit donc d'une période de crise, *"moment de production intense de texte et [...] puissant 'révélateur' des codes utilisés"* [Del77][12].

Quatre quotidiens ont été retenus, dont deux sont conservateurs (La Croix et La Libre Parole) et deux se situent plutôt au centre (Le Matin et Le Petit Journal). Un journal de gauche, comme La Petite République, n'a malheureusement pas pu être pris en compte, faute de temps. Voici un résumé de ce qui est dit au sujet des quatre quotidiens dans *L'histoire générale de la presse française* [BGGT76] :

• **Le Matin :** quotidien à large diffusion lancé en 1884 par un américain commandité par deux anglais ; il est plutôt original avec des articles courts et surtout des nouvelles ; Alfred Edwards, son rédacteur en chef, est compromis par le scandale de Panama qui révèle qu'il avait touché plus de 200 000 Francs de la Compagnie ; au début, il est de toutes les campagnes antidreyfusardes avant de se dégager aux alentours de 1898 d'une cause devenue trop compromettante et de devenir peu à peu dreyfusard.

• **Le Petit Journal :** c'est un phénomène dans la presse mondiale du fait de son très fort tirage ; il est lancé à l'origine contre les panamistes et surtout contre Clémenceau ; nationaliste lié aux milieux de la Patrie Française, le rédacteur en chef Ernest Judet lance ensuite son journal, sans retenue, dans la lutte antidreyfusarde et l'y maintient après le procès de Rennes.

• **La Libre Parole :** lancé en 1892 par Edouard Drumont, il doit son succès considérable à la campagne contre la Banque de France menée par le marquis de Morès en 1892 et à la révélation du scandale de Panama ; ce quotidien s'illustre par ses nombreuses campagnes antisémites, ses dénonciations de scandales financiers, qu'ils soient réels ou totalement imaginaires, et ses attaques virulentes contre le gouvernement et les parlementaires ; il est bien entendu profondément antidreyfusard.

• **La Croix :** lancé une première fois en 1880 puis repris en 1893, ce journal catholique populaire est géré par les services de la Bonne Presse qui cachent en réalité le mouvement des assomptionnistes ; il possède une vision très conservatrice et dénonce le progrès avilissant, la décadence des mœurs ; il exhalte au contraire les vertus de l'ordre et du travail ; il est

[12] Cette citation est extraite de l'ouvrage de M. Souchard sur les discours de presse [Sou92].

violemment antidreyfusard et antisémite.

Il nous faut remarquer que le choix de cette problématique n'a pas été le fruit du hasard. Tout d'abord, et comme nous l'avons déjà suggéré, la presse était à cette époque le média par excellence qui pesait d'un poids considérable dans l'opinion publique [Bre83][Mol94]. Son influence n'est plus à démontrer à l'époque de l'Affaire Dreyfus, le point culminant étant certainement la publication, dans L'Aurore, du "J'accuse" d'Emile Zola. D'un autre côté, l'Affaire Dreyfus est un sujet qui a été maintes fois traité et il n'est pas exagéré de dire que des centaines d'ouvrages ont été écrits à son sujet (voir à ce propos l'étonnante bibliographie de L'Affaire [Bre83]). Il est ainsi d'autant plus aisé de comparer les résultats de nos travaux avec les observations des historiens qui ont étudié cette période. Nous regrettons toutefois de n'avoir pu développer davantage nos expérimentations réelles avec un deuxième jeu de données tiré d'une période-clef de l'affaire, comme la publication de l'article de Zola ou le procès de Rennes.

4.2.2.2 Langage de représentation

Le langage utilisé pour traduire les articles comporte 33 attributs et a été élaboré en étudiant à la fois les ouvrages concernant cette période (ceux que nous avons cités plus haut mais aussi beaucoup d'autres) et la "matière première", c'est-à-dire les articles eux-mêmes[13]. Nous avons essayé de les rendre les moins ambigus possible en leur attribuant une définition claire et en ne leur affectant une valeur que si le texte de l'article était suffisamment explicite. Ce langage ne se prévaut bien entendu pas d'une objectivité indubitable mais tâche de traduire au mieux le discours tenu dans les articles de presse étudiés. La plupart des attributs sont binaires (oui ou non), 4 ont une modalité supérieure à deux et 2 sont ordonnés. Parmi ceux-ci :

- *Politique :* Le groupement auquel l'homme politique appartient.
- *Patriotisme :* Indique si l'homme politique est un patriote.
- *Corruption :* Indique s'il s'agit d'un homme politique corrompu.

Le détail des attributs est donné dans le tableau situé en annexe page 175.

4.2.2.3 Traduction des articles

Après avoir choisi la période analysée, au regard de la problématique étudiée, puis sélectionné le support de notre étude, c'est-à-dire les quatre quotidiens que nous venons de citer, il reste à choisir les articles et à les traduire dans le formalisme utilisé. Nous évoquons en quelques mots ces deux étapes particulièrement délicates. La figure 4.9 illustre le processus général d'extraction des stéréotypes à partir des articles de presse.

La sélection des articles, tout d'abord, n'est pas un exercice facile. Ainsi, les extraits intéressants se cachent souvent au milieu d'articles plus généraux concernant la politique ou même des sujets de société qui semblent de prime abord très éloignés du sujet. Il n'est donc

[13] On note ainsi la présence de certains attributs qui peuvent paraître surprenants, comme Implication-FrancMacons ou LieClemenceau.

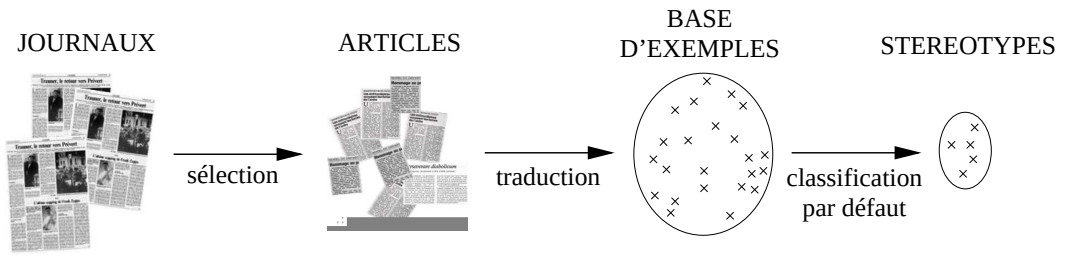


FIG. 4.9 – Processus général d'extraction des stéréotypes.

pas possible d'effectuer une sélection sommaire des articles en ne se basant, par exemple, que sur leur titre ou leur emplacement dans le journal. Il est nécessaire que le critère de sélection soit particulièrement clair pour ne laisser aucune ambiguïté face à des articles souvent imbriqués les uns dans les autres, repris à plusieurs endroits dans le quotidien, commentés, etc. Dans notre cas, par exemple, nous avons sélectionné tous les extraits dans lequel un homme politique au sens large (membre du gouvernement, député, candidat à la députation, sénateur) est impliqué, d'une façon ou d'une autre, dans un événement à connotations négatives (trafic de voix, corruption, diffamation, duel, problèmes de santé) susceptibles d'expliquer le désordre de la scène politique française.

La traduction, ensuite, est certainement l'étape la plus difficile. Le premier problème est celui de connaître le mieux possible la période étudiée, que ce soit au niveau des événements historiques, du système politique, de la distribution des grands médias, et ainsi de suite. En effet, il faut savoir interpréter certains mots ou groupes de mots pour leur associer l'attribut et le descripteur correspondant, utiliser intelligemment les synonymes, lever certaines ambiguïtés, etc. Cela est particulièrement saillant avec les attributs ordonnés dont il faut parvenir à déterminer le découpage adéquat (celui concernant le niveau de la violence par exemple). Un autre point est qu'il est parfois nécessaire d'extraire plusieurs exemples à partir d'un même article. Où trouver alors les frontières entre l'un et l'autre exemple ? Ensuite, et il s'agit peut-être du point le plus important, il faut être en mesure de gérer le mieux possible les incertitudes inhérentes au langage naturel. Ainsi "nous pensons fortement que" ou "il est probable que" ne proposent pas des traductions univoques. Lorsque le choix est trop difficile, nous avons délibéremment choisi de laisser l'attribut indéterminé pour ne pas trop biaiser les résultats. Enfin il faut déterminer au plus juste le poids associé à chaque exemple. Pour nos expérimentations, nous avons utilisé une échelle grossière (de 1 à 3) qui permet de pratiquer des critères d'attribution très simples (page de garde ou non, taille du titre, etc.). Une fois notre ensemble d'apprentissage construit, il ne nous reste plus qu'à lancer notre algorithme et à interpréter les stéréotypes découverts pour les différents quotidiens.

La figure 4.10 donne un exemple de traduction d'un article (donné ici sous la forme d'un résumé) dans le langage de description utilisé pour nos expérimentations. Les quatre jeux de données tirés de la presse ne sont pas donnés dans leur totalité à cause de la longueur excessive des fichiers XML correspondant. Seules les données tirées du quotidien Le Matin peuvent être trouvées dans les annexes page 179.

Maquignonnage : infect trafic électoral, le gouvernement approuve. Paul Lafargue, socialiste révolutionnaire et internationaliste court à une défaite certaine face au libéral Loyer, sauf si les républicains votent pour Lafargue. En échange, les socialistes soutiendraient Dron, républicain, dans une autre circonscription. Il s'agit d'une coalition cynique et anticléricale approuvée par le préfet du Nord, donc par le gouvernement ! C'est un véritable scandale, une trahison. Les socialistes révolutionnaires flétrissent le gouvernement quand il s'agit d'exiger le respect de la loi.

↓

{ (Socialiste=oui), (Revolutionnaire=oui), (Internationaliste=oui),
(ImplicationGouvernementale=oui), Clericalisme=anticlerical), (RespectLoi=non),
(Impunité=oui), (Action=trafic-electoral) }

FIG. 4.10 – Exemple de traduction dans le formalisme attribut-valeur.

Chapitre 5

Expérimentations et résultats

Ce chapitre présente les résultats obtenus à l'aide du programme basé sur le modèle que nous avons décrit aux chapitres 2 et 3. la première partie précise les choix qui ont été faits, d'une part pour l'implémentation de l'algorithme de classification par défaut, d'autre part pour la comparaison avec d'autres algorithmes de clustering. La deuxième partie présente les résultats liés aux jeux de données artificiels et donne une première idée des avantages procurés par notre approche. La dernière partie traite de jeux de données réels extraits d'articles de journaux. Elle montre que notre modèle peut être utilisé pour donner une représentation du discours de presse.

5.1 Choix d'implémentation

5.1.1 PRESS

Un programme nommé PRESS a été réalisé afin de mener à bien nos expérimentations. Cet acronyme signifie Programme de Reconstruction d'Ensembles de StéréotypeS et implémente l'algorithme de classification par défaut dans le formalisme attribut-valeur. Le programme a été écrit en Java dans sa version 1.5 et exécuté dans un environnement de type UNIX. Il est capable de lire en entrée des fichiers décrivant les données à classer dans le format XML et dans le format ARFF utilisé par la plate-forme WEKA [Gar95]. Il peut également générer des jeux de données artificiels suivant le processus décrit dans la partie 4.2.1. Nous ne détaillons pas les spécifications du programme qui est encore à l'état de prototype.

5.1.2 Les k-modes, EM et COBWEB

Nous avons choisi de comparer les performances de notre algorithme avec trois algorithmes de clustering que sont les k-modes, EM et COBWEB. Nous utilisons pour cela la plate-forme WEKA [Gar95] implémentée en Java et passons par le format ARFF pour représenter les données à classer. Nous discutons du choix de ces trois algorithmes parmi la multitude d'algorithmes de classification automatique :

• **Les k-modes :** Il s'agit de la version catégorielle de l'algorithme des c-moyennes[1], largement reconnu et utilisé dans de nombreux domaines de recherche. Il nous semblait indispensable de comparer les performances de PRESS avec ce "pilier" de la classification automatique.

Avec les k-modes, le nombre de clusters k doit être précisé comme paramètre d'entrée. Lorsque cela est possible, c'est-à-dire dans le cas des jeux de données artificiels, il nous suffit de poser $k = n_I$. Dans le cas contraire, nous fixons k suivant le nombre de stéréotypes découverts par notre algorithme de classification par défaut. Cela nous a semblé la démarche la plus raisonnable à adopter.

• **EM :** Nous utilisons un algorithme de type *Expectation-Maximization*[1] basé sur un mélange de distributions adaptées aux données symboliques. EM nous intéresse à deux titres : tout d'abord parce que c'est un algorithme adapté aux données à caractère lacunaire ; ensuite parce qu'il s'agit d'une méthode d'imputation couramment utilisée, basée sur le critère de maximum de vraisemblance. Sa programmation sous WEKA permet de lancer l'algorithme en utilisant un mécanisme de validation croisée afin de découvrir le nombre optimal de clusters.

• **COBWEB :** Le choix de ce troisième algorithme peut surprendre car COBWEB a été élaboré au départ pour construire un clustering hiérarchique[1], et non une partition de l'ensemble des exemples.

Il nous intéresse à deux titres. Tout d'abord, la fonction qu'il utilise pour construire sa hiérarchie de concepts repose sur la notion de *cue validity* proposée en théorie de la catégorisation[2]. L'idée de pouvoir prédire la valeur des attributs est également très présente dans notre propre travail. Ensuite, il s'agit historiquement de la première fonction d'évaluation à laquelle nous avons pensé pour découvrir les stéréotypes à travers l'espace des hypothèses. Les résultats s'étant révélés médiocres, en grande partie du fait de la difficulté d'utiliser la mesure de *Category Utility* dans un contexte non-incrémental, mais aussi à cause de problèmes de temps de calcul, nous avons préféré ne pas en faire état dans le présent document. Nous souhaitions cependant observer quelles sorties COBWEB propose dans un cadre lacunaire.

Une partition des exemples peut être extraite à partir des feuilles de l'arbre et sa programmation dans WEKA permet de faire varier un paramètre de *cutoff* afin de modifier le nombre de clusters obtenu en sortie. Comme il n'était pas envisageable de déterminer à la main la meilleure valeur pour chaque jeu de données utilisé[3], nous avons conservé la valeur par défaut proposée par l'interface.

5.2 Jeux de données artificiels

L'objectif principal de cette première série d'expérimentations est de comparer, d'une part, les quatre fonctions d'évaluation proposées, et, d'autre part, les performances de PRESS

[1] Les trois algorithmes de clustering considérés sont décrits dans la section 1.1.1.4.
[2] Au sujet de la *cue validity*, voir la section 1.2.1.3 page 37.
[3] On imagine le travail pour 100 jeux artificiels variant selon le pourcentage de données manquantes.

relativement à d'autres algorithmes de clustering. Nous proposons également une étude succinte des performance de notre algorithme en temps machine. Nous rappelons que les jeux artificiels sont générés suivant un processus décrit dans le chapitre 4 à la section 4.2.1.

Si le contraire n'est pas précisé, les valeurs par défaut des variables caractérisant le langage attribut-valeur employé, la taille de la liste taboue et le critère d'aspiration (c'est-à-dire $n_{\mathcal{A}}$, $|T|$ et n_{aspi}) sont respectivement 30, 20 et 10. De plus, nous fixons un poids minimum de $\rho_{pds} = 3$. Précisons que, dans les tableaux de résultat, les valeurs des fonctions d'évaluation et de la plupart des indices (autres que le nombre de descriptions et les scores de compacité-séparation) sont donnés sous la forme de pourcentages. Enfin, les algorithmes de clustering sont systématiquement exécutés sur n_{exe} jeux de données aux caractéristiques similaires afin de pouvoir calculer des moyennes et des variances sur les résultats.

5.2.1 Comparaison des fonctions d'évaluation

L'objectif de ces expérimentations est de comparer les résultats obtenus avec PRESS suivant la fonction utilisée pour évaluer la qualité de l'ensemble de stéréotypes recherchés. Rappelons que ces fonctions, définies dans la section 3.1.6.1, sont chacune basées sur l'une des quatre mesures de comparaison proposées dans la section 3.1.4. Ces mesures sont inspirées de celles proposées dans la littérature pour traiter le cas des données catégorielles[4].

5.2.1.1 Protocole d'expérimentations

Deux types de jeux de données artificiels sont considérés dans ces expérimentations :

- **ART-1 :** $n_I = 3$ – $dup = 50$ – $m = 80\%$ – $n_E = 150$.
- **ART-2 :** $n_I = 5$ – $dup = 50$ – $m = 90\%$ – $n_E = 250$.

Chacun de ces "patrons" est utilisé pour générer un ensemble de jeux de données aux caractéristiques similaires sur lesquels sont lancés les algorithmes de clustering. La recherche est effectuée en faisant varier la fonction d'évaluation q_N utilisée. Le nombre d'exécution est fixé à $n_{exe} = 50$ et le nombre maximum d'itérations à $maxIter = 300$.

La comparaison est effectuée en fonction des indices suivant :

- Le nombre de descriptions et le taux de couverture : n_D et $couv$.
- La capacité prédictive et l'adéquation relative à E : $pred$ et $adeq_E$.
- L'erreur de classification : err_C.
- L'adéquation, la contradiction et la perte relatives à I : $adeq_I$, $cont_I$ et $perte_I$.
- La compacité des clusters : cmp_1 et cmp_2.

Notons qu'à ce stade les indices de séparation sep_1 et sep_2 ne sont pas pertinents dans le sens où la contrainte de non redondance nous garantit justement une parfaite séparation entre les stéréotypes.

[4] Au sujet des mesures de ressemblance pour les données catégorielles, voir la section 1.1.3.3 page 24.

	ART-1				ART-2			
	q_{N_1}	q_{N_2}	q_{N_3}	q_{N_4}	q_{N_1}	q_{N_2}	q_{N_3}	q_{N_4}
n_D	3.04	25.6	24.2	3.66	5.57	37.12	37.82	5.8
couv	97.91	100	100	98.01	97.77	100	99.99	96.74
pred	65.92	6.54	9.66	61.85	44.1	4.03	4.55	43.26
$adeq_E$	67.13	28.22	32.34	64.26	57.05	37.52	39.5	54.12
err_C	7.3	24.22	22.24	9.16	65.77	50.64	50.28	67.56
$adeq_I$	99.42	7.94	11.58	91.2	42.9	5.44	5.85	38.58
$cont_I$	0	0.12	0.06	0.72	5.85	0.02	0.11	8.66
$perte_I$	0.58	92.04	88.36	8.73	51.25	94.54	94.04	52.76
cmp_1	0.8702	0.9437	0.9355	0.8755	0.9435	0.9625	0.9605	0.9462
cmp_2	0.8281	0.6958	0.6794	0.8245	0.8794	0.5797	0.5773	0.8782

FIG. 5.1 – Résultats obtenus par PRESS avec les fonctions d'évaluation q_{N_1} à q_{N_4}.

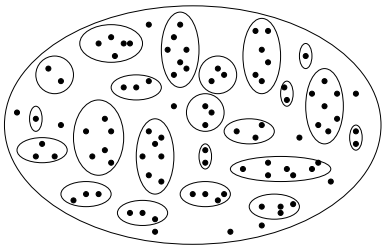


FIG. 5.2 – Type de configuration obtenue en utilisant q_{N_2} et q_{N_3}.

5.2.1.2 Résultats

Le tableau 5.1 présente la moyenne des résultats obtenus par PRESS avec les jeux de données ART-1 et ART-2. Les résultats complets, incluant les valeurs de variance, sont donnés en annexe page 192.

• **Deux familles de fonction :** On constate tout d'abord une dichotomie nette entre les résultats obtenus, d'une part, avec les fonctions q_{N_1} et q_{N_4}, et, d'autre part, avec les fonctions q_{N_2} et q_{N_3}. Alors que les premières permettent d'extraire un nombre de stéréotypes n_D proche de celui des descriptions initiales (3.04 et 3.66 pour $n_I = 3$; 5.57 et 5.8 pour $n_I = 5$), les secondes multiplient le nombre de stéréotypes (25.6 et 24.2 pour $n_I = 3$; 37.12 et 37.82 pour $n_I = 5$). En sus, les indices d'adéquation $adeq_I$ et $adeq_E$, ainsi que l'indice de prédiction *pred*, indiquent que ces stéréotypes sont pauvres, non seulement pour représenter les données E mais également pour rendre compte des descriptions initiales (on note des scores de prédiction et d'adéquation relative à I nettement inférieurs). Les fonctions q_{N_2} et q_{N_3} amènent à construire un grand nombre de clusters de très petite taille dont le stéréotype associé est très proche de quelques exemples et présente en conséquence une description bien pauvre. Cette situation est illustrée par la figure 5.2.

Le score bien meilleur obtenu par l'indice de compacité $comp_2$ s'explique du fait qu'il

se base, tout comme la fonction q_{N_3}, sur une mesure de type Jaccard. Cette mesure ne prenant en compte que les attributs décrits par les deux descriptions comparées, elle favorise naturellement ce genre de configuration. C'est pourquoi nous écartons dès à présent les indices de séparation et de compacité sep_2 et cmp_2 dans l'étude de nos résultats. Nous nous apercevons également que la mesure de Kendall amène au même type de conclusion car elle défavorise les stéréotypes aux descriptions trop riches.

• **Comparaison q_{N_1}/q_{N_4} :** Intéressons-nous à présent aux résultats obtenus entre les fonctions q_{N_1} et q_{N_4}. Ils montrent clairement que la première fonction est préférable en tout point à la seconde, et ce quel que soit le type de jeu de données. En effet, les stéréotypes obtenus à partir de la fonction basée sur M_{N_1} sont plus riches ($pred$), plus représentatifs des données ($adeq_E$ et cmp_1) et davantage fidèles aux descriptions d'origine ($adeq_I$, $cont_I$ et $perte_I$). De plus, l'erreur de classification (err_C) est moindre.

5.2.1.3 Conclusions

Ces premières expérimentations nous amènent à deux conclusions :

1. Tout d'abord, la fonction d'évaluation q_{N_1} présente clairement des résultats plus satisfaisants que les trois autres et sera dorénavant systématiquement utilisée pour les expérimentations.
2. Ensuite, nous pouvons affirmer que la mesure définie à partir de la mesure de Jaccard n'est pas adaptée au cadre des données lacunaires, en tout cas si l'objectif est d'obtenir quelques descriptions très riches et non une multitudes de descriptions ne représentant qu'une poignée d'exemples. C'est pourquoi nous pensons légitime d'écarter de nos futures expérimentations les mesures de séparation et de compacité sep_2 et cmp_2.

5.2.2 Comparaison des techniques T_1, T_2, T_3 et T_4

L'objectif de ces nouvelles expérimentations est de comparer, à partir d'une même base d'exemples et d'un même algorithme de clustering, les quatre techniques d'extraction présentées dans la partie 4.1.4. Rappelons qu'il s'agit de techniques permettant d'extraire, à partir d'une catégorisation donnée en sortie d'un algorithme comme EM ou les k-modes, un ensemble de descriptions étiquetant les clusters. Nous parlons également de "représentants" du clustering ou des données.

5.2.2.1 Protocole d'expérimentations

Nous utilisons une nouvelle fois les deux types de jeux de données ART-1 et ART-2 tels qu'ils ont été définis dans la partie précédente, auquel nous ajoutons un nouveau jeu de données :

- **ART-1 :** $n_I = 3$ – $dup = 50$ – $m = 80\%$ – $n_E = 150$.
- **ART-2 :** $n_I = 5$ – $dup = 50$ – $m = 90\%$ – $n_E = 250$.

	ART-1				ART-2			
	T_1	T_2	T_3	T_4	T_1	T_2	T_3	T_4
$adeq_E$	88.18	66.58	68.29	66.05	69.77	15.54	64.06	14.57
$pred$	99.98	66.26	67.37	65.5	99.93	15.41	91.48	14.08
$cont_E$	11.82	0	2.77	0	30.23	0	29.11	0
$cont_s$	46.36	0	14.44	0	56.72	0	54.22	0
$cont_p$	23.36	0	17.41	0	48.92	0	48.21	0
$adeq_I$	99.99	97.76	90.95	97.45	68.99	17.77	65.05	17.01
$cont_I$	0	0	0	0	30.91	1.45	28.43	1.21
$perte_I$	0.01	2.24	9.05	2.55	0.1	80.78	6.51	81.79
sep_1	0.799	0.6114	0.6065	0.6065	0.6794	0.6164	0.6065	0.6065
cmp_1	0.8311	0.8713	0.8682	0.8723	0.9317	0.9847	0.9372	0.9857

FIG. 5.3 – Comparaison des techniques d'extraction T_1 à T_4 avec l'algorithme EM.

- **ART-3 :** $n_I = 3$ – $dup = 50$ – $n_E = 150$.

Ce dernier est utilisé afin d'observer l'évolution des scores de cohérence en fonction du pourcentage m de données manquantes. La comparaison est effectuée, pour chacun des trois algorithmes EM, COBWEB et k-modes, en utilisant les quatre techniques T_1, T_2, T_3 et T_4 permettant d'extraire des représentants D.

La comparaison est effectuée en fonction des indices suivant :

- Les scores d'adéquation relative à E et de prédiction : $adeq_E$ et $pred$.
- Les scores de cohérence relatifs à E : $cont_E$, $cont_s$ et $cont_p$.
- L'adéquation, la contradiction et la perte relatives à I : $adeq_I$, $cont_I$ et $perte_I$.
- Les scores de séparation et de compacité : sep_1 et cmp_1.

Des indices comme le nombre de représentants n_D ou l'erreur de classification err_C ne sont pour le moment pas pertinents car nous ne comparons pas les algorithmes de clustering entre eux.

5.2.2.2 Résultats

Le tableau 5.3 présente la moyenne des résultats obtenus par EM avec les jeux de données ART-1 et ART-2. Les résultats complets obtenus par les trois algorithmes sont donnés en annexe pages 193 à 195.

- **La cohérence :** Le premier résultat remarquable concerne les indices de cohérence $cont_E$, $cont_s$ et $cont_p$. En effet, les techniques T_2 et T_4 impliquent nécessairement une absence de contradiction entre D et E. *A contrario*, les techniques T_1 et T_3 entraînent des scores de contradiction qui peuvent devenir très élevés. L'exemple le plus frappant est celui de l'algorithme des k-modes et de la technique T_1 avec lesquels 83% des exemples ont presque la moitié de leur description en contradiction avec le centroïde correspondant. Cela signifie que les descriptions qui ont été extraites à partir de la partition fournie par l'algorithme ne sont pas cohérentes avec les données qu'elles couvrent. Même si on peut avoir l'impression que la

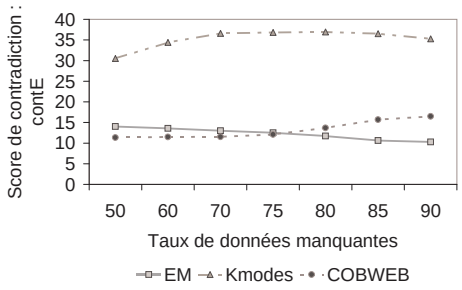


FIG. 5.4 – Variation de $cont_E$ en fonction de m pour ART-3 et T_1.

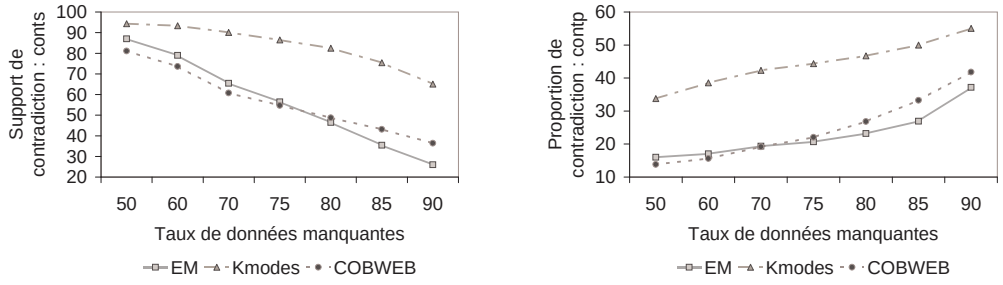


FIG. 5.5 – Variation de $cont_s$ et $cont_p$ en fonction de m pour ART-3 et T_1.

technique T_3 apporte moins de contradictions que T_1 (2.77% contre 11.82% en considérant ART-1), cet avantage paraît s'effondrer lorsque le taux de données manquantes et le bruit sont plus élevés (29.11% contre 30.23% avec ART-2).

