Extraction de Connaissances à partir de Données Numériques et Textuelles

Jérôme Azé

N° d'ordre : 7442

Thèse

de

L'Université Paris–Sud

présentée en vue de l'obtention du

Doctorat de l'Université Paris–Sud

Spécialité : Informatique

Par

Jérôme Azé

Extraction de connaissances à partir de données numériques et textuelles

Soutenue le 16 décembre 2003 devant la commission d'examen :

Mme	Marie-Christine Rousset	Professeur	Examinatrice
M.	Jean-Marc Petit	Maître de Conférences	Examinateur
M.	Israël-César Lerman	Professeur Émérite	Rapporteur
M.	Jean-Daniel Zucker	Professeur	Rapporteur
M.	Yves Kodratoff	Directeur de recherche C.N.R.S.	Directeur de thèse

Laboratoire de Recherche en Informatique, U.M.R. CNRS 8623,
Université Paris-Sud, 91405 Orsay Cedex, France

Remerciements

Je remercie M. Yves Kodratoff pour m'avoir offert l'opportunité de réaliser ma thèse sous sa direction, dans l'équipe « Inférence et Apprentissage ». Le cadre de travail offert par l'équipe I&A et la liberté que m'a laissé Yves Kodratoff m'ont permis d'explorer des domaines de recherche connexes à ma thèse et d'établir des collaborations avec d'autres chercheurs. Dans le cadre de ces collaborations, je tiens à remercier Sylvie Guillaume pour tout le temps qu'elle m'a consacré.

Je remercie M^{me} Marie-Christine Rousset pour m'avoir fait l'honneur de présider mon jury de thèse.

Je remercie mes rapporteurs, M. Israël-César Lerman et M. Jean-Daniel Zucker tant pour leurs commentaires pertinents que pour avoir accepté des délais de relecture relativement courts.

Je remercie M. Jean-Marc Petit de m'avoir fait l'honneur de faire partie de mon jury de thèse et pour ses commentaires avertis sur mes travaux de recherche.

Je tiens aussi à remercier les membres du groupe de travail GafoQualité de l'action spécifique STIC pour les conversations et débats sur les mesures de qualité et leurs utilisations.

Je remercie également Michèle Sebag pour m'avoir incité à étendre mon domaine de compétences au problème de la fouille de données médicales. Dans ce contexte, nos travaux communs réalisés en collaboration avec Noël Lucas sur le problème médical de l'athérosclérose nous ont permis d'obtenir des résultats significatifs dans le domaine novateur de la fouille de données médicales.

Je tiens à remercier particulièrement Mathieu Roche et Alexandre Termier pour leurs multiples relectures de mon manuscrit et pour la patience dont ils ont su faire preuve pendant la période de rédaction de ma thèse.

Table des matières

CHAPITRE I

Introduction

1 Comparaison de l'Apprentissage Automatisé et de la Fouille de Données

L'Apprentissage Symbolique Automatique a fait son apparition dès la fin des années 1970 alors que la Fouille de Données n'est apparue qu'au début des années 90.

Actuellement, la Fouille de Données peut se définir comme la conjonction des recherches en Bases de Données et en **Apprentissage Automatisé** qui est un domaine en cours de création rassemblant l'Apprentissage Symbolique Automatique, l'Analyse de Données, la Reconnaissance des Formes, les Réseaux Neuronaux et les approches bayésiennes.

De façon assez surprenante et malgré les liens qui unissent ces deux domaines, la différence la plus remarquable entre la Fouille de Données et l'Apprentissage Automatisé est que la première a été adoptée avec enthousiasme par les milieux industriels alors que l'Apprentissage Automatisé a toujours été un peu boudé par ces milieux. L'analyse des différences entre ces deux domaines montre, comme nous allons le voir, que la cause profonde du succès de la Fouille de Données semble être liée à sa capacité d'innovation contrairement au traitement plus classique de l'Apprentissage Automatisé.

Les principales caractéristiques des différents domaines constituant l'Apprentissage Automatisé seront présentées dans les sections suivantes. Nous décrirons chaque domaine de manière relativement succinte en présentant les principes généraux et les applications les plus frappantes de ces différents domaines. Puis, nous présenterons une liste de conseils suggérée par la « Oracle Data Mining Suite » pour bien réussir la Fouille de Données et nous comparerons les choix faits en Fouille de Données à ceux faits en Apprentissage Automatisé dans le cadre de ces conseils. Enfin nous conclurons cette section en tentant d'expliquer le succès observé de la Fouille de Données auprès du monde industriel contrairement à l'Apprentissage Automatisé.

1.1 Cadre général

Considérons des données constituées d'un ensemble d'individus (patients pour un médecin, clients d'une banque, produits finis pour une entreprise, *etc.*) où chaque individu est décrit par un ensemble de caractéristiques, par exemple : taille, poids, âge, sexe, taux de cholestérol, *etc.*

L'exploitation de telles données dans le cadre de la Fouille de Données ou de l'Apprentissage Automatisé peut être réalisée selon deux approches différentes : une approche supervisée ou une

approche non supervisée.

1.1.1 Approche supervisée

Dans le cadre de l'approche supervisée, les données sont constituées d'un ensemble de caractéristiques décrivant chaque individu (caractéristiques appelées *variables exogènes*) et chaque individu possède une caractéristique particulière (appelée *classe* ou *variable endogène*). L'objectif de la fouille de données supervisée est de trouver des relations entre les variables exogènes permettant d'expliquer et/ou de prévoir le comportement de la variable endogène.

Par exemple, dans le cadre de la fouille de données médicales, les individus sont les patients, les variables exogènes représentent l'ensemble des informations relatives à chaque patient et la classe représente l'état de santé du patient (bonne santé ou malade). La découverte supervisée de connaissances dans de telles données peut donc se caractériser par la recherche de corrélations entre les variables exogènes des patients appartenant à une classe donnée.

Un des moyens d'assister le médecin est de lui proposer un modèle des données permettant de prédire au mieux la valeur de la variable endogène pour un nouvel individu. Ainsi, étant donné un nouveau patient, le médecin pourra utiliser le modèle pour prédire l'état du patient et comparer cette prédiction avec son propre diagnostic ou pour l'aider à choisir le médicament ou le traitement le mieux adapté.

Étant donné cet objectif, la fouille de données supervisée se décompose en deux étapes : apprentissage d'un modèle sur une partie des données et validation du modèle sur l'autre partie des données. Les données sont donc divisées en deux ensembles : l'ensemble d'apprentissage et l'ensemble de test. Ce découpage est réalisé de manière aléatoire pour ne pas biaiser l'apprentissage du modèle.

À partir de l'ensemble d'apprentissage, un modèle décrivant au mieux les données est appris. Puis, l'ensemble de test est utilisé pour vérifier la capacité prédictive du modèle. Pour chaque individu de l'ensemble de test, seules les valeurs des variables exogènes sont fournies au modèle qui les utilise pour prédire la valeur de la variable endogène. Cette valeur est comparée à la véritable valeur de la variable endogène de l'individu et un taux d'erreur est alors calculé pour le modèle appris.

Les connaissances apprises par le modèle peuvent prendre différentes formes : arbres de décision, ensemble de règles, « boîte noire prédictive », *etc.*

L'un des principaux avantages de la fouille de données supervisée réside dans la connaissance de la variable endogène. Cette variable permet de « guider » la recherche vers la découverte de connaissances utiles pour expliquer au mieux la variable endogène. Ainsi, dans le cadre médical, les connaissances obtenues permettent idéalement de séparer les patients malades des patients sains.

Cet aspect de l'approche supervisée constitue aussi son plus grand défaut, en ce sens qu'il faut disposer de données étiquetées (c'est-à-dire dont la valeur de la variable endogène est connue) pour pouvoir fonctionner. Or l'obtention de telles données est très coûteuse car la détermination de la valeur de la classe d'un individu doit impérativement être réalisée par un expert du domaine.

1.1.2 Approche non supervisée

Dans le cadre de l'approche non supervisée, la variable endogène n'est pas explicitée dans les données.

L'objectif de la fouille de données non supervisée est donc de trouver des relations entre caractéristiques (variables exogènes) suffisamment significatives et permettant d'augmenter les connaissances du domaine étudié.

Considérons l'étude d'une base de données bancaires dans laquelle les individus sont les clients de la banque et les caractéristiques représentent l'ensemble des informations détenues par la banque sur ces clients (âge, sexe, statut familial, nombre de comptes en banques, état des comptes, *etc.*). Le banquier peut être intéressé par la découverte de comportements fréquents dans ses données (c'est-à-dire des comportements suivis par un nombre suffisamment important de clients) pour pouvoir adapter au mieux ses services à sa clientèle. Le banquier ne connait pas *a priori* la nature des connaissances contenues dans ses données mais il espère y trouver des informations permettant de l'aider dans la gestion de sa banque.

Par exemple, si nous supposons que des informations décrivant l'aspect vestimentaire des clients sont stockées dans la base, il serait possible de trouver des règles du type : **si** *le client porte des chaussures non cirées* **et** *qu'il demande un prêt* **alors** *il ne le remboursera pas.*

Cette règle, en supposant qu'elle soit fondée, pourrait être utilisée par le banquier pour l'aider à attribuer ses prêts.

1.2 L'Apprentissage Symbolique Automatique

Les débuts de l'Apprentissage Symbolique Automatique sont assimilés aux travaux de Michalski [Michalski et Chilausky 1980]. La plupart des travaux ont été réalisés en apprentissage supervisé et ont engendré de nombreux systèmes dont quelques uns sont couramment utilisés dans le monde industriel. L'un des exemples les plus connus est celui des arbres de décisions qui prennent en entrée des données décrites par un ensemble de descripteurs continus ou discrets et des classes impérativement discrètes. Les descripteurs continus sont alors discrétisés par rapport aux classes et ces systèmes engendrent des arbres de classement qui peuvent être interprétés comme une description en intention des classes. Le plus célèbre de ces systèmes, C4.5[1] [Quinlan 1993] engendre, à partir des arbres de décision, un ensemble de règles souvent plus compréhensibles pour l'expert.

Une des bases de l'apprentissage est la généralisation.

L'espace des généralisations possibles a été appelé l'espace des versions [Mitchell 1977, Mitchell 1997] et de nombreuses méthodes se caractérisent par leur façon de se déplacer dans cet espace.

1.2.1 La Programmation Logique Inductive

La Programmation Logique Inductive est précisément une façon de se déplacer dans l'espace des versions relationnel. En Programmation Logique Inductive, les individus sont décrits par des descripteurs qui peuvent être $n-aires$ (c'est-à-dire décrire une relation entre n objets ou individus) contrairement aux autres méthodes où les descripteurs sont toujours unaires.

Par exemple, les propriétés de deux objets O_1 et O_2 peuvent être décrites avec les valeurs de descripteurs unaire (O_1 est vert, O_2 est rouge) et la distance séparant les deux objets peut être décrite par un descripteur binaire ($distance(O_1, O_2) = 56$). À partir de ce type d'informations, la Programmation Logique Inductive peut apprendre des lois générales sur la distance telle que *il n'existe pas d'objets situés à une distance supérieure à* 60 *de* O_2. La taille de l'espace des versions est alors très importante de part la puissance descriptive de la Programmation Logique Inductive. Ainsi, toute cette puissance descriptive est absorbée par la complexité des calculs nécessaires pour vérifier les hypothèses permettant de construire un modèle pour expliquer les données.

Face à ce problème, la Programmation Logique Inductive tend actuellement à s'orienter vers des techniques dites de **propositionnalisation** où les descriptions $n-aires$ sont simplement remplacées par des relations unaires. Le principe est de créer autant de descriptions qu'il existe d'appariements

[1] commercialisé sous le nom C5 ou See5, http ://www.rulequest.com/

possibles. L'explosion combinatoire en temps de calcul est remplacée par une explosion combinatoire en espace. Le gain de cette approche est lié au fait que seule une partie des descriptions « astucieusement choisies » sont conservées [Flach et Lavrac 2003, Alphonse et Matwin 2002].

1.2.2 Le système COBWEB

La principale contribution de l'Apprentissage Symbolique Automatique en apprentissage non supervisé et plus précisément en classification est le système COBWEB [Fisher 1987]. Ce système utilise un critère d'optimisation appelé l'**utilité**. L'utilité d'une classe C contenant le descripteur de valeur v se mesure par le produit des probabilités $P(A = v)$, $P(A = v|C)$ et $P(C|A = v)$. $P(A = v)$ est la probabilité que le descripteur A prenne la valeur v, $P(A = v|C)$ est la probabilité que A prenne la valeur v dans la classe C et $P(C|A = v)$ est la probabilité de rencontrer la classe C sachant que $A = v$. L'expression de l'utilité, après reformulation grâce à la loi de Bayes, est

$$U = \frac{1}{n} \sum_{C \in \mathcal{C}} P(C) \left(\sum_{A \in \mathcal{A}} \sum_{v \in dom(A)} P(A = v|C)^2 - P(A = v)^2 \right)$$

où $\begin{cases} n \text{ est le nombre de classes} \\ \mathcal{C} \text{ l'ensemble de toutes les classes} \\ \mathcal{A} \text{ l'ensemble de tous les descripteurs} \\ dom(A) \text{ l'ensemble des valeurs (ou domaine) du descripteur } A \end{cases}$

L'utilité est calculée pour chaque configuration possible ce qui n'est réalisable que grâce au calcul incrémental du gain d'utilité. Ainsi, COBWEB est très lent mais il a la particularité d'être incrémental est donc très bien adapté aux problèmes demandant justement une mise à jour régulière. Ce système est adapté à différents types de données (mixtes, continues et discrètes). Dans le cas de données continues, les sommes se transforment évidemment en intégrales. Malgré toutes ces qualités, ce système n'a pas connu le succès commercial auquel il aurait pu s'attendre.