Considérons à présent le jeu de données ART-3 en utilisant uniquement la technique d'extraction T_1. L'évolution du score de contradiction relative aux exemples $adeq_E$ en fonction de m (voir figure 5.4) confirme ce résultat incohérent quelle que soit la valeur de m mais reste difficile à interpréter. Par contre, sa décomposition suivant les scores $cont_s$ et $cont_p$ (voir figure 5.5) se révèle beaucoup plus parlant. On constate effectivement que le support de la contradiction diminue pour tous les algorithmes de clustering, mais qu'en même temps la proportion de descripteurs contradictoires au sein de ces exemples augmente. Ce phénomène est identique quel que soit l'algorithme : plus le pourcentage de valeurs manquantes est important, moins le nombre d'exemples possédant au moins une contradiction est important ; mais ceux qui restent dans le support sont de plus en plus contradictoires. Ceci n'est évidemment pas souhaitable.

• **La séparation :** Intéressons-nous maintenant au score de séparation sep_1. La méthode de partage des descripteurs entre les différentes descriptions de D permet d'obtenir avec T_4 une séparation parfaite indiquée par le score minimal 0.6065. D'un autre côté, la technique

T_2 affiche des scores plus importants traduisant une plus grande redondance de l'information entre les représentants des clusters. Il est clair que la contrainte de non-redondance choisie pour notre algorithme rapproche les résultats obtenus avec la technique T_4 des hypothèses de notre modèle sur lequel est basé PRESS. Cependant, nous préférons conserver la technique T_2 dans les prochaines expérimentations afin d'observer, à titre de comparaison, ce qu'il advient lorsque la contrainte de redondance n'est pas vérifiée.

5.2.2.3 Conclusions

Ces expérimentations nous amènent à tirer deux conclusions :

1. Les descriptions obtenues à l'aide des techniques d'extraction T_1 et T_3 ne sont pas cohérentes avec les données qu'elles sont sensées représenter. De plus, même si la proportion des exemples exhibant des contradictions $cont_s$ diminue avec le taux de données manquantes m, la proportion de contradiction $cont_p$ augmente considérablement dans le même temps. Cette notion de cohérence entre les représentants et les données étant fondamentale dans notre modèle, nous ne pouvons conserver ces techniques si nous souhaitons réaliser une réelle comparaison entre PRESS et les autres algorithmes de clustering.
2. La technique T_2 produit des descriptions possédant un certain taux de redondance illustré par le score de séparation sep_1. Nous continuons cependant d'utiliser les deux techniques T_2 et T_4 dans la suite de nos expérimentations.

5.2.3 Comparaison des algorithmes de clustering

A présent que seule la fonction d'évaluation q_{N_1} est utilisée dans PRESS et que deux des quatre techniques d'extraction ont été mises de côté, nous cherchons à comparer effectivement les performance de notre algorithme au regard de trois algorithmes classiques de clustering : EM, COBWEB et les k-modes. Nous comparons également les résultats obtenus par PRESS avec l'ensemble des descriptions "idéales" I qui sont à l'origine des données artificielles.

5.2.3.1 Protocole d'expérimentations

Nous reprenons le jeu de données ART-3 afin de faire varier le taux de valeurs manquantes, auquel nous rajoutons un nouveau jeu ART-4 afin de tester l'influence du nombre des exemples initiaux :

- **ART-3 :** $n_I = 3$ – $dup = 50$ – $n_E = 150$.
- **ART-4 :** $n_I = 5$ – $dup = 50$ – $n_E = 250$.

La recherche est effectuée en faisant varier le taux de données manquantes m. Pour chaque algorithme de clustering, nous utilisons les techniques T_2 et T_4 afin d'extraire les descriptions D associées aux clusters. Le nombre d'exécution est fixé à $n_{exe} = 100$ et le nombre maximum d'itérations à $maxIter = 300$.

La comparaison est effectuée en fonction des indices suivant :

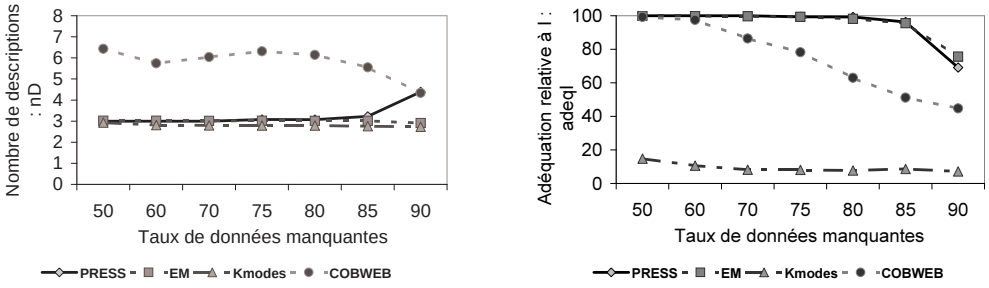


FIG. 5.6 – Variation de n_D et $adeq_I$ en fonction de m pour ART-3 et T_2.

- Le nombre de descriptions : n_D.
- L'adéquation relative à E : $adeq_E$.
- L'adéquation relative à I : $adeq_I$.
- La compacité et la séparation des clusters : cmp_1 et sep_1.
- L'erreur de classification : err_C.

Les indices n_D, $adeq_I$ et err_C permettent de rendre compte de la capacité à retrouver les descriptions initiales I à l'origine des données artificielles. D'un autre côté, $adeq_E$, cmp_1 et sep_1 donnent une évaluation de la qualité des stéréotypes indépendamment d'une classification a priori.

5.2.3.2 Résultats avec ART-3

Les tableaux présentant les résultats détaillés sont donnés en annexe à partir de la page 196. Nous présentons ici les résultats, pour une raison évidente de clarté, sous la forme de graphiques pour lesquels l'abscisse correspond au pourcentage de données manquantes et l'ordonnée à l'indice étudié. Nous commençons par considérer la technique T_2 qui ne garantit pas une séparation parfaite entre les descriptions de D.

- **Utilisation de T_2 :** Les graphiques des figures 5.6, 5.7 et 5.8 présentent l'évolution des différents indices en fonction du pourcentage m de données manquantes.

Le premier graphique (voir la figure 5.6) montre que, par rapport à n_D, les algorithmes les plus robustes sont EM, les k-modes puis PRESS. Rappelons que le nombre de clusters à trouver est l'un des paramètres des k-modes et, par conséquent, que le résultat associé est prévisible. Le processus de cross-validation utilisé pour EM semble lui très efficace puisqu'il permet de trouver quasiment à tous les coups le bon nombre de clusters. PRESS est également très robuste jusque 85% de données manquantes, mais il tend ensuite à surestimer le nombre de clusters. Soulignons que ce phénomène n'est pas nécessairement néfaste lorsque l'on cherche à prédire la valeur d'attributs. De son côté, l'algorithme COBWEB surestime largement le nombre de clusters pour toutes les valeurs de m. Cela ne l'empêche pas d'obtenir de meilleurs résultats que les k-modes avec presque tous les indices. Seule le score de

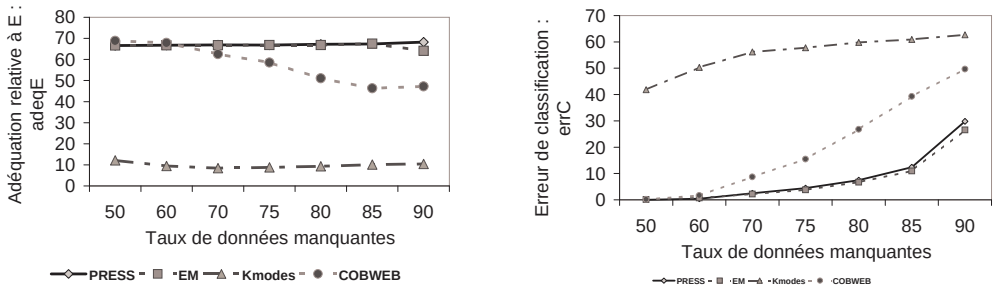


FIG. 5.7 – Variation de $adeq_E$ et err_C en fonction de m pour ART-3 et T_2.

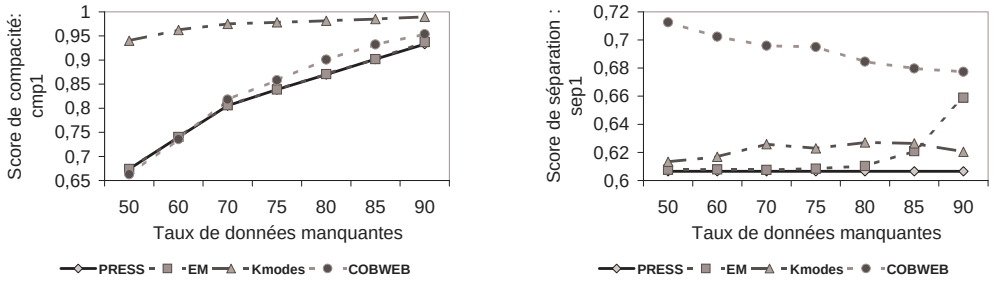


FIG. 5.8 – Variation de cmp_1 et sep_1 en fonction de m pour ART-3 et T_2.

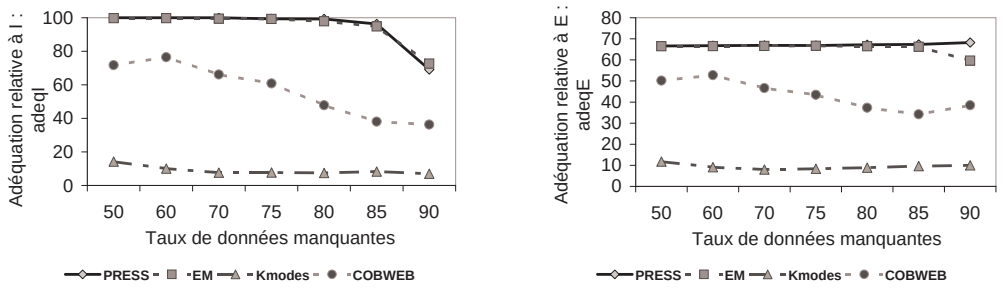


FIG. 5.9 – Variation de $adeq_I$ et $adeq_E$ en fonction de m pour ART-3 et T_4.

séparation est, comme on pouvait s'y attendre, plus mauvais du fait du nombre important de clusters.

Les graphiques des figures 5.6 et 5.7 permettent de nous rendre compte que le nombre de clusters, et donc de représentants, ne détermine pas totalement la qualité du résultat. Ainsi, les scores d'adéquation $adeq_I$ obtenus avec EM et PRESS dépassent largement ceux des k-modes et de COBWEB. Les descripteurs provenant des descriptions initiales I sont très bien retrouvés jusqu'à 85% et le résultat se détériore pour 90%, tout en demeurant acceptable. Ces résultats sont conformes à ceux obtenus pour l'erreur de classification err_C. Concernant l'adéquation $adeq_E$, on remarque que les résultats restent très stables, autour de 70% de descripteurs, et que le nombre plus élevés de clusters n'est pas un handicap concernant leur homogénéité interne.

Les résultats de compacité (voir la figure 5.8) confirment ce qui a été obtenu avec les scores d'adéquation. Le score obtenu par PRESS pour la séparation est constant et égale à 0.6065 du fait de la contrainte de non-redondance. En comparaison, les descriptions trouvées par les autres algorithmes ont une information plus redondante. C'est particulièrement le cas avec EM dont le score augmente fortement à partir de 80% de données manquantes. Il faut considérer la technique d'extraction T_4 si l'on souhaite comparer PRESS avec des descriptions respectant la même contrainte.

• **Utilisation de T_4 :** Les figures 5.9 et 5.10 donnent les résultats obtenus lorsque l'on considère la technique d'extraction T_4 qui permet d'assurer une séparation parfaite entre les descriptions de D.

On constate une très légère baisse des performance de EM et des k-modes si l'on prend en compte les indices d'adéquation $adeq_I$ et $adeq_E$ (voir la figure 5.9). Ceux de COBWEB sont, par contre, sérieusement dégradés. La compacité, elle-aussi, subit une augmentation importante de presque 0.1 (voir la figure 5.10). Cela est dû, une fois de plus, au nombre de représentants n_D qui est plus important que chez les autres algorithmes de clustering.

• **Comparaison avec les descriptions initiales I :** Il s'agit ici de comparer les résultats obtenus par les ensembles de stéréotypes découverts par PRESS avec ceux obtenus par les

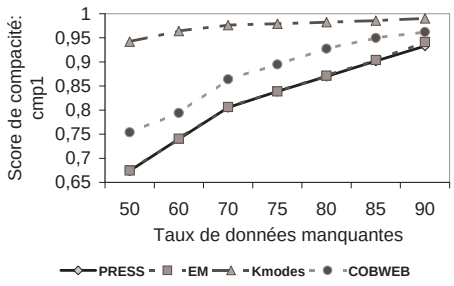


FIG. 5.10 – Variation de cmp_1 en fonction de m pour ART-3 et T_4.

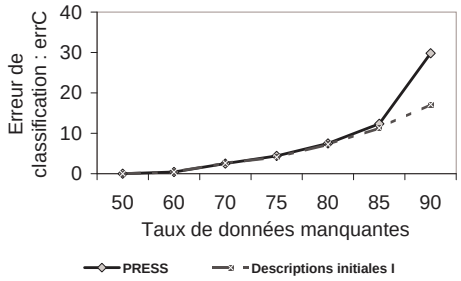


FIG. 5.11 – Comparaison de err_C pour ART-3.

descriptions initiales I. L'ensemble I est considéré comme l'ensemble "idéal" car il a généré les données artificielles. Mais attention, cela ne signifie pas qu'il s'agit de l'ensemble qui obtient le meilleur score avec notre fonction d'évaluation. C'est pourquoi, même si la valeur de certains indices sont nécessairement fixés par avance ($adeq_I = 100\%$, $n_D = 3$ et $sep_1 = 0.6065$), les valeurs pour l'erreur de classification err_C, l'adéquation $adeq_E$ et la compacité cmp_1 doivent être calculés. Les descriptions initiales ayant été dégradés (complétion avec des descripteurs aléatoires, suppression de descripteurs pour obtenir un taux m de données manquantes), nous ne pouvons espérer obtenir des scores parfaits (comme 0% pour err_C) même si $D = I$. Par contre, ces valeurs peuvent être considérées, d'une certaine manière, comme un bon repère auquel comparer les valeurs obtenues par notre propre algorithme.

Les résultats sont donnés par les graphiques 5.11 et 5.12. On constate que ces résultats sont quasiment identiques jusqu'à 85%. A 90%, les résultats diffèrent par rapport aux indices err_C et $adeq_E$, ce qui est probablement dû à la surestimation du nombre de clusters. Par contre, l'homogénéité interne aux clusters, représentée par le score de compacité, est, pour ainsi dire, identique quelle que soit la valeur de m.

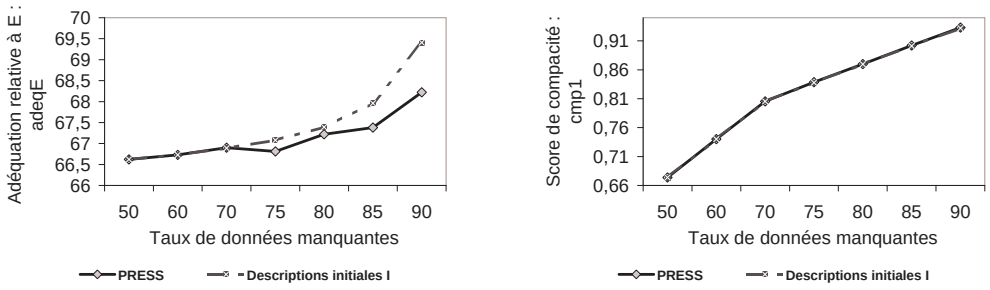


FIG. 5.12 – Comparaison de $adeqE$ et cmp_1 pour ART-3.

5.2.3.3 Résultats avec ART-4

Les tableaux présentant les résultats détaillés sont donnés en annexe à partir de la page 198. Ils sont illustrés par les figures 5.13, 5.14, 5.15 pour la technique T_2 et 5.16, 5.17 pour la technique T_4. Nous comparons d'abord les algorithmes suivant ces deux techniques avant de regarder les résultats de PRESS relativement aux descriptions idéales I, résultats illustrés par les figures 5.18 et 5.19.

• **Comparaison des algorithmes :** De par la méthode de construction des jeux de données artificiels, ART-4 est une base plus difficile à traiter car présentant davantage de descriptions initiales. Les résultats obtenus laissent constater que PRESS obtient de bien meilleurs scores d'adéquation, aussi bien relativement aux descriptions initiales I qu'aux données elles-mêmes E. Un événement singulier mérite d'être ici souligné. Alors que le premier indice $adeq_I$ décroît régulièrement jusqu'à parvenir à une valeur proche de 43%, le deuxième indice $adeq_E$ croît de son côté jusqu'à presque 60% (voir les figures 5.13 et 5.14). Autrement dit, plus le taux de données manquantes est important, plus les stéréotypes s'éloignent des descriptions initiales pour se rapprocher des données. Cette robustesse est confirmée par le nombre le score de compacité cmp_1 (figure 5.15) qui montre que nous obtenons des clusters bien homogènes, ainsi que par l'erreur de classification err_C (figure 5.14).

Notons les résultats obtenus concernant le nombre de représentants n_D (voir la figure 5.13). Alors que PRESS découvre un nombre de stéréotypes légèrement plus important, COBWEB a tendance à surestimer largement ce nombre lorsque m est faible et à réduire l'écart avec son accroissement. Ce dernier retrouve d'ailleurs plutôt bien les descriptions initiales si l'on s'intéresse à l'indice $adeq_I$ et à de fortes valeurs de m. EM, quant à lui, présente de manière plus marquée le défaut à peine visible sur ART-3 de sous-estimation du nombre de clusters. Cela se ressent également sur les mauvais résultats qu'il obtient concernant les scores d'adéquation $adeq_E$ et $adeq_I$. Comme précédemment, rappelons que le nombre de clusters voulu est donné en paramètre de l'algorithmes des k-modes.

Une fois encore, le score de séparation sep_1 (voir la figure 5.15) souligne le principal défaut de l'algorithme COBWEB qui surestime le nombre de descriptions n_D. Les graphiques des

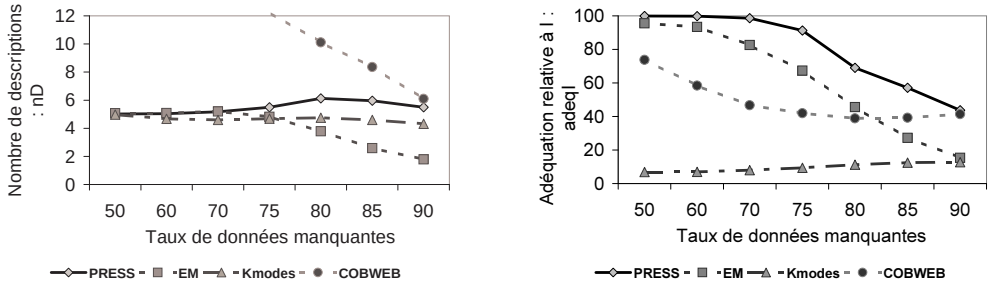


FIG. 5.13 – Variation de n_D et $adeq_I$ en fonction de m pour ART-4 et T_2.

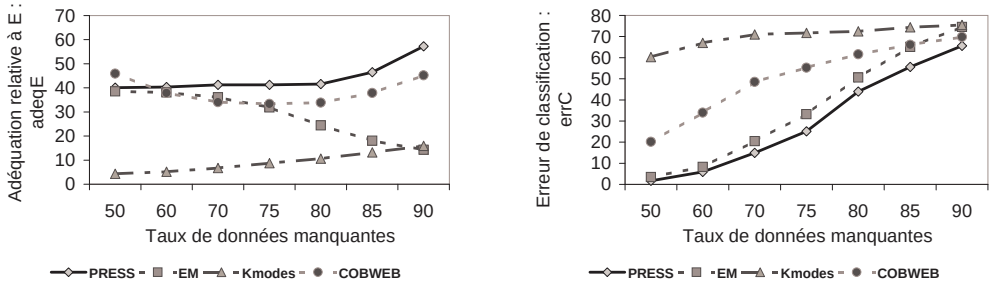


FIG. 5.14 – Variation de $adeq_E$ et err_C en fonction de m pour ART-4 et T_2.

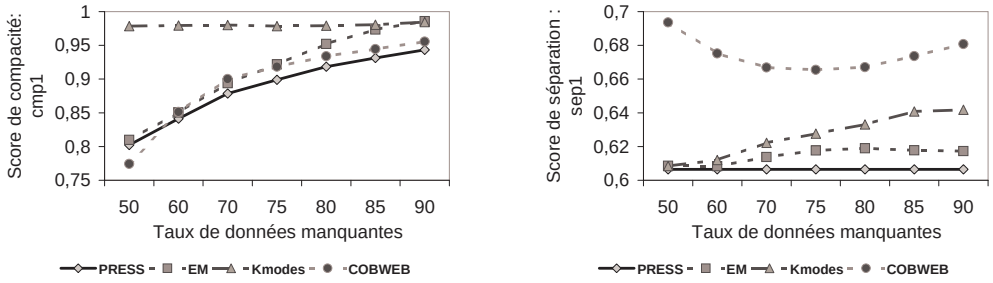


FIG. 5.15 – Variation de cmp_1 et sep_1 en fonction de m pour ART-4 et T_2.

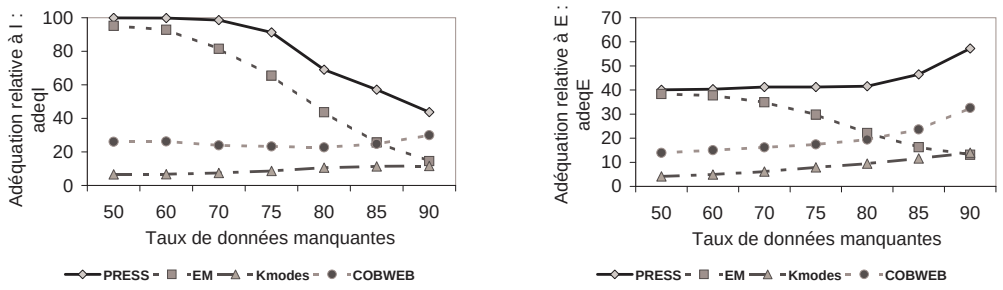


FIG. 5.16 – Variation de $adeq_I$ et $adeq_E$ en fonction de m pour ART-4 et T_4.

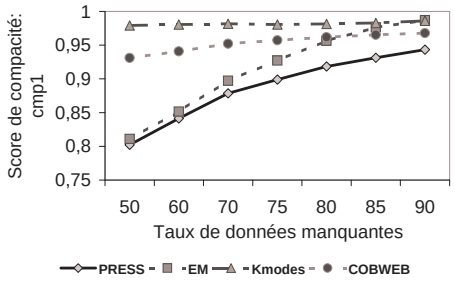


FIG. 5.17 – Variation de cmp_1 en fonction de m pour ART-4 et T_4.

figures 5.16 et 5.17, où la technique d'extraction T_4 remplace T_2, permettent d'arriver à la même conclusion que pour ART-3 : une forte dégradation des résultats de COBWEB dû au nombre de clusters et une très légère baisse des performances pour EM et les k-modes.

- **Comparaison avec les descriptions initiales I :**

Les résultats constatés sur ART-3 concernant la comparaison entre les stéréotypes de PRESS et les descriptions initiales sont confirmés, avec cependant une différence qui survient plus un peu plus tôt au niveau de l'erreur de classification (voir la figure 5.18). Par contre, l'adéquation $adeq_E$ et surtout la compacité cmp_1 suivent un comportement remarquablement similaire (voir la figure 5.19).

5.2.3.4 Conclusions

Les résultats obtenus à la fois pour ART-3 et ART-4 permettent d'arriver aux conclusions suivantes :

1. Les algorithmes COBWEB et les k-modes peuvent tout-de-suite être disqualifiés, à la fois par les expérimentations réalisées sur ART-3, et par celles réalisées sur ART-4. COBWEB a tendance à surestimer largement le nombre de clusters, ce qui se ressent notamment sur le score de séparation. Ce résultat est bien entendu lié au paramètre

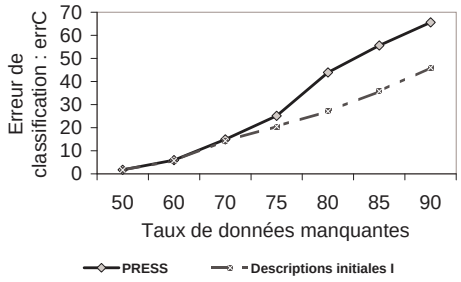


FIG. 5.18 – Comparaison de err_C pour ART-4.

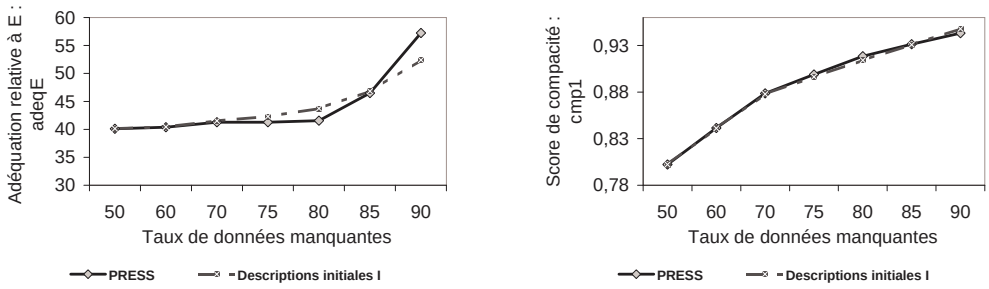


FIG. 5.19 – Comparaison de $adeqE$ et cmp_1 pour ART-4.

de *cut-off* fixé par défaut pour obtenir une partition à partir de la hiérarchie proposée par l'algorithme. Si l'on force la séparation par la technique T_4, les résultats concernant l'adéquation $adeq_I$ et $adeq_E$ sont fortement dégradés. L'algorithme des k-modes obtient nettement les plus mauvais résultats malgré un nombre de clusters donné à l'avance.

2. L'algorithme EM obtient de très bons résultats avec ART-3 mais est mis en défaut par ART-4 où l'on pressent sa tendance à sous-estimer le nombre de clusters. En comparaison, PRESS marque de bien meilleurs résultats en terme d'adéquation et de compacité. De plus, il semble plutôt robuste pour découvrir le bon nombre de représentants.

5.2.4 Variation du nombre de descriptions initiales

Les résultats obtenus dans la partie précédente ont conduit à cette nouvelle série d'expérimentations basée sur la variation du nombre de descriptions initiales n_I. Elle doit permettre de tester la robustesse de notre algorithme au bruit et à l'augmentation du nombre de classes initiales.

5.2.4.1 Protocole d'expérimentations

Nous utilisons le type de jeux de données suivant :

- **ART-5 :** $m = 80\%$ – $dup = 50$.

La recherche est effectuée en faisant varier le nombre de descriptions initiales n_I. Pour chaque algorithme de clustering, nous utilisons les techniques T_2 et T_4 afin d'extraire les descriptions D associées aux clusters. Le nombre d'exécution est fixé à $n_{exe} = 100$ et le nombre maximum d'itérations à $maxIter = 300$.

La comparaison est effectuée en fonction des indices suivant :

- Le nombre de descriptions : n_D.
- L'adéquation relative à E : $adeq_E$.
- L'adéquation relative à I : $adeq_I$.
- La compacité et la séparation des clusters : cmp_1 et sep_1.
- L'erreur de classification : err_C.

5.2.4.2 Résultats

Les tableaux présentant les résultats détaillés sont donnés en annexe à partir de la page 201. Ils sont illustrés par les figures 5.21 à 5.26.