1.3 La Reconnaissance des Formes

Avant l'apparition de l'Apprentissage Symbolique Automatique, les chercheurs en reconnaissance des formes ont proposé des systèmes d'apprentissage dont le plus connu et le plus utilisé est le séparateur linéaire (appelé perceptron) [Rosenblatt 1958]. Un perceptron est capable de séparer deux ensembles d'exemples répartis dans deux classes indicées 0 et 1 en un nombre fini d'étapes de calcul (dans la mesure où il existe une séparation linéaire). La démonstration de cette propriété est donnée par le théorème important de Novikoff.

1.4 Les Réseaux Neuronaux

Les Réseaux Neuronaux sont nés du besoin d'aller au delà des séparateurs linéaires tels que le perceptron. Le succès rencontré par les Réseaux Neuronaux est essentiellement dû à leurs capacités de traiter, aussi bien en entrée qu'en sortie, des descripteurs numériques, discrets ou mixtes. De plus, les sorties d'un réseau de neurones peuvent être multiples. Ces deux propriétés traduisent un réel besoin des données industrielles d'où le succès rencontré par cette méthode.

1.5 Les approches bayésiennes

L'effort principal des statistiques bayésiennes est relatif au développement de méthodes de raisonnement déductif permettant de prendre en compte les indépendances conditionnelles de variables discrètes.

Une méthode d'apprentissage supervisé appelée « Bayes Naïf » a été développée. Dans cette méthode, tous les descripteurs dépendent de la classe à reconnaître et sont supposés conditionnellement indépendants si leur classe est connue. L'apprentissage est réduit, dans ce cas, à la prise en considération des probabilités d'occurrence des événements observés. Cet apprentissage, bien que très simple, présente l'avantage d'être l'un des plus efficace en précision. Notons toutefois que le modèle engendré par cette méthode est absolument incompréhensible pour l'utilisateur.

Dans le cas non supervisé, et étant donnée une structure de réseau, il est possible d'apprendre les tables de probabilités conditionnelles à partir des données. Le réseau, beaucoup plus compact que les données, permet alors d'expliquer entièrement les données. Ainsi, la génération automatique des structures de grands réseaux bayésiens représente une avancée majeure dans le domaine du raisonnement inductif [Heckerman et al. 1995]. Le critère d'optimisation utilisé est le principe de longueur minimale de description. La structure de réseau induite à partir des données présente l'avantage d'être compréhensible pour l'expert contrairement au modèle obtenu avec l'approche « Bayes Naïf ».

1.6 La Fouille de Données

La naissance de la Fouille de Données est assimilée au premier atelier sur « Knowledge Discovery in Data Bases » organisé par G. Piatetsky-Shapiro en 1989 [Frawley et Piatetsky-Shapiro 1991]. La Fouille de Données est à l'origine de trois types de méthodes qui sont toutes intégrées dans de nombreux systèmes commerciaux :

- La recherche des associations, en particulier la découverte des théorèmes incertains confirmés par les données et les multiples mesures d'intérêt qui sont associées au choix des associations. Ce type de méthode sera détaillé dans la suite de cette thèse qui s'incrit dans cette thématique de recherche.
- La découverte de séries temporelles qui correspond à une extension de la première méthode et qui représente l'un des plus grands succès de la Fouille de Données.
- Enfin, et ce sous l'influence du monde industriel, la Fouille de Données a développé de multiples méthodes de nettoyage et de segmentation des données.

1.6.1 Conseils de *Oracle Data Mining Suite*

Voici les douze conseils pour bien réussir sa Fouille de Données, d'après la *Oracle Data Mining Suite*.

1. **Extraire à partir de plus de données**

   L'Apprentissage Automatisé tend à favoriser l'étude approfondie de petite bases de données, alors que la Fouille de Données se concentre plus sur l'étude des grandes bases de données (qui correspondent plus aux réalités actuelles du monde industriel).

2. **Créez de nouvelles variables pour mieux faire parler vos données**

   L'Apprentissage Automatisé s'est intéressé à un apprentissage dit « constructif » et à des techniques de sélection de variables [Liu et Motoda 1998]. L'effort de l'Apprentissage Automatisé a essentiellement consisté à justifier les modifications apportées aux descripteurs alors que la

Fouille de Données est prête à se contenter d'une justification *a posteriori*, par l'amélioration des résultats obtenus, plutôt que par une justification préalable.

L'approche de la Fouille de Données serait plutôt de conserver les divers modèles induits et de fournir à l'expert du domaine les moyens de combiner les modèles ou de choisir entre eux.

3. **Utilisez une stratégie « en surface d'abord »**
   Dans un contexte purement universitaire, une telle approche n'est jamais recommandée. Cependant, bien des erreurs peuvent être évitées par un examen superficiel des données avant de les analyser.
4. **Construisez rapidement plusieurs modèles explicatifs**
   La Fouille de Données n'hésite pas à construire plusieurs modèles explicatifs des données alors que l'Apprentissage Automatisé cherche à construire le modèle optimal. La vérification et la sélection des différents modèles fourni par la Fouille de Données relèvent des compétences de l'expert.
5. **Segmentez d'abord vos clients, et construisez des modèles multi-buts**
   L'Apprentissage Automatisé est largement dominé par l'approche supervisée, alors que la Fouille de Données est plutôt dominée par l'approche non supervisée. Un des buts de l'apprentissage non supervisé est de construire des classes d'individus en les segmentant. Lorsqu'une telle segmentation des données est obtenue alors des méthodes de recherche de règles peuvent être appliquées sur chacun des segments obtenus.

   Il peut sembler anodin de donner ce conseil mais lorsque des méthodes de détection de formes sont appliquées sur l'intégralité des données, le système cherche des lois générales qui s'appliqueraient sur tous les individus. Cette démarche tend souvent à ne détecter que des lois triviales pour tous les enregistrements.

   Lorsque les données ont été préalablement segmentées, le problème consiste alors à chercher des formes valides pour des sous-ensembles d'individus qui peuvent s'avérer très intéressantes pour peu que la segmentation ait engendré des sous-populations significatives.
6. **Construction automatique des modèles**
   Aussi bien l'Apprentissage Automatisé que la Fouille de Données sont concernés par ce conseil qui se traduit par l'utilisation de l'induction pour construire les modèles. Les deux domaines ne diffèrent donc pas sur ce point mais il permet de mettre en évidence non seulement un problème universitaire intéressant mais aussi un réel besoin du monde industriel.
7. **Interprétez vos résultats en termes du domaine d'application par rétro ingénierie**
   Nous savons que des techniques telles que les Réseaux Neuronaux et le modèle « Bayes Naïf » fournissent des résulats interprétables par l'expert. Ces techniques ne sont pas les seules à fournir des résultats intelligibles et la Fouille de Données tend à les utiliser aussi bien que d'autres méthodes. Par contre, ce qui n'est pas acceptable, c'est de fournir à l'expert des résultats qui ne sont pas compréhensibles (par exemple les résultats bruts fourni par les systèmes). Il est donc important d'interpréter les résultats dans un langage intelligible pour l'utilisateur.
8. **Complétez les données manquantes par des modèles prédictifs**
   Ce problème est bien connu en Apprentissage Automatisé et en Fouille de Données surtout lorsque le problème de l'étude de données réelles est abordé. Les solutions adoptées pour résoudre ce problème sont de deux types :
   - si les données manquantes sont « naturellement » manquantes (par exemple étude des cancers sur une population d'hommes et de femmes et seuls les hommes seront concernés par les problèmes de prostates et inversement seules les femmes seront concernées par les

problèmes d'uterus) alors la valeur absente est remplacée par une valeur prédéfinie et les algorithmes sont adaptés pour en tenir compte.
- si les données manquantes correspondent à un manque réel d'information alors les valeurs sont, soit remplacées par une valeur moyenne calculée sur l'ensemble des exemples ou sur ceux de la même classe, soit le poids du descripteur concerné est diminué lors de l'apprentissage du modèle.

  Pour ce type de données manquantes, le comportement de l'Apprentissage Automatisé diffère de celui de la Fouille de Données. L'Apprentissage Automatisé effectue une seule passe sur les données pour régler le problème des valeurs manquantes alors que la Fouille de Données intègre l'expert du domaine dans cette phase pour optimiser les résultats.

9. **Utilisez plusieurs modèles de prédiction à la fois et consevez un modèle prédictif s'appuyant sur la coopération d'experts**

   De nombreuses approches de génération de modèles ont été développées dans le cadre de l'Apprentissage Automatisé, notamment des approches incluant un vote des différents modèles jusqu'à ce qu'un des modèle finisse par l'emporter. Cependant, la notion de coopération entre experts est peu abordée et les travaux allant dans ce sens sont assez récents.

10. **Laissez tomber les pratiques traditionelles d'hygiène en matière de données**

    Les données réelles sont souvent peu conformes aux attentes des spécialistes des systèmes de bases de données. En effet, il n'est pas rare d'avoir des doublons, des données absentes, des valeurs différentes mais ayant la même sémantique pour un descripteur, *etc.*

    Il faut être prêt à travailler avec ce type de données car elles représentent une partie des données réelles sur lesquelles les systèmes de Fouille de Données devront travailler.

11. **Enrichissez vos données de données extérieures**

    En Apprentissage Automatisé, les données sont considérées comme fournies une fois pour toutes et non pas comme obtenues par un processus avec lequel il est possible d'entrer en interaction. Si le système est capable de s'apercevoir que certaines nouvelles données, ou des données supplémentaires, pourraient résoudre un problème alors une solution impossible à trouver pourrait émerger grâce à l'apport de ces nouvelles données.

12. **Variez les sources de données en fonction de vos modèles**

    Ce conseil est très proche du conseil précédent. La différence notable est que le modèle obtenu à une itération du système peut être utilisé pour rechercher les données les mieux adaptées pour l'itération suivante.

1.6.2 Succès de la Fouille de Données

La principale raison du succès de la Fouille de Données par rapport à l'Apprentissage Automatisé réside simplement dans le fait que les créateurs du domaine de la Fouille de Données ont su prendre en considération les problèmes des industriels contrairement aux chercheurs en Apprentissage Automatisé qui sont restés centrés sur les problématiques scientifiques. Ces problématique sont, bien que très intéressantes, malheureusement souvent déconnectées des applications et des préocupations des industriels. La majorité des travaux et des publications issues du monde de l'Apprentissage Automatisé concernent des données non réelles témoignant par là même de l'« isolement académique » de l'Apprentissage Automatisé par rapport aux problèmes industriels.

2 Cadre général de cette thèse : la Fouille de Données

Cette thèse s'inscrit dans le cadre général de la fouille de données.

De nombreux domaines (la médecine, les banques, les entreprises, *etc.*) disposent de bases de données peu ou pas exploitées.

Ainsi, l'exploitation efficace de données médicales concernant une maladie complexe telle que l'athérosclérose peut permettre aux médecins de tester des hypothèses sur les données et d'apprendre le modèle le plus adapté à celles-ci. Ce type de fouille de données peut apporter d'une part une meilleure compréhension de la maladie étudiée et de ses facteurs aggravants (le tabac, le poids, le cholestérol, *etc.*) et peut d'autre part fournir une aide précieuse au médecin en lui suggérant des tests médicaux adaptés aux données étudiées.

2.1 Approche retenue

Les domaines d'applications de ces deux approches (supervisée, non supervisée) sont nombreux et variés : aide à la décision, détection de fraudes, de pannes, explications de phénomènes complexes, *etc.*

Il existe de nombreuses bases de données pour lesquelles il est difficile de déterminer la variable endogène. Certaines des applications étudiées dans cette thèse sont liées à la fouille de textes et plus précisément à la recherche de connaissances nouvelles dans des textes. Les connaissances recherchées étant par définition préalablement inconnues de l'expert, il ne lui est pas possible de définir une variable endogène.

Nous nous sommes focalisés, dans le cadre de cette thèse, sur la découverte **non supervisée** de connaissances dans les données. L'extraction de connaissances, nouvelles et utiles, dans de telles données peut se comparer à la recherche de pépites d'or dans un fleuve.

2.1.1 Des rivières remplies de pépites ...

Mettons nous quelques instants dans la peau de chercheurs d'or du XIX^{ème} siècle. Ces orpailleurs qui, face à une rivière inconnue, espéraient trouver des pépites d'or et faire très rapidement fortune étaient souvent bien incapables de déterminer précisément le meilleur endroit où commencer leurs recherches. De même, il leur était probablement difficile de choisir le bon tamis pour fouiller la rivière. En effet, l'utilisation d'un tamis au maillage trop fin ne filtre pas assez les scories (déchets) et impose un grand travail pour fouiller les éléments retenus. Par contre, l'utilisation d'un tamis au maillage trop important risque de ne pas retenir les pépites tant recherchées.

2.1.2 ... aux pépites de connaissances des bases de données

Dans le cadre de la fouille de données non supervisée, nous sommes confrontés au même dilemne : trouver le meilleur compromis entre :

- efficacité de la recherche des nouvelles connaissances
- et volume de nouvelles connaissances trouvées.

Les masses de données existant actuellement représentent toutes des rivières contenant potentiellement des quantités de pépites d'or inimaginables aux yeux des experts du domaine. Les domaines concernés sont multiples : médecine, communication, commerce, sécurité, productivité, *etc.*Chacun de ces domaines détient des quantités de données souvent inexploitées car sous-estimées ou tout simplement par faute de temps et de moyens. Pourtant, certaines des connaissances contenues dans ces données représentent de véritables mines d'or pour leurs experts. Par exemple, dans le monde

Identifiant	Transaction
1	beurre fruits lait pain
2	fruits lait pain
3	beurre fromage pain pâtes viande vin
4	fromage fruits lait légumes pain pâtes poisson
5	beurre fruits lait légumes pain pâtes poisson viande
6	beurre fromage légumes pain pâtes viande vin
7	beurre fromage lait légumes pain pâtes viande vin
8	fruits légumes poisson
9	beurre fromage lait pain pâtes viande vin
10	beurre fromage fruits lait légumes pain poisson viande

TAB. I.1 – Quelques tickets de caisse d'un magasin.

industriel, la détection de pannes dans des systèmes de production, réalisée à partir de données collectées pendant la fabrication d'un produit devrait permettre, d'une part de comprendre la source des pannes et d'autre part de diminuer le nombre de rebuts produits par le système.

Un de nos objectif est de créer un orpailleur automatique pouvant travailler avec des données respectant le format détaillé dans la section suivante.

3 Nature des données étudiées

Nous nous sommes focalisés sur la fouille de données transactionnelles. Une base de données transactionnelles est composée d'un ensemble d'enregistrements. Chaque enregistrement est décrit par un ensemble d'attributs. Chaque attribut est à valeur booléenne ou discrète. Voici quelques exemples classiques de données transactionnelles : tickets de caisse d'un supermarché, résultats des recensements, réponses à un sondage d'opinions, *etc.*

3.1 Exemple de données transactionnelles

Le Tableau I.1 présente un exemple de données transactionnelles.

Cet exemple va nous permettre d'introduire une partie des mesures utilisées en fouille de données. La première mesure importante en fouille de données est le **support** (ou fréquence) des attributs. Sur l'exemple du Tableau I.1, le support de l'attribut « lait » est égal à 7/10, c'est-à-dire que 70% des transactions (enregistrements) contiennent l'attribut « lait ». Nous pouvons aussi exprimer le support d'un ensemble d'attributs, ainsi, le support de l'ensemble des attributs « pain », « lait » et « fromage » est égal à 4/10.

L'extraction de connaissances à partir de ces données peut prendre différentes formes :

- recherche des attributs les plus fréquemment présents dans les transactions (pour l'exemple donné et par ordre décroissant de support, nous obtenons : « pain », « lait », « beurre », « pâtes », *etc.*) ;
- recherche des transactions contenant un ou plusieurs attributs donnés ;
- recherche de relations entre attributs permettant d'expliquer le comportement des acheteurs (par exemple : « tous les clients achetant du fromage achètent aussi du pain ») ;
- *etc.*

Chacune de ces formes d'analyse des données est pertinente. Le choix du type de connaissances recherchées est lié à l'utilisation que l'on souhaite en faire.

Règle d'association	Support	Confiance
beurre → pain	70%	100 %
pain → beurre	70%	77,7 %
vin → beurre	40%	100 %
poisson viande → lait	20%	100 %
fromage pâtes → vin	40%	80 %

TAB. I.2 – Quelques règles d'association observées sur la base commerciale.

3.1.1 Exemple d'exploitation lucrative de données transactionnelles

Nous pouvons aisément imaginer que le gérant du magasin souhaite augmenter la rentabilité de son commerce et donc mettre en rayon les produits qui se vendent le plus. Il sera donc certainement intéressé, dans un premier temps, par le premier type de connaissances que nous pouvons obtenir : la liste des produits qui se vendent le plus.

Ayant mis en évidence (dans les rayons de son magasin) et en quantité suffisante les produits qui se vendent le plus, et ainsi réalisé de substantiels gains en termes de chiffres d'affaire, notre gérant aura très certainement envie de continuer à accroître ses bénéfices. Le second type de connaissances peut ainsi lui permettre de trouver toutes les transactions contenant le produit le moins cher, puis à partir de cet ensemble réduit de transactions, trouver les produits les plus fréquemment achetés avec ce produit bon marché. Ayant cette information, notre gérant pourra créer des offres promotionnelles incluant le produit le moins cher et un des produits les plus fréquements achetés avec celui-ci (exemples d'offres liés à nos données : « pâtes + vin » ou « pâtes + viande »). Ceci lui permet donc, à moindre frais, d'augmenter le volume des ventes de l'ensemble des produits ciblés.

Enfin, le troisième type de connaissances peut être utilisé pour réorganiser le magasin en s'appuyant sur des règles telles que « les clients achetant du pain et du vin achètent toujours du fromage ». L'agencement des rayons du magasin peut donc être modifié de manière à, soit rapprocher ces trois produits pour aider le client dans ses achats, soit éloigner les produits pour obliger le client à traverser le plus de rayons possibles et ainsi l'inciter à acheter d'autres produits se trouvant sur son passage.

3.2 Connaissances recherchées

Dans le cadre de cette thèse, nous nous sommes focalisés sur l'extraction de relations entre attributs dans le but d'essayer de trouver des lois permettant de mieux comprendre les données étudiées. Ces relations sont appelées des **règles d'association** et peuvent s'exprimer sous la forme suivante : $attribut_1$ & $attribut_4$ → $attribut_{10}$. Cette règle d'association est composée

- d'une **prémisse**, $attribut_1$ & $attribut_4$,
- et d'une **conclusion** : $attribut_{10}$.

La règle nous indique que lorsqu'une transaction contient les attributs $attribut_1$ et $attribut_4$ alors l'attribut $attribut_{10}$ est souvent rencontré. Une nouvelle mesure, la **confiance**, nous permet d'apprécier la qualité de la règle trouvée. La confiance indique le pourcentage de transactions contenant prémisse et conclusion parmi celles qui contiennent la prémisse. Une règle est toujours fournie avec les deux mesures que nous venons de voir : le support et la confiance.

Le Tableau I.2 présente quelques unes des 8555 règles trouvées à partir de la base du magasin (voir Tableau I.1), et vérifiant les contraintes suivantes : $support \geq 10\%$ et $confiance \geq 80\%$.

Il existe plusieurs méthodes permettant de trouver toutes les règles d'association contenues dans une base de données transactionnelles. La plus simple d'entre elles consiste à énumérer tous les

Symbole	Signification
C_k	Ensemble des k-itemsets candidats
L_k	Ensemble des k-itemsets fréquents
Apriori_Gen	fonction qui engendre les itemsets candidats

TAB. I.3 – Notations utilisées dans l'algorithme APRIORI.

ensembles d'attributs appelés *itemsets*, puis à partir de ces itemsets, de calculer toutes les règles d'association possibles.

Le nombre total d'itemsets, pour une base de données contenant n attributs booléens, est égal à 2^n. Cette méthode très naïve est bien évidemment inapplicable aux données réelles car le nombre d'itemsets candidats devient très rapidement supérieur aux capacités de traitement d'un ordinateur.

Par exemple, pour notre base commerciale, le nombre total d'itemsets pouvant être obtenus est égal à $2^{10} = 1024$. Dans un centre commercial, le nombre d'articles mis en rayon est plus proche de 1000 que de 10. Dans ce cas, le nombre d'itemsets pouvant être obtenu est proche de 10^{300}.

Les règles d'association recherchées sont obtenues à partir des itemsets fréquents trouvés. Pour un itemset de taille k, appelé **k-itemset** le nombre possible de règles est $\sum_{i=2}^{k} C_k^i$. Ainsi, pour un itemset de taille 6 (c'est-à-dire un panier contenant seulement 6 articles) le nombre de règles possible est égal à $\sum_{i=2}^{6} C_6^i = 57$. Ce nombre de règles peut paraître faible mais le problème est que le nombre d'itemsets de taille 6 est égal à $C_{1000}^6 \sim 10^{18}$.

Cette première méthode est donc inexploitable pour des données transactionnelles réalistes. Un des défauts de cette méthode est lié au fait que de nombreux itemsets sont calculés pour finalement ne pas conduire à des règles intéressantes (support ou confiance inférieurs aux seuils choisis).

Une méthode beaucoup moins naïve et nettement plus efficace consiste à calculer les itemsets ayant un support supérieur à un seuil fixé par l'expert. Ces itemsets sont appelés **FrequentItemSets (FIS)**. L'algorithme APRIORI [Agrawal et al. 1993] est l'algorithme de référence implantant cette méthode. Il existe de nombreuses améliorations de cet algorithme mais leur principe général est le même et est fondé sur la *génération des itemsets par niveaux* :

- calcul des 1-itemset fréquents (**1-FIS**)
- puis utilisation des **(n-1)-FIS** pour calculer les **n-FIS** candidats.

La réduction significative du nombre d'itemsets engendrés (par rapport à la méthode naïve) est due à la propriété mathématique d'*anti-monotonie* du support qui assure que si un FIS de taille k n'est pas fréquent alors tous les ensembles de taille $k+1$ pouvant être obtenus à partir de ce FIS ne sont pas fréquents. Il n'est donc pas nécessaire d'engendrer les FIS de taille n si ceux de taille $n-1$ ne sont pas fréquents.

Cette approche reste complète, c'est-à-dire que tout les FIS vérifiant la contrainte du support minimal sont trouvés.

Le pseudo-code d'APRIORI est donné dans l'algorithme 1. Les notations utilisées sont définies dans le Tableau I.3.

Considérons la base de données $\mathcal{B}$ présentée dans le Tableau I.4. La recherche des FIS sur ces données, avec l'algorithme APRIORI (seuil minimal de support fixé à 2/6), nous fournit 23 itemsets fréquents (voir Figure I.1). Le nombre d'itemsets non calculés est égal à 7.

À partir de ces itemsets fréquents, nous pouvons obtenir 32 règles d'association ayant une confiance supérieure ou égale à 80% (voir Tableau I.5).

Notons que selon nous l'avantage majeur de l'algorithme APRIORI est auss son plus grand défaut. En effet, si nous nous plaçons dans l'optique de la découverte de pépites de connaissance, il est souvent difficile, voire impossible pour l'expert, de définir le seuil de support tel qu'aucune

Algorithme 1 APRIORI $(\mathcal{B}, s_0)$

Entrée :

$\mathcal{B}$: une base de données transactionnelles

s_0 : un seuil de fréquence absolue

Sortie :

$\cup_k L_k$: L'ensemble de tous les itemsets fréquents de $\mathcal{B}$

Début

$L_1 \leftarrow \{1\text{-itemsets fréquents}\}$

Pour $(k \leftarrow 2\ ;\ L_{k-1} \neq \emptyset\ ;\ k++)$ **Faire**

-- Apriori_Gen calcule les itemsets candidats de taille k à partir de ceux de taille $k-1$

$C_k \leftarrow Apriori\_Gen(L_{k-1})$

Pour tout $(t \in \mathcal{B})$ **Faire**

$C_t \leftarrow Subset(C_k, t)$ *-- Sélectionne les itemsets de C_k contenus dans la transaction t*

Pour tout $(c \in C_t)$ **Faire**

$support(c) \leftarrow support(c) + 1$

Fin Pour

Fin Pour

-- Sélection des itemsets vérifiant la contrainte de support s_0

$L_k \leftarrow \{c \in C_k | support(c) \geq s_0\}$

Fin Pour

Retourner $\cup_k L_k$

Fin

Identifiant	**Transaction**
1	A B C D E
2	A B C E
3	C
4	B C
5	A B C D
6	A B C

TAB. I.4 – Une base de données transactionnelles $\mathcal{B}$.

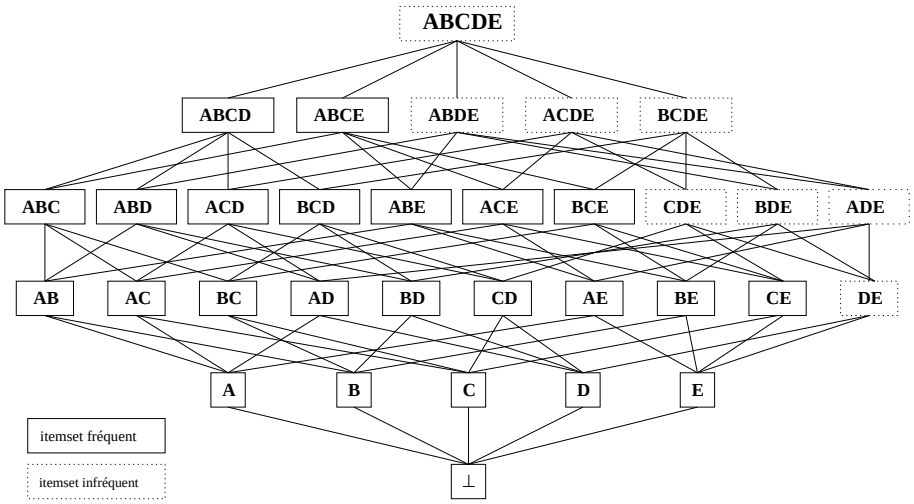


FIG. I.1 – Treillis des Itemsets obtenus à partir de la base $\mathcal{B}$.

connaissance nouvelle ne possède un support inférieur au seuil choisi.