- **Comparaison des algorithmes :** La figure 5.20 montre que PRESS parvient à retrouver approximativement le bon nombre de classes, tandis que COBWEB continue à surestimer ce nombre et que EM confirme la tendance inverse. On peut supposer que ces mauvais résultats obtenus par EM proviennent de la technique de cross-validation qui n'est pas adaptée à la situation. Nous avons cependant pu vérifier à plusieurs reprises que, tout comme l'algorithme des k-modes, préciser le bon nombre de clusters dès le départ ne conduit pas nécessairement

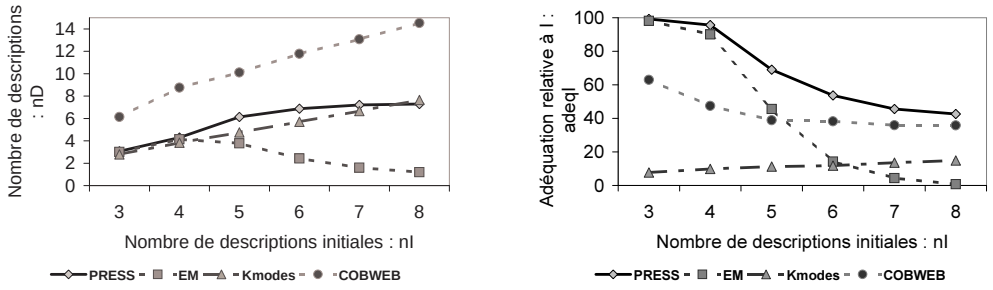


FIG. 5.20 – Variation de n_D et $adeq_I$ en fonction de n_I avec T_2.

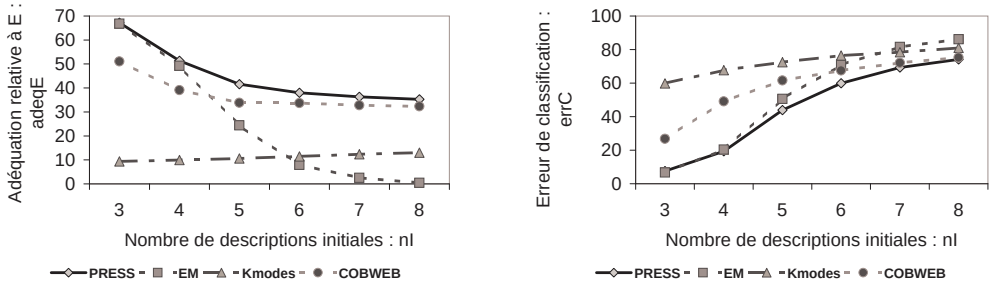


FIG. 5.21 – Variation de $adeq_E$ et err_C en fonction de n_I avec T_2.

à une amélioration des performances. En effet, il se trouve que de nombreux clusters donnés en sortie sont vides, ce qui revient finalement à obtenir les mêmes résultats.

Les figures 5.21 et 5.22 montrent que PRESS est beaucoup plus robustes que les autres algorithmes présentés en terme d'adéquation et d'homogénéité, et ce malgré une proportion élevée de descripteurs bruités.

Si maintenant on force la séparation, on obtient les résultats présentés dans les figures 5.23 et 5.24. Comme précédemment, cette opération est surtout néfaste concernant les performances de l'algorithme COBWEB qui, sans cela, est presque aussi robuste que PRESS avec un taux de données manquantes élevé.

• **Comparaison avec les descriptions initiales I :** Les résultats sont illustrés par les graphiques 5.25 et 5.26.

Excepté l'erreur de classification qui, une fois encore, est bien plus élevée avec les stéréotypes de PRESS qu'avec I, les résultats sont quasiment identiques en ce qui concerne les indices d'adéquation $adeq_E$ et la compacité cmp_1. Même si la distribution initiale des classes n'est pas toujours bien retrouvée, les stéréotypes découverts impliquent une homogénéité presque idéale relativement à l'ensemble des descriptions à l'origine des données artificielles.

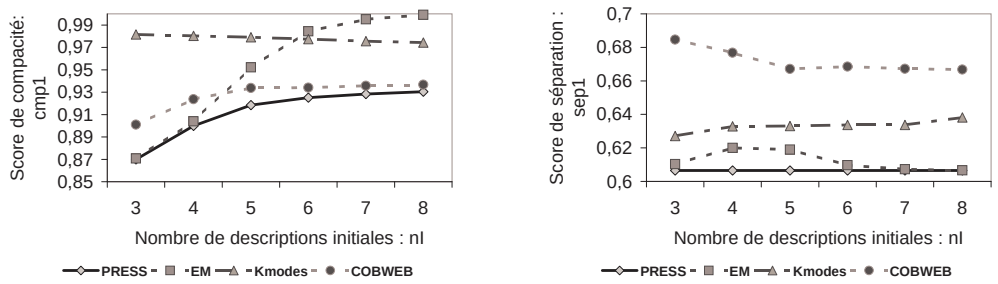


FIG. 5.22 – Variation de cmp_1 et sep_1 en fonction de n_I avec T_2.

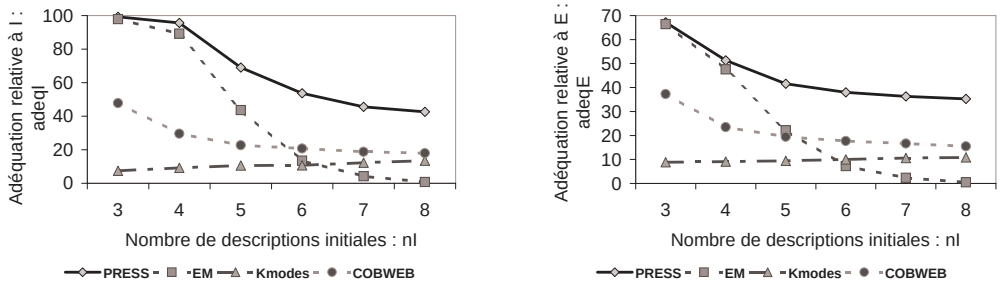


FIG. 5.23 – Variation de $adeq_I$ et $adeq_E$ en fonction de n_I avec T_4.

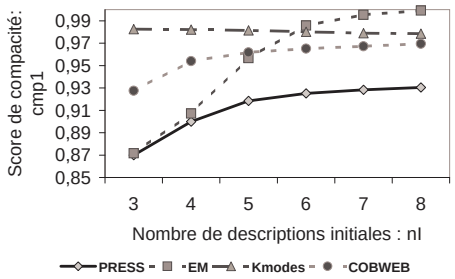


FIG. 5.24 – Variation de cmp_1 en fonction de n_I avec T_4.

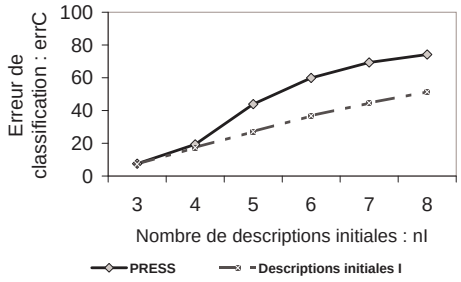


FIG. 5.25 – Comparaison de err_C pour ART-5.

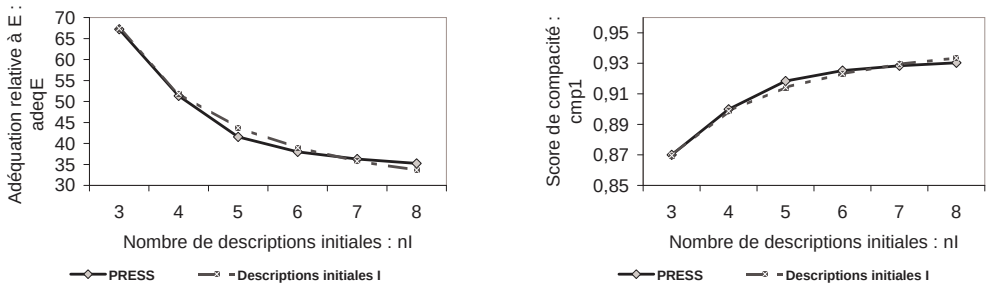


FIG. 5.26 – Comparaison de $adeqE$ et cmp_1 pour ART-5.

5.2.4.3 Conclusions

Ces expérimentations montrent que l'algorithme de classification par défaut est très robuste, relativement au taux de données manquantes, au nombre de descriptions initiales et au bruit provoqué par l'ajout de descripteurs aléatoires. Si l'on compare ces résultats aux résultats "parfaits" associés à l'ensemble I, on peut même ajouter que les stéréotypes obtenus sont quasiment les meilleurs possibles au vu de la dégradation des données. En tout cas, il s'agit de descriptions cohérentes associées à des clusters homogènes et caractérisées par une séparation parfaite.

5.2.5 Comparaison avec une méthode simple d'imputation

Dans cette partie, nous utilisons une méthode d'imputation basée sur les k-plus proches voisins. Elle consiste à calculer, pour chaque exemple, les k exemples les plus similaires (les voisins) et à les utiliser afin de compléter les valeurs qui manquent dans sa description. Si $k = 1$, il s'agit de la méthode appelée *hot-deck imputation*. Si k correspond à tous les autres exemples de E, c'est-à-dire si $k = n_E - 1$, cela revient à utiliser les descripteurs les plus fréquents pour compléter les "trous". Nous comparons ensuite les résultats obtenus à l'aide de PRESS, d'une part à partir de la base initiale, et, d'autre part, à partir de la base complétée avec cette méthode d'imputation.

5.2.5.1 Protocole d'expérimentations

Nous utilisons à nouveau les deux types de jeux de données ART-2 et ART-4 :

- **ART-2 :** $n_I = 5$ – $dup = 50$ – $m = 90\%$ – $n_E = 250$.
- **ART-4 :** $n_I = 5$ – $dup = 50$ – $n_E = 250$.

Le premier est utilisé afin d'étudier les résultats lorsque l'on fait varier la taille k du voisinage entre 1 (*hot-desk imputation*) et 249 (fréquence sur E tout entier). Le second est utilisé afin, une fois la taille du voisinage fixée à $k = 10$, d'observer l'évolution des différents indices en fonction du pourcentage de données manquantes. Le nombre d'exécution pour PRESS est fixé à $n_{exe} = 50$ et le nombre maximum d'itérations à $maxIter = 300$.

La comparaison est effectuée en fonction des indices suivant :

- Le nombre de descriptions : n_D.
- L'erreur de classification : err_C.
- L'adéquation relative à I : $adeq_I$.
- L'adéquation relative à E : $adeq_E$.

5.2.5.2 Résultats

Le tableau 5.27 présente la moyenne des résultats obtenus en faisant varier la valeur de k. Les résultats complets indiquant les variances sont donnés en annexe page 204.

Nous pouvons constater que ces méthodes basiques d'imputation amènent à une dégradation des résultats obtenus, et ce quel que soit l'indice considéré et quelle que soit la taille du

	sans	k=1	k=2	k=5	k=10	k=50	k=249
n_D	5.5	6.64	8.72	7.78	7.24	5.88	7.04
err_C	65.6	67.34	67.5	69.9	70.38	70.96	72.22
$adeq_I$	43.73	37.31	30.4	34.45	33.68	35.36	32.12
$adeq_E$	57.24	48.25	34.94	37.67	41.72	43.47	45.98

FIG. 5.27 – Résultats obtenus avec PRESS en utilisant les k-plus-proches voisins.

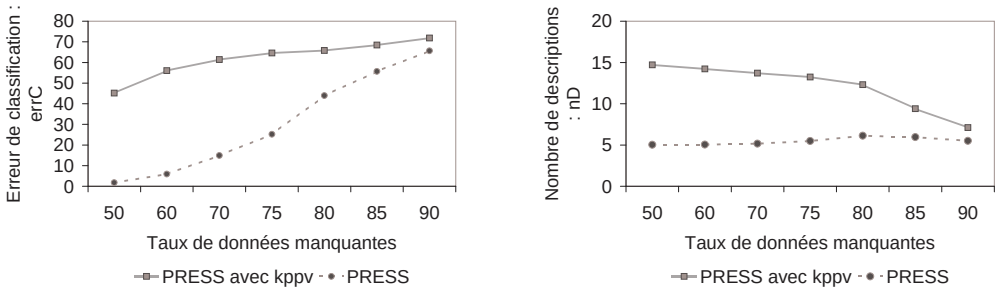


FIG. 5.28 – Variation de err_C et n_D en fonction de m.

voisinage choisie. Fixons maintenant la taille du voisinage à $k = 10$ et observons le comportement de ces indices en fonction du taux de données manquantes m. Le résultat est présenté dans les figures 5.28 et 5.29 qui confrontent les résultats obtenus, d'une part en utilisant les k-plus-proches voisins, d'autre part sans utiliser cette méthode d'imputation.

Ces graphiques confirment la dégradation obtenue lorsque l'on applique la méthode des k-plus-proches voisins avant d'utiliser notre algorithme de classification par défaut. Notons la diminution considérable de l'écart entre les deux courbes lorsque m devient élevé. Tout porte à croire que, moins il y a de descripteurs, moins la méthode d'imputation peut s'appliquer et plus cela revient à utiliser directement PRESS sans compléter la description des exemples à l'avance.

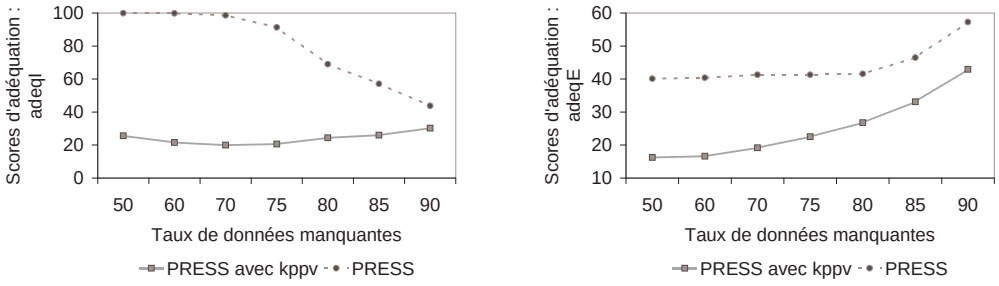


FIG. 5.29 – Variation de $adeq_I$ et $adeq_E$ en fonction de m.

	0	5	10	15	20	25
q_{N_1}	5.58	5.63	5.72	5.77	5.79	5.78
n_D	5.62	5.71	5.56	5.35	5.37	5.38
err_C	66.81	65.98	65.56	65.87	65.12	65.39
$adeq_I$	42.9	44.42	44.92	44.32	44.85	45.2

FIG. 5.30 – Résultats obtenus avec PRESS en faisant varier la taille de la liste taboue.

5.2.5.3 Conclusions

Les résultats obtenus confirment ce que l'on peut habituellement trouver dans la littérature relative au traitement des valeurs manquantes. Les méthodes basiques d'imputation, comme les k-proches-voisins ou la méthode de la moyenne (la fréquence dans le cas des données catégorielles), ne sont pas adaptées à un fort taux de données lacunaires et, par conséquent, sont à proscrire.

5.2.6 Analyse sur la recherche taboue

Les expérimentations suivantes ont été réalisées afin de mettre en évidence l'apport de l'utilisation d'une liste taboue dans le processus de recherche local. En plus de permettre un gain considérable en temps de calcul, cette implémentation de la recherche taboue, même si elle reste très sommaire, doit également nous permettre d'échapper aux optima locaux pour trouver de meilleurs résultats pour la fonction d'évaluation utilisée.

5.2.6.1 Protocole d'expérimentations

Nous utilisons une nouvelle fois le jeu de données ART-2 :

- **ART-2 :** $n_I = 5$ – $dup = 50$ – $m = 90\%$ – $n_E = 250$.

Le nombre d'exécution pour PRESS est fixé à $n_{exe} = 100$ et le nombre maximum d'itérations à $maxIter = 300$. La recherche est effectuée en faisant varier la taille $|T|$ de la liste taboue. Remarquons qu'une taille de 0 correspond à une stratégie de type *hill-climbing*.

La comparaison est effectuée en fonction des indices suivant :

- Le score obtenu par la fonction d'évaluation : q_{N_1}.
- Le nombre de descriptions : n_D.
- L'erreur de classification : err_C.
- L'adéquation relative à I : $adeq_I$.

5.2.6.2 Résultats

Le tableau 5.30 présente la moyenne des résultats obtenus en faisant varier la valeur de $|T|$. Ils sont illustrés par les figures 5.31 et 5.32 qui montre l'évolution de chacun des indices. Les résultats complets indiquant les variances sont donnés en annexe page 204.

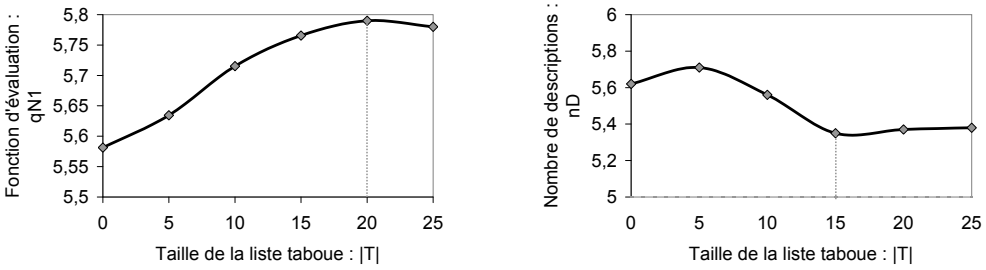


FIG. 5.31 – Variation de q_{N_1} et n_D en fonction de $|T|$.

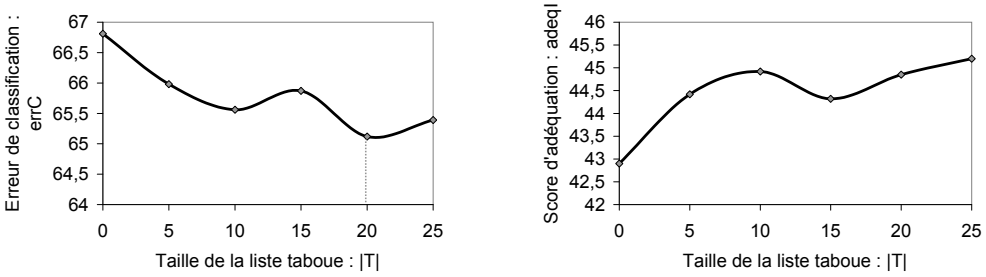


FIG. 5.32 – Variation de err_C et $adeq_I$ en fonction de $|T|$.

La première constatation est que nous parvenons effectivement à trouver des valeurs supérieures pour la fonction d'évaluation par rapport à celles découvertes par une recherche locale simple ($|T| = 0$). Cependant, le gain réalisé est en deça de ce que nous espérions. Le problème n'est peut-être pas suffisamment complexe pour tirer pleinement partie de la stratégie de recherche taboue concernant une amélioration du résultat.

La deuxième constatation est qu'il s'agit de trouver une bonne valeur pour $|T|$. En effet, une valeur trop basse revient à privilégier l'intensification dans une région connue, alors qu'une valeur trop élevée conduit à une diversification importante qui ne mène pas aux résultats espérés. Bien entendu, cette valeur est clairement dépendante de la taille de notre langage de description. Nous avons choisi d'utiliser la valeur $|T| = 20$ qui semble raisonnable pour un langage de 30 attributs. Une étude plus approfondie du choix de cette variable peut faire l'objet de futurs travaux. Quelques expérimentations, réalisées en ce sens en considérant une stratégie de recherche taboue réactive[5], nous ont semblé prometteuses mais ne sont pas suffisamment abouties pour figurer dans cette thèse.

5.2.6.3 Conclusions

La recherche taboue permet effectivement de trouver des résultats au-delà des optima locaux découverts par une recherche locale basique. Les résultats ne montrent cependant pas un gain important de la fonction d'évaluation et posent le problème du choix de la taille de la liste. L'intérêt de l'heuristique réside donc essentiellement dans sa capacité à éliminer un grand nombre de calculs (20 attributs sur 30 lorsque $|T| = 20$ et $n_{\mathcal{A}} = 30$) à chaque étape de la recherche, et donc à améliorer considérablement les performances en temps machine de PRESS.

5.2.7 Etude des performances

Nous souhaitons maintenant évaluer notre algorithme en fonction, d'une part, de la variation du nombre d'exemples à classifier, et, d'autre part, de la taille du langage utilisé, c'est-à-dire du nombre d'attribut utilisés pour coder les exemples. Ces expérimentations ont été réalisées sur un ordinateur doté d'un processeur Pentium IV 3.4Ghz et d'une mémoire vive de 2Ghz.

5.2.7.1 Protocole d'expérimentations

Deux types de jeux de données artificiels sont considérés dans ces expérimentations :

- **ART-6 :** $n_I = 3$; $m = 80\%$.
- **ART-7 :** $n_I = 3$; $dup = 50$; $m = 80\%$; $n_E = 150$.

La recherche est effectuée sur ART-6 en faisant varier la variable *dup*, et donc le nombre d'exemples n_E. Pour ART-7, on fait varier le nombre d'attributs $n_{\mathcal{A}}$. Le nombre d'exécution

[5] Cette extension de la recherche taboue a été proposée par R. Battiti et G. Tecchiolli [BT94].

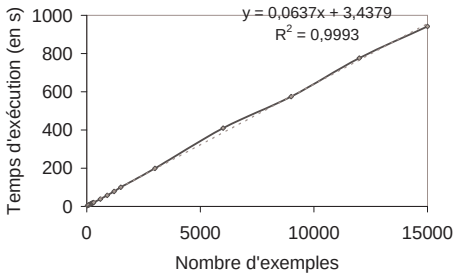


FIG. 5.33 – Temps moyen d'une exécution en fonction de n_E.

est fixé à $n_{exe} = 10$ et le nombre maximum d'itérations à $maxIter = 200$. La comparaison est effectuée en fonction du temps d'exécution de l'algorithme exprimé en secondes.

5.2.7.2 Résultats

Les tableaux de résultats complets sont donnés en annexe page 205.

- **En fonction du nombre d'exemples :**

La figure 5.33 montre le temps moyen (en secondes) mis pour exécuter notre algorithme en fonction du nombre d'exemples n_E à classifier.

Nous pouvons constater que le temps est linéaire par rapport au nombre d'exemples. Cela confirme les calculs de complexité réalisé dans la partie 3.1.6.5. La droite n'est pas parfaite du fait des conditions d'expérimentations[6] et du faible nombre d'exécutions réalisées. Malgré cela, les résultats correspondent bien à ce que nous attendions.

- **En fonction du nombre d'attributs :**

La figure 5.34 montre le temps moyen (en secondes) mis pour exécuter notre algorithme en fonction du nombre d'attributs $n_{\mathcal{A}}$ utilisés pour coder notre langage.

La courbe obtenue suggère un temps basé sur le cube du nombre d'attributs. Cela correspond, là aussi, à la complexité théorique calculée dans la partie 3.1.6.5.

5.2.7.3 Conclusions

Les expérimentations réalisées, à la fois en faisant varier le nombre d'exemples n_E et le nombre d'attributs $n_{\mathcal{A}}$, donnent des résultats correspondant aux calculs théoriques qui ont été faits sur la complexité de notre algorithme à la section 3.1.6.5.

[6] Comment être sûr, en effet, qu'aucun autre processus ne vient se superposer au calcul de notre algorithme ?

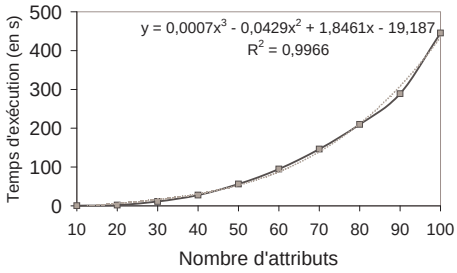


FIG. 5.34 – Temps moyen d'une exécution en fonction de $n_{\mathcal{A}}$.

5.3 Expérimentations sur les données réelles

Nous utilisons dans cette partie les quatre jeux de données tirés d'articles de la presse française à la fin du XIX^{e} siècle : Le Matin, La Libre Parole, La Croix et Le Petit Journal. Voici une brève description de ces bases utilisant nos indices[7] :

- **Le Matin :** $n_{\mathcal{A}} = 33$ – $m = 87.01\%$ – $n_E = 50$ – $\rho(E) = 63$.
- **La Libre Parole :** $n_{\mathcal{A}} = 33$ – $m = 86.3\%$ – $n_E = 63$ – $\rho(E) = 98$.
- **La Croix :** $n_{\mathcal{A}} = 33$ – $m = 89.29\%$ – $n_E = 85$ – $\rho(E) = 114$.
- **Le Petit Journal :** $n_{\mathcal{A}} = 33$ – $m = 84.91\%$ – $n_E = 28$ – $\rho(E) = 53$.

Nous fixons les valeurs suivantes pour les paramètres de notre algorithme :

- Fonction d'évaluation utilisée : q_{N_1}.
- Nombre maximum d'itérations : $maxIter = 200$.
- Fréquence d'aspiration : $n_{aspi} = 10$.
- Taille de la liste taboue : $|T| = 20$.
- Poids minimum des clusters : $\rho_{pds} = 3$.

5.3.1 Comparaison des techniques d'extraction sur la cohérence

La première expérimentation a pour objectif de confirmer les résultats obtenus avec les jeux de données artificiels concernant les quatre techniques d'extraction T_1, T_2, T_3 et T_4. Cela nous permettra d'éliminer les techniques qui mènent à de mauvais résultats pour les prochaines expérimentations.

5.3.1.1 Protocole d'expérimentations

Nous utilisons uniquement le jeu de données Le Matin en comparant les quatre algorithmes de clustering PRESS, EM, COBWEB et les k-modes. La comparaison est effectuée en fonction des indices suivant :

- L'indice global de contradiction : $cont_E$.

[7] Une description détaillée des jeux de données réels est donnée dans la section 4.2.2 page 111.

	PRESS	EM				COBWEB				KModes			
		T_1	T_2	T_3	T_4	T_1	T_2	T_3	T_4	T_1	T_2	T_3	T_4
$cont_E$	0	16	0	16	0	19	0	20	0	13	0	14	0
$cont_s$	0	47	0	47	0	55	0	55	0	30	0	31	0
$cont_p$	0	57	0	59	0	52	0	52	0	46	0	47	0

FIG. 5.35 – Comparaison des indices de cohérence à partir de la base Le Matin.

- Le support de contradiction : $cont_s$.
- La proportion de contradiction au sein du support : $cont_p$.

5.3.1.2 Résultats

Le tableau 5.35 donne les résultats obtenus[8].

Ces résultats sont tout à fait cohérents avec ceux obtenus en utilisant les jeux artificiels. La proportion des descripteurs contradictoires est cependant plus beaucoup plus importante, ce qui nous conforte dans l'idée que les techniques ne prenant pas en compte le critère de cohérence doivent être abandonnées. Remarquons que des résultats similaires sont obtenus avec les autres jeux de données La Libre Parole, La Croix et Le Petit Journal.

5.3.1.3 Conclusions

Les résultats obtenus avec les données artificielles sont confirmés et nous choisissons de ne pas utiliser les techniques T_1 et T_3 dans le cadre de nos expérimentations sur les jeux de données réels.

5.3.2 Analyse comparative

Nous faisons à présent une analyse comparative des résultats obtenus à l'aide des différents algorithmes de clustering. Cette analyse se base, dans un premier temps, sur les indices proposés dans le chapitre 4 afin d'effectuer une comparaison la plus objective possible. Une analyse qualitative des résultats obtenus par PRESS est proposée dans la partie suivante.

5.3.2.1 Protocole d'expérimentations

Nous utilisons cette fois les quatre jeux de données Le Matin, La Libre Parole, La Croix et Le Petit Journal en comparant les quatre algorithmes de clustering PRESS, EM, COBWEB et les k-modes. Les techniques d'extraction T_2 et T_4 sont utilisées pour EM, COBWEB et les k-modes.

La comparaison est effectuée en fonction des indices suivant :

- Le nombre de représentants et le taux de couverture : n_D et $couv$.
- Les capacités prédictives : $pred$.

[8] Ces résultats ont été arrondis à l'unité pour des raisons de lisibilité.

	PRESS	EM		COBWEB		KModes	
		T_2	T_4	T_2	T_4	T_2	T_4
n_D	5	2	2	2	2	5	5
couv	96.83	100	100	100	100	100	100
pred	47.43	55.99	55.84	51.08	45.02	41.85	33.77
$adeq_E$	78.52	62.8	62	56.4	47.2	57.6	50
$perte_E$	21.48	37.2	38	43.6	52.8	42.4	50
sep_1	0.6065	0.6249	0.6065	0.7676	0.6065	0.6266	0.6065
cmp_1	0.9545	0.9749	0.9759	0.9819	0.9932	0.9803	0.9893

FIG. 5.36 – Comparaison des quatre algorithmes de clustering avec Le Matin.

	PRESS	EM		COBWEB		KModes	
		T_2	T_4	T_2	T_4	T_2	T_4
n_D	4	2	2	1	1	4	4
couv	91.84	100	100	100	100	100	100
pred	65.58	68.18	68.18	63.64	63.64	67.53	64.16
$adeq_E$	80.81	63.45	63.45	48.28	48.28	73.1	60.23
$perte_E$	19.19	36.55	36.55	51.72	51.72	26.9	39.77
sep_1	0.6065	0.6065	0.6065	0.6065	0.6065	0.6667	0.6065
cmp_1	0.9583	0.9781	0.9781	1	1	0.9647	0.9828

FIG. 5.37 – Comparaison des quatre algorithmes de clustering avec La Libre Parole.

- Les scores d'adéquation relatifs à E : $adeq_E$ et $perte_E$.
- Les scores de séparation et de compacité : sep_1 et cmp_1.

Nous n'avons pas besoin ici de prendre en compte les indices de cohérence. *cont* étant nécessairement égale à 0, les scores $adeq_E$ et $perte_E$ suffisent à rendre compte de l'adéquation des descriptions de D avec les données E.

5.3.2.2 Résultats

Les tableaux 5.36, 5.37, 5.38 et 5.39 donnent respectivement les résultats obtenus avec les bases Le Matin, La Libre Parole, La Croix et Le Petit Journal.

La première remarque concerne le taux de couverture *couv* qui est nécessairement de 100% pour les trois algorithmes EM, COBWEB et k-modes. En effet, ceux-ci classent nécessairement tous les exemples dans un cluster et ne donne pas la possibilité d'écarter certains exemples afin d'obtenir de meilleurs résultats. PRESS obtient dans trois cas un bon taux de couverture dépassant 90%. Par contre, le taux n'est que de 84.91% en ce qui concerne Le Petit Journal. Si l'on regarde les exemples non classés de plus près, on se rend compte qu'il s'agit d'un ensemble de cas particuliers dont le poids ne justifient pas la création d'un stéréotype à part entière.

La deuxième remarque s'intéresse au nombre de descriptions trouvées n_D qui est systéma-

	PRESS	EM		COBWEB		KModes	
		T_2	T_4	T_2	T_4	T_2	T_4
n_D	4	2	2	2	2	4	4
$couv$	97.37	100	100	100	100	100	100
$pred$	68.98	66.35	65.63	60.45	56.46	59.28	58.21
$adeq_E$	77.92	68.67	67.17	60.15	55.14	63.66	61.4
$perte_E$	22.08	31.33	32.83	39.85	44.86	36.34	38.6
sep_1	0.6065	0.6615	0.6065	0.7505	0.6065	0.6248	0.6065
cmp_1	0.9813	0.9842	0.9859	0.9932	0.9984	0.9896	0.9921

FIG. 5.38 – Comparaison des quatre algorithmes de clustering avec La Croix.

	PRESS	EM		COBWEB		KModes	
		T_2	T_4	T_2	T_4	T_2	T_4
n_D	2	2	2	1	1	2	2
$couv$	84.91	100	100	100	100	100	100
$pred$	39.22	47.68	46.31	39.39	39.39	41.05	41.05
$adeq_E$	84.47	80.77	79.23	40	40	47.69	47.69
$perte_E$	15.53	19.23	20.77	60	60	52.31	52.31
sep_1	0.6065	0.6432	0.6065	0.6065	0.6065	0.6065	0.6065
cmp_1	0.9409	0.9391	0.9414	1	1	0.9882	0.9882

FIG. 5.39 – Comparaison des quatre algorithmes de clustering avec Le Petit Journal.

tiquement faible avec les algorithmes EM et COBWEB. Le nombre trouvé par les k-modes ne peut être pris en compte car il s'agit d'un paramètre d'entrée de l'algorithme. Notons toutefois qu'il est nécessaire de préciser un grand nombre de *seed* de départ (jusqu'à 10 pour LeMatin) si l'on souhaite obtenir en sortie la quantité de clusters (non vides) désirée. Comparativement, PRESS trouve de lui-même un nombre de clusters généralement plus élevé (de 2 pour Le Petit Journal à 5 pour Le Matin).

Une fois passé ce préambule, nous pouvons constater que les exemples sont bien mieux représentés par les stéréotypes découverts par PRESS que par les descriptions produites par EM, COBWEB ou les k-modes. En effet, la proportion des descripteurs retrouvés dans les stéréotypes $adeq_E$ est systématiquement plus important et l'homogénéité interne calculée par $comp_1$ le confirme. De plus, la contrainte de non-redondance implique une séparation parfaite entre les stéréotypes. Remarquons que les résultats obtenus par les autres algorithmes se détériorent légèrement si on force la séparation à l'aide de la technique T_4, comme dans le cas des jeux des données artificiels.

5.3.2.3 Conclusions

Les indices que nous avons définis afin de juger de la validité des résultats le plus objectivement possible nous indiquent que les stéréotypes obtenus avec PRESS sont intéressants car ils assurent à la fois la cohérence entre les descriptions, une bonne homogénéité interne aux clusters et une séparation parfaite entre les stéréotypes. De surcroît, les ensembles de stéréotypes obtenus sont les seuls à vérifier la contrainte de cohésion cognitive qui assure un lien entre les descriptions des exemples.

5.3.3 Analyse qualitative

Nous faisons à présent une analyse plus qualitative des résultats obtenus par PRESS pour les quatre jeux de données considérés. Celle-ci est effectuée à la lumière de nombreuses lectures sur la période [Bre83][Tag02][Mol94][BGGT76] en essayant de rester le plus objectif possible. Bien entendu, il ne s'agit que d'une interprétation qui ne remplace pas les mesures réalisées dans la partie précédente mais donne quelques pistes concernant les stéréotypes découverts. Enfin, nous n'hésitons pas à revenir, lorsque cela s'avère nécessaire, aux sources des stéréotypes, c'est-à-dire aux exemples eux-mêmes.

5.3.3.1 Remarques d'ordre général

Tout d'abord, le nombre de stéréotypes se situe entre 2 (pour Le Petit Journal) à 5 (pour Le Matin). L'un de ces stéréotypes est toujours prépondérant vis-à-vis des autres (entre 60.38% pour Le Petit Journal et 85.96% pour La Croix). Il est intéressant de noter que les deux quotidiens obtenant les valeurs les plus élevées sont les journaux conservateurs, c'est-à-dire La Croix et La Libre Parole. Leur caractère dogmatique renforce probablement le déséquilibre de la distribution entre les différents stéréotypes.

Ensuite, nous pouvons constater que le pourcentage de descripteurs retrouvés au sein des stéréotypes, représentés par le score d'adéquation $adeq_E$, est toujours très élevé (entre 77.92% pour La Croix et 84.47% pour Le Petit Journal). Un taux de couverture plus faible n'implique clairement pas une adéquation plus faible (l'exemple de La Croix est suffisamment parlant). La classification par défaut permet donc bien de construire des stéréotypes qui utilisent au maximum les descriptions des exemples afin d'être le plus "explicatifs" possible. Cela se retrouve d'ailleurs dans la taille des stéréotypes majeurs qui est toujours importante vis-à-vis du langage de description (de 19 descripteurs pour Le Petit Journal à 27 pour La Libre Parole). Les stéréotypes de petite taille possèdent, quant à eux, une description systématiquement moins riche.

En plus de ces remarques d'ordre structurel, nous pouvons constater un dénominateur commun aux stéréotypes majeurs des quatre quotidiens. Il s'agit de la description : socialiste, lié à l'étranger, implication du gouvernement, antipatriote, non honnête homme, non respecteux de la loi, corrompu, dénué de sens moral, mêlé à des affaires d'argent et lié à des actions violentes. Cela pourrait correspondre à une perception de la vie politique probablement très répandue parmi les quotidiens du "centre-droite" de cette époque. L'apparition du descripteur 'socialiste' par exemple s'explique plutôt aisément. En effet, au début des années 1880 parviennent en France les théories initiées par Karl Marx, un allemand, qui effraient une très grande partie de la population. Il semble prévisible que bon nombre de quotidiens transposent cette crainte éprouvée à l'égard du socialisme en cette fin de siècle. De même, les notions de patriotiste, d'honnête homme et de sens moral étaient très populaires à cette époque et les contrastes associés (patriote/antipatriote, traître/non traître par exemple) se retrouvent à plusieurs reprises dans nos stéréotypes.

La description exacte de tous les stéréotypes obtenus est donnée en annexe page 207[9]. Nous en donnons une description plus succinte ci-dessous.

5.3.3.2 Stéréotypes du Matin

Le graphique de la figure 5.40 présente la distribution des exemples à travers les cinq stéréotypes proposés par PRESS à partir des données tirées du Matin. Notons que 3.17% des exemples ne sont associés à aucun stéréotype.

Le stéréotype le plus important (65.08% des exemples) présente un socialiste révolutionnaire, d'origine étrangère, anticlérical, antipatriote, traître, en relation avec l'étranger et les franc-maçons, dénué de sens moral, dangereux, incompétent, etc. Il est bien illustré par le cas de "sans-patrie" socialistes ayant reconnu avoir profité de l'argent allemand "au mépris de la France". Il est intéressant de savoir que ce quotidien subissait des attaques régulières concernant l'origine de ses fonds (nous rappelons que les deux commanditaires sont anglais). L'impression laissée par ce premier stéréotype est celle de chercher un bouc émissaire dans une personne politique corrompu par l'argent allemand. Le thème du socialiste internationaliste, traître, antipatriote... se retrouve dans de nombreux articles et s'explique par la peur

[9] Les stéréotypes ont été ordonnés, non en fonction de leur poids mais de la richesse de leurs descriptions.

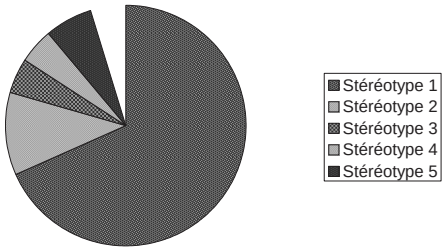


FIG. 5.40 – Distribution des stéréotypes du Matin.

provoquée par les théories de Marx.

Le second stéréotype proposé est radicalement opposé au premier et ne couvre que 11.11% des exemples de la base. Il s'agit d'un opportuniste républicain, clérical, qui n'est ni socialiste, ni traître envers sa patrie, ni lié à des puissances étrangères, etc. Nous pouvons citer le cas du républicain Chabrand, très estimé et loyal, qui s'est fait sauter la cervelle après avoir manqué à sa parole. Cet homme politique est clairement présenté comme en rupture avec le premier stéréotype majoritaire. D'une certaine manière, il peut être vu comme une victime de cet être "monstrueux" qui est rendu responsable de la plus grande partie des maux de la scène politique française.

Les trois derniers stéréotypes sont de descriptions pauvres et ne couvrent qu'un faible nombre d'exemples (respectivement 9.52%, 4.76% et 6.35%). Les articles relatifs à ces événements s'attachent surtout à décrire l'action et non les acteurs eux-mêmes. Le troisième stéréotype présente un homme politique impliqué dans un incident violent dans lequel il se trouve en position de victime. Il s'agit, par exemple, de Clémenceau et Winter qui sont aggressés par deux individus. Le quatrième et le dernier stéréotype proposent, d'une part, le duelliste qui préfère la salle d'arme à la scène politique et, d'autre part, un homme politique poussé à la démission (les maires Ferry et Clavier et le préfet Poubelle).

5.3.3.3 Stéréotypes de La Libre Parole

Le graphique de la figure 5.41 présente la distribution des exemples à travers les quatre stéréotypes proposés par PRESS à partir des données tirées de La Libre Parole. La proportion d'exemples non couverts est plus important que pour Le Matin avec un taux de 8.16%.

Le poids du stéréotype principal est beaucoup important ici (78.57% des exemples). Il dresse le portrait d'un républicain socialiste révolutionnaire, juif, entrenant des relations avec l'étranger, les protestants et les franc-maçons, antipatriote, anticlérical, corrompu, etc. Il s'agit, par exemple, du président Carnot qui souhaite opérer la laïcisation des églises au profit du "régime judéo-maçonnique", ou encore du candidat Bompart "allié à l'oligarchie financière". Ce stéréotype fait clairement référence à la théorie du complot "judéo-protestant-maçonnique" tel que le décrit Drumont, rédacteur en chef du quotidien, dans son ouvrage

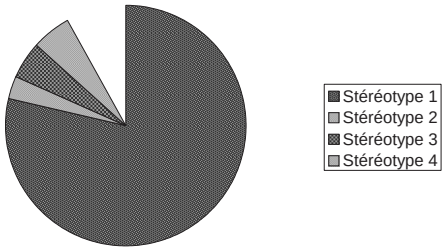


Fig. 5.41 – Distribution des stéréotypes de La Libre Parole.

La France Juive [Dru86]. Ce stéréotype découle sans ambiguïté des orientations xénophobes et antisémites de La Libre Parole. Notons le descripteur 'républicain' qui fait référence à la tendance conservatrice du journal.

Le deuxième stéréotype, qui ne couvre que quelques exemples (3.06%), présente un catholique pourvu de sens moral et victime des diverses diffamations proférées par ses adversaires politiques. L'exemple le plus saillant est le comte de Mun, "le grand orateur catholique" aux qualités morales certaines, qui est "injustement" accusé par son vainqueur.

Le troisième stéréotype, quant à lui, couvre 5.1% des exemples et présente un homme politique sans lien avec les juifs mais impliqué dans un incident emprunt de violence. Ce stéréotype est surtout lié au "guet-apens" tendu par Rotschield sur Drumont à Amiens où ce dernier était élu.

Enfin, le dernier stéréotype (5.1% des exemples) présente un homme politique qui trompe la population en faisant croire qu'il est souffrant. Il fait essentiellement référence à la maladie illusoire du président Carnot qui rencontre à cette époque un écho étonnant à travers la presse (allant même jusqu'à atteindre le Japon !).

5.3.3.4 Stéréotypes de La Croix

Le graphique de la figure 5.42 présente la distribution des exemples à travers les quatre stéréotypes proposés par PRESS à partir des données tirées de La Croix. La proportion d'exemples non couverts est de seulement 2.63%.

Le déséquilibre entre les stéréotypes est davantage marqué que dans le cas de La Libre Parole. En effet, le premier stéréotype couvre une très large majorité des exemples avec un score de 85.96%. Il campe un républicain protestant et révolutionnaire, socialiste, entretenant des liens avec l'étranger, les juifs et les franc-maçons mais pas avec le milieu catholique, corrompu, dénué de sens moral, dangereux, etc. La notion de complot est ici aussi présente même si elle est plutôt associée au "triangle maçonnique" qui est l'ennemi juré des catholiques. Ce stéréotype est illustré par les politiques manipulés par les franc-maçons ayant innocenté

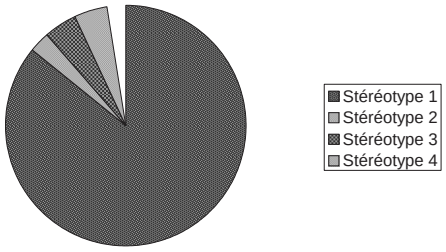


FIG. 5.42 – Distribution des stéréotypes de La Croix.

les principaux coupables lors du scandale de Panama[10]. Notons que le descripteur 'protestant' ne provient que d'un unique exemple relatant la venue en Martinique de deux ministres pour y propager leurs "funestes doctrines". De plus, on retrouve comme dans La Libre Parole le descripteur 'républicain' qui fait référence à la tendance conservatrice de La Croix.

Le deuxième stéréotype ne couvre que 2.63% des exemples et présente (une fois encore) un catholique clérical poussé à la démission. Il s'agit, par exemple, du conseiller général Mazurel-Jonglez qui démissionne suite à la victoire des radicaux alliés aux socialistes et aux révolutionnaires.

Le troisième stéréotype couvre quelques exemples de plus (4.39%) et campe un homme politique usant de favoritisme et victime d'un incident. L'exemple saillant est celui de Clémenceau victime d'une tentative d'assassinat (on a tiré sur sa voiture) après une réunion avec Jourdan.

Pour finir, le dernier stéréotype (4.39%) présente un homme politique non révolutionnaire et non internationaliste victime de diffamation. Il s'agit par exemple de Judet calomnié par Clémenceau et qui réclame un duel pour laver l'affront ou du préfet du Gers traduisant Cassagnac, qui l'accusait d'avoir falsifié le scrutin du 20 août, devant la cour d'assises.

Nous rappelons brièvement que ce quotidien était profondément catholique et présentait de fortes tendances monarchistes. Il est intéressant de noter que le pape Léon XIII avait condamné en 1891 la théorie marxiste de la lutte des classes, contraire à la fraternité religieuse des croyants et à la fraternité naturelle des hommes. Les socialistes étaient vus comme de dangereux agitateurs prêts à tous les compromis pour soulever les masses, surtout dans les campagnes où La Croix possédait un large tirage. A ce sujet, une campagne avait d'ailleurs été menée courant septembre 1893 pour dénoncer ces pratiques.

5.3.3.5 Stéréotypes du Petit Journal

Le graphique de la figure 5.42 présente la distribution des exemples à travers les deux stéréotypes proposés par PRESS à partir des données tirées du Petit Journal. 15.09% des

[10] Voir en page 173 pour un résumé de cette affaire.

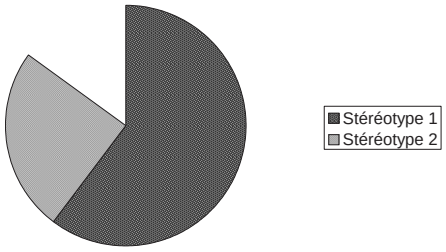


FIG. 5.43 – Distribution des stéréotypes du Petit Journal.

exemples ne sont pas couverts par les stéréotypes, ce qui est un taux élevé.

Ce quotidien se caractérise par le faible nombre de stéréotypes qui en sont extraits. Le plus important couvre 60.38% de la base et présente un républicain radical-socialiste, lié à l'étranger, antipatriote et traître, corrompu, incompétent, etc. Il s'attaque clairement à Clémenceau, alors directeur de La Justice, qui est l'objet d'une très grande partie des articles. En effet, le "commandité de Cornélius Hertz" est accusé d'agir pour le compte de l'Angleterre et de la Triple Alliance et d'être un perturbateur dans la politique du pays. Judet, le directeur du Petit Journal, ne mâche pas ses mots en le traitant de "détrousseur de grand-chemin" et en associant le "clémenciste" à une "maladie politique". Remarquons qu'ici, le descripteur 'républicain' est associé au parti radical, ce qui différencie son utilisation de celle faite par les deux journaux conservateurs.

Le second stéréotype, quant à lui, couvre 24.53% des exemples et donne un portrait qui contraste avec le précédent : un patriote, honnête homme, non traître et victime de diffamation. Il s'oppose d'autant plus au premier qu'il est bien précisé qu'il ne fait pas partie du "comité clémenciste". Il s'agit par exemples des "honnêtes citoyens" Sandt, Lavocat et Métivier, insultés par le clémenciste Maujan.

5.3.3.6 Conclusions

Les stéréotypes construits à partir des données tirées de quatre quotidiens en cette période d'élection électorale font ressortir des caractéristiques qui nous semblent tout à fait cohérentes et intéressantes. Ce résultat est en partie due au fait qu'une très grande partie des descripteurs provenant des exemples est utilisée pour construire les stéréotypes. Cela est rendu possible par l'utilisation d'une mesure adaptée et d'une fonction d'évaluation basée sur cette mesure. Ce résultat est également due à la prise en compte de nouvelles contraintes, liées à la cohérence de la couverture et à la cohésion interne des catégories. Finalement, nous obtenons bien des descriptions riches et constratées qui peuvent être analysées à la lumière des informations historiques. A partir de ces descriptions, un travail d'interprétation, encore impossible à réaliser avec une machine, est indispensable pour tirer toute la signification de ces représentations.

Conclusion et perspectives

Les recherches effectuées au sein de l'équipe ACASA sur la découverte scientifique [Cor96] [CG97] sont à l'origine de cette thèse. L'idée initiale proposée par J.G. Ganascia d'appliquer ces techniques d'induction au domaine des sciences sociales et à l'étude du sens commun nous a conduit à proposer un modèle à base de stéréotypes. Ce modèle de représentation nous a permis d'élaborer un algorithme d'apprentissage non-supervisé adapté à des données symboliques et lacunaires. Nous soulignons qu'il repose sur la constatation du caractère lacunaire de l'information, à la fois dans les travaux précédents, mais également dans nos propres expérimentations sur les articles de presse.

Les algorithmes permettant d'effectuer une tâche de classification automatique non-supervisée sont très nombreux et leur nombre ne cesse de croître chaque année. Dans ce contexte, notre objectif n'était pas d'élaborer un algorithme permettant de traiter de grandes bases de données ou manipulant à la fois des données numériques et symboliques. Le temps machine demandé par le processeur pour exécuter les calculs ou la place mémoire pour stocker l'information temporaire n'était pas non plus au cœur de nos préoccupations. Notre objectif était plutôt de proposer un modèle général adapté à la problématique des données lacunaires, ainsi qu'un algorithme de classification par défaut basé sur ce modèle. Cet algorithme pouvait ensuite être comparé à des algorithmes de clustering plus classiques.

Les stéréotypes que nous avons obtenu au fil de nos expérimentations correspondent à l'idée que nous avions de cet objet souvent insaisissable, associé aussi bien à la théorie de la catégorisation qu'à celle des représentations sociales. Ils sont une forme de représentations des données, très proche des exemples dans le sens où elle utilise le maximum d'information disponible, qui prend explicitement en compte la séparation entre les catégories et répond à des critères de cohérence (par l'intermédiaire de la nouvelle relation de subsomption par défaut) et de cohésion (avec la contrainte de cohésion cognitive). Les descriptions que nous obtenons peuvent être utilisées pour faciliter l'interprétation des données et pour prédire les valeurs manquantes, que ce soit au sein même des données traitées, mais également à propos de nouveaux objets observés. Nous pensons que cette technique est une alternative intéressante à l'analyse du discours de presse et qu'elle peut être appliquée à d'autres domaines comme celui de la modélisation utilisateur. Elle peut également être considérée comme une manière d'aborder l'étude du sens commun.

Les techniques actuellement utilisée en fouille de données pour traiter les corpus de textes, comme l'algorithme TF-IDF [SAB94], ne sont pas adéquates pour fournir le type de données

manipulé par notre algorithme de classification. C'est pourquoi la traduction des articles de journaux a été jusqu'à présent réalisée manuellement. Nous pensons cependant que les années à venir vont voir le développement d'outils prenant davantage en compte la sémantique des textes, c'est-à-dire le sens du discours qui y est tenu. Dans ce cadre, des formalismes plus évolué, à base de graphes conceptuels [Sow84], par exemple, ou d'objets comme ceux manipulés en analyse de données symboliques [BD00][Did00], pourraient être utilisés. La représentation de textes sous forme de stéréotypes pourrait alors trouver une place au sein de programmes plus complexes traitant de l'analyse du contenu de la presse.

Perspectives

Plusieurs sujets mériteraient d'être approfondis suite à cette thèse. Nous en détaillons quelques-uns :

Etude synchronique et diachronique des représentations

Comme nous l'avons suggéré dans notre conclusion, le modèle à base de stéréotypes nous semble adapté à l'étude du contenu de la presse, champ de recherche qui appartient au domaine des sciences sociales. En effet, les stéréotypes peuvent être vus comme une partie de ce savoir "naïf" acquis à travers les médias. L'étude peut être effectuée à deux niveaux.

Tout d'abord, l'étude peut être réalisée de manière synchronique, c'est-à-dire en considérant les stéréotypes construits à partir de différents journaux à une époque donnée concernant un même objet social. Les ensembles de stéréotypes peuvent alors constituer le support d'une comparaison des représentations induites à partir de ces journaux, et donc donner une idée générale sur le message qu'ils font passer à leurs lecteurs. Le faible nombre de descriptions et leur caractère contrasté permet alors de rendre plus aisée l'interprétation des résultats.

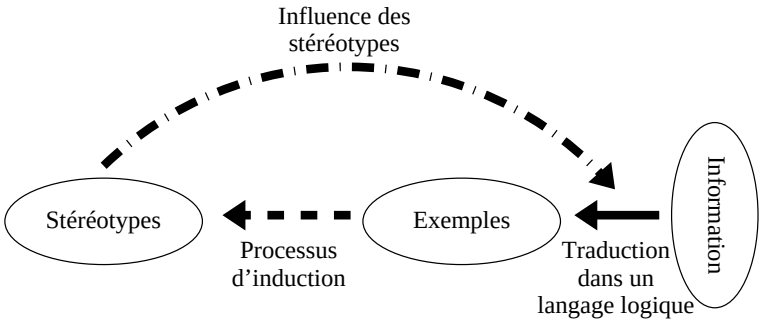
Ensuite, l'étude peut être réalisée de manière diachronique, c'est-à-dire en considérant les stéréotypes fournis par un même journal à plusieurs époques différentes. Une application de cette approche, qui n'a pu être réalisée dans cette thèse mais qui nous semble tout à fait intéressante, serait de prendre des périodes-clefs de l'Affaire Dreyfus (révélation de l'affaire, publication de "J'accuse", procès de Rennes) et d'observer l'évolution des stéréotypes qui sont induits à chaque fois. Des stéréotypes se verraient transformés, alors que certains disparaîtraient et que de nouveaux apparaîtraient. Une analyse comparée, suivant plusieurs quotidiens, de l'évolution de ces représentations pourrait également se révéler riche en enseignements.

Etude de la dynamique des représentations

Revenons aux travaux qui ont conduit à cette recherche. Les résultats concernant la découverte du scorbut permettent de comprendre le pouvoir explicatif étonnamment élevé de certains attributs en prenant en compte les connaissances des médecins de l'époque sur la maladie. Le même type de phénomène nous semble avoir lieu dans une problématique de

représentations sociales dans laquelle les théorie implicites biaisent le raisonnement au sujet de l'objet social.

Les stéréotypes construits à l'aide de la classification par défaut constituent, selon nous, une forme de théorie, naïve et peu structurée, qui peut être utilisée pour filtrer l'information perçue à travers les différents médias, comme la presse ou la télévision. De nouveau stéréotypes peuvent ensuite être construits à partir de cette information filtrée, et le processus de se répéter indéfiniment. Cela nous conduit au schéma suivant, inspiré directement des travaux réalisés par J.G. Ganascia et V. Corruble :



Ce phénomène circulaire mériterait, selon nous, une étude plus approfondie, notamment à l'aide de techniques proposées en intelligence artificielle.

Validation psychologique des stéréotypes

Les études sur la catégorisation ont montré, depuis les années 70, la validité de la théorie du protype. Bien sûr, les interprétations des résultats obtenus empiriquement ont évolué et le prototype est passé du statut de "premier exemplaire" de la catégorie à celui de simple "effet prototypique" indépendant de la structure de la catégorie. Nous pensons que le stéréotype, du moins tel que nous l'avons défini dans cette thèse, est également un effet des catégories, effet conditionné par le manque d'information disponible sur un sujet donné.

L'individu, au contact de sources aussi diverses que la presse, la télévision, les discussions inter-personnelles, les affiches... construit une représentation avec l'information dont il dispose afin de pouvoir raisonner sur le sujet, même si cela amène à effectuer des erreurs de jugement. Le terme "erreur" que nous employons ici, peut-être à la légère, signifie que cette représentation du monde est nécessairement biaisée par l'intermédiaire des médias. Conscient de la difficulté de la tâche (comment en effet être capable de connaître la totalité de l'information perçue par un individu sur un thème donné ?), nous pensons qu'une validation empirique d'ordre psychologique pourrait être menée au sujet du type de structure que nous décrivons dans cette thèse. De plus, la contrainte de cohésion cognitive que nous définissons, inspirée de la théorie de la catégorisation, pourrait faire l'objet d'une réelle validation psychologique.

Relâcher la contrainte sur la redondance

Comme nous l'avons écrit à plusieurs reprises dans ce document, la contrainte de non-redondance renvoie à la notion de discrimination qui est associée au concept de stéréotype.

Il nous semble pourtant que cette contrainte peut être relâchée en considérant une mesure de séparation calculée entre les stéréotypes. Cette approche permettrait de tolérer quelques redondances et d'obtenir des résultats qui dépendent de l'objectif que l'on souhaite atteindre. En effet, nous pensons fortement qu'il n'existe pas une seule catégorisation "idéale", mais bien plusieurs adaptées à des visions parfois très différentes.