Face à ce dilemne, déjà évoqué précédemment, nous proposons d'étudier directement les caractéristiques des pépites recherchées et de nous focaliser sur les meilleures pépites dont la qualité est déterminée par une ou plusieurs mesures de qualité.

Sur notre exemple, si l'on considère que seules les règles indiquées en gras dans le Tableau I.5 sont intéressantes alors plusieurs questions se posent :

- comment isoler ces règles parmi l'ensemble des règles obtenues ?
- doit on proposer toutes les règles à l'expert ?
  - si oui : comment les présenter à l'utilisateur pour qu'il puisse toutes les analyser efficacement ?
  - si non : peut-on éviter de calculer les règles que nous ne proposerons pas à l'expert ?

Règle	Support	Confiance
$D \rightarrow A$	33,3%	100%
$D \rightarrow B$	33,3%	100%
$\boldsymbol{D \rightarrow C}$	33,3%	100%
$E \rightarrow A$	33,3%	100%
$E \rightarrow B$	33,3%	100%
$E \rightarrow C$	33,3%	100%
$\boldsymbol{A \rightarrow B}$	66,7%	100%
$B \rightarrow A$	66,7%	80%
$A \rightarrow C$	66,7%	100%
$B \rightarrow C$	83,3%	100%
$C \rightarrow B$	83,3%	83,3%
$DA \rightarrow B$	33,3%	100%
$DB \rightarrow A$	33,3%	100%
$DA \rightarrow C$	33,3%	100%
$DC \rightarrow A$	33,3%	100%
$DB \rightarrow C$	33,3%	100%

Règle	Support	Confiance
$DC \rightarrow B$	33,3%	100%
$\boldsymbol{EA \rightarrow B}$	33,3%	100%
$EB \rightarrow A$	33,3%	100%
$EA \rightarrow C$	33,3%	100%
$EC \rightarrow A$	33,3%	100%
$EB \rightarrow C$	33,3%	100%
$EC \rightarrow B$	33,3%	100%
$AB \rightarrow C$	66,7%	100%
$AC \rightarrow B$	66,7%	100%
$BC \rightarrow A$	66,7%	80%
$DAB \rightarrow C$	33,3%	100%
$DAC \rightarrow B$	33,3%	100%
$DBC \rightarrow A$	33,3%	100%
$EAB \rightarrow C$	33,3%	100%
$EAC \rightarrow B$	33,3%	100%
$EBC \rightarrow A$	33,3%	100%

TAB. I.5 – Règles d'association obtenues à partir de la base $\mathcal{B}$.

4 Plan de la thèse

Pour filtrer l'ensemble des règles intéressantes parmi toutes les règles obtenues, nous pouvons utiliser des mesures de qualité autres que la mesure de confiance. Nous verrons dans le chapitre II un ensemble de mesures de qualité permettant d'ordonner les règles obtenues selon différents critères.

Cependant, il demeure un problème essentiel lié à la nature de l'approche utilisée. Que se passe-t-il si les informations intéressantes et utiles ont un support et/ou une confiance inférieurs aux seuils fixés par l'utilisateur ? Les algorithmes utilisés ne seront pas en mesure de trouver ces informations et donc les connaissances proposées à l'utilisateur ne lui seront pas utiles.

Nous proposons, dans le cadre de cette thèse, une nouvelle approche permettant à l'utilisateur de fouiller ses données sans *a priori* sur le support et la confiance des connaissances recherchées. En contrepartie, la recherche de toutes les pépites de connaissance (ou règles d'association) devient difficile (voire impossible) car la plupart des mesures qui permettent d'évaluer la qualité des pépites trouvées ne possèdent plus les propriétés mathématiques permettant d'assurer un élagage efficace de l'espace de recherche. Ainsi, nous ne serons pas en mesure de proposer à l'utilisateur la totalité des pépites contenues dans les données mais juste celles ayant les « meilleures » caractéristiques (évaluées par la ou les mesures de qualité utilisées). Cette approche sera présentée dans le chapitre III.

Si l'on considère que nous disposons d'un outil capable de fournir à l'utilisateur les « meilleures » règles d'association contenues dans ses données, quelle est la résistance de cet outil lorsque les données manipulées sont bruitées ?

En effet, dès qu'un tel outil est utilisé sur des données réelles, la notion de « perfection » des données ne doit plus être considérée comme un acquis. Les données sont souvent altérées lors de leur acquisition et/ou de leur enregistrement dans la base (imperfection des appareils de mesures, fautes de frappe lors de la saisie des données, erreurs d'observations, *etc.*). Il est donc impératif de disposer d'outils peu sensibles au bruit et pour lesquels cette sensibilité peut être évaluée. Nous verrons dans le chapitre IV différents protocoles expérimentaux permettant d'évaluer la sensibilité d'un outil et nous proposerons une méthode permettant de minimiser l'impact du bruit sur les connaissances obtenues.

Dans le dernier chapitre de cette thèse, nous présentons une chaîne d'outils permettant d'obtenir des connaissances du type règles d'association à partir de textes. L'outil de fouille de textes que nous présentons a été testé sur plusieurs corpus différents et une partie des résultats obtenus a été validée par un expert.

Chapitre II

Mesures de qualité

1 Introduction

Dans ce chapitre, nous considérons qu'un algorithme du type Apriori (voir l'algorithme 1) fournit un ensemble de règles d'association.

L'utilisation de mesures de qualité permet de proposer à l'utilisateur les règles d'association les mieux adaptées à sa recherche. Pour pouvoir ordonner et/ou filtrer les règles ne correspondant pas aux attentes de l'utilisateur, il est indispensable de comprendre ce que l'utilisateur recherche dans les données. Idéalement, nous devrions adapter le système de fouille de données aux besoins de chaque utilisateur. Malheureusement, ceci est difficilement réalisable, d'une part car cette adaptation représente un coût important en terme de travail et d'autre part les utilisateurs ont souvent beaucoup de problèmes pour exprimer clairement la nature des connaissances qu'ils recherchent dans les données.

Par exemple, dans le cadre de cette thèse, un ensemble d'introductions d'articles scientifiques du domaine de la fouille de données a été utilisé pour effectuer de la recherche de règles d'association (voir le chapitre V). Ces articles, écrits en anglais, sont répartis en deux sous-ensembles : les articles écrits par des anglophones et ceux écrits par des francophones. L'un des objectifs fixé par l'utilisateur est de détecter des formes particulières dans les articles permettant aux auteurs francophones de mieux rédiger leurs articles en anglais. Cet objectif facilement compréhensible pour un humain est difficilement interprétable pour la machine. Nous avons donc choisi, après avoir représenté les textes sous une forme matricielle adaptée à l'extraction de règles d'association, de nous focaliser sur la détection des règles d'association les moins contredites par les données, en espérant que les résultats trouvés puissent répondre aux attentes de l'utilisateur. Pour trouver ces règles, nous avons utilisé deux mesures de qualité différentes : la moindre contradiction et l'intensité d'implication normalisée. Ces mesures sont respectivement présentées dans la section 5.13 de ce chapitre et dans la section 5 du chapitre III.

Les règles obtenues ont montré que le style des auteurs francophones rédigeant en anglais est très différent du style des auteurs anglophones. Les règles d'association ont permis de mettre en évidence certaines propriétés des textes anglais rédigés par des anglophones qui pourraient aider les auteurs francophones. Ces résultats sont présentés dans le chapitre V.

Nous avons aussi utilisé une approche plus classique pour extraire des connaissances de ces mêmes corpus, à savoir, l'algorithme Apriori et la mesure de confiance. Les résultats obtenus ont

été moins concluants que ceux obtenus avec les mesures précédentes. La satisfaction de l'utilisateur étant un critère essentiel, nous pouvons conclure que, pour ce problème particulier, les mesures de qualité retenues étaient adaptées au probème posé.

Cet exemple montre aussi que, face à une demande de l'utilisateur, le choix de la mesure de qualité à utiliser pour trouver les connaissances satisfaisant cette demande n'est pas trivial.

Une des particularités de l'approche choisie (la détection des règles les moins contredites par les données) est de fournir généralement peu de règles. Et, dans toutes les situations, notre approche fournit un ensemble de règles nettement moins volumineux que celui obtenu avec l'algorithme Apriori avant « nettoyage ».

Ce faible nombre de règles est appréciable car lors de l'étape de validation, sollicitant le plus l'expert, il est préférable de pouvoir évaluer toutes les règles obtenues.

Ainsi, l'existence d'un ensemble de mesures de qualité ayant des propriétés clairement définies permet à l'utilisateur de choisir celles qui correspondent le plus à ses attentes. L'utilisation de ces mesures permet donc de proposer à l'expert des connaissances qui correspondent aux critères qu'il a choisis comme étant ceux qui correspondent le plus à ses propres critères de qualité.

L'objectif de ce chapitre est donc de présenter un ensemble de propriétés requises ou souhaitées pour les connaissances recherchées. Nous n'avons pas la prétention de dresser une liste exhaustive des propriétés existantes mais plutôt de présenter les propriétés les plus utilisées et les plus pertinentes pour évaluer la qualité des connaissances obtenues lors de la fouille de données.

Nous verrons dans le chapitre III comment utiliser directement les mesures de qualité pour obtenir un sous-ensemble de règles d'association vérifiant au mieux les critères de qualité choisis par l'utilisateur.

2 Définitions

Nous avons vu dans l'introduction de cette thèse que les données sur lesquelles nous nous sommes focalisés sont des données transactionnelles. Ces données peuvent être représentées sous la forme d'un tableau d'incidence ou d'existence à double entrée de dimension $n * p$ croisant un ensemble $\mathcal{O} = \{o_i | 1 \leq i \leq n\}$ de n objets avec un ensemble $\mathcal{A} = \{a^j | 1 \leq j \leq p\}$ de p attributs booléens.

Nous noterons α_i^j la valeur $Vrai$ ou $Faux$ de l'attribut a^j sur l'objet o_i :

$$\alpha_i^j = a^j(o_i), 1 \leq i \leq n, 1 \leq j \leq p$$

La Figure II.1 illustre les notations retenues pour les données manipulées.

La Figure II.2 présente la correspondance entre la représentation transactionnelle d'une base $\mathcal{B}$ composée de 6 objets décrits par 5 attributs booléens et la représentation matricielle de cette même base. Nous adopterons la seconde représentation pour la suite de cette thèse.

On peut supposer sans restreindre la généralité que la valeur $Vrai$ d'un attribut est sémantiquement plus signifiante que la valeur $Faux$ de cet attribut. Cela, le plus souvent, se traduit statistiquement par le fait que le nombre d'objets où l'attribut est à $Vrai$ est inférieur au nombre d'objets où l'attribut est à $Faux$.

Pour éviter la surcharge d'indices et faciliter la lecture des notations, nous désignerons, dans la suite de cette thèse, les éléments de $\mathcal{A}$ par des lettres capitales A,B,C plutôt que par a^1, a^2, a^3. De plus, ces notations sont couramment utilisées dans le domaine de la fouille de données [Agrawal et al. 1993, Guillaume 2000, Lehn 2000, Bastide et al. 2002].

Notre principal objectif est la détection de relations entre les objets de $\mathcal{O}$. Chaque objet étant décrit par un sous-ensemble des attributs booléens de $\mathcal{A}$, nous nous sommes intéressés aux relations pouvant exister entre les attributs de $\mathcal{A}$ et étant vérifiées par les objets de $\mathcal{O}$.

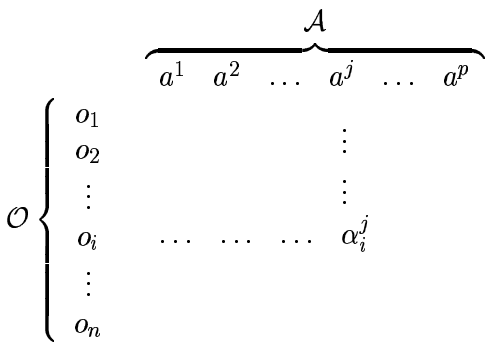


FIG. II.1 – Illustration des notations pour les données.

$\mathcal{O}$	**Transaction**
o_1	A B C D E
o_2	A B C E
o_3	C
o_4	B C
o_5	A B C D
o_6	A B C

$\mathcal{O}$ \ $\mathcal{A}$	A	B	C	D	E
o_1	1	1	1	1	1
o_2	1	1	1	0	1
o_3	0	0	1	0	0
o_4	0	1	1	0	0
o_5	1	1	1	1	0
o_6	1	1	1	0	0

FIG. II.2 – Deux représentations d'une base de données transactionnelles $\mathcal{B}$.

Définition 1 *La relation $A \rightarrow B$ est appelée* **règle d'association** *entre A et B ($A \neq B$) avec $A \subseteq \mathcal{A}$, $B \subseteq \mathcal{A}$ et $A \cap B = \emptyset$.*

A est appelée la **prémisse** *de la règle et B la* **conclusion**.