Données lacunaires et jeux de données artificiels

Les jeux de données artificiels que nous utilisons permettent de montrer des résultats qui seraient impossibles à obtenir à partir de données réelles. En ce sens, ils nous aident à valider en partie notre modèle bien qu'ils ne se prévalent pas d'une légitimité excessive. Ainsi serait-il intéressant d'essayer d'autres stratégies de génération afin de tester notre modèle dans d'autres situations, voire à partir d'autres hypothèses. Le fait de produire des descriptions initiales contrastées correspond bien entendu à un modèle basé sur l'absence de redondance entre les stéréotypes. En relâchant cette contrainte, comme nous le suggérons dans la section précédente, nous pourrions considérer des descriptions exhibant un certain taux de redondance initial. De plus, une étude plus spécifique sur la robustesse de notre algorithme vis-à-vis du bruit, par exemple en introduisant des *outliers*, pourrait être menée. Remarquons que la manière dont sont générés les jeux de données artificiels, utilisés dans la plus grande majorité des recherches sur le clustering, est très rarement explicitée, en dehors peut-être de certains domaines.

Concernant plus spécifiquement le cas des données lacunaires, il est clair que le fait de retirer les descripteurs de manière aléatoire ne recouvre pas toutes les situations et place nos expérimentations dans le cas de données du type MCAR. Cela est d'ailleurs le cas dans la plupart des travaux traitant des données manquantes à partir d'exemples artificiels [FKP04][IL01]. Mais comment croire que, dans le cas des articles de journaux, les oublis, les erreurs, les rajouts, sont le fruit du hasard ? Une simulation plus fine de ces processus altérant les données initiales devrait mener à des résultats très intéressants, permettant peut-être de mettre en relation la manière dont l'information est modifiée avec la manière dont cette information peut être représentée.

Extraction d'exemples-pivots

L'idée de base concernant les exemples-pivots découle de la définition de la contrainte de cohésion cognitive que nous avons détaillée dans les sections 2.2.7 et 3.1.5.2. En effet, l'intuition nous porte à croire que dans la vérification de la contrainte, les différents exemples considérés ne jouent pas tous le même rôle. Ainsi, certains exemples semblent importants dans le maintien de la cohésion interne aux catégories, alors que d'autres sont plus anecdotiques parce que l'information qu'ils fournissent est redondante.

Les exemples-pivots sont les exemples de chaque catégorie qui, s'ils sont retirés, amène la contrainte de cohésion cognitive à ne plus être respectée. Cela signifie que, sans eux, la catégorie perd de sa cohésion et doit être divisée en deux ou plusieurs catégories. Tout porte

à croire que ces exemples spécifiques jouent un rôle important dans la construction d'une représentation sous la forme de stéréotypes. Le temps nous a malheureusement manqué pour effectuer de réelles expérimentations à ce sujet, mais des résultats préliminaires donnent à penser qu'une analyse tirerait avantage de l'étude de ces exemples-pivots.

Bibliographie

[Abr03] J.-C. Abric. *Pratiques sociales et représentations*. PUF, 2003.

[And73] M.R. Anderberg. *Cluster analysis for applications*. Academic Press, New York, 1973.

[AS95] K. Al-Sultan. A Tabu Search Approach to the Clustering Problem. *Pattern Recognition*, 28(9) :1443–1451, 1995.

[ASK96] K. Al-Sultan and M. Khan. Computational experience on four algorithms for the hard clustering problem. *Pattern Recognition Letters*, 17(3) :295–308, 1996.

[Ba73] Benzécri and al. *L'analyse des données*, volume 2. Dunod, 1973.

[BB95] V. Batagelj and M. Bren. Comparing Resemblance Measures. *Journal of Classification*, 12(2) :73–90, 1995.

[BBD00] P.S. Bradley, K.P. Bennett, and A. Demiriz. Constrained k-means clustering, 2000. Technical Report MSR-TR-2000-65.

[BD00] H.-H. Bock and E. Diday. *Analysis of symbolic data : exploratory methods for extracting statistical information from complex data*. Springer-Verlag, Heidelberg, 2000. Seconde édition.

[Bea64] L.R. Beach. Cue probabilism and inference behavior. *Psychological Monographs*, 78, 1964. (Whole No.582).

[Ben03] S. Benati. *Categorical data fuzzy clustering : an analysis of local search heuristics*. Universita degli Studi di Trento, 2003. Internal report.

[BGGT76] C. Bellanger, J. Godechot, P. Guiral, and F. Terrou. *Histoire générale de la presse française. Tome 3 : de 1871 à 1940*. PUF, 1976.

[BK99] E. Burke and G. Kendall. Comparison of Meta-Heuristic Algorithms for Clustering Rectangles. *Computers and Industrial Engineering*, 37(1–2), 1999.

[Bol98] D. Boley. Principal direction divisive partitioning. *Data Mining and Knowledge Discovery*, 2(4) :325–344, 1998.

[Bre83] J.-D. Bredin. *L'Affaire*. Julliard, 1983.

[Bru78] P. Brucker. On the Complexity of Clustering Problems. In *Lecture Notes in Economics and Mathematical Systems*, volume 157, pages 45–54, 1978.

[BT94] R. Battiti and G. Tecchiolli. The Reactive Tabu Search. *Journal on Computing*, 6(2) :126–140, 1994.

[Buh02] J. Buhmann. Data Clustering and Learning. In *Handbook of Brain Theory and Neural Networks*, pages 308–312. MIT Press, Cambridge, MA, 2002. 2nd edition.

[CFC+01] T. Chiu, DP. Fang, J. Chen, Y. Wang, and C. Jeris. A robust and scalable clustering algorithm for mixed type attributes in large database environment. In *Proceedings of KDD*, pages 263–268, 2001.

[CG97] V. Corruble and J.-G. Ganascia. Induction and the discovery of the causes of scurvy : a computational reconstruction. *Artificial Intelligence Journal*, 91(2) :205–223, 1997.

[Cha03] F. Chateauraynaud. *Prospéro : une technologie littéraire pour les sciences humaines*. CNRS Editions, 2003.

[CK81] L. Coleman and P. Kay. Prototype Semantics : the english word LIE. *Langage*, 57 :26–44, 1981.

[CLR92] T. Cormen, C. Leiserson, and R. Rivest. *Introduction à l'algorithmique*. Dunod, 1992.

[CM02] A. Cornuéjols and L. Miclet. *Apprentissage Artificiel : concepts et algorithmes*. Editions Eyrolles, France, 2002.

[CMB03] L. Chaudron, N. Maille, and M. Boyer. The CUBE lattice model and its applications. *Applied Artificial Intelligence*, 17(3) :207–242, 2003.

[Cor96] V. Corruble. *Une approche inductive de la découverte en médecine : les cas du scorbut et de la lèpre*. Rapport interne LAFORIA TH96/18, thèse de l'Université Pierre et Marie Curie, 1996.

[CSL03] C.H. Chou, M.C. Su, and E. Lai. A New Cluster Validity Measure for Clusters with Different Densities. In *Proceedings of International Conference on Intelligent Systems & Control*, pages 276–281, Salzburg, Austria, 2003.

[Del77] Y. Delahaye. *La frontière et le texte*. Payot, Paris, 1977.

[DIC] Le Trésor de la Langue Française Informatisé. `http ://atilf.atilf.fr/`.

[Did79] E. et al. Diday. *Optimisation en classification automatique*. INRIA, 1979.

[Did87] E. Diday. The symbolic approach in clustering and related methods of Data Analysis. In H. Bock, editor, *Classification and Related Methods of Data Analysis*, Aachen, Germany, 1987. Proc. IFCS.

[Did00] E. Diday. Knowledge Discovery from Symbolic Data and the Sodas Software. In *Proceedings of the Workshop of Symbolic Data Analysis : Theory, Software and Applications for Knowledge Mining*, 2000.

[DK91] E. Diday and Y. Kodratoff. *Induction symbolique et numérique à partir de données.* CEPADUES, 1991.

[DLD77] A.P. Dempster, N.M. Laird, and Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm. *Royal Statistical Society*, 1 :1–38, 1977.

[DLPT84] E. Diday, J. Lemaire, J. Poujet, and F. Testu. *Eléments d'Analyse des données.* Dunod, Paris, 1984.

[DM04] E. Davis and L. Morgenstern. Introduction : progress in formal commonsense reasoning. *Artificial Intelligence. Special issue on logical formalizations and commonsense reasoning*, 153(1–2) :1–12, 2004.

[DRR93] D. Dubois and P. Resche-Rigon. Prototypes ou stéréotypes : productivité et figement d'un concept. In *Lieux Communs, topoï, stéréotypes, clichés*, pages 372–389. 1993.

[Dru86] E. Drumont. *La France Juive.* Palmé,V., Paris, 1886.

[Dub86] D. Dubois. *La compréhension des phrases : représentation, sémantique et processus.* PhD thesis, thèse de Doctorat d'Etat de l'Université Paris VIII, 1986.

[Dun74] J.C. Dunn. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4 :95–104, 1974.

[Far98] C. Faron. *Connaissances taxinomiques : Représentation de taxinomies comportant des exceptions et construction d'hypermédias à base de connaissances taxinomiques.* rapport interne LIP6, thèse de l'Université Pierre et Marie Curie, 1998.

[FBJ71] E. Feigenbaum, Buchanan B., and Ledergerg J. On generality and problem Solving : a Case Study Using the DENDRAL Program. In *Machine Intelligence*, volume 6. Edinburgh University Press, Edinburgh, 1971.

[Fis87] D.H. Fisher. Knowledge Acquisition Via Incremental Conceptual Clustering. *Machine Learning*, (2) :139–172, 1987.

[FKP04] A. Farhangfar, L.A. Kurgan, and W. Pedrycz. Experimental analysis of methods for imputation of missing values in databases. *Proceedings of SPIE*, 5421 :172–182, 2004.

[FL00] F. Farnstrom and G. Lewis. *Fast, Single-pass k means algorithms.* Department of Computer Science and Engineering, University of California, 2000. Internal report.

[Gan] J.-G. Ganascia. Deriving the learning bias from rule properties. In J.E. Hayes, D Michie, and E. Tyugu, editors, *Machine intelligence*, volume 12.

[Gan90] J.-G. Ganascia. CHARADE : apprentissage de bases de connaissances. In *Apprentissage symbolique et numérique.* CEPADUES, Toulouse, 1990. Textes réunis par Y. Kodratoff et E. Diday.

[Gan93a] J.-G. Ganascia. Algebraic Structure of Some Learning Systems. In *Proceedings of the International Workshop, ALT'93*, Tokyo, Japan, 1993.

[Gan93b] J.-G. Ganascia. TDIS : An algebraic generalization. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, Chambéry, France, 1993.

[Gan02] J.-G. Ganascia. Aperçus sur la Découverte Scientifique : Modèles de la créativité. *Revue d'Intelligence Artificielle*, (Volume 16) :101–122, 2002.

[Gan04] J.-G. Ganascia. Rational Reconstruction of Wrong Theories. In L. Hajek and D. Westerstahl, editors, *Proceedings of the 12th LMPS.* Elsevier - North, 2004.

[Gar95] S.R. Garner. WEKA : The waikato environment for knowledge analysis. pages 57–64, 1995.

[GC85] M.A. Gluck and J.E. Corter. Information, uncertainty, and the utility of categories. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, pages 283–287, 1985.

[Gee] D. Geeraerts. Les données stéréotypiques, prototypiques et encyclopédiques dans le dictionnaire. *Cahiers de lexicologie*, 46(1) :28–43.

[Giv86] T. Givon. Prototypes : Between Plato and Wittgenstein. In C.G. Craig, editor, *Noun, Classes and Categorisation.* John Benjamins Publishing Company, Amsterdam, 1986.

[GJ94] Z. Ghahramani and M.I. Jordan. Supervised learning from incomplete data via an EM approach. In J.D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6, pages 120–127. Morgan Kaufmann Publishers, Inc., 1994.

[GKR98] D. Gibson, J.M. Kleinberg, and P. Raghavan. Clustering Categorical Data : An Approach Based on Dynamical Systems. In *VLDB'98, Proceedings of 24rd International Conference on Very Large Data Bases*, pages 311–322, New York City, New York (USA), 1998. Morgan Kaufmann.

[GL97] F. Glover and M.S Laguna. *Tabu Search*. Kluwer Academic Publishers, 1997.

[GLF92] J.H. Gennari, P. Langley, and D. Fisher. Models of incremental concept formation. In *Machine Learning : Paradigms and Methods*. MIT/Elsevier, 1992.

[Glo86] F. Glover. Future paths for Integer Programming and Links to Artificial Intelligence. *Computers and Operations Research*, (5) :533–549, 1986.

[Gor96] A.D. Gordon. A survey of constrained classification. *Computational Statistics and Data Analysis*, 21(1) :17–29, 1996.

[GRS00] S. Guha, R. Rastogi, and K. Shim. ROCK : A Robust Clustering Algorithm for Categorical Attributes. *Information Systems*, 25(5) :345–366, 2000.

[GS98] D. Genest and E. Salvat. A platform allowing typed nested graphs : How cogito became cogitant. *Proceedings of the Sixth International Conference on Conceptual Structures*, 1998.

[GV04] J.-G. Ganascia and J. Velcin. Clustering of conceptual graphs with sparse data. In *Proceedings of the 12^{th} International Conference on Conceptual Structures*, Huntsville, USA, 2004. Springer-Verlag.

[HBV02a] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Cluster Validity Methods : Part i. *Special Interest Groups on Management Of Data*, 2002.

[HBV02b] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Cluster Validity Methods : Part ii. *Special Interest Groups on Management Of Data*, 2002.

[HK01] J. Han and M. Kamber. *Data Mining : Concepts and Techniques*. Morgan Kaufmann publishers, 2001.

[HL83] R. Harré and R. Lamb. *The Encyclopedic Dictionnary of Psychology*. MIT Press, Cambridge, 1983.

[HL04] C.-C. Huang and H.-M. Lee. A Grey-Based Nearest Neighbor Approach for Missing Attribute Value Prediction. *Applied Intelligence*, 20(3) :239–252, 2004.

[HM01] P. Hansen and N. Mladenovic. J-MEANS : a new local search heuristic for minimum sum of squares clustering. *Pattern Recognition*, 34(2) :405–413, 2001.

[How66] R. Howard. Classifying a Population into Homogeneous Groups. In *Operational Research in the Social Sciences*. Tavistock Publ., London, 1966.

[HTTS02] J. He, A.-H. Tan, C.-L. Tan, and S.-Y. Sung. On Qualitative Evaluation of Clustering Systems. *Information Retrieval and Clustering*, 2002.

[Hua97a] Z. Huang. Clustering Large Data Sets with Mixed Numeric and Categorical Values. In *Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Singapore, 1997. World Scientific.

[Hua97b] Z. Huang. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. In *DMKD*, volume 8, 1997.

[HVB00] M. Halkidi, M. Vazirgiannis, and I. Batistakis. Quality scheme assessment in the clustering process. In *Proceedings of the 4^{th} PKDD Conference*, Lyon, 2000.

[IL01] W. Iba and P. Langley. Unsupervised Learning of Probabilistic Concept Hierarchies. *Machine learning and its applications*, 2001.

[Jan66] R.C. Jancey. Multidimensional Group Analysis. *Australian Journal on Botany*, 14 :127–130, 1966.

[JMF99] A.K. Jain, M.N. Murty, and P.J. Flynn. Data Clustering : A Review. *ACM Computing Surveys*, 31(3), 1999.

[Jod89] D. Jodelet. *Représentations sociales : un domaine en expansion.* PUF, 1989.

[Kin67] B. King. Step-wise clustering procedures. *Journal of the American Statistical Association*, 69 :86–101, 1967.

[Kle88] G. Kleiber. Prototype, Stéréotype : un air de famille ? *DRLAV : revue linguistique*, 38 :1–66, 1988.

[Kle90] G. Kleiber. *La sémantique du prototype : catégories et sens lexical.* PUF, 1990.

[Kol83] J.L. Kolodner. Reconstructive memory : A computer model. *Cognitive Science*, 7, 1983.

[Kot03] Joël et Dan Kotek. *Au nom de l'antisionisme : l'image des Juifs et d'Israël dans la caricature depuis la seconde Intifada.* Editions complexe, Bruxelles, 2003.

[KR95] R.E. Kass and A.E. Raftery. Bayes factors. *Journal of American Statistical Association*, 90 :773–795, 1995.

[Lak76] I. Lakatos. *Proofs and Refutations.* Cambridge University Press, Cambridge, 1976.

[Lak87] G. Lakoff. *Women, Fire and Dangerous Things : What categories reveal about mind.* The university of Chicago Press, 1987.

[Lan87] R.W. Langacker. *Foundations of Cognitive Grammar*, volume 1. Stanford University Press, Stanford, 1987.

[LBS83] P. Langley, G.L. Bradshaw, and H.A. Simon. Rediscovering Chemistry with the Bacon System. In *Machine Learning : An Artificial Intelligence Approach 1*, pages 307–330. Springer-Verlag, Palo Alto (CA), 1983.

[Leb87] M. Lebowitz. Experiments with Incremental Concept Formation : UNIMEM. *Machine Learning*, (2) :103–138, 1987.

[Len95] D.B. Lenat. CYC : A Large-Scale Investment in Knowledge Infrastructure. In *Communications of the ACM*, volume 38, pages 33–38, 1995.

[LG90] D.B. Lenat and R.V. Guha. *Building Large KnowledgeBased Systems : Representation and Inference in the CYC Project.* Addison-Wesley, 1990.

[Lip22] W. Lippman. *Public Opinion.* Ed. MacMillan, NYC, 1922.

[LJ89] P. Langley and Zytkow J. Data-Driven Approaches to Empirical Discovery. *Artificial Intelligence*, (40) :283–312, 1989.

[LR02] R. Little and D. Rubin. *Statistical analysis with missing data.* Wiley-Interscience publication, 2002.

[Mac67] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, Berkeley, Califonia, 1967. University of California Press.

[McC80] J. McCarthy. Circumscription : a form of non-monotonic reasoning. *Artificial Intelligence*, 13 :27–39,171–172, 1980.

[MD80] D. McDermott and J. Doyle. Nonmonotonic logic 1. *Artificial Intelligence*, 13 :41–72, 1980.

[Mic80] R.S. Michalski. Knowledge acquisition through conceptual clustering : A theoretical framework and algorithm for partitioning data into conjunctive concepts. *International Journal of Policy Analysis and Information Systems*, (4) :219–243, 1980.

[Min75] M. Minsky. A framework for representing knowledge. New York : McGraw-Hill, 1975.

[Mit80] T.M. Mitchell. The need for biases in learning generalizations. In *Machine Learning*, pages 184–191. 1980.

[ML79] J. McCarthy and V. Lifschitz. *Formalizing Common Sense : papers.* Norwood, N.J., 1979.

[Mol94] J.-Y. Mollier. La bataille de l'imprimé. In *Les représentations de l'Affaire Dreyfus dans la presse en France et à l'étranger*. Publication de l'Université François-Rabelais, Tours, 1994. Numéro spécial - Hors-série. Actes du collogque de Saint-Cyr-sur-Loire.

[Mos76] S. Moscovici. *La psychanalyse : son image et son public*. PUF, Paris, 1976. 2nd edition.

[MS83] R.S. Michalski and R.E. Stepp. Learning from Observation : Conceptual Clustering. In *Machine Learning : An Artificial Intelligence Approach, I*, pages 331–363. 1983.

[MSD81] R.S. Michalski, R.E. Stepp, and E. Diday. A recent advance in data analysis : clustering objects into classes characterized by conjunctive concepts. *Pattern Recognition*, 1 :33–55, 1981.

[Nar01] J.-P. Narboux. Ressemblances de famille, caractères, critères. In *Wittgenstein : métaphysique et jeux de langage*, pages 69–95. PUF, 2001.

[NW02] M.K. Ng and J.C. Wong. Clustering categorical data sets using tabu search techniques. *Pattern Recognition*, 35(12) :2783–2790, 2002.

[PB97] N.R. Pal and J. Biswas. Cluster Validation using graph theoretic concepts. *Pattern Recognition*, 30(6), 1997.

[PDP98] P. Plante, L. Dumas, and A. Plante. *Nomino version 2. Environnement de programmation dédié à la conception de systèmes d'analyse linguistique, d'extraction d'informations dans les textes et de mise au point de progiciels à base de connaissances*. WEB UQAM-ATO, 1998.

[Pla99] J.-L. Plane. Les figures de la typicité : vers une approche formelle des stéréotypes et prototypes, 1999. Rapport de stage du D.E.A. en Sciences Cognitives.

[Pop73] K. Popper. *Logique de la découverte scientifique (Logik der forschung*. 1973.

[Pos86] M. Posner. Empirical Studies of Prototypes. In C.G. Craig, editor, *Noun, Classes and Categorisation*, pages 53–61. John Benjamins Publishing Company, Amsterdam, 1986.

[Put75] H. Putnam. The meaning of 'meaning'. pages 215–271. Cambridge University Press, Cambridge, 1975.

[Put90] H. Putnam. Peirce the Logician. In *Realism with a Human Face*, pages 252–260. Harvard University Press, 1990. `http ://www.jfsowa.com/peirce/putnam.htm`.

[Ree72] S.K. Reed. Pattern recognition and categorization. *Cognitive Psychology*, 3 :382–407, 1972.

[Rei80] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, (13) :81–132, 1980.

[Rei86] M. Reinert. Un logiciel d'analyse lexicale [ALCESTE]. *Les cahiers de l'analyse de données*, 4, 1986.

[RG90] S.J. Russel and B.N. Grossof. Declarative Bias : An Overview. In D.P. Benjamin, editor, *Change of Representation and inductive bias.* Kluwer Academic Publisher, 1990.

[Ric79] E. Rich. User Modeling via Stereotypes. *Cognitive Science*, 3 :329–354, 1979.

[Rif96] M. Rifqi. *Mesures de comparaison, typicalité et classification d'objets flous : théorie et pratique.* rapport interne LIP6, thèse de l'Université Pierre et Marie Curie, 1996.

[RLR98] R. Rezaee, B.P.F. Lelieveldt, and J.H.C. Reiber. A new cluster validity index for the fuzzy c-mean. *Pattern Recognition Letters*, 19 :237–246, 1998.

[RM75] E. Rosch and C.B. Mervis. Family Resemblances : Studies in the Internal Structure of Categories. *Cognitive Psychology*, 7 :573–605, 1975.

[RMG+76] E. Rosch, C.B. Mervis, W.D. Gray, D.M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8 :382–439, 1976.

[Ros73] E. Rosch. On the internal structure of perceptual and semantic categories. In T.E. Moore, editor, *Cognitive Development and the Acquisition of Language.* Academic Press, NewYork, 1973.

[Ros75] E. Rosch. Cognitive representations of semantic categories. *Journal of Experimental Psychology : General*, 104 :192–232, 1975.

[Ros78] E. Rosch. Principles of categorization. In E. Rosch and B. Lloyd, editors, *Cognition and Categorization*, pages 27–48. NJ : Lawrence Erlbaum, Hillsdale, 1978.

[Rub76] D.B. Rubin. Inference and missing data. *Biometrika*, 63 :581–592, 1976.

[RV00] F. Rossi and F. Vautrain. Expert Constrained Clustering : A Symbolic Approach. In *Proceedings of the 4^{th} European Conference on Principles of Data Mining and Knowledge Discovery*, London, UK, 2000. Springer-Verlag.

[RW84] R. Redner and H. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2), 1984.

[SA77] R.C. Schank and R.P. Abelson. *Goals, Plans, Scripts and Understanding : An Enquiry into Human Knowledge Structures.* Lawrence Erlbaum, 1977.

[SAB94] G. Salton, J. Allan, and C. Buckley. Automatic structuring and retrieval of large text files. *Communications of the ACM*, 37(2) :97–108, 1994.

[Sap90] G. Saporta. *Probabilités, Analyse des Données et Statistiques.* Technip, Paris, 1990.

[Sch85] Ch. Schwarz. Lexique et compréhension textuelle. 1985.

[Sem89] G.R. Semin. Prototypes et représentations sociales. In *Les Representations Sociales.* PUF, Paris, 1989.

[Sha04] L. Shawver. Commentary on Wittgenstein's Philosophical Investigations, 2004. `http ://users.rcn.com/rathbone/lw65-69c.htm`.

[SJ00] C.S. Sung and H.W. Jin. A tabu-search-based heuristic for clustering. *Pattern Recognition*, 33(5) :849–858, 2000.

[SM81] E.E. Smith and D.L. Medin. *Categories and concepts.* MA : Harvard University Press, Cambridge, 1981.

[SM05] P. Soucy and G.W. Mineau. Beyond TFIDF Weighting for Text Categorization in the Vector Space Model. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 1130–1135, 2005.

[Sou92] M. Souchard. *Le discours de la presse : l'image des syndicats au Québec.* Gallimard, 1992.

[Sow] J.F. Sowa. Semantic Networks. `http ://www.jfsowa.com/pubs/semnet.htm`.

[Sow84] J.F. Sowa. *Conceptual Structures : Information Processing in Mind and Machine.* The Systems Programming Series. Addison-Wesley Publishing Company, Massachusetts, 1984.

[Spä80] H. Späth. *Cluster Analysis Algorithms for data reduction and classification of objects.* Ellis Horwood Limited, Great Britain, 1980.

[SS63] R.R. Sokal and P.H. Sneath. *Principles of numerical taxonomy.* W.H. Freeman and CO, 1963.

[SS73] P.H. Sneath and R.R. Sokal. *Numerical taxonomy.* W.H. Freeman, San Francisco, 1973.

[Tag02] P.-A. Taguieff. Déterminisme racial, antisémitisme et nationalisme : de Drumont à Soury. In *La couleur et le sang.* Mille et une nuits, 2002.

[TNLH01] A.K.H. Tung, R.T. Ng, L.V.S. Lakshmanan, and J. Han. Constraint-Based Clustering in Large Databases. In *Proceedings of the 8th ICDT*, London, UK, 2001.

[Tve77] A. Tversky. Features of similarity. *Psychological Review*, 84 :327–352, 1977.

[Vel02] J. Velcin. Reconstruction rationnelle des mentalités collectives : deux études sur la xénophobie, 2002. Rapport de DEA IARFA, LIP6.

[VG04] J. Velcin and J.-G. Ganascia. Modeling default induction with conceptual structures. In Lu, Atzeni, Chu, Zhou, and Ling, editors, *Proceedings of ER 2004*, Shangai, China, 2004. Springer-Verlag.

[VG05a] J. Velcin and J.-G. Ganascia. Default Clustering from Sparse Data Sets. In *Proceedings of the 8^{th} European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Barcelona, Spain, 2005. Springer-Verlag.

[VG05b] J. Velcin and J.-G. Ganascia. Stereotype Extraction with Default Clustering. In *Proceedings of the 19^{th} International Joint Conference on Artificial Intelligence*, Edinburgh, Scotland, 2005.

[Vol81] M. Volle. *L'Analyse des données (2^e éd.)*. Economica, 1981.

[War63] J.H.Jr Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58 :236–244, 1963.

[WBF98] J.B. Weinberg, G. Biswas, and D. Fisher. ITERATE : A conceptual clustering algorithm for data mining. In *Proceedings of IEEE Transactions on Systems, Man and Cybernetics*, 1998.

[WBK92] J.B. Weinberg, G. Biswas, and G.R. Koller. Conceptual Clustering with Systematic Missing Values. In *Proceedings of 9^{th} Int. Workshop on Machine Learning*, pages 464–469, 1992.

[WCRS01] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[Wit53] L. Wittgenstein. *Philosophical Investigations*. Blackwell, Oxford, UK, 1953.

[XP05] N. Xue and M. Palmer. Automatic Semantic Role Labeling for Chinese Verbs. In *Proceedings of the 19^{th} International Joint Conference on Artificial Intelligence*, pages 1160–1165, 2005.

[Zad65] L.A. Zadeh. Fuzzy sets. *Informatic Control*, 8 :338–353, 1965.

[ZZLY02] J. Zhong, H. Zhu, J. Li, and Y. Yu. Conceptual Graph Matching for Semantic Search. In *Proceedings of the 10^{th} International Conference on Conceptual Structures : Integration and Interfaces*, pages 92–106. Spring-Verlag, 2002.