Une règle est caractérisée par son **support** *et sa* **confiance**.

Avant de définir plus précisément le support et la confiance, nous devons introduire des notations permettant de rendre compte des interactions entre les attributs A et B.

Définition 2 *Notons $n_A = card(\mathcal{O}(A))$ (resp. n_B) le nombre d'objets de $\mathcal{O}$ où A (resp. B) est à $Vrai$.*

Définition 3 *Notons $n_{AB} = card(\mathcal{O}(A \cap B))$ le nombre d'objets de $\mathcal{O}$ où A et B sont tous les deux à $Vrai$. Cette valeur est appelée* **indice « brut » d'association** *entre A et B.*

Définition 4 *Notons $\overline{A}$ (resp. $\overline{B}$) la négation de A (resp. B). Nous avons $\mathcal{O}(A) \cup \mathcal{O}(\overline{A}) = \mathcal{O}$ (resp. $\mathcal{O}(B) \cup \mathcal{O}(\overline{B}) = \mathcal{O}$). $\mathcal{O}(A)$ (resp. $\mathcal{O}(B)$) représente l'ensemble des objets de $\mathcal{O}$ où A (resp. B) est à $Faux$.*

La figure II.3 illustre les notations définies précédemment pour une règle d'association $A \rightarrow B$ et un ensemble $\mathcal{O}$ d'objets.

Ayant introduit ces différentes notations, nous pouvons maintenant donner la définition du support et de la confiance.

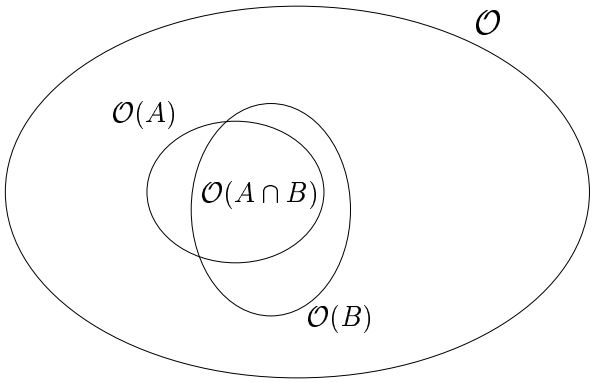


FIG. II.3 – Illustration des notations $\mathcal{O}$, $\mathcal{O}(A)$, $\mathcal{O}(B)$ et $\mathcal{O}(A \cup B)$ pour une règle d'association $A \to B$.

Définition 5 *Le* **support** *ou taux de couverture d'une règle d'association $A \to B$ est défini par*

$$support(A \to B) = \frac{n_{AB}}{n}$$

avec $n = card(\mathcal{O})$. Cette mesure représente le pourcentage d'objets (transactions) vérifiant la règle.

Définition 6 *La* **confiance** *d'une règle d'association $A \to B$ est définie par*

$$confiance(A \to B) = \frac{n_{AB}}{n_A} \approx P(B|A)$$

Cette mesure représente le pourcentage d'objets vérifiant la conclusion de la règle parmi celles qui vérifient la prémisse. $P(B|A)$ représente la probabilité conditionnelle qu'une transaction contienne l'attribut B sachant qu'elle contient l'attribut A. On considère que cette valeur est une approximation satisfaisante de $P(B|A)$. Dans la suite de cette thèse nous utiliserons indifféremment la notation $P(B|A)$ et $confiance(A \to B)$.

Définition 7 *Étant donnée une règle d'association $A \to B$ les* **exemples** *de cette règle sont les transactions vérifiant AB et les* **contre-exemples** *sont les transactions vérifiant $A\overline{B}$.*

Étant donnée une règle d'association $A \to B$ les observations dont nous disposons sont présentées dans le Tableau II.1.

	A	$\overline{A}$	$\sum$
B	n_{AB}	$n_{\overline{A}B}$	n_B
$\overline{B}$	$n_{A\overline{B}}$	$n_{\overline{A}\overline{B}}$	$n_{\overline{B}}$
$\sum$	n_A	$n_{\overline{A}}$	n

	A	$\overline{A}$	$\sum$
B	$P(AB)$	$P(\overline{A}B)$	$P(B)$
$\overline{B}$	$P(A\overline{B})$	$P(\overline{A}\overline{B})$	$P(\overline{B})$
$\sum$	$P(A)$	$P(\overline{A})$	1

TAB. II.1 – Observations disponibles étant donnée une règle $A \to B$ et $n = card(\mathcal{O})$.

À partir de ces observations, nous pouvons créer de nombreuses mesures de qualité permettant d'évaluer un ensemble de règles d'association.

Les mesures de support et de confiance définies précédemment permettent d'évaluer une certaine forme de l'intérêt des règles d'association.

3 Support et Confiance

Ces deux mesures initialement utilisées par [Agrawal et al. 1993] dans un contexte d'extraction de règles d'association avaient déjà été introduites par [Hájek et al. 1966, Hájek 2001] dans un cadre de génération automatique d'hypothèses à partir de données.

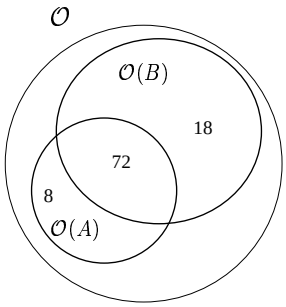
Comme nous l'avons déjà abordé dans l'introduction, ces mesures permettent de ne retenir que les règles d'association vérifiant les conditions imposées par l'utilisateur. Ces conditions garantissent que chaque règle trouvée possède un support et une confiance supérieurs aux seuils min_{sup} et min_{conf} fixés par l'utilisateur avant l'utilisation de l'algorithme d'extraction des règles.

Limites de l'approche Support-Confiance

[Brin et al. 1997b, Lallich et Teytaud 2004] montrent que l'utilisation seule de ces deux mesures ne suffit pas à garantir la qualité des règles détectées. En effet, comme le montre l'exemple du Tableau II.2 (extrait de [Lallich et Teytaud 2004]), la règle $A \rightarrow B$ possède un support élevé (si l'on considère $n = 100$) et une confiance élevée : $support(A \rightarrow B) = 72\%$ et $confiance(A \rightarrow B) = 90\%$. Cependant, la confiance de cette règle est égale à la probabilité d'observer la conclusion de la règle dans les données (indépendemment de la prémisse, c'est-à-dire que $confiance(A \rightarrow B) = P(B|A) = P(B)$). Et donc, nous avons $P(A) \times P(B) = P(AB)$, donc la probabilité que des individus de $\mathcal{O}$ possèdent les attributs A et B à $Vrai$ est purement fortuite car les attributs sont totalement indépendants. Précisons que deux attributs X et Y sont considérés comme indépendants si $|P(X \cap Y) - P(X) \times P(Y)| = 0$ [Kodratoff 2000].

Pour cet exemple, bien que la règle vérifie les conditions minimales de support ($min_{sup} = 0.3$) et de confiance ($min_{conf} = 0.9$), elle n'apporte aucune connaissance nouvelle à l'utilisateur. Cette règle n'est pas fondée car les attributs sont indépendants. Ainsi, sur cet exemple, la seule utilisation de la confiance comme mesure de qualité entraine la sélection d'une règle non pertinente.

Il est donc impératif d'utiliser d'autres mesures de qualité (telle que la dépendance sur cet exemple) pour évaluer l'intérêt des règles d'association.



	A	$\overline{A}$	$\sum$
B	72	18	90
$\overline{B}$	8	2	10
$\sum$	80	20	100

TAB. II.2 – Inconvénient de l'approche Support-Confiance.

4 Étude des critères de qualité

Il est impossible de concevoir une mesure de qualité permettant de statisfaire tous les critères de qualité de l'utilisateur.

[Lallich et Teytaud 2004] présentent une liste de critères permettant d'apprécier la qualité d'une mesure m. Nous reprenons ici leurs travaux en les étendant avec l'étude réalisée dans [Guillaume 2000] sur différentes mesures de qualité. Nous complétons cette étude en y ajoutant quelques nouveaux critères de qualité.

Voici donc une liste de quelques critères pouvant être utilisés pour concevoir des mesures de qualité :

1. Compréhensibilité de la mesure pour l'utilisateur et utilisation directe de celle-ci
2. Nature des règles ciblées par la mesure
3. sensibilité à l'apparition des exemples et des contre-exemples
4. sens de variation de la mesure
5. nature de la variation : linéaire/non linéaire
6. comportement par rapport à la taille de la prémisse et de la conclusion
7. sensibilité à la taille des données
8. caractère discriminant de la mesure
9. utilisation d'un seuil d'élagage controlé par l'utilisateur
10. classement induit par une mesure
11. comportement par rapport au contexte des règles étudiées
12. contradiction des connaissances *a priori* de l'utilisateur
13. sensibilité au bruit

Chacun de ces critères est détaillé dans les sections suivantes. Dans un premier temps, une présentation générale des critères est effectuée. C'est-à-dire que les critères sont détaillés pour une mesure de qualité m quelconque. Puis, dans la section 5, nous présentons plusieurs mesures de qualité vérifiant tout ou partie de ces critères.

4.1 Compréhensibilité de la mesure pour l'utilisateur et utilisation directe de celle-ci

Certaines mesures telles que le support et la confiance sont directement interprétables par l'utilisateur. Ces mesures possèdent donc un sens « concret » et elles permettent à l'utilisateur de situer les règles les unes par rapport aux autres.

Considérons deux règles R_1 et R_2 ayant même support et telles que $confiance(R_1) = 2 \times confiance(R_2)$. La seule connaissance de la confiance permet à l'utilisateur de savoir que la règle R_1 est deux fois plus fiable que la règle R_2 car indépendamment de la taille de la conclusion, la règle R_1 indique que parmi les individus vérifiant sa prémisse, la proportion d'individus vérifiant aussi la conclusion est deux fois plus importante que dans le cas de la règle R_2.

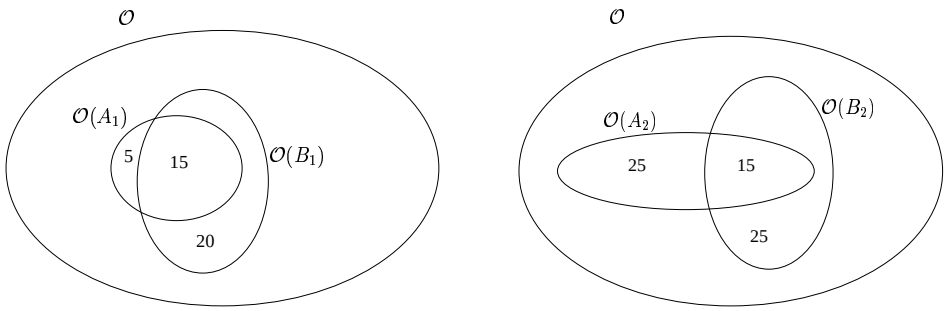
La figure II.4 illustre les règles R_1 et R_2, sur un exemple.

La plupart des mesures de qualité ne possèdent pas cette propriété, ce qui rend ainsi difficile l'interprétation de leur valeur pour un utilisateur.

4.2 Nature des règles ciblées par la mesure

Lors de l'étude de la nature de la liaison entre deux attributs (ou ensemble d'attributs) A et B, il convient de distinguer plusieurs cas :

– recherche d'un lien orienté ou non entre A et B,



$R_1 = A \rightarrow B$ $\qquad$ $R_2 = A_1 \rightarrow B_2$

FIG. II.4 – $support(R_1) = support(R_2)$ et $confiance(R_1) = 2 \times confiance(R_2)$.

– capacité de la mesure à distinguer les cas $A \rightarrow B$ et $A \rightarrow \overline{B}$.

Pour le premier cas, si l'orientation du lien entre A et B n'est pas significative, il convient d'utiliser des mesures évaluant de la même façon $A \rightarrow B$ et $B \rightarrow A$. Par contre, si l'orientation est importante, alors l'utilisation de telles mesures est à prohiber au profit de mesures évaluant différemment ces deux règles.

Pour le deuxième cas, l'utilisation de mesures ne distinguant pas les exemples de la règle $A \rightarrow B$ de ceux de la règle $A \rightarrow \overline{B}$ conduit à évaluer de la même manière ces deux règles qui sont de nature fort différente. Lorsque de telles mesures sont utilisées, l'expert doit avoir en mémoire ce comportement pour pouvoir interpréter correctement les résultats obtenus.

4.3 Sensibilité à l'apparition des contre-exemples

L'évaluation de l'intérêt d'une règle peut se mesurer en fonction du nombre (élevé) d'exemples de la règle ou en fonction du nombre (faible) de ses contre-exemples.

Dans le premier cas, le comportement de la mesure doit rendre compte de l'apparition des exemples par une augmentation de sa valeur. Et dans le second cas, l'apparition des contre-exemples doit se traduire par une diminution de la valeur de la mesure.

Rappelons que l'observation n_{AB} correspond au nombre d'exemples de la règle $A \rightarrow B$ et que l'observation $n_{A\overline{B}}$ correspond au nombre de ses contre-exemples.

Certaines mesures prennent en considération de manière implicite $n_{A\overline{B}}$ dans leurs évaluations. Par exemple, la confiance prend de manière implicite le nombre de contre-exemples en considération puisque les observations n_{AB} et $n_{A\overline{B}}$ sont liées par la relation $n_A = n_{AB} + n_{A\overline{B}}$.

L'observation n_{AB}, correspondant au nombre d'exemples de la règle $A \rightarrow B$, est au cœur de toutes les mesures de qualité. Ainsi, puisque $n_{A\overline{B}} = n_{AB} - n_A$, toutes les mesures de qualité prennent de manière implicite la quantité $n_{A\overline{B}}$ en considération.

Quelques mesures prenant de manière explicite le nombre de contre-exemples dans l'évaluation de la mesure font intervenir n_{AB} et $n_{A\overline{B}}$, ainsi qu'éventuellement d'autres observations.

Les mesures prenant explicitement en considération $n_{A\overline{B}}$ ont une sémantique faisant intervenir cette quantité. Alors que la prise en considération implicite de la quantité $n_{A\overline{B}}$ indique fort souvent que les contre-exemples sont absents de la sémantique de la mesure.

4.4 Sens de variation de la mesure

L'utilisation de mesures de qualité prenant des valeurs positives pour les règles intéressantes permet de se rapprocher des *a priori* de l'utilisateur sur la notion de qualité. De plus, pour deux

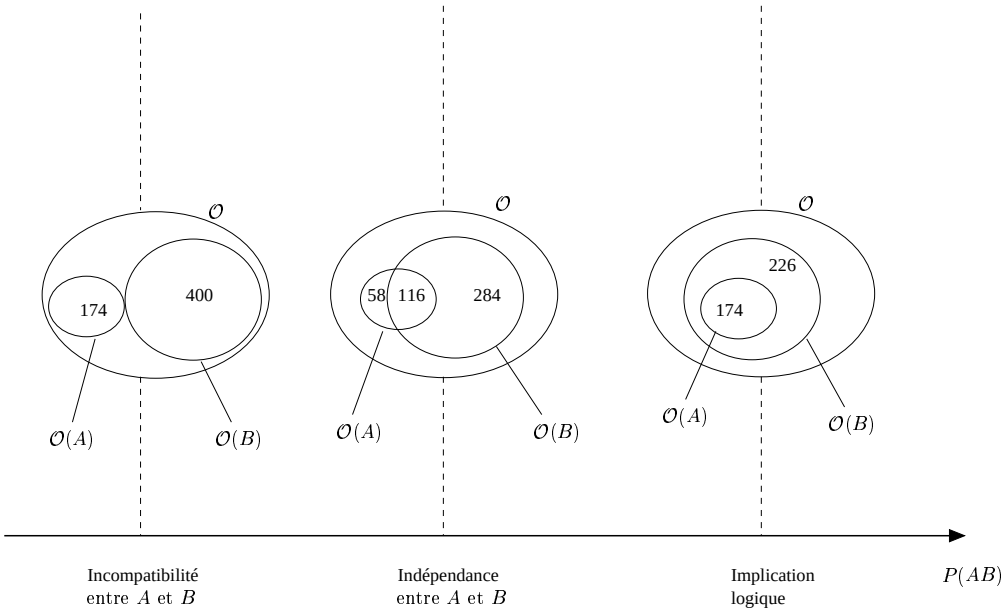


FIG. II.5 – Exemple illustrant les différentes zones d'intérêt.

règles R_1 et R_2, si la qualité de la règle R_1 est supérieure à celle de R_2 alors, la mesure de qualité m doit assurer que $m(R_1) > m(R_2)$.

Selon [Piatetsky-Shapiro 1991], une bonne mesure m de la qualité de la règle $A \rightarrow B$ doit être :
- $m(A \rightarrow B) < 0$ en cas de répulsion : $P(AB) < P(A) \times P(B)$
- $m(A \rightarrow B) = 0$ en cas d'indépendance de A et B : $P(AB) = P(A) \times P(B)$
- $m(A \rightarrow B) > 0$ en cas d'attraction : $P(AB) > P(A) \times P(B)$

Rappelons que la notion d'indépendance entre deux items A et B est à comparer avec la notion d'indépendance utilisée en statistique où deux évenements X et Y sont considérés comme indépendant si $P(X \cap Y) = P(X) \times P(Y)$.

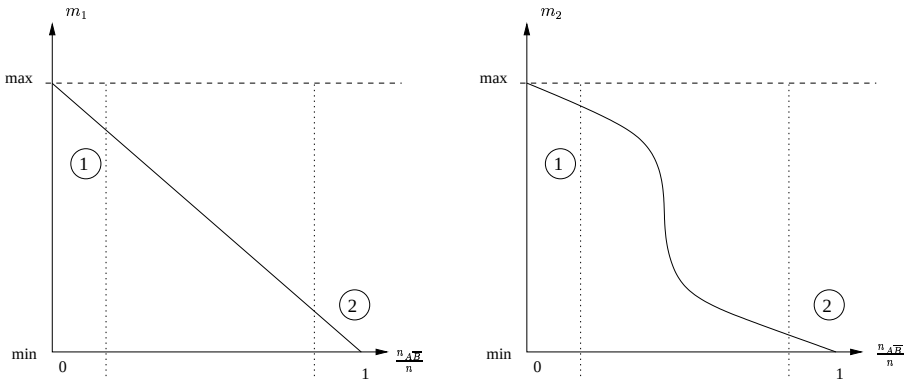
La définition de ces deux zones permet de capturer un certain aspect de la qualité des règles d'association : le comportement par rapport à l'indépendance.

L'exemple de la Figure II.5 (où $n_A = 174$, $n_B = 400$ et $n = 600$), extrait de la thèse de S. Guillaume [Guillaume 2000], illustre les zones d'intérêt pour une mesure quelconque.

4.5 Nature de la variation : linéaire/non linéaire

La mesure peut varier linéairement en fonction de $n_{A\overline{B}}$ ou bien avoir un comportement permettant de rendre compte de l'apparition progressive des contre-exemples. La mesure ayant alors une tendance à décroître lentement lorsque peu de contre-exemples apparaissent puis de plus en plus rapidement jusqu'à atteindre une valeur minimale (idéalement nulle). La figure II.6 illustre les deux types de comportement d'une mesure m en fonction de la proportion de contre-exemples.

Une mesure m variant linéairement en fonction du nombre d'exemples (resp. contre-exemples) de la règle $A \rightarrow B$ est *a priori* plus sensible au bruit qu'une mesure ayant une variation non linéaire. En effet, si les données sont bruitées (voir chapitre IV), pour une règle $A \rightarrow B$, quelques exemples vont se transformer en contre-exemples et inversement. Les mesures de qualité ayant un comportement non linéaire en fonction des exemples permettent donc d'accepter quelques contre-exemples sans pour autant diminuer la valeur de la mesure.



m_1 : comportement linéaire m_2 : comportement non linéaire

FIG. II.6 – Différents comportements d'une mesure de qualité m.

Dans les zones 1 et 2 de la figure II.6, nous pouvons voir que la mesure m_2 se montre moins sensible à l'apparition des contre-exemples que la mesure m_1. En effet, si la règle n'a aucun contre-exemple, les deux mesures sont maximales pour la règle. Pour la mesure m_1, il suffit de très peu de contre-exemples, pour que sa valeur chute significativement alors qu'intuitivement, la règle reste fiable même en présence de peu de contre-exemples. Par exemple, la règle « les êtres humains ont une durée de vie inférieure à 100 ans » est vraie pour la majorité de la population mondiale même si elle admet quelques contre-exemples.

Dans une telle situation, la mesure m_2 conserve une valeur élevée lorsque peu de contre-exemples infirment la règle. Elle diminue de plus en plus rapidement dès que le nombre de contre-exemples dépasse un seuil. Puis de nouveau, elle a un comportement plus modéré lorsque de nombreux contre-exemples infirment la règle.

Nous verrons dans le chapitre IV qu'il existe d'autres facteurs ayant une influence sur le « bon » comportement des mesures de qualité en présence de données bruitées.

4.6 Comportement par rapport à la taille de la prémisse et de la conclusion

L'influence des observations n_A et n_B dans l'évaluation de la mesure de qualité m permet de contrôler le comportement de la mesure par rapport aux tailles des prémisses et des conclusions des règles étudiées.

Considérons que $n_A \leq n_B$. La figure II.7 illustre trois cas pour lesquels le nombre d'exemples et les confiances des règles sont tous identiques. La seule observation qui varie est la taille de la conclusion : n_B.

Lors de l'étude des règles $A \rightarrow B$, l'utilisation de n_B dans la mesure de qualité permet de distinguer chacun de ces cas qui du point de vue du support et de la confiance sont identiques.

Dans l'exemple de la Figure II.7, l'intérêt de la règle $A \rightarrow B$ décroit au fur et à mesure où la taille de $\mathcal{O}(B)$ augmente. La première situation, celle où $\mathcal{O}(A)$ et $\mathcal{O}(B)$ sont de tailles comparables correspond à une règle nettement plus intéressante que celles issues des autres situations.

L'utilisation de n_A permet dans certains cas de distinguer des situations similaires où certaines mesures évaluent de manière identique la qualité de règles qui manifestement ne sont pas de même qualité.

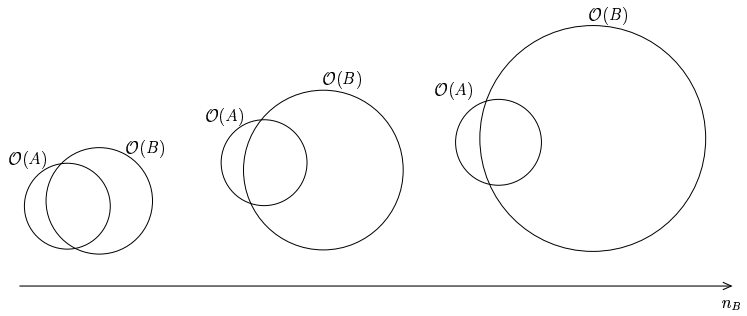


FIG. II.7 – Illustration de l'influence de n_B.

4.7 Sensibilité à la taille des données

Nous nous intéressons ici à une règle $A \rightarrow B$, telle que A et B sont deux items d'une base de données de taille n.

Augmentons artificiellement la taille des données en augmentant la valeur de n. La question est de savoir comment la règle $A \rightarrow B$ est évaluée dans la nouvelle base de données « dilatée », si on suppose que ni la taille de A, ni la taille de B ne sont changées.

La figure II.8 illustre ce phénomène. Nous avons $n_{A_1} = n_{A_2}$, $n_{B_1} = n_{B_2}$, $n_{A_1B_1} = n_{A_2B_2}$.

L'utilisation d'une mesure de qualité sensible à la taille des données permet d'évaluer de manière statistique l'intérêt des règles étudiées. D'un pur point de vue statistique, l'intérêt de la règle $A_1 \rightarrow B_1$ n'est pas le même que celui de la règle $A_2 \rightarrow B_2$. En général, la règle $A_2 \rightarrow B_2$ est plus intéressante que la règle $A_1 \rightarrow B_1$ car elle peut ne pas être connue de l'utilisateur.

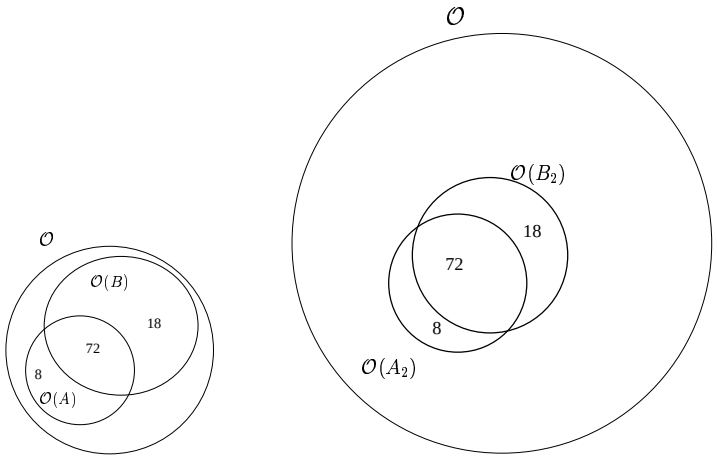


FIG. II.8 – Illustration de l'influence de la taille des données.

Soit m la mesure de qualité étudiée. Cette mesure peut combiner tout ou partie des observations liées à une règle $A \rightarrow B$ (voir Tableau II.1).

Si la mesure est sensible à la taille des données, n doit intervenir dans l'évaluation de la mesure m. Notons la mesure m_n pour rendre compte de l'influence de n sur m. Ainsi si m_n est bornée, à valeurs dans $[\alpha, \beta]$, alors m_n doit atteindre ses bornes lorsque n croit. Notons R_1 une règle d'association $A_1 \rightarrow B_1$ et R_2 une règle d'association $A_2 \rightarrow B_2$.

$$\exists R_1 \; tq \lim_{n \to +\infty} m_n(R_1) = \alpha$$

$$\exists R_2 \; tq \lim_{n \to +\infty} m_n(R_2) = \beta$$

Et donc, dès que n devient suffisamment élevé, m_n perd son pouvoir discriminant et la mesure ne permet plus que de distinguer deux familles de règles $\mathcal{E}_\alpha$ et $\mathcal{E}_\beta$ telles que :

$$\forall R_1 \in \mathcal{E}_\alpha \; m_n(R_1) = \alpha$$

$$\forall R_2 \in \mathcal{E}_\beta \; m_n(R_2) = \beta$$

Inversement, une mesure m insensible à la taille des données ne fait par intervenir n dans son évaluation. Une telle mesure, à caractère non statistique, évalue toujours de la même manière une règle $A \to B$ quelque soit la taille des données desquelles elle est issue. Une telle mesure demeure donc discriminante quelle que soit la taille des données étudiées.