ANNEXES

Annexe A

Tableau récapitulatif des indices d'évaluation

Nom	Description	Formule
Indices associés aux descriptions D :		
n_D	Le nombre de descriptions découvertes par l'algorithme de classification.	$card(\{d \in D / d \neq \top\})$
sep_1	Le score de séparation basé sur une mesure de type Russel et Rao.	$\frac{\sum_{d \in D} \sum_{d' \in D/\ d' \neq d} \exp - \frac{dis^2(d,d')}{2\ n_{\mathcal{A}}}}{n_D \times (n_D - 1)}$
sep_2	Le score de séparation basé sur une mesure de type Jaccard.	$\frac{\sum_{d \in D} \sum_{d' \in D/\ d' \neq d} \exp - \frac{dis^2(d,d')}{2\ card(\{X \in \mathcal{A}/\ D_X(d) \text{ ou } D_X(d')\})}}{n_D \times (n_D - 1)}$
Indices associés aux données E :		
n_E	Le nombre d'exemples que l'on cherche à classer.	$card(E)$ ou $n_I \times dup$ si E est artificiel
n_d	Le nombre total de descripteurs décrits par E.	$\sum_{e \in E} \lvert\delta(e)\rvert \times \rho(e)$
m	La proportion de descripteurs manquants dans E.	$1 - [\frac{n_d(E)}{n_{\mathcal{A}} \times \rho(E)}]$ ou m si E est artificiel
Indices associés à la couverture des données (A) :		
$couv$	Le taux de couverture des exemples.	$\frac{1}{\rho(E)} \times \sum_{e \in E / C_D(e) \neq \top} \rho(e)$
cmp_1	Le score de compacité basé sur une mesure de type Russel et Rao.	$\frac{1}{\rho(E)} \sum_{d_i \in D} \frac{v_1(E_i,d_i)}{v_1(E,d_G)} \rho(E_i)$ où $v_1(E,d) = \sqrt{\frac{1}{\rho(E)} \sum_{e \in E} \frac{dis^2(\delta(e),d)}{n_{\mathcal{A}}} \rho(e)}$
cmp_2	Le score de compacité basé sur une mesure de type Jaccard.	$\frac{1}{\rho(E)} \sum_{d_i \in D} \frac{v_2(E_i,d_i)}{v_2(E,d_G)} \rho(E_i)$ où $v_2(E,d) =$ $\sqrt{\frac{1}{\rho(E)} \sum_{e \in E} \frac{dis^2(\delta(e),d)}{card(\{X \in \mathcal{A}/D_X(\delta(e)) \text{ ou } D_X(d)\})} \rho(e)}$

$adeq_E$	La proportion des descripteurs de E en adéquation avec D.	$\frac{\sum_{e\in E} card(\{X\in\mathcal{A}/\ D_X(\delta(e),C_D(e))\text{ et }\tau_{C_D(e)}(\delta(e)_{\mid X})\})\times\rho(e)}{n_d(E)}$
$cont_E$	La proportion des descripteurs de E en contradiction avec D.	$\frac{\sum_{e\in E} card(\{X\in\mathcal{A}/\ D_X(\delta(e),C_D(e))\text{ et }\neg\tau_{C_D(e)}(\delta(e)_{\mid X})\})\times\rho(e)}{n_d(E)}$
$perte_E$	La proportion des descripteurs de E non présents dans D.	$\frac{\sum_{e\in E} card(\{X\in\mathcal{A}/\ D_X(\delta(e))\text{ et }\neg D_X(C_D(e))\})\times\rho(e)}{n_d(E)}$
$pred$	Les capacités prédictives de D par rapport à E.	$\frac{1}{n_{\mathcal{A}}\times\rho(E)}\times\sum_{d\in D}\lvert d\rvert\times\rho_{D,E}(d)$
$cont_s$	La proportion des exemples de E exhibant au moins une contradiction avec la description qui les couvre.	$\frac{1}{\rho(E)}\times\sum_{e\in E/\neg\tau(\delta(e),C_D(e))}\rho(e)$
$cont_p$	La proportion des descripteurs contradictoires (support) au sein des exemples exhibant au moins une contradiction avec la description qui les couvre.	$[cont_{sup}(D,E)\times\rho(E)]$ $\times$ $\sum_{e\in E/\neg\tau(\delta(e),C_S(e))}[\rho(e)$ $\times\frac{card(\{X\in\mathcal{A}/\ D_X(\delta(e),C_D(e))\text{ et }\neg\tau_{C_D(e)}(\delta(e)_{\mid X}))\})}{\lvert\delta(e)\rvert}]$

Indices associés aux descriptions initiales I et à la génération des données artificielles (B) :

n_I	Le nombre de descriptions initiales générées.	fixé
dup	Le nombre de duplication de chaque description initiale de I.	fixé
m	La proportion de descripteurs manquants fixée pour les jeux de données artificielles.	fixé

Indices d'adéquation entre D et I (C) :

$adeq_I$	La proportion des descripteurs de I retrouvés dans D.	$\frac{1}{\rho_E(D)}\times\sum_{d\in D/d\neq\top}\frac{sim(d,\mu_I(d))}{\lvert\mu_I(s)\rvert}\times\rho_{D,E}(d)$
$cont_I$	La proportion des descripteurs de I contradictoires avec D.	$\frac{1}{\rho_E(D)}\times\sum_{d\in D/d\neq\top}\frac{cont(d,\mu_I(d))}{\lvert\mu_I(d)\rvert}\times\rho_{D,E}(d)$
$perte_I$	La proportion des descripteurs de I qui n'apparaissent pas dans D.	$\frac{1}{\rho_E(D)}\times\sum_{d\in D/d\neq\top}\frac{abst(d,\mu_I(d))}{\lvert\mu_I(s)\rvert}\times\rho_{D,E}(d)$

Erreur de classification inférée par D :

err_C	La proportion d'exemples dont la classe initiale est mal prédite.	$\frac{1}{\rho(E)}\times\sum_{e\in E/\mu(e)\neq\mu_{D,I,E}(C_D(e))}\rho(e)$

Annexe B

Deux grandes affaires de la fin du XIX^{e} siècle

B.1 Le scandale de Panama

En 1878, le gouvernement de Colombie octroie à la France, dans l'isthme de Panama, un canal inter océanique. Ferdinand de Lesseps, créateur du canal de Suez en 1869, se fit confier l'ouvrage. Mais les obstacles techniques mirent la compagnie de Panama en difficulté et la contraignirent à faire appel à l'épargne française. L'emprunt fut confié à de grands financiers comme Cornélius Hertz ou le baron Jacques de Reinach. Dix ans après le début des travaux, le choix technique primitif se révéla un échec. Lesseps dut faire appel à l'ingénieur Gustave Eiffel pour concevoir un canal de l'écluse.

En 1888, à court d'argent, la compagnie tenta d'obtenir l'autorisation d'émettre un emprunt à lots (une loterie récompensant certains épargnants) pour lequel le vote d'une loi était nécessaire. Le suffrage d'une partie des parlementaires et l'appui de certains journaux furent alors obtenus par la corruption. Toutefois, l'emprunt n'empêchant pas la faillite en 1889 de la compagnie de Panama, le canal fut alors confié aux Etats-Unis. Plusieurs dizaines de milliers de souscripteurs furent ruinés et une instruction judiciaire s'ouvrit en 1891.

Le scandale fut rendu public en 1892 lorsque La Libre Parole d'Édouard Drumont et la presse boulangiste dénoncèrent les députés compromis. Le 20 novembre 1892, le baron de Reinach mourut subitement : une commission d'enquête parlementaire fut ouverte et l'autopsie demandée. La campagne contre "les chéquards", les révélations successives compromettant des députés tels que Maurice Rouvier, Charles Floquet et surtout Georges Clemenceau entraînèrent une crise ministérielle.

En 1893, le procès contre les administrateurs aboutit à un verdict léger ; parmi les parlementaires, seul le ministre des travaux publics Charles Baihart fut condamné à cinq ans de détention. La révélation de la corruption des députés frappa plus l'opinion que celle de la vénalité de la presse. Si le scandale de Panama n'ébranla pas la République comme l'auraient souhaité les boulangistes, il laissa cependant des traces profondes.

Aux élections de 1893, Clemenceau ne fut pas réélu. En outre, le scandale de Panama

favorisa un mouvement d'opinion anti-parlementaire, anti-capitaliste et violemment antisémite, dénonçant les financiers juifs. L'épargne française demeura insuffisante pour couvrir les investissements, et la troisième République s'orienta vers un capitalisme rentré. L'attention soupçonneuse portée par l'opinion aux liens entre le monde des affaires et le parlement demeura par la suite un trait caractéristique de la vie politique française.

B.2 L'Affaire Dreyfus

En 1894, le capitaine Alfred Dreyfus (1859-1935), israélite alsacien, fut accusé d'espionnage et condamné par un tribunal militaire à la dégradation et à la déportation dans l'île du Diable. Deux ans plus tard, il fut prouvé que le jugement était fondé sur des documents falsifiés et l'on eut de sérieuses raisons de penser qu'un officier criblé de dettes, le commandant Esterhazy (1847-1923), était le vrai coupable.

Celui-ci, après un simulacre de procès, fut néanmoins acquitté. C'est alors que Clémenceau publia dans son journal, *L'Aurore*, un article d'Emile Zola intitulé "J'accuse", qui faisait peser contre l'état-major de très lourdes charges. Il apparut de plus en plus clairement que certains militaires, cléricaux et antisémites s'efforçaient d'empêcher une révision du procès.

L'Affaire devint politique, partagea la France en deux camps (dreyfusards et antidreyfusards) et faillit ébranler la République, cependant que l'opinion internationale s'indignait de l'injustice commise. Anatole France puis Jean Jaurès défendirent Alfred Dreyfus avec ardeur. En 1899, celui-ci fut renvoyé devant le tribunal militaire de Rennes et de nouveau déclaré coupable. Il fut amnistié la même année, mais ce n'est qu'en 1906 qu'il fut complètement réhabilité.

Annexe C

Le langage de description

Attribut	Description	Valeurs
Politique	Le groupement politique auquel l'Agent fait partie. On ne parlait pas encore de "parti" politique à la fin du XIX^{e} siècle. Les groupements "radical" et "opportuniste" sont républicains, alors que "boulangistes" et "monarchistes" sont conservateurs.	radical, opportuniste, boulangiste, monarchiste, centre
Tendance	Il s'agit des principales tendances politiques auxquelles l'Agent peut être rattaché.	republicain, conservateur, liberal, anarchiste
Révolutionnaire	L'Agent prône le recours à la révolution pour résoudre les problèmes de la République.	oui – non
Socialiste	L'Agent adhère à la doctrine socialiste. En effet, le socialisme n'était pas encore un groupement politique à proprement parler à la fin du XIX^{e} siècle.	oui – non
OrigineEtrangere	L'Agent est d'origine étrangère. Cela peut être explicitement écrit dans le texte ou déduit du nom de famille s'il est suffisamment caractéristique.	oui – non
Religion	La religion pratiquée par l'Agent.	catholique, protestant, juif

RelationJuifs	L'Agent entretient des relations avec des personnes de confession juive, qu'il soit lui-même juif ou non.	oui – non
Relation Protestants	L'Agent entretient des relations avec des personnes de confession protestante, qu'il soit lui-même protestant ou non.	oui – non
Relation Catholiques	L'Agent entretient des relations avec des personnes de confession catholique, qu'il soit lui-même catholique ou non.	oui – non
RelationEtranger	L'Agent entretient des relations dans un pays étranger, que ce soit en Angleterre, en Allemagne, en Russie, etc.	oui – non
Implication FrancsMacons	Le désordre de la scène politique suppose l'implication de membres de la franc-maçonnerie.	oui – non
Implication Gouvernementale	Le désordre de la scène politique suppose l'implication de membres du gouvernement (président, ministre, etc.).	oui – non
Patriotisme	L'Agent est profondément attaché et dévoué à sa patrie ou au contraire est un "sans-patrie".	patriote, antipatriote
Clericalisme	L'Agent est partisant d'une d'une forte influence du clergé, dans le domaine temporel et plus spécialement dans le domaine politique, ou au contraire est anticlérical.	clerical, anticlerical
Internationaliste	L'Agent est pour une fédération des nations de l'Europe, pour une entraide entre les partis politiques des différents pays. Attention, à la fin du XIX^{e} siècle, ce terme ne s'oppose pas nécessairement à celui de patriote.	oui – non
HonneteHomme	L'Agent est considéré comme un "honnête homme" au sens du XIX^{e} siècle, c'est-à-dire une personne loyale, honorable, honnête, etc.	oui – non

RespectLoi	L'Agent respecte ou non la loi. S'il a été condamné par un tribunal, la réponse est bien évidemment non.	oui – non
Traitre	L'Agent est considéré comme ayant trahi son pays au profit d'une puissance étrangère.	oui – non
Favoritisme	L'Agent pratique le favoritisme, c'est-à-dire qu'il n'hésite pas à utiliser sa fonction pour donner des avantages à ses proches.	oui – non
Corruption	L'Agent est corrompu. En échanges de promesses, pots-de-vins ou d'autres avantages, il se laisse facilement acheter.	oui – non
SensMoral	L'Agent possède un sens aigu des valeurs morales.	oui – non
Impunite	L'Agent est à l'abri des lois, en haut de sa "tour d'ivoire".	oui – non
Sentiment Superieur	L'Agent est dédaigneux, suffisant. Il se sent manifestement supérieur à ses contemporains.	oui – non
Incompetence	L'Agent n'est pas considéré comme suffisamment compétent pour assumer sa fonction politique. Ceci peut être dû à un manque d'éducation, parce qu'il ne prend pas le temps nécessaire, etc.	oui – non
PbViePrivee	L'Agent rencontre des problèmes dans sa vie privée.	oui – non
PbSante	L'Agent connaît des problèmes de santé.	oui – non
Dangereux	L'Agent est considéré comme mauvais, au sens de néfaste ou dangereux.	oui – non
MeleAffaire DArgent	L'Agent est mêlé à une ou plusieurs affaires louches mettant en jeu de l'argent : ressources occultes, trafics, subvention d'un parti par l'étranger, etc.	oui – non

Action	Il s'agit de l'action particulière dont laquelle l'Agent est impliqué.	demission, suicide, diffamation, incident, duel, trafic-electoral, soins, trahison-electorale, affront-religieux, tromperie, avantages, injures, prosélithisme
VictimeViolence (ordonné)	L'Agent est victime de violences verbales, d'incivilités (un peu), ou de blessures physiques, de tentatives de meurtre (beaucoup).	pas, unpeu, beaucoup
SourceViolence (ordonné)	L'Agent est la source de violences verbales, d'incivilités (un peu), ou de blessures physiques, de tentatives de meurtre (beaucoup).	pas, unpeu, beaucoup
Complot	L'article affirme l'existence d'un vaste complot dans lequel l'Agent est impliqué.	oui – non
LieClemenceau	L'Agent est lié, d'une façon ou d'une autre, à G. Clémenceau.	oui – non

Annexe D

Les données du quotidien Le Matin

```
<?xml version="1.0" encoding="ISO-8859-1" ?>

<racine>

*** Le Matin, 1er septembre 1893

<instance> 0
  <poids>2</poids>
  <jour>1</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Socialiste"> <valeur>oui</valeur> </attribut>
  <attribut nom="Revolutionnaire"> <valeur>oui</valeur> </attribut>
  <attribut nom="Internationaliste"> <valeur>oui</valeur> </attribut>
  <attribut nom="ImplicationGouvernementale"> <valeur>oui</valeur> </attribut>
  <attribut nom="Clericalisme"> <valeur>anticlerical</valeur> </attribut>
  <attribut nom="RespectLoi"> <valeur>non</valeur> </attribut>
  <attribut nom="HonneteHomme"> <valeur>non</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
  <attribut nom="Impunite"> <valeur>oui</valeur> </attribut>
  <attribut nom="Action"> <valeur>trafic-electoral</valeur> </attribut>
</instance>

<instance> 1
  <poids>2</poids>
  <jour>1</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Tendance"> <valeur>republicain</valeur> </attribut>
  <attribut nom="Clericalisme"> <valeur>anticlerical</valeur> </attribut>
  <attribut nom="RespectLoi"> <valeur>non</valeur> </attribut>
  <attribut nom="ImplicationGouvernementale"> <valeur>oui</valeur> </attribut>
```

```
  <attribut nom="HonneteHomme"> <valeur>non</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
  <attribut nom="Action"> <valeur>trafic-electoral</valeur> </attribut>
</instance>

<instance> 2
  <poids>2</poids>
  <jour>1</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Traitre"> <valeur>oui</valeur> </attribut>
  <attribut nom="RespectLoi"> <valeur>non</valeur> </attribut>
  <attribut nom="ImplicationGouvernementale"> <valeur>oui</valeur> </attribut>
  <attribut nom="HonneteHomme"> <valeur>non</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
  <attribut nom="Action"> <valeur>trafic-electoral</valeur> </attribut>
</instance>

<instance> 3
  <jour>1</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="PbSante"> <valeur>oui</valeur> </attribut>
  <attribut nom="ImplicationGouvernementale"> <valeur>oui</valeur> </attribut>
  <attribut nom="Action"> <valeur>soins</valeur> </attribut>
</instance>

<instance> 4
  <jour>1</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="PbSante"> <valeur>oui</valeur> </attribut>
  <attribut nom="Action"> <valeur>incident</valeur> </attribut>
</instance>

<instance> 5
  <jour>1</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Patriotisme"> <valeur>antipatriote</valeur> </attribut>
  <attribut nom="Traitre"> <valeur>oui</valeur> </attribut>
  <attribut nom="Patriotisme"> <valeur>antipatriote</valeur> </attribut>
  <attribut nom="Corruption"> <valeur>oui</valeur> </attribut>
  <attribut nom="OrigineEtrangere"> <valeur>oui</valeur> </attribut>
  <attribut nom="Favoritisme"> <valeur>oui</valeur> </attribut>
  <attribut nom="RelationEtranger"> <valeur>oui</valeur> </attribut>
```

```
  <attribut nom="HonneteHomme"> <valeur>non</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
  <attribut nom="MeleAffaireDArgent"> <valeur>oui</valeur> </attribut>
  <attribut nom="SourceViolence"> <valeur>unpeu</valeur> </attribut>
  <attribut nom="Action"> <valeur>incident</valeur> </attribut>
  <attribut nom="LieClemenceau"> <valeur>oui</valeur> </attribut>
</instance>

<instance> 6
  <jour>1</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="VictimeViolence"> <valeur>beaucoup</valeur> </attribut>
  <attribut nom="Action"> <valeur>incident</valeur> </attribut>
  <attribut nom="LieClemenceau"> <valeur>oui</valeur> </attribut>
</instance>

*** Le Matin, 2 septembre 1893

<instance> 7
  <poids>2</poids>
  <jour>2</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Action"> <valeur>duel</valeur> </attribut>
  <attribut nom="SourceViolence"> <valeur>beaucoup</valeur> </attribut>
</instance>

<instance> 8
  <jour>2</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Patriotisme"> <valeur>antipatriote</valeur> </attribut>
  <attribut nom="RelationEtranger"> <valeur>oui</valeur> </attribut>
  <attribut nom="LieClemenceau"> <valeur>oui</valeur> </attribut>
</instance>

<instance> 9
  <jour>2</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="PbSante"> <valeur>oui</valeur> </attribut>
</instance>

<instance> 10
  <jour>2</jour>
```

```
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="ImplicationGouvernementale"> <valeur>oui</valeur> </attribut>
  <attribut nom="PbSante"> <valeur>oui</valeur> </attribut>
  <attribut nom="Action"> <valeur>soins</valeur> </attribut>
</instance>

<instance> 11
  <jour>2</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="PbViePrivee"> <valeur>oui</valeur> </attribut>
  <attribut nom="SourceViolence"> <valeur>unpeu</valeur> </attribut>
</instance>

<instance> 12
  <jour>2</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="RespectLoi"> <valeur>non</valeur> </attribut>
  <attribut nom="Action"> <valeur>diffamation</valeur> </attribut>
  <attribut nom="HonneteHomme"> <valeur>non</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
  <attribut nom="SourceViolence"> <valeur>unpeu</valeur> </attribut>
</instance>

<instance> 13
  <jour>2</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="VictimeViolence"> <valeur>unpeu</valeur> </attribut>
  <attribut nom="Action"> <valeur>diffamation</valeur> </attribut>
</instance>

<instance> 14
  <jour>2</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="VictimeViolence"> <valeur>unpeu</valeur> </attribut>
  <attribut nom="Action"> <valeur>diffamation</valeur> </attribut>
</instance>

<instance> 15
  <jour>2</jour>
  <mois>9</mois>
  <an>1893</an>
```

```
  <attribut nom="Socialiste"> <valeur>oui</valeur> </attribut>
  <attribut nom="RespectLoi"> <valeur>non</valeur> </attribut>
  <attribut nom="HonneteHomme"> <valeur>non</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
  <attribut nom="Action"> <valeur>diffamation</valeur> </attribut>
  <attribut nom="SourceViolence"> <valeur>unpeu</valeur> </attribut>
</instance>

*** Le Matin, 3 septembre 1893

<instance> 16
  <jour>3</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Tendance"> <valeur>republicain</valeur> </attribut>
  <attribut nom="HonneteHomme"> <valeur>oui</valeur> </attribut>
  <attribut nom="Action"> <valeur>suicide</valeur> </attribut>
</instance>

<instance> 17
  <jour>3</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Socialiste"> <valeur>oui</valeur> </attribut>
  <attribut nom="Incompetence"> <valeur>oui</valeur> </attribut>
  <attribut nom="Dangereux"> <valeur>oui</valeur> </attribut>
</instance>

<instance> 18
  <jour>3</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Politique"> <valeur>boulangiste</valeur> </attribut>
  <attribut nom="Tendance"> <valeur>conservateur</valeur> </attribut>
  <attribut nom="Incompetence"> <valeur>oui</valeur> </attribut>
  <attribut nom="Dangereux"> <valeur>oui</valeur> </attribut>
</instance>

<instance> 19
  <jour>3</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Tendance"> <valeur>anarchiste</valeur> </attribut>
  <attribut nom="Incompetence"> <valeur>oui</valeur> </attribut>
  <attribut nom="Dangereux"> <valeur>oui</valeur> </attribut>
</instance>
```

```
<instance> 20
  <jour>3</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Action"> <valeur>incident</valeur> </attribut>
</instance>

*** Le Matin, 4 septembre 1893

<instance> 21
  <jour>4</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="RespectLoi"> <valeur>non</valeur> </attribut>
  <attribut nom="HonneteHomme"> <valeur>non</valeur> </attribut>
  <attribut nom="RelationEtranger"> <valeur>oui</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
  <attribut nom="ImplicationGouvernementale"> <valeur>oui</valeur> </attribut>
  <attribut nom="MeleAffaireDArgent"> <valeur>oui</valeur> </attribut>
  <attribut nom="Impunite"> <valeur>oui</valeur> </attribut>
  <attribut nom="Action"> <valeur>trafic-electoral</valeur> </attribut>
  <attribut nom="LieClemenceau"> <valeur>oui</valeur> </attribut>
</instance>

<instance> 22
  <jour>4</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Socialiste"> <valeur>oui</valeur> </attribut>
  <attribut nom="ImplicationGouvernementale"> <valeur>oui</valeur> </attribut>
  <attribut nom="Incompetence"> <valeur>oui</valeur> </attribut>
  <attribut nom="Dangereux"> <valeur>oui</valeur> </attribut>
</instance>

<instance> 23
  <jour>4</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Socialiste"> <valeur>oui</valeur> </attribut>
  <attribut nom="Dangereux"> <valeur>oui</valeur> </attribut>
</instance>

*** Le Matin, 5 septembre 1893

<instance> 24
```

```
  <poids>2</poids>
  <jour>5</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Politique"> <valeur>centre</valeur> </attribut>
  <attribut nom="Socialiste"> <valeur>non</valeur> </attribut>
  <attribut nom="Incompetence"> <valeur>oui</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
  <attribut nom="LieClemenceau"> <valeur>oui</valeur> </attribut>
</instance>

<instance> 25
  <poids>2</poids>
  <jour>5</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Socialiste"> <valeur>oui</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
</instance>

<instance> 26
  <jour>5</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Socialiste"> <valeur>oui</valeur> </attribut>
  <attribut nom="HonneteHomme"> <valeur>non</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
</instance>

<instance> 27
  <jour>5</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="HonneteHomme"> <valeur>non</valeur> </attribut>
  <attribut nom="ImplicationGouvernementale"> <valeur>oui</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
  <attribut nom="Impunite"> <valeur>oui</valeur> </attribut>
  <attribut nom="LieClemenceau"> <valeur>oui</valeur> </attribut>
</instance>

<instance> 28
  <jour>5</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Politique"> <valeur>radical</valeur> </attribut>
  <attribut nom="Tendance"> <valeur>republicain</valeur> </attribut>
```

```
  <attribut nom="Clericalisme"> <valeur>anticlerical</valeur> </attribut>
  <attribut nom="HonneteHomme"> <valeur>non</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
  <attribut nom="ImplicationFrancsMacons"> <valeur>oui</valeur> </attribut>
  <attribut nom="Action"> <valeur>trahison-electorale</valeur> </attribut>
</instance>

<instance> 29
  <jour>5</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Action"> <valeur>duel</valeur> </attribut>
  <attribut nom="SourceViolence"> <valeur>beaucoup</valeur> </attribut>
</instance>

<instance> 30
  <jour>5</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="RespectLoi"> <valeur>non</valeur> </attribut>
  <attribut nom="ImplicationGouvernementale"> <valeur>oui</valeur> </attribut>
  <attribut nom="HonneteHomme"> <valeur>non</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
  <attribut nom="Impunite"> <valeur>oui</valeur> </attribut>
</instance>

*** Le Matin, 6 septembre 1893

<instance> 31
  <jour>6</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Impunite"> <valeur>oui</valeur> </attribut>
  <attribut nom="Corruption"> <valeur>oui</valeur> </attribut>
  <attribut nom="ImplicationGouvernementale"> <valeur>oui</valeur> </attribut>
  <attribut nom="HonneteHomme"> <valeur>non</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
  <attribut nom="Incompetence"> <valeur>oui</valeur> </attribut>
  <attribut nom="PbSante"> <valeur>oui</valeur> </attribut>
</instance>

<instance> 32
  <jour>6</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Socialiste"> <valeur>oui</valeur> </attribut>
```

```
  <attribut nom="Revolutionnaire"> <valeur>oui</valeur> </attribut>
  <attribut nom="Patriotisme"> <valeur>antipatriote</valeur> </attribut>
  <attribut nom="RelationEtranger"> <valeur>oui</valeur> </attribut>
  <attribut nom="MeleAffaireDArgent"> <valeur>oui</valeur> </attribut>
</instance>

<instance> 33
  <jour>6</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Politique"> <valeur>opportuniste</valeur> </attribut>
  <attribut nom="Tendance"> <valeur>republicain</valeur> </attribut>
  <attribut nom="Clericalisme"> <valeur>clerical</valeur> </attribut>
  <attribut nom="HonneteHomme"> <valeur>non</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
  <attribut nom="LieClemenceau"> <valeur>oui</valeur> </attribut>
</instance>

*** Le Matin, 7 septembre 1893

<instance> 34
  <poids>2</poids>
  <jour>7</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Socialiste"> <valeur>oui</valeur> </attribut>
  <attribut nom="Revolutionnaire"> <valeur>oui</valeur> </attribut>
  <attribut nom="RelationEtranger"> <valeur>oui</valeur> </attribut>
  <attribut nom="MeleAffaireDArgent"> <valeur>oui</valeur> </attribut>
</instance>

<instance> 35
  <poids>2</poids>
  <jour>7</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Action"> <valeur>incident</valeur> </attribut>
</instance>

<instance> 36
  <poids>2</poids>
  <jour>7</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Tendance"> <valeur>republicain</valeur> </attribut>
  <attribut nom="VictimeViolence"> <valeur>unpeu</valeur> </attribut>
```

```
  <attribut nom="RelationEtranger"> <valeur>non</valeur> </attribut>
  <attribut nom="MeleAffaireDArgent"> <valeur>non</valeur> </attribut>
  <attribut nom="Traitre"> <valeur>non</valeur> </attribut>
  <attribut nom="HonneteHomme"> <valeur>oui</valeur> </attribut>
  <attribut nom="LieClemenceau"> <valeur>oui</valeur> </attribut>
</instance>

<instance> 37
  <poids>2</poids>
  <jour>7</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Tendance"> <valeur>republicain</valeur> </attribut>
  <attribut nom="Incompetence"> <valeur>oui</valeur> </attribut>
</instance>

<instance> 38
  <jour>7</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Action"> <valeur>demission</valeur> </attribut>
</instance>

<instance> 39
  <jour>7</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="PbSante"> <valeur>oui</valeur> </attribut>
  <attribut nom="Action"> <valeur>soins</valeur> </attribut>
  <attribut nom="ImplicationGouvernementale"> <valeur>oui</valeur> </attribut>
</instance>

*** Le Matin, 8 septembre 1893

<instance> 40
  <poids>2</poids>
  <jour>8</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Action"> <valeur>demission</valeur> </attribut>
</instance>

<instance> 41
  <jour>8</jour>
  <mois>9</mois>
  <an>1893</an>
```

```
  <attribut nom="Socialiste"> <valeur>oui</valeur> </attribut>
  <attribut nom="Clericalisme"> <valeur>anticlerical</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
  <attribut nom="Action"> <valeur>affront-religieux</valeur> </attribut>
</instance>

*** Le Matin, 9 septembre 1893

<instance> 42
  <poids>3</poids>
  <jour>9</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Socialiste"> <valeur>oui</valeur> </attribut>
  <attribut nom="Patriotisme"> <valeur>antipatriote</valeur> </attribut>
  <attribut nom="Internationaliste"> <valeur>oui</valeur> </attribut>
  <attribut nom="Traitre"> <valeur>oui</valeur> </attribut>
  <attribut nom="HonneteHomme"> <valeur>non</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
  <attribut nom="RelationEtranger"> <valeur>oui</valeur> </attribut>
  <attribut nom="MeleAffaireDArgent"> <valeur>oui</valeur> </attribut>
  <attribut nom="Dangereux"> <valeur>oui</valeur> </attribut>
</instance>

<instance> 43
  <jour>9</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Patriotisme"> <valeur>antipatriote</valeur> </attribut>
  <attribut nom="Internationaliste"> <valeur>oui</valeur> </attribut>
</instance>

<instance> 44
  <jour>9</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Action"> <valeur>demission</valeur> </attribut>
</instance>

<instance> 45
  <jour>9</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Politique"> <valeur>opportuniste</valeur> </attribut>
  <attribut nom="Tendance"> <valeur>republicain</valeur> </attribut>
  <attribut nom="HonneteHomme"> <valeur>non</valeur> </attribut>
```

```
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
  <attribut nom="Socialiste"> <valeur>non</valeur> </attribut>
  <attribut nom="Action"> <valeur>trahison-electorale</valeur> </attribut>
</instance>

<instance> 46
  <jour>9</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="PbSante"> <valeur>oui</valeur> </attribut>
  <attribut nom="MeleAffaireDArgent"> <valeur>oui</valeur> </attribut>
  <attribut nom="Action"> <valeur>soins</valeur> </attribut>
</instance>

*** Le Matin, 10 septembre 1893

<instance> 47
  <jour>10</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="Socialiste"> <valeur>oui</valeur> </attribut>
  <attribut nom="Tendance"> <valeur>republicain</valeur> </attribut>
  <attribut nom="Dangereux"> <valeur>oui</valeur> </attribut>
</instance>

<instance> 48
  <jour>10</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="OrigineEtrangere"> <valeur>oui</valeur> </attribut>
  <attribut nom="HonneteHomme"> <valeur>non</valeur> </attribut>
  <attribut nom="PbSante"> <valeur>non</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
  <attribut nom="MeleAffaireDArgent"> <valeur>oui</valeur> </attribut>
  <attribut nom="ImplicationGouvernementale"> <valeur>oui</valeur> </attribut>
  <attribut nom="Action"> <valeur>tromperie</valeur> </attribut>
</instance>

<instance> 49
  <jour>10</jour>
  <mois>9</mois>
  <an>1893</an>
  <attribut nom="MeleAffaireDArgent"> <valeur>oui</valeur> </attribut>
  <attribut nom="RespectLoi"> <valeur>non</valeur> </attribut>
  <attribut nom="HonneteHomme"> <valeur>non</valeur> </attribut>
  <attribut nom="SensMoral"> <valeur>non</valeur> </attribut>
```

```
</instance>

</racine>
```