4.8 Utilisation d'un seuil d'élagage

Les mesures de qualité retenues pour extraire les règles d'association doivent pouvoir être utilisées avec un seuil d'élagage de manière à éliminer toutes les règles qui n'intéressent pas l'utilisateur.

Les mesures ayant un sens concret pour l'utilisateur, ainsi que les mesures normalisées et ayant un caractère statistique se prêtent bien à la détermination d'un seuil d'élagage.

Ce seuil peut être fixé par l'utilisateur soit avant la phase d'extraction des règles d'association, soit lors d'une phase de post-élagage des règles. Lorsque le seuil est déterminé *a priori* par l'utilisateur, ce seuil ne prend pas en considération la nature des données et peut donc conduire à des résultats ne représentant pas toujours les données.

L'utilisation de seuils d'élagage calculés directement à partir des données peut permettre d'éviter ce problème. De plus, ceci évite de soliciter l'utilisateur en lui demandant de déterminer *a priori* des seuils associés à des mesures dont la compréhension peut lui échapper.

De tels seuils peuvent être obtenus à partir des valeurs moyennes observées sur les données. Une méthode classique en fouille de données [Lerman 1984, Daudé 1992] consiste à centrer et réduire les valeurs observées par rapport à la moyenne et à l'écart-type de la mesure.

4.9 Classement induit par une mesure

Considérons deux mesures de qualité m_1 et m_2. Si les deux mesures sont comparables au sens des critères précédemment énoncés, comment aider l'utilisateur à choisir une des deux mesures ?

Un des critères de comparaison pouvant être utilisé est l'ordre induit par ces deux mesures sur les règles obtenues. Soit $\mathcal{E}$ l'ensemble des règles obtenues en utilisant un algorithme indépendant de m_1 et de m_2. Les deux mesures sont équivalentes du point de vue de l'ordre induit sur les règles ssi :

$$\forall R, R' \in \mathcal{E} \; m_1(R) > m_1(R') \; \Leftrightarrow \; m_2(R) > m_2(R')$$

C'est-à-dire que pour tout couple de règles (R, R') trouvées par l'algorithme, les mesures m_1 et m_2 classent toujours R et R' de la même manière l'une par rapport à l'autre.

Étant donné ce critère, il devient alors possible de créer des familles de mesures induisant le même ordre sur un ensemble de règles. L'utilisateur pourra alors choisir indifféremment une mesure parmi toutes celles d'une famille donnée, après avoir choisi les critères de qualité qui l'intéressent.

Bien que cette propriété soit relativement compréhensible, nous préférons l'illuster par un exemple relativement simple. Soit l'ensemble $\mathcal{E}$ réduit aux cinq règles d'association R_1, R_2, R_3, R_4, R_5. Soient trois mesures de qualité m_1, m_2 et m_3 qui fournissent les classements suivants pour les règles de $\mathcal{E}$.

Mesure	**Règles triées selon les valeurs croissantes de la mesure**				
m_1	R_4	R_2	R_3	R_1	R_5
m_2	R_2	R_5	R_1	R_3	R_4
m_3	R_4	R_2	R_3	R_1	R_5

TAB. II.3 – Classement induit par trois mesures de qualité différentes sur un ensemble de cinq règles.

Comme les mesures m_1 et m_3 induisent exactement le même ordre sur les règles de $\mathcal{E}$, ces deux mesures peuvent donc être associées à la même famille de mesures de qualité, pour ces règles. Notons que cela peut ne pas être vérifié pour d'autres règles.

4.10 Prise en considération du contexte

Nous appelons **contexte** un ensemble de règles d'association obtenu par un algorithme tel qu'APRIORI. Les règles de ce contexte vérifient un ensemble de contraintes imposées sur les mesures de qualité utilisées par l'algorithme d'extraction (support et confiance minimaux pour APRIORI par exemple). Les règles de ce contexte sont considérées comme des règles *valides*.

Soit une règle d'association $R = A \rightarrow B$ située dans un contexte $\mathcal{E}_1$ ou $\mathcal{E}_2$. Une mesure m est dite sensible au contexte si elle évalue différemment la règle R en fonction du contexte duquel R est extraite.

Notons $m_{\mathcal{E}}(R)$ une règle d'association, issue d'un ensemble de règles valides $\mathcal{E}$, et dont la qualité est évaluée avec la mesure m. Nous considérons qu'une règle R se situe dans un contexte $\mathcal{E}$ si $R \in \mathcal{E}$.

Si m est sensible au contexte et si $\mathcal{E}_1 \neq \mathcal{E}_2$ alors $m_{\mathcal{E}_1}(R) \neq m_{\mathcal{E}_2}(R)$. Si R est une règle intéressante pour l'utilisateur et si $\mathcal{E}_1$ contient moins de règles intéressantes que $\mathcal{E}_2$ alors

$$m_{\mathcal{E}_1}(R) \geq m_{\mathcal{E}_2}(R)$$

En effet, si la règle R se situe dans un contexte $\mathcal{E}_1$ contenant peu de règles intéressantes alors le contraste entre les règles intéressantes et les autres est élevé. Inversement, si la règle R se situe dans un contexte $\mathcal{E}_2$ contenant beaucoup de règles intéressantes, alors la mesure m lui associera une valeur plus faible que celle associée à R dans $\mathcal{E}_1$.

La prise en considération du contexte permet d'introduire un aspect statistique dans la mesure m. Le contexte peut intervenir de différentes façons dans la mesure : utilisation des valeurs moyennes de m observées sur $\mathcal{E}_i$, de l'écart-type, de la médiane.

Ce critère de qualité sera développé en détail dans le chapitre III. Nous proposerons un algorithme permettant d'extraire des pépites de connaissance et utilisant le contexte des règles d'association pour sélectionner les pépites.

4.11 Contradiction des connaissances *a priori* de l'utilisateur

Ce critère permet de comparer les règles d'association obtenues à un ensemble de connaissances fournies par l'utilisateur et considérées comme valides pour le domaine étudié.

Si, à partir des données, nous obtenons les règles d'association suivantes :

Règle	Confiance
$A \rightarrow B$	75%
$B \rightarrow A$	18%

Et si l'utilisateur nous a présenté la règle $B \rightarrow A$ comme étant une connaissance *a priori* du domaine, alors la règle $A \rightarrow B$ est considérée comme intéressante car elle indique qu'il existe bien une relation entre A et B mais que le sens de la relation est inversé par rapport aux *a priori* du domaine. En ce sens, la règle $A \rightarrow B$ contredit des connaissances *a priori* de l'utilisateur.

Ce critère de qualité, non développé dans cette thèse, a été étudié dans [Suzuki 1997] et [Suzuki et Kodratoff 1998

4.12 Sensibilité au bruit

L'étude de données réelles est souvent synonyme de travail sur des données bruitées. En effet, il est rare de trouver des données réelles « parfaites » et l'étude de ces données ne doit pas produire des résultats trop éloignés de ceux obtenus sur les mêmes données non bruitées.

Le bruit peut prendre différentes formes : valeur absente dans les données ou remplacée par une valeur par défaut ; erreur de réglage d'un appareil de mesure utilisé pour obtenir les données ; fautes de frappe lors de la création de la base. Cette liste non exhaustive est suffisante pour mesurer l'importance du bruit et de son étude dans un système d'extraction de connaissances.

La prise en considération du bruit dans une mesure de qualité est difficile car le bruit est par nature mal défini.

Le chapitre IV de cette thèse présente une étude détaillée sur le bruit dans les données et son impact sur les connaissances obtenues.

Idéalement, un système d'extraction de règles d'association ne doit pas fournir à l'expert des règles erronnées même lorsque les données sont bruitées.

La prise en considération du bruit peut intervenir soit dans une phase de pré-traitement des données, soit au niveau de la mesure de qualité, soit dans une phase de post-filtrage des règles obtenues.

5 Différentes mesures de qualité

Nous proposons dans ce paragraphe d'associer à chaque mesure de qualité une sémantique pouvant être utilisée pour présenter la mesure à l'expert. Pour chaque mesure et si cela s'avère pertinent, nous la situerons par rapport aux différents critères de qualité précédemment introduits.

Nous présentons dans le Tableau II.6 la liste des mesures étudiées, ainsi que leur définition en fonction des observations effectuées sur les données.

5.1 Support

$$support(A \rightarrow B) = P(AB) \tag{II.1}$$

Le support [Agrawal et al. 1993] d'une règle $A \rightarrow B$ indique la proportion d'objets vérifiant à la fois la prémisse et la conclusion de la règle.

Cette mesure est symétrique, i.e. $support(A \rightarrow B) = support(B \rightarrow A)$. Elle ne suffit pas à évaluer l'intérêt d'une règle. Elle est souvent utilisée pour élaguer les règles d'association non

intéressantes. Il suffit à l'utilisateur de fixer un seuil d'élagage associé à cette mesure. Grâce à la propriété d'anti-monotonie, toutes les règles issues d'une règle $A \rightarrow B$ telle que son support est inférieur au seuil retenu par l'utilisateur seront supprimées de l'ensemble des règles proposées à l'expert.

Cette mesure est intéressante dans certains domaines d'applications particuliers tels que les domaines commerciaux par exemple. Dans ces domaines, les experts souhaitent détecter les comportements les plus fréquents parmis leurs clients pour optimiser leurs offres en touchant la clientèle la plus large possible.

5.2 Confiance

$$confiance(A \rightarrow B) = P(B|A) = \frac{P(A \wedge B)}{P(A)} \tag{II.2}$$

La confiance [Agrawal et al. 1993] indique la proportion d'objets vérifiant la conclusion parmi ceux vérifiant la prémisse. Cette mesure est non symétrique et insensible à la taille des données. En effet, quelque soit la taille de $\mathcal{O}$, l'ensemble des objets étudiés, la confiance d'une règle $A \rightarrow B$ est constante. Cette mesure doit donc être utilisée avec le support pour pouvoir palier à ce problème.

En revanche, un des défauts de la confiance est d'évaluer de manière identique les trois situations présentées sur la figure II.9.

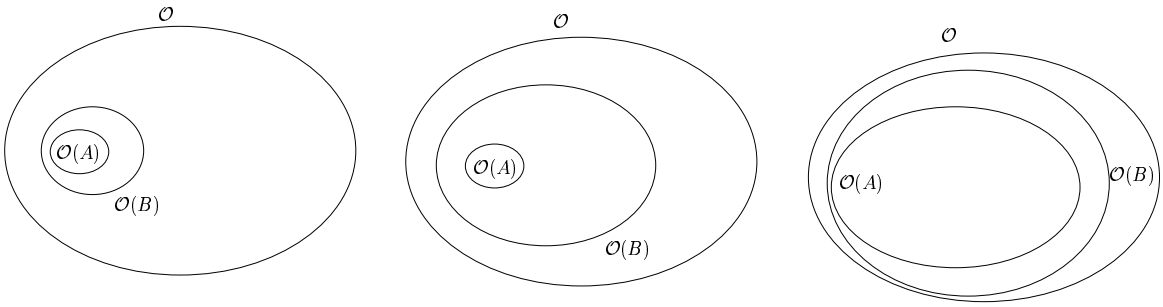


FIG. II.9 – Trois cas intéressants pour la confiance. Cas où la confiance est égale à 1.

Or, il est évident que ces situations doivent être interprétées différemment. Les premier et deuxième cas correspondent soit à du bruit, soit à des pépites de connaissance. Selon nous, la probabilité que le premier cas soit dû au bruit est moins importante que pour le second cas. En effet, comme nous le verrons dans le chapitre IV, si A et B sont deux attributs ayant un faible support, il est peu probable que le bruit, entrainant l'apparition et la disparition de transactions pouvant contenir A et B, ne fasse apparaître que des transactions contenant $A \wedge B$ et fasse disparaître toutes les autres. En revanche, si le support de B (resp. A) est nettement supérieur à celui de A (resp. B), alors le bruit aura tendance à faire disparaître des transactions contenant A et ne contenant pas B et à faire apparaître des transactions contenant $A \wedge B$.

Le troisième cas correspond à des connaissances fortes du domaine mais qui sont probablement déjà connues des experts.

De nombreuses mesures permettent de différencier ces trois situations, la confiance centrée fait partie de celles-ci.

5.3 Confiance centrée

$$confiance_{centree}(A \rightarrow B) = P(B|A) - P(B) \tag{II.3}$$

La confiance centrée [Lallich et Teytaud 2004] permet de prendre en considération la taille de la conclusion de la règle $A \rightarrow B$.

Elle permet donc de relativiser la confiance d'une règle par rapport à la taille de sa conclusion.

Sur l'exemple de la Figure II.9, la confiance centrée permet d'ordonner les trois cas par « intérêt » croissant. L'intérêt du premier cas étant supérieur à celui du deuxième, lui même supérieur à celui du dernier.

De plus, l'introduction de $P(B)$ dans cette mesure permet de la rendre sensible à la taille des données. Il est aisé de voir que lorsque la taille des données augmente et si les marges n_A,n_B et n_{AB} restent constantes, la confiance centrée de la règle $A \rightarrow B$ augmente et tend vers la confiance de la règle.