Annexe E

Résultats avec les jeux de données artificiels

E.1 Comparaison des fonctions d'évaluation

E.1.1 Résultats obtenus à partir de ART-1

	PRESS							
	q_{N_1}		q_{N_2}		q_{N_3}		q_{N_4}	
	μ	σ	μ	σ	μ	σ	μ	σ
n_D	3.04	± 0.20	25.6	± 1.54	24.2	± 1.56	3.66	± 0.65
$couv$	97.91	± 1.22	100	± 0	100	± 0	98.01	± 1.22
$pred$	65.92	± 0.91	6.54	± 0.33	9.66	± 1.03	61.85	± 4.65
$adeq_E$	67.13	± 1.51	28.22	± 0.95	32.34	± 1.59	64.26	± 3.68
err_C	7.3	± 2.44	24.22	± 3.19	22.24	± 3.41	9.16	± 3.87
$adeq_I$	99.42	± 1.43	7.94	± 0.79	11.58	± 1.93	91.2	± 9.64
$cont_I$	0	± 0	0.12	± 0.05	0.06	± 0.11	0.72	± 0.24
$perte_I$	0.58	± 1.43	92.04	± 0.79	88.36	± 1.91	8.73	± 9.51
cmp_1	0.8702	± 0.0026	0.9437	± 0.0017	0.9355	± 0.003	0.8755	± 0.0067
cmp_2	0.8281	± 0.0049	0.6958	± 0.008	0.6794	± 0.005	0.8245	± 0.0066
q_{N_1}	13.43	± 0.3	5.64	± 0.19	6.47	± 0.32	12.85	± 0.74
q_{N_2}	17.92	± 0.53	32.13	± 0.96	35.27	± 0.61	18.28	± 0.71
q_{N_3}	39.02	± 1.09	84.74	± 0.14	83.28	± 0.46	42.16	± 3.71
q_{N_4}	68.49	± 2.03	34.84	± 1.18	41.48	± 1.08	66.91	± 2.75

E.1.2 Résultats obtenus à partir de ART-2

	PRESS							
	q_{N_1}		q_{N_2}		q_{N_3}		q_{N_4}	
	μ	σ	μ	σ	μ	σ	μ	σ
n_D	5.57	± 0.69	37.12	± 1.76	37.82	± 1.67	5.8	± 0.69
$couv$	97.77	± 1.14	100	± 0	99.99	± 0.05	96.74	± 1.66
$adeq_E$	57.05	± 1.1	37.52	± 0.73	39.5	± 0.69	54.12	± 1.4
$pred$	44.1	± 2.86	4.03	± 0.11	4.55	± 0.17	43.26	± 3.56
err_C	65.77	± 2.93	50.64	± 1.84	50.28	± 1.97	67.56	± 3.31
$adeq_I$	42.9	± 7.96	5.44	± 0.85	5.85	± 0.99	38.58	± 6.96
$cont_I$	5.85	± 3.72	0.02	± 0.05	0.11	± 0.11	8.66	± 3.92
$perte_I$	51.25	± 7.84	94.54	± 0.85	94.04	± 1.01	52.76	± 6.24
cmp_1	0.9435	± 0	0.9625	± 0	0.9605	± 0	0.9462	± 0.001
cmp_2	0.8794	± 0.0083	0.5797	± 0.0053	0.5773	± 0.0048	0.8782	± 0.0106
q_{N_1}	5.71	± 0.11	3.76	± 0.07	3.96	± 0.69	5.52	± 0.14
q_{N_2}	12.36	± 0.88	43.83	± 0.67	45.23	± 0.59	12.45	± 1.12
q_{N_3}	55.34	± 3.43	93.48	± 0.06	93.34	± 0.1	55	± 4.59
q_{N_4}	62.3	± 1.17	45.12	± 0.77	47.96	± 0.61	64.59	± 1.25

E.2 Comparaison des techniques T_1, T_2, T_3 et T_4 pour EM

E.2.1 ART-1

	EM							
	T_1		T_2		T_3		T_4	
	μ	σ	μ	σ	μ	σ	μ	σ
$adeq_E$	88.18	± 0.97	66.58	± 2.17	68.29	± 1.46	66.05	± 2.05
$pred$	99.98	± 0.14	66.26	± 1.78	67.37	± 1.3	65.5	± 1.64
$cont$	11.82	± 0.97	0	± 0	2.77	± 1.2	0	± 0
$cont_{sup}$	46.36	± 3.78	0	± 0	14.44	± 5.53	0	± 0
$cont_{prop}$	23.36	± 1.59	0	± 0	17.41	± 2.81	0	± 0
$adeq_I$	99.99	± 0.08	97.76	± 2.64	90.95	± 4.04	97.45	± 3.05
$cont_I$	0	± 0	0	± 0	0	± 0	0	± 0
$perte_I$	0.01	± 0.08	2.24	± 2.64	9.05	± 4.04	2.55	± 3.05
sep_1	0.799	± 0.0022	0.6114	± 0.0137	0.6065	± 0	0.6065	± 0
cmp_1	0.8311	± 0.0017	0.8713	± 0.0041	0.8682	± 0.0026	0.8723	± 0.0039

E.2.2 ART-2

	EM							
	T_1		T_2		T_3		T_4	
	μ	σ	μ	σ	μ	σ	μ	σ
$adeq_E$	69.77	± 12.59	15.54	± 19.38	64.06	± 7.34	14.57	± 17.66
$pred$	99.93	± 0.25	15.41	± 18.87	91.48	± 14.52	14.08	± 16.58
$cont$	30.23	± 12.59	0	± 0	29.11	± 14.4	0	± 0
$cont_{sup}$	56.72	± 17.33	0	± 0	54.22	± 21.81	0	± 0
$cont_{prop}$	48.92	± 9.43	0	± 0	48.21	± 10.51	0	± 0
$adeq_I$	68.99	± 17.9	17.77	± 21.53	65.05	± 16.11	17.01	± 20.38
$cont_I$	30.91	± 18	1.45	± 2.82	28.43	± 19.84	1.21	± 2.55
$perte_I$	0.1	± 0.38	80.78	± 23.2	6.51	± 11.39	81.79	± 21.63
sep_1	0.6794	± 0.0869	0.6164	± 0.021	0.6065	± 0	0.6065	± 0
cmp_1	0.9317	± 0.0123	0.9847	± 0.019	0.9372	± 0.0073	0.9857	± 0.0173

E.3 Comparaison des techniques T_1, T_2, T_3 et T_4 pour COBWEB

E.3.1 ART-1

	COBWEB							
	T_1		T_2		T_3		T_4	
	μ	σ	μ	σ	μ	σ	μ	σ
$adeq_E$	86.15	± 2.75	50.1	± 10.31	49.36	± 7.97	36.22	± 7.72
$pred$	97.77	± 1.47	54.67	± 9.14	50.22	± 11.42	37.18	± 7.15
$cont$	13.85	± 2.75	0	± 0	6.04	± 2.46	0	± 0
$cont_{sup}$	48.86	± 7.09	0	± 0	25.24	± 8.11	0	± 0
$cont_{prop}$	26.87	± 3.12	0	± 0	24.38	± 4.59	0	± 0
$adeq_I$	89.52	± 7.23	62.35	± 13.88	56.4	± 15.12	47.53	± 13.5
$cont_I$	8.16	± 6.64	3.85	± 3.16	1.15	± 2.05	1.070	± 1.38
$perte_I$	2.32	± 1.59	33.8	± 12.44	42.45	± 14.7	51.4	± 12.72
sep_1	0.8185	± 0.0244	0.6838	± 0.0268	0.6065	± 0	0.6065	± 0
cmp_1	0.8344	± 0.005	0.9028	± 0.0197	0.9046	± 0.0149	0.9295	± 0.0147

E.3.2 ART-2

	COBWEB							
	T_1		T_2		T_3		T_4	
	μ	σ	μ	σ	μ	σ	μ	σ
$adeq_E$	83.76	± 2.04	45.22	± 6.24	48.58	± 3.1	32.55	± 2.67
$pred$	97.1	± 1.58	56.12	± 5.78	50.49	± 7.51	37.54	± 3.55
$cont$	16.24	± 2.04	0	± 0	9.46	± 2.55	0	± 0
$cont_{sup}$	34.78	± 4.25	0	± 0	23.2	± 5.55	0	± 0
$cont_{prop}$	40.15	± 1.91	0	± 0	38.27	± 3.02	0	± 0
$adeq_I$	65.48	± 7.11	40.1	± 7.66	38.56	± 6.99	28.92	± 6.72
$cont_I$	31.68	± 7.21	17.59	± 5.22	13.18	± 5.76	10.74	± 4.51
$perte_I$	2.84	± 2.22	42.31	± 7.35	48.26	± 6.99	60.34	± 6.85
sep_1	0.81	± 0.038	0.681	± 0.0185	0.6065	± 0	0.6065	± 0
cmp_1	0.9177	± 0.0017	0.9555	± 0.0061	0.9519	± 0.0028	0.9679	± 0.0024

E.4 Comparaison des techniques T_1, T_2, T_3 et T_4 pour les k-modes

E.4.1 ART-1

	K-modes							
	T_1		T_2		T_3		T_4	
	μ	σ	μ	σ	μ	σ	μ	σ
$adeq_E$	62.34	± 2.98	8.4	± 4.16	58.26	± 1.6	7.98	± 3.95
$pred$	99.3	± 0.51	6.18	± 4.15	96.15	± 2.81	5.83	± 3.98
$cont$	37.66	± 2.98	0	± 0	37.59	± 3.32	0	± 0
$cont_{sup}$	82.98	± 3.18	0	± 0	82.82	± 3.87	0	± 0
$cont_{prop}$	47.14	± 1.91	0	± 0	47.22	± 2.05	0	± 0
$adeq_I$	63.8	± 13.56	6.68	± 5.01	60.96	± 12.83	6.42	± 4.97
$cont_I$	35.5	± 13.4	0.34	± 0.71	35.27	± 13.49	0.31	± 0.69
$perte_I$	0.7	± 0.54	92.97	± 5.09	3.76	± 2.77	93.26	± 5.04
sep_1	0.7913	± 0.057	0.6247	± 0.0213	0.6065	± 0	0.6065	± 0
cmp_1	0.8796	± 0.0055	0.9835	± 0.008	0.8874	± 0.0026	0.9843	± 0.0076

E.4.2 ART-2

	K-modes							
	T_1		T_2		T_3		T_4	
	μ	σ	μ	σ	μ	σ	μ	σ
$adeq_E$	65.05	± 2.86	14.76	± 5.49	59.23	± 1.31	13.08	± 4.43
$pred$	96.95	± 1.39	12.4	± 5.45	92.61	± 3.63	11.05	± 4.41
$cont$	34.95	± 2.86	0	± 0	34.6	± 3.3	0	± 0
$cont_{sup}$	65.06	± 4.21	0	± 0	64.44	± 4.89	0	± 0
$cont_{prop}$	54.87	± 1.75	0	± 0	55.07	± 1.68	0	± 0
$adeq_I$	60.44	± 12.4	12.23	± 7.16	57.75	± 12.09	11.34	± 6.92
$cont_I$	35.59	± 12.51	2.09	± 2.37	35.12	± 12.78	1.790	± 2.34
$perte_I$	2.97	± 1.5	85.67	± 8.15	7.13	± 3.5	86.86	± 7.77
sep_1	0.7421	± 0.0562	0.6393	± 0.0189	0.6065	± 0	0.6065	± 0
cmp_1	0.9363	± 0.0026	0.9856	± 0.0053	0.9419	± 0.001	0.9872	± 0.0042

E.5 Comparaison des algorithmes avec ART-3

E.5.1 Utilisation de la technique T_2

- **Adéquation relative à I ($adeq_I$) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
50	100	± 0	100	± 0	99.98	± 0.16	14.78	± 7.91	99.21	± 1.75
60	100	± 0	100	± 0	99.84	± 0.59	10.57	± 5.29	97.47	± 3.56
70	100	± 0	99.99	± 0.15	99.61	± 0.95	8.14	± 3.61	86.42	± 12.49
75	100	± 0	99.37	± 2.16	99.33	± 1.04	8.09	± 4.03	78.27	± 12.97
80	100	± 0	99.28	± 1.84	98.1	± 2.17	7.74	± 4.85	62.94	± 13.13
85	100	± 0	96.29	± 7.31	95.51	± 2.96	8.62	± 5.78	51.19	± 10.78
90	100	± 0	69.09	± 14.16	75.67	± 16.26	7.12	± 4.29	44.71	± 8.44

- **Erreur de classification (err_C) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
50	0	± 0	0	± 0	0.09	± 0.23	41.83	± 6.53	0.19	± 0.43
60	0.43	± 0.62	0.43	± 0.64	0.69	± 0.6	50.31	± 6.25	1.67	± 1.46
70	2.64	± 1.34	2.51	± 1.31	2.19	± 1.19	56.21	± 4.68	8.7	± 5.72
75	4.1	± 1.81	4.42	± 1.78	3.77	± 1.59	57.79	± 4.23	15.48	± 6.7
80	7.26	± 2.11	7.49	± 2.19	6.71	± 1.99	59.82	± 3.3	26.81	± 8.08
85	11.31	± 2.69	12.36	± 3.99	10.98	± 2.5	60.95	± 3.33	39.31	± 7.38
90	17.14	± 3.21	29.79	± 8.48	26.54	± 9.12	62.69	± 1.97	49.69	± 5.32

- **Nombre de descriptions (n_D) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
50	3	± 0	3	± 0	3.02	± 0.14	2.93	± 0.29	6.43	± 1.44
60	3	± 0	3	± 0	3.02	± 0.14	2.81	± 0.44	5.74	± 1.35
70	3	± 0	3	± 0	3.01	± 0.1	2.8	± 0.47	6.04	± 1.67
75	3	± 0	3.08	± 0.27	3.02	± 0.14	2.79	± 0.43	6.31	± 1.59
80	3	± 0	3.07	± 0.29	3.01	± 0.1	2.79	± 0.43	6.14	± 1.48
85	3	± 0	3.23	± 0.51	3.01	± 0.1	2.76	± 0.49	5.55	± 1.19
90	3	± 0	4.39	± 0.81	2.91	± 0.4	2.73	± 0.53	4.34	± 0.85

- **Adéquation relative à E ($adeq_E$) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
50	66.62	± 0.64	66.62	± 0.64	66.63	± 0.67	12.12	± 6.42	68.83	± 1.89
60	66.73	± 0.92	66.73	± 0.92	66.67	± 1.06	9.48	± 4.61	67.83	± 2.87
70	66.89	± 1.05	66.9	± 1.05	66.75	± 1.23	8.45	± 3.24	62.57	± 8.44
75	67.08	± 1.17	66.81	± 1.56	66.84	± 1.31	8.78	± 3.35	58.84	± 7.84
80	67.39	± 1.37	67.22	± 1.55	66.84	± 1.84	9.36	± 4.07	51.05	± 9.11
85	67.96	± 1.67	67.38	± 2.56	67.5	± 2.29	10.16	± 4.94	46.33	± 8.26
90	69.41	± 2.15	68.22	± 2.83	64.09	± 12.19	10.46	± 4.97	47.24	± 7.14

- **Compacité (cmp_1) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
50	0.6738	± 0.0032	0.6738	± 0.0032	0.6738	± 0.0033	0.94	± 0.0317	0.6631	± 0.0092
60	0.74	± 0.0036	0.74	± 0.0036	0.7403	± 0.0041	0.9625	± 0.0181	0.7358	± 0.011
70	0.8054	± 0.003	0.8054	± 0.003	0.806	± 0.0036	0.975	± 0.0095	0.818	± 0.0242
75	0.8378	± 0.0026	0.8386	± 0.0036	0.8386	± 0.003	0.9784	± 0.0082	0.8583	± 0.0187
80	0.8696	± 0.0024	0.87	± 0.0028	0.8708	± 0.0035	0.9816	± 0.0079	0.901	± 0.0174
85	0.9011	± 0.002	0.9019	± 0.0035	0.9018	± 0.0032	0.9851	± 0.0071	0.9323	± 0.012
90	0.9318	± 0.002	0.9329	± 0.0024	0.9372	± 0.0119	0.9898	± 0.0047	0.9536	± 0.0069

- **Séparation (sep_1) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
50	0.6065	± 0	0.6065	± 0	0.6077	± 0.0084	0.6133	± 0.0145	0.7126	± 0.026
60	0.6065	± 0	0.6065	± 0	0.6079	± 0.01	0.6169	± 0.0206	0.7023	± 0.0286
70	0.6065	± 0	0.6065	± 0	0.6075	± 0.0069	0.6259	± 0.0278	0.6959	± 0.0313
75	0.6065	± 0	0.6065	± 0	0.6084	± 0.0095	0.6229	± 0.0173	0.695	± 0.0232
80	0.6065	± 0	0.6065	± 0	0.6103	± 0.0103	0.6271	± 0.02	0.6846	± 0.0228
85	0.6065	± 0	0.6065	± 0	0.6209	± 0.0101	0.6264	± 0.0195	0.6797	± 0.0195
90	0.6065	± 0	0.6065	± 0	0.6589	± 0.025	0.6204	± 0.0144	0.6774	± 0.0203

E.5.2 Utilisation de la technique T_4

- **Adéquation relative à I ($adeq_I$) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
50	100	± 0	100	± 0	99.73	± 1.86	14.26	± 7.99	71.72	± 12.16
60	100	± 0	100	± 0	99.72	± 0.98	10.04	± 5.08	76.48	± 11.15
70	100	± 0	99.99	± 0.15	99.45	± 1.68	7.64	± 3.59	66.1	± 14.48
75	100	± 0	99.37	± 2.16	99.29	± 1.12	7.76	± 3.96	60.8	± 14.7
80	100	± 0	99.28	± 1.84	97.9	± 2.49	7.41	± 4.74	47.8	± 13.09
85	100	± 0	96.29	± 7.31	94.79	± 3.38	8.25	± 5.4	38.08	± 9.63
90	100	± 0	69.09	± 14.16	72.66	± 15.21	6.94	± 4.18	36.26	± 7.81

- **Adéquation relative à E ($adeq_E$) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
50	66.62	± 0.64	66.62	± 0.64	66.47	± 1.12	11.73	± 6.46	50.21	± 7.44
60	66.73	± 0.92	66.73	± 0.92	66.56	± 1.17	9.1	± 4.46	52.78	± 6.72
70	66.89	± 1.05	66.9	± 1.05	66.62	± 1.43	7.97	± 3.03	46.59	± 8.45
75	67.08	± 1.17	66.81	± 1.56	66.71	± 1.27	8.36	± 3.23	43.41	± 8.37
80	67.39	± 1.37	67.22	± 1.55	66.44	± 1.76	8.85	± 3.83	37.29	± 7.51
85	67.96	± 1.67	67.38	± 2.56	66.21	± 2.07	9.59	± 4.53	34.16	± 5.48
90	69.41	± 2.15	68.22	± 2.83	59.68	± 10.22	9.97	± 4.63	38.45	± 4.48

- **Compacité (cmp_1) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
50	0.6738	± 0.0032	0.6738	± 0.0032	0.6745	± 0.0055	0.9419	± 0.0319	0.7538	± 0.0361
60	0.74	± 0.0036	0.74	± 0.0036	0.7407	± 0.0046	0.964	± 0.0175	0.7942	± 0.0259
70	0.8054	± 0.003	0.8054	± 0.003	0.8063	± 0.0041	0.9764	± 0.0089	0.8641	± 0.0242
75	0.8378	± 0.0026	0.8386	± 0.0036	0.8388	± 0.003	0.9794	± 0.0079	0.8947	± 0.0199
80	0.8696	± 0.0024	0.87	± 0.0028	0.8716	± 0.0033	0.9826	± 0.0074	0.9275	± 0.0144
85	0.9011	± 0.002	0.9019	± 0.0035	0.9037	± 0.0028	0.9859	± 0.0066	0.9499	± 0.0078
90	0.9318	± 0.002	0.9329	± 0.0024	0.9414	± 0.0099	0.9902	± 0.0044	0.9622	± 0.0042

E.6 Comparaison des algorithmes avec ART-4

E.6.1 Utilisation de la technique T_2

- **Adéquation relative à I ($adeq_I$) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
50	100	± 0	99.88	± 0.82	95.53	± 8.14	6.87	± 3.9	73.78	± 5.99
60	100	± 0	99.77	± 1.13	93.35	± 7.08	7	± 3.55	58.4	± 7.3
70	100	± 0	98.57	± 3.35	82.58	± 12.59	7.96	± 4.14	46.7	± 7.41
75	100	± 0	91.3	± 13.5	67.31	± 25.18	9.38	± 4.5	41.93	± 7.5
80	100	± 0	69.04	± 16.79	45.54	± 29.67	11.28	± 5.8	38.85	± 9.03
85	100	± 0	57.11	± 12.59	27.27	± 28.09	12.5	± 7.19	39.25	± 7.25
90	100	± 0	43.73	± 8.9	15.36	± 20.66	12.75	± 6.3	41.37	± 7.92

- **Erreur de classification (err_C) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
50	1.66	± 0.85	1.73	± 0.87	3.5	± 3.53	60.28	± 4.2	20.16	± 3.74
60	5.72	± 1.52	5.92	± 1.68	8.26	± 2.72	66.98	± 4.16	3.99	± 4.32
70	14.34	± 2.1	14.92	± 2.38	20.44	± 5.96	70.96	± 3.04	48.52	± 4.35
75	20.52	± 2	25.15	± 8.22	33.22	± 14.03	71.7	± 3.09	55.25	± 4.38
80	27.12	± 2.72	43.93	± 9.25	50.63	± 16.07	72.5	± 2.62	61.57	± 3.9
85	35.76	± 2.64	55.62	± 5.96	65.06	± 13.04	74.37	± 2.2	66.13	± 2.67
90	46.04	± 3.26	65.6	± 3.27	74.54	± 7.11	75.51	± 1.58	69.8	± 2.18

- **Nombre de descriptions (n_D) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
50	5	± 0	5.02	± 0.14	5.07	± 0.26	4.97	± 0.17	27.92	± 2.93
60	5	± 0	5.04	± 1.2	5.08	± 0.27	4.67	± 0.55	19.84	± 2.64
70	5	± 0	5.17	± 0.47	5.19	± 5.54	4.6	± 0.68	14.5	± 1.83
75	5	± 0	5.49	± 0.7	4.82	± 1.31	4.69	± 0.58	12.26	± 2.08
80	5	± 0	6.13	± 0.7	3.78	± 1.74	4.75	± 0.57	10.11	± 1.76
85	5	± 0	5.96	± 0.6	2.59	± 1.52	4.6	± 0.6	8.36	± 1.49
90	5	± 0	5.5	± 0.69	1.79	± 1.09	4.32	± 0.75	6.09	± 1.18

- **Adéquation relative à E ($adeq_E$) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
50	40.14	± 0.48	40.11	± 0.53	38.53	± 3.01	4.33	± 2.13	45.88	± 2.86
60	40.43	± 0.74	40.37	± 0.86	38.07	± 2.68	5.2	± 2.05	37.88	± 3.9
70	41.52	± 0.84	41.27	± 1	36.05	± 4.74	6.72	± 1.93	34.03	± 3.86
75	42.3	± 0.88	41.28	± 2.18	31.89	± 11.47	8.73	± 2.33	33.36	± 4.83
80	43.74	± 1.2	41.55	± 2.05	24.39	± 15.77	10.61	± 2.79	33.84	± 5.51
85	46.83	± 1.41	46.5	± 1.34	18.05	± 18.66	13.2	± 4.42	37.89	± 5.36
90	52.41	± 1.53	57.24	± 1.08	14.3	± 19.75	15.9	± 5.68	45.19	± 6.19

- **Compacité (cmp_1) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
50	0.8022	± 0.0022	0.8023	± 0.0024	0.81	± 0.0148	0.9785	± 0.0105	0.7743	± 0.14
60	0.8411	± 0.0026	0.8414	± 0.0032	0.8503	± 0.0104	0.9794	± 0.0081	0.8514	± 0.0152
70	0.8778	± 0.0022	0.8787	± 0.0028	0.8939	± 0.0139	0.98	± 0.0057	0.9001	± 0.0112
75	0.8964	± 0.002	0.8989	± 0.0051	0.9219	± 0.028	0.9784	± 0.0057	0.9184	± 0.0117
80	0.9142	± 0.002	0.9184	± 0.0037	0.9522	± 0.0308	0.9791	± 0.0054	0.9338	± 0.0106
85	0.9309	± 0.0017	0.9314	± 0.0017	0.9735	± 0.074	0.9805	± 0.0065	0.9444	± 0.0077
90	0.948	± 0.001	0.9433	± 0	0.9859	± 0.0194	0.9844	± 0.0055	0.9555	± 0.006