Le support et la confiance sont deux mesures très simples à comprendre pour l'utilisateur par contre, la confiance centrée est plus difficile à appréhender. La principale difficulté de cette mesure est due au fait qu'elle combine deux quantités $P(B|A)$ et $P(B)$. La première, $P(B|A)$, est indépendante du nombre d'objets étudiés alors que la seconde, $P(B)$, est sensible au nombre d'objets. Il existe donc plusieurs configurations différentes pouvant conduire à des valeurs identiques de la confiance centrée. Le Tableau II.4 présente trois règles différentes ayant la même confiance centrée : $0,5$. La connaissance de cette seule valeur ne permet donc pas d'apprécier directement la nature d'une règle.

	A	$\overline{A}$	$\sum$
B	20	30	50
$\overline{B}$	0	50	50
$\sum$	20	80	100

	A	$\overline{A}$	$\sum$
B	8	22	30
$\overline{B}$	2	68	70
$\sum$	10	90	100

	A	$\overline{A}$	$\sum$
B	15	5	20
$\overline{B}$	10	70	80
$\sum$	25	75	100

TAB. II.4 – Trois règles $A \rightarrow B$ différentes ayant la confiance centrée.

5.4 Rappel

$$rappel(A \rightarrow B) = P(A|B) \tag{II.4}$$

Le rappel (appelé sensibilité par [Lavrac et al. 1999]) permet d'évaluer la proportion d'objets vérifiant la prémisse de la règle parmi ceux vérifiant la conclusion. Cette mesure possède le même défaut que la confiance, à savoir être indépendante de la taille de la prémisse. Pour résoudre ce problème, nous pouvons, comme pour la confiance, introduire le rappel centré.

5.5 Lift

$$lift(A \rightarrow B) = \frac{P(AB)}{P(A) \times P(B)} \tag{II.5}$$

Le lift [IBM 1996] représente le rapport à l'indépendance de la règle $A \rightarrow B$. Pour cette règle, l'indépendance entre les attributs A et B est égale à $P(A) \times P(B)$. L'indice brut d'association

réellement observé est lui égal à $P(AB)$. Le lift permet donc d'apprécier simplement, pour une règle $A \rightarrow B$, sa « distance » à l'indépendance.

Par exemple, une règle $A \rightarrow B$ ayant un lift égal à 2 indique que les individus ayant la propriété A ont deux fois plus de chances d'avoir la propriété B que les individus en général.

Cette mesure est symétrique et ne permet donc pas de distinguer les règles $A \rightarrow B$ et $B \rightarrow A$.

5.6 Pearl

$$pearl(A \rightarrow B) = P(A) \times |P(B|A) - P(B)| \tag{II.6}$$

La mesure de Pearl [Pearl 1988] permet d'évaluer l'intérêt d'une règle $A \rightarrow B$ par rapport à l'hypothèse d'indépendance entre la prémisse et la conclusion de la règle. L'utilisation de cette mesure permet de vérifier que la règle n'est pas purement le fruit du hasard et qu'elle apporte vraiment une connaissance nouvelle sur les données. En effet, si la mesure de Pearl de la règle est proche de 0, cela nous indique que considérer les attributs A et B liés par une relation est aussi intéressant que de les considérer indépendants.

La mesure de dépendance, introduite dans [Kodratoff 2000], est identique à la mesure de Pearl.

5.7 Corrélation

$$correlation(A \rightarrow B) = \frac{P(AB) - P(A)P(B)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}} \tag{II.7}$$

La corrélation [Lerman 1981] d'une règle d'association $A \rightarrow B$ est évaluée par l'écart à l'indépendance de l'indice associatif brut normalisé par le produit des marges intervenant dans l'étude de l'association liant A et B.

Cette mesure est symétrique et ne permet donc pas de différencier la règle $A \rightarrow B$ de la règle $B \rightarrow A$. Elle est sensible à la taille des données et lorsque les marges n_A, n_B et n_{AB} sont fixées alors la mesure tend vers $\alpha \times n \times n_{AB}$ lorsque la valeur de n augmente, avec α négligeable devant n. Cette mesure tend donc à trouver des relations entre tous les attributs étudiés dès que les données sont suffisament volumineuses.

5.8 Indice d'implication

L'indice d'implication [Lerman et al. 1981] permet d'évaluer la petitesse du nombre de contre-exemples à la règle $A \rightarrow B$ par rapport à la quantité attendue sous l'hypothèse d'indépendance.

Le principe de l'indice d'implication est de comparer la valeur observée du nombre d'exemples $P(AB)$ au nombre attendu pour deux variables indépendantes X et Y de mêmes cardinalités que A et B respectivement. La Figure II.10 illustre le principe de l'indice d'implication.

L'approche retenue par [Lerman et al. 1981] consiste à proposer une normalisation par rapport à un modèle probabiliste d'absence de liaison de la forme

$$(\mathcal{O}(A), \mathcal{O}(B), \mathcal{O}) \rightarrow (\mathcal{X}, \mathcal{Y}, \Omega) \tag{II.8}$$

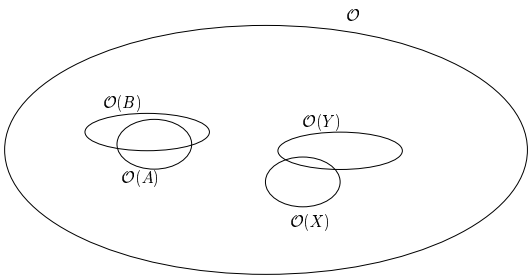


FIG. II.10 – Principe de l'indice d'implication.

où Ω est sinon $\mathcal{O}$, un ensemble aléatoire associé ; et, pour Ω fixé, $\mathcal{X}$ et $\mathcal{Y}$ sont deux parties aléatoires indépendantes de Ω respectivement associées à $\mathcal{O}(A)$ et $\mathcal{O}(B)$.

Le modèle noté $\mathcal{N}$, est construit de telle façon que Ω, $\mathcal{X}$ et $\mathcal{Y}$ respectent sinon exactement, du moins en espérance mathématique les cardinaux n, n_A et n_B.
La partie aléatoire $\mathcal{X}$ (resp. $\mathcal{Y}$) peut également être notée $\mathcal{O}(A^*)$ (resp. $\mathcal{O}(B^*)$) où A^* (resp. B^*) est l'attribut booléen aléatoire associé à A (resp. B). Dans ces conditions, en reprenant nos notations habituelles, à $s = n_{AB} = card[\mathcal{O}(A) \cap \mathcal{O}(B)]$, nous associons sous le modèle (II.8), une variable aléatoire

$$\mathcal{S} = n(A^* \wedge B^*) = card(\mathcal{X} \cap \mathcal{Y}) \tag{II.9}$$

où (A^*, B^*) est le couple d'attributs booléens aléatoires indépendants correspondant à (A, B). $\mathcal{S}$ est l'indice brut aléatoire.
La première forme de normalisation consiste à centrer et à réduire s au moyen de l'espérance mathématique et de l'écart type de $\mathcal{S}$. On obtient dans ces conditions le coefficient

$$q(A, B) = \frac{s - \mathcal{E}(\mathcal{S})}{\sqrt{var(\mathcal{S})}} \tag{II.10}$$

où $\mathcal{E}(\mathcal{S})$ et $var(\mathcal{S})$ désignent l'espérance mathématique et la variance de $\mathcal{S}$.
L'indice probabiliste « local » de la vraisemblance du lien définit la deuxième forme de normalisation. Il s'écrit

$$\mathcal{I}(A, B) = P\{\mathcal{S} \leq s|\mathcal{N}\} = Pr\{q(A^*, B^*) \leq q(A, B)|\mathcal{N}\} \tag{II.11}$$

Dans un tel indice, le degré d'association entre A et B est évalué à partir du degré d'invraisemblance de la grandeur de s, eu égard à l'hypothèse d'absence de liaison $\mathcal{N}$. Bien que l'indice (II.11) correspond au complément à l'unité d'un seuil critique au sens des tests d'hypothèses d'indépendance, il ne s'agit nullement ici d'un test conditionnel [Bernard et Charron 1996] mais d'une évaluation probabiliste conditionnée par les tailles $n(A)$ et $n(B)$.
[Lerman et al. 1981] ont mis en évidence trois formes fondamentales de l'hypothèse d'absence de liaison $\mathcal{N}$ que nous notons $\mathcal{N}_1$, $\mathcal{N}_2$ et $\mathcal{N}_3$. Elles se distinguent dans la manière d'associer à un sous-ensemble $\mathcal{O}(c)$ de $\mathcal{O}$, une partie aléatoire $\mathcal{L}$ d'un ensemble Ω. Désignons ici par $\mathcal{P}(\Omega)$ l'ensemble des parties de Ω organisé en niveaux à partir de la relation d'inclusion.
Pour $\mathcal{N}_1$, $\Omega = \mathcal{O}$ et $\mathcal{L}$ est un élément aléatoire pris uniformément au hasard sur le niveau $n(c)$ de

$\mathcal{P}(\Omega)$.
Pour $\mathcal{N}_2$, $\Omega = \mathcal{O}$; mais le modèle du choix $\mathcal{L}$ est à deux pas. Le premier consiste en le choix aléatoire d'un niveau et le second consiste en le choix aléatoire et uniformément réparti d'un élément de ce niveau. Précisons que le choix du niveau $k, 0 \leq k \leq n$, s'effectue avec la probabilité binomiale

$$C_n^k P(c)^k P(\overline{c})^{n-k}$$

où $P(c) = \frac{n(c)}{n}$ et $P(\overline{c}) = \frac{n(\overline{c})}{n}$.
$\mathcal{N}_3$ est un modèle aléatoire à trois pas. Le premier consiste à associer à $\mathcal{O}$ un ensemble aléatoire Ω d'objets - pour fixer les idées, de même nature que ceux observés - la seule caractéristique aléatoire de Ω est sa cardinalité $\mathcal{N}$ qui est supposée suivre une loi de Poisson de paramètre $n = card(\mathcal{O})$. Les deux autres pas sont analogues à ceux du modèle $\mathcal{N}_2$. Pour $\mathcal{N} = m$ fixé et Ω_0 un ensemble de taille m, $\mathcal{L}$ est une partie aléatoire de Ω_0. $\mathcal{L}$ n'est définie que pour $m \geq n(c)$ et dans ce cas on pose γ le rapport $\frac{n(c)}{m}$. Dans ces conditions, le choix du niveau k de $\mathcal{P}(\Omega_0)$ se fait avec la probabilité binomiale $C_m^k \gamma^k (1-\gamma)^{n-k}$. Pour un niveau choisi, le choix de $\mathcal{L}$ se fait alors uniformément au hasard sur ce niveau.
[Lerman et al. 1981] ont démontré que la distribution de l'indice brut aléatoire $\mathcal{S}$ [*cf.* (II.9)] est :

- hypergéométrique de paramètres $(n, n(A), n(B))$ sous l'hypothèse $\mathcal{N}_1$;
- binomiale de paramètres $(n, P(A) \times P(B))$ sous l'hypothèse $\mathcal{N}_2$;
- de Poisson de paramètres $(n, n \times P(A) \times P(B))$ sous l'hypothèse $\mathcal{N}_3$.

Les trois indices suivants sont alors obtenus

$$q_1(A,B) = \sqrt{n} \times \frac{P(A \wedge B) - P(A) \times P(B)}{\sqrt{P(A) \times P(B) \times P(\overline{A}) \times P(\overline{B})}} \tag{II.12}$$

$$q_2(A,B) = \sqrt{n} \times \frac{P(A \wedge B) - P(A) \times P(B)}{\sqrt{P(A) \times P(B) \times [1 - P(A) \times P(B)]}} \tag{II.13}$$

et

$$q_3(A,B) = \sqrt{n} \times \frac{P(A \wedge B) - P(A) \times P(B)}{\sqrt{P(A) \times P(B)}} \tag{II.14}$$

L'indice $q_1(A,B)$ est parfaitement symétrique dans ce sens où $q_1(A,B) = q_1(\overline{A}, \overline{B})$
On suppose que les attributs booléens (sous leur forme positive) ont été établis de telle façon que la proportion relative d'objets où un même attribut est à $Vrai$ est inférieure à $0,5$ (voir l'introduction). Dans ces conditions, on peut démontrer les inégalités suivantes :

$$q_2(A,B) > q_2(\overline{A}, \overline{B}) \tag{II.15}$$

et

$$q_3(A,B) > q_3(\overline{A}, \overline{B}) \tag{II.16}$$

La dernière inégalité étant plus prononcée que celle qui précède, nous nous limiterons à considérer les deux indices les plus différenciés que sont $q_1(A,B)$ et $q_3(A,B)$.
Nous ne retiendrons que l'indice q_3 pour la suite de cette thèse.

5.9 Intensité d'implication

[Gras 1979] a proposé d'évaluer la petitesse du nombre de contre-exemples à la règle $A \rightarrow B$ par rapport à la quantité attendue sous l'hypothèse d'indépendance.

Cet indice fait intervenir explicitement le nombre de contre-exemples dans son expression.

Le principe de l'intensité d'implication est donc de comparer la valeur observée du nombre de contre exemples $P(A\overline