- **Séparation (sep_1) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
50	0.6065	± 0	0.6065	± 0	0.6086	± 0.0066	0.6084	± 0.003	0.6936	± 0.0073
60	0.6065	± 0	0.6065	± 0	0.6083	± 0.0057	0.612	± 0.0062	0.6752	± 0.0097
70	0.6065	± 0	0.6065	± 0	0.6138	± 0.0098	0.6222	± 0.0111	0.6669	± 0.0101
75	0.6065	± 0	0.6065	± 0	0.6177	± 0.0132	0.6277	± 0.0133	0.6655	± 0.0125
80	0.6065	± 0	0.6065	± 0	0.619	± 0.0163	0.6331	± 0.0128	0.6671	± 0.0123
85	0.6065	± 0	0.6065	± 0	0.6178	± 0.0175	0.6408	± 0.0134	0.6736	± 0.014
90	0.6065	± 0	0.6065	± 0	0.6173	± 0.0225	0.6418	± 0.02	0.6808	± 0.0173

E.6.2 Utilisation de la technique T_4

- **Adéquation relative à I ($adeq_I$) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
50	100	± 0	99.88	± 0.82	95.13	± 9.1	6.58	± 3.78	26.04	± 4.26
60	100	± 0	99.77	± 1.13	92.83	± 7.67	6.67	± 3.48	26.26	± 5.38
70	100	± 0	98.57	± 3.35	81.48	± 13.06	7.46	± 4.1	23.88	± 5.34
75	100	± 0	91.3	± 13.5	65.47	± 24.7	8.65	± 4.49	23.21	± 5.23
80	100	± 0	69.04	± 16.79	43.63	± 28.06	10.56	± 5.76	22.74	± 5.53
85	100	± 0	57.11	± 12.59	25.6	± 26.17	11.35	± 6.85	24.7	± 5.49
90	100	± 0	43.73	± 8.9	14.55	± 19.44	11.61	± 5.96	29.99	± 7.16

- **Adéquation relative à E ($adeq_E$) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
50	40.14	± 0.48	40.11	± 0.53	38.34	± 3.38	4.16	± 2.03	13.93	± 1.55
60	40.43	± 0.74	40.37	± 0.86	37.75	± 2.98	4.92	± 1.91	15.01	± 1.6
70	41.52	± 0.84	41.27	± 1	34.9	± 4.75	6.13	± 1.77	16.22	± 1.4
75	42.3	± 0.88	41.28	± 2.18	29.77	± 10.49	7.89	± 2.18	17.44	± 1.36
80	43.74	± 1.2	41.55	± 2.05	22.12	± 13.66	9.43	± 2.55	19.44	± 1.72
85	46.83	± 1.41	46.5	± 1.34	16.23	± 16.23	11.59	± 3.78	23.65	± 1.64
90	52.41	± 1.53	57.24	± 1.08	13.16	± 17.64	14.01	± 4.53	32.54	± 2.79

- **Compacité (cmp_1) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
50	0.8022	± 0.0022	0.8023	± 0.0024	0.811	± 0.0166	0.9793	± 0.01	0.9311	± 0.0075
60	0.8411	± 0.0026	0.8414	± 0.0032	0.8516	± 0.0116	0.9805	± 0.0075	0.9408	± 0.0062
70	0.8778	± 0.0022	0.8787	± 0.0028	0.8973	± 0.0139	0.9818	± 0.0052	0.9521	± 0.004
75	0.8964	± 0.002	0.8989	± 0.0051	0.9271	± 0.0257	0.9805	± 0.0053	0.9572	± 0.0032
80	0.9142	± 0.002	0.9184	± 0.0037	0.9566	± 0.0267	0.9814	± 0.005	0.9618	± 0.0032
85	0.9309	± 0.0017	0.9314	± 0.0017	0.9761	± 0.0238	0.9829	± 0.0055	0.9652	± 0.0022
90	0.948	± 0.001	0.9433	± 0	0.987	± 0.173	0.9862	± 0.0044	0.9679	± 0.0026

E.7 Comparaison des algorithmes avec ART-5

E.7.1 Utilisation de la technique T_2

- **Adéquation relative à I ($adeq_I$) :**

m	I		PRESS		EM		KModes		COBWEB	
	n_I	σ	μ	σ	μ	σ	μ	σ	μ	σ
3	100	± 0	99.28	± 1.85	98.1	± 2.17	7.74	± 4.85	62.94	± 13.13
4	100	± 0	95.63	± 8.47	90.01	± 6.74	9.87	± 5.27	47.46	± 8.76
5	100	± 0	69.04	± 16.79	45.54	± 29.67	11.28	± 5.8	38.85	± 9.03
6	100	± 0	53.72	± 12.6	14.2	± 18.96	11.77	± 5.9	38.19	± 6.4
7	100	± 0	45.61	± 10.44	4.35	± 11.34	13.54	± 7.05	35.9	± 7.37
8	100	± 0	42.62	± 8.6	0.73	± 2.54	14.88	± 7.29	35.81	± 7.14

- **Erreur de classification (err_C) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
3	7.26	± 2.11	7.49	± 2.19	6.71	± 1.99	59.82	± 3.3	26.81	± 8.08
4	17.33	± 2.72	19.32	± 5.12	20.41	± 3.42	67.68	± 2.91	49.12	± 5.1
5	27.12	± 2.72	43.93	± 9.25	50.63	± 16.07	72.5	± 2.62	61.57	± 3.9
6	36.76	± 3.12	59.95	± 5.85	70.86	± 11.05	76.42	± 2.37	67.51	± 2.52
7	44.61	± 2.51	69.3	± 3.61	81.55	± 6.99	78.45	± 1.92	72.13	± 2.08
8	51.4	± 1.97	74.17	± 2.63	86.18	± 3.27	80.98	± 1.53	75.21	± 1.72

- **Nombre de descriptions (n_D) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
3	3	± 0	3.07	± 0.29	3.01	± 0.1	2.79	± 0.43	6.14	± 1.48
4	4	± 0	4.3	± 0.64	4.14	± 0.35	3.83	± 3.38	8.76	± 1.33
5	5	± 0	6.13	± 0.7	3.78	± 1.74	4.75	± 0.57	10.11	± 1.76
6	6	± 0	6.87	± 0.58	2.43	± 1.58	5.7	± 0.59	11.78	± 1.81
7	7	± 0	7.21	± 0.45	1.61	± 1.22	6.66	± 0.57	13.08	± 2.18
8	8	± 0	7.29	± 0.48	1.2	± 0.55	7.65	± 0.59	14.53	± 2.32

- **Adéquation relative à E ($adeq_E$) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
3	67.39	± 1.37	67.22	± 1.55	66.84	± 1.84	9.36	± 4.07	51.05	± 9.11
4	51.91	± 1.15	51.34	± 2.05	49.22	± 3.11	9.94	± 2.9	39.07	± 5.88
5	43.74	± 1.2	41.55	± 2.05	24.39	± 15.77	10.61	± 2.79	33.84	± 5.51
6	39.01	± 0.86	37.99	± 1.1	7.93	± 11.19	11.46	± 2.63	33.73	± 4.18
7	35.77	± 0.92	36.3	± 0.66	2.52	± 6.67	12.34	± 2.42	32.76	± 4.39
8	33.64	± 0.71	35.26	± 0.6	0.48	± 1.83	12.98	± 2.27	32.31	± 4.34

- **Compacité (cmp_1) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
3	0.8696	± 0.0024	0.87	± 0.0028	0.8708	± 0.0035	0.9816	± 0.0079	0.901	± 0.0174
4	0.8987	± 0.002	0.8999	± 0.0037	0.904	± 0.0059	0.9804	± 0.0057	0.9238	± 0.0113
5	0.9142	± 0.002	0.9184	± 0.0037	0.9522	± 0.0308	0.9791	± 0.0054	0.9338	± 0.0106
6	0.9232	± 0.0014	0.9252	± 0.0017	0.9844	± 0.0219	0.9774	± 0.0051	0.934	± 0.0081
7	0.9294	± 0.0014	0.9284	± 0.001	0.9951	± 0.0131	0.9756	± 0.0047	0.9358	± 0.0085
8	0.9335	± 0.001	0.9304	± 0	0.9991	± 0.0035	0.9744	± 0.0044	0.9367	± 0.0084

- **Séparation (sep_1) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
3	0.6065	± 0	0.6065	± 0	0.6103	± 0.0103	0.6271	± 0.02	0.6846	± 0.0228
4	0.6065	± 0	0.6065	± 0	0.62	± 0.0136	0.6329	± 0.0156	0.6768	± 0.0136
5	0.6065	± 0	0.6065	± 0	0.619	± 0.0163	0.6331	± 0.0128	0.6671	± 0.0123
6	0.6065	± 0	0.6065	± 0	0.6097	± 0.0101	0.6337	± 0.011	0.6685	± 0.0109
7	0.6065	± 0	0.6065	± 0	0.6073	± 0.0035	0.6337	± 0.0104	0.6672	± 0.012
8	0.6065	± 0	0.6065	± 0	0.6066	± 0	0.6382	± 0.0108	0.6668	± 0.0117

E.7.2 Utilisation de la technique T_4

- **Adéquation relative à I ($adeq_I$) :**

m	I		PRESS		EM		KModes		COBWEB	
	n_I	σ	μ	σ	μ	σ	μ	σ	μ	σ
3	100	± 0	99.28	± 1.85	97.9	± 2.49	7.41	± 4.74	47.8	± 13.09
4	100	± 0	95.63	± 8.47	89.17	± 7.09	9.23	± 5.12	29.56	± 6.48
5	100	± 0	69.04	± 16.79	43.63	± 28.06	10.56	± 5.76	22.74	± 5.53
6	100	± 0	53.72	± 12.6	13.46	± 17.54	10.69	± 5.79	20.78	± 5.24
7	100	± 0	45.61	± 10.44	4.14	± 10.54	12.29	± 7.09	18.77	± 4.84
8	100	± 0	42.62	± 8.6	0.7	± 2.41	13.48	± 7.16	17.92	± 5.76

- **Adéquation relative à E ($adeq_E$) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
3	67.39	± 1.37	67.22	± 1.55	66.44	± 1.76	8.85	± 3.83	37.29	± 7.51
4	51.91	± 1.15	51.34	± 2.05	47.57	± 3.17	9.08	± 2.68	23.48	± 2.33
5	43.74	± 1.2	41.55	± 2.05	22.12	± 13.66	9.43	± 2.55	19.44	± 1.72
6	39.01	± 0.86	37.99	± 1.1	7.22	± 9.52	10.02	± 2.43	17.71	± 1.17
7	35.77	± 0.92	36.3	± 0.66	2.32	± 5.88	10.63	± 2.13	16.64	± 0.95
8	33.64	± 0.71	35.26	± 0.6	0.46	± 1.73	10.85	± 2.11	15.5	± 1.03

- **Compacité (cmp_1) :**

m	I		PRESS		EM		KModes		COBWEB	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
3	0.8696	± 0.0024	0.87	± 0.0028	0.8716	± 0.0033	0.9826	± 0.0074	0.9275	± 0.0144
4	0.8987	± 0.002	0.8999	± 0.0037	0.9072	± 0.0061	0.9821	± 0.0052	0.954	± 0.0044
5	0.9142	± 0.002	0.9184	± 0.0037	0.9566	± 0.0267	0.9814	± 0.005	0.9618	± 0.0032
6	0.9232	± 0.0014	0.9252	± 0.0017	0.9858	± 0.0187	0.9802	± 0.0047	0.9652	± 0.002
7	0.9294	± 0.0014	0.9284	± 0.001	0.9954	± 0.0115	0.979	± 0.0041	0.9673	± 0.0017
8	0.9335	± 0.001	0.9304	± 0	0.9991	± 0.0032	0.9785	± 0.004	0.9695	± 0.0017

E.8 Résultats de PRESS en utilisant les k-plus-proches voisins

	sans		k=1		k=2		k=5	
	μ	σ	μ	σ	μ	σ	μ	σ
n_D	5.5	± 0.69	6.64	± 0.79	8.72	± 1.08	7.78	± 2.43
err_C	65.6	± 3.27	67.34	± 2.66	67.5	± 2.56	69.9	± 2.6
$adeq_I$	43.73	± 8.9	37.31	± 7.21	30.4	± 9.07	34.45	± 7.6
$adeq_E$	57.24	± 1.08	48.25	± 3.54	34.94	± 2.23	37.67	± 2.27

	sans		k=10		k=50		k=249	
	μ	σ	μ	σ	μ	σ	μ	σ
n_D	5.5	± 0.69	7.24	± 2.82	5.88	± 3.13	7.04	± 2.58
err_C	65.6	± 3.27	70.38	± 2.63	70.96	± 2.25	72.22	± 2.06
$adeq_I$	43.73	± 8.9	33.68	± 8.34	35.36	± 7.65	32.12	± 6.97
$adeq_E$	57.24	± 1.08	41.72	± 1.7	43.47	± 0.83	45.98	± 0.9

E.9 Résultats de PRESS en faisant varier $|T|$

	0		5		10	
	μ	σ	μ	σ	μ	σ
q_{N_1}	5.58	± 1.3	5.63	± 0.11	5.72	± 0.11
n_D	5.62	± 0.83	5.71	± 0.84	5.56	± 0.68
err_C	66.81	± 3.08	65.98	± 3.46	65.56	± 3.13
$adeq_I$	42.9	± 8.76	44.42	± 9.18	44.92	± 8.3

	15		20		25	
	μ	σ	μ	σ	μ	σ
q_{N_1}	5.77	± 0.11	5.79	± 0.1	5.78	± 0.1
n_D	5.35	± 0.59	5.37	± 0.58	5.38	± 0.56
err_C	65.87	± 3.7	65.12	± 3.52	65.39	± 3.28
$adeq_I$	44.32	± 9.47	44.85	± 9.53	45.2	± 8.66

E.10 Etude des performances de PRESS

E.10.1 Temps moyen en fonction du nombre d'exemples

dup	n_E	temps (en s)	
		μ	σ
10	30	4.1	± 0.83
20	60	6.8	± 1.08
30	90	8.1	± 0.54
40	120	10.2	± 1.17
50	150	11.4	± 1.28
60	180	13.1	± 0.94
70	140	15.7	± 1.68
80	160	17.4	± 1.02
90	180	18.8	± 0.87
100	300	20.5	± 0.92
200	600	38.5	± 1.63
300	900	57.8	± 2.36
400	1200	78.1	± 2.47
500	1500	100.4	± 6.07
1000	3000	198.8	± 9.81
2000	6000	409.7	± 27.14
3000	9000	574.4	± 27.6
4000	12000	776	± 41.22
5000	15000	942.9	± 45.11

E.10.2 Temps moyen en fonction du nombre d'attributs

$n_{\mathcal{A}}$	temps (en s)	
	μ	σ
10	0.3	± 0.46
20	2.2	± 0.4
30	11.4	± 1.28
40	27.9	± 1.51
50	56.2	± 1.25
60	94.8	± 11.27
70	146.3	± 6.99
80	209.9	± 12.67
90	289	± 14.18
100	445.3	± 42.58

Annexe F

Résultats avec les données réelles

F.1 Résultat quantitatifs de PRESS sur les jeux de données réels

F.1.1 Le Matin

	PRESS	EM		COBWEB		KModes	
		T_2	T_4	T_2	T_4	T_2	T_4
n_D	5	2	2	2	2	5	5
$couv$	96.83	100	100	100	100	100	100
$pred$	47.43	55.99	55.84	51.08	45.02	41.85	33.77
$adeq_E$	78.52	62.8	62	56.4	47.2	57.6	50
$perte_E$	21.48	37.2	38	43.6	52.8	42.4	50
sep_1	0.6065	0.6249	0.6065	0.7676	0.6065	0.6266	0.6065
cmp_1	0.9545	0.9749	0.9759	0.9819	0.9932	0.9803	0.9893

F.1.2 La Libre Parole

	PRESS	EM		COBWEB		KModes	
		T_2	T_4	T_2	T_4	T_2	T_4
n_D	4	2	2	1	1	4	4
$couv$	91.84	100	100	100	100	100	100
$pred$	65.58	68.18	68.18	63.64	63.64	67.53	64.16
$adeq_E$	80.81	63.45	63.45	48.28	48.28	73.1	60.23
$perte_E$	19.19	36.55	36.55	51.72	51.72	26.9	39.77
sep_1	0.6065	0.6065	0.6065	0.6065	0.6065	0.6667	0.6065
cmp_1	0.9583	0.9781	0.9781	1	1	0.9647	0.9828

F.1.3 La Croix

	PRESS	EM		COBWEB		KModes	
		T_2	T_4	T_2	T_4	T_2	T_4
n_D	4	2	2	2	2	4	4
$couv$	97.37	100	100	100	100	100	100
$pred$	68.98	66.35	65.63	60.45	56.46	59.28	58.21
$adeq_E$	77.92	68.67	67.17	60.15	55.14	63.66	61.4
$perte_E$	22.08	31.33	32.83	39.85	44.86	36.34	38.6
sep_1	0.6065	0.6615	0.6065	0.7505	0.6065	0.6248	0.6065
cmp_1	0.9813	0.9842	0.9859	0.9932	0.9984	0.9896	0.9921

F.1.4 Le Petit Journal

	PRESS	EM		COBWEB		KModes	
		T_2	T_4	T_2	T_4	T_2	T_4
n_D	2	2	2	1	1	2	2
$couv$	84.91	100	100	100	100	100	100
$pred$	39.22	47.68	46.31	39.39	39.39	41.05	41.05
$adeq_E$	84.47	80.77	79.23	40	40	47.69	47.69
$perte_E$	15.53	19.23	20.77	60	60	52.31	52.31
sep_1	0.6065	0.6432	0.6065	0.6065	0.6065	0.6065	0.6065
cmp_1	0.9409	0.9391	0.9414	1	1	0.9882	0.9882

F.2 Résultats complets de PRESS sur les jeux de données réels

F.2.1 Le Matin

- **Ensemble de stéréotypes :**

▷ Stéréotype 0 = (33 ex. / poids 41 (65,08%)) : Revolutionnaire=oui ∧ Socialiste=oui ∧ OrigineEtrangere=oui ∧ RelationEtranger=oui ∧ ImplicationFrancsMacons=oui ∧ ImplicationGouvernementale=oui ∧ Patriotisme=antipatriote ∧ Clericalisme=anticlerical ∧ Internationaliste=oui ∧ HonneteHomme=non ∧ RespectLoi=non ∧ Traitre=oui ∧ Favoritisme=oui ∧ Corruption=oui ∧ SensMoral=non ∧ Impunite=oui ∧ Incompetence=oui ∧ PbViePrivee=oui ∧ Dangereux=oui ∧ MeleAffaireDArgent=oui ∧ SourceViolence=unpeu ∧ LieClemenceau=oui

Exemples : 0 - 1 - 2 - 3 - 5 - 8 - 10 - 11 - 12 - 15 - 17 - 18 - 19 - 21 - 22 - 23 - 25 - 26 - 27 - 28 - 30 - 31 - 32 - 34 - 37 - 39 - 41 - 42 - 43 - 46 - 47 - 48 - 49

▷ Stéréotype 1 = (6 ex. / poids 7 (11,11%)) : Politique=opportuniste ∧ Tendance=republicain ∧ Socialiste=non ∧ RelationEtranger=non ∧ Clericalisme=clerical ∧ Traitre=non ∧ MeleAffaireDArgent=non ∧ VictimeViolence=unpeu

Exemples : 13 - 14 - 16 - 33 - 36 - 45

▷ Stéréotype 2 = (5 ex. / poids 6 (9,52%)) : PbSante=oui ∧ Action=incident ∧ VictimeViolence=beaucoup

Exemples : 4 - 6 - 9 - 20 - 35

▷ Stéréotype 3 = (2 ex. / poids 3 (4,76%)) : Action=duel ∧ SourceViolence=beaucoup

Exemples : 7 - 29

▷ Stéréotype 4 = (3 ex. / poids 4 (6,35%)) : Action=demission

Exemples : 38 - 40 - 44

▷ Stéréotype vide (1 ex. / poids 2 (3,17%) :

Exemples : 24

- **Indices :**

▷ Nombre d'exemples : 50.

▷ Poids total des exemples : 63.

▷ Pourcentage de données manquantes : 87,01%.

▷ Nombre de stéréotypes découverts : 5.

▷ Score de séparation : 0,6065.

▷ Couverture des exemples : 96,83%.

▷ Score de compacité : 0,9545.

▷ Adéquation relative à E : 78,52%.

▷ Contradiction relative à E : 0%.

▷ Perte relative à E : 21,48%.

▷ Capacité prédictive : 47,43%.

▷ Temps d'exécution : 17s.

F.2.2 La Libre Parole

- **Ensemble de stéréotypes :**

▷ Stéréotype 0 = (48 ex. / poids 77 (78,57%)) : Tendance=republicain ∧ Revolutionnaire=oui ∧ Socialiste=oui ∧ Religion=juif ∧ RelationJuifs=oui ∧ RelationProtestants=oui ∧ RelationEtranger=oui ∧ ImplicationFrancsMacons=oui ∧ ImplicationGouvernementale=oui ∧ Patriotisme=antipatriote ∧ Clericalisme=anticlerical ∧ Internationaliste=oui ∧ HonneteHomme=non ∧ RespectLoi=non ∧ Favoritisme=oui ∧ Corruption=oui ∧ SensMoral=non ∧ Impunite=oui ∧ SentimentSuperieur=oui ∧ Incompetence=oui ∧ PbSante=non ∧ Dangereux=oui ∧ MeleAffaireDArgent=oui ∧ VictimeViolence=unpeu ∧ SourceViolence=unpeu ∧ Complot=oui ∧ LieClemenceau=oui

Exemples : 0 - 1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 10 - 11 - 13 - 14 - 15 - 16 - 17 - 19 - 20 - 21 - 22 - 23 - 24 - 25 - 26 - 27 - 28 - 29 - 30 - 32 - 35 - 36 - 37 - 39 - 40 - 41 - 42 - 43 - 44 - 45 - 48 - 49 - 50 - 55 - 56 - 57 - 59 - 60 - 62

▷ Stéréotype 1 = (2 ex. / poids 3 (3,06%)) : Religion=catholique ∧ RelationCatholiques=oui ∧ SensMoral=oui ∧ Action=diffamation

Exemples : 52 - 61

▷ Stéréotype 2 = (3 ex. / poids 5 (5,1%)) : RelationJuifs=non ∧ Action=incident ∧ VictimeViolence=beaucoup ∧ SourceViolence=beaucoup

Exemples : 12 - 31 - 54

▷ Stéréotype 3 = (3 ex. / poids 5 (5,1%)) : PbSante=oui ∧ Action=tromperie

Exemples : 33 - 46 - 51

▷ Stéréotype vide (7 ex. / poids 8 (8,16%)) :

Exemples : 9 - 18 - 34 - 38 - 47 - 53 - 58

- **Indices :**

▷ Nombre d'exemples : 63.

▷ Poids total des exemples : 98.

▷ Pourcentage de données manquantes : 86,3%.

▷ Nombre de stéréotypes : 4.

▷ Score de séparation : 0,6065.

▷ Couverture des exemples : 91,84%.

▷ Score de compacité : 0,9583.

▷ Adéquation relative à E : 80,81%.

▷ Contradiction relative à E : 0%.

▷ Perte relative à E : 19,19%.

▷ Capacité prédictive : 65,58%.

▷ Temps d'exécution : 16s.

F.2.3 La Croix

- **Ensemble de stéréotypes :**

▷ Stéréotype 0 = (72 ex. / poids 98 (85,96%)) : Tendance=republicain ∧ Revolutionnaire=oui ∧ Socialiste=oui ∧ Religion=protestant ∧ RelationJuifs=oui ∧ RelationProtestants=oui ∧ RelationCatholiques=non ∧ RelationEtranger=oui ∧ ImplicationFrancsMacons=oui ∧ ImplicationGouvernementale=oui ∧ Patriotisme=antipatriote ∧ Clericalisme=anticlerical ∧ Internationaliste=oui ∧ HonneteHomme=non ∧ RespectLoi=non ∧ Corruption=oui ∧ SensMoral=non ∧ Impunite=oui ∧ SentimentSuperieur=oui ∧ Incompetence=oui ∧ PbSante=oui ∧ Dangereux=oui ∧ MeleAffaireDArgent=oui ∧ VictimeViolence=unpeu ∧ unpeu<=SourceViolence<=beaucoup ∧ Complot=oui

Exemples : 0 - 1 - 2 - 3 - 4 - 5 - 7 - 8 - 9 - 10 - 11 - 12 - 14 - 16 - 17 - 18 - 19 - 20 - 21 - 23 - 24 - 25 - 26 - 28 - 29 - 30 - 31 - 32 - 33 - 34 - 35 - 36 - 37 - 39 - 40 - 41 - 42 - 43 - 45 - 47 - 48 - 49 - 51 - 52 - 53 - 54 - 55 - 56 - 57 - 58 - 60 - 61 - 62 - 63 - 64 - 65 - 66 - 67 - 68 - 69 - 70 - 72 - 73 - 74 - 75 - 76 - 77 - 78 - 80 - 81 - 82 - 84

▷ Stéréotype 1 = (2 ex. / poids 3 (2,63%)) : Religion=catholique ∧ RelationCatholiques=oui ∧ Clericalisme=clerical ∧ Action=demission

Exemples : 46 - 79

▷ Stéréotype 2 = (4 ex. / poids 5 (4,39%)) : Favoritisme=oui ∧ Action=incident ∧ VictimeViolence=beaucoup ∧ LieClemenceau=oui

Exemples : 15 - 22 - 27 - 71

▷ Stéréotype 3 = (5 ex. / poids 5 (4,39%)) : Revolutionnaire=non ∧ Internationaliste=non ∧ Action=diffamation

Exemples : 6 - 13 - 44 - 50 - 83

▷ Stéréotype vide (2 ex. / poids 3 (2,63%)) :

Exemples : 38 - 59

- **Indices :**

▷ Nombre d'exemples : 85.

▷ Poids total des exemples : 114.

▷ Pourcentage de données manquantes : 89,29%.

▷ Nombre de stéréotypes : 4.

▷ Score de séparation : 0,6065.

▷ Couverture des exemples : 97,37%.

▷ Score de compacité : 0,9813.

▷ Adéquation relative à E : 77,92%.

▷ Contradiction relative à E : 0%.

▷ Perte relative à E : 22,08%.

▷ Capacité prédictive : 68,98%.

▷ Temps d'exécution : 20s.

F.2.4 Le Petit Journal

- **Ensemble de stéréotypes :**

▷ Stéréotype 0 = (16 ex. / poids 32 (60,38%)) : Politique=radical ∧ Tendance=republicain ∧ Socialiste=oui ∧ RelationEtranger=oui ∧ ImplicationGouvernementale=oui ∧ Patriotisme=antipatriote ∧ HonneteHomme=non ∧ RespectLoi=non ∧ Traitre=oui ∧ Favoritisme=oui ∧ Corruption=oui ∧ SensMoral=non ∧ Incompetence=oui ∧ PbSante=oui ∧ Dangereux=oui ∧ MeleAffaireDArgent=oui ∧ VictimeViolence=beaucoup ∧ SourceViolence=unpeu ∧ LieClemenceau=oui

Exemples : 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - 13 - 15 - 18 - 21 - 22 - 24 - 25 - 26

▷ Stéréotype 1 = (7 ex. / poids 13 (24,53%)) : Patriotisme=patriote ∧ HonneteHomme=oui ∧ Traitre=non ∧ Action=diffamation ∧ VictimeViolence=unpeu ∧ LieClemenceau=non

Exemples : 1 - 11 - 12 - 14 - 16 - 17 - 20

▷ Stéréotype vide (5 ex. / poids 8 (15,09%)) :

Exemples : 0 - 10 - 19 - 23 - 27

- **Indices :**

▷ Nombre d'exemples : 28.

▷ Poids total des exemples : 53.

- ▷ Pourcentage de données manquantes : 84,91%.
- ▷ Nombre de stéréotypes : 2 stéréotypes.
- ▷ Score de séparation : 0,6065.
- ▷ Couverture des exemples : 84,91%.
- ▷ Score de compacité : 0,9409.
- ▷ Adéquation relative à E : 84,47%.
- ▷ Contradiction relative à E : 0%.
- ▷ Perte relative à E : 15,53%.
- ▷ Capacité prédictive : 39,22%.
- ▷ Temps d'exécution : 5s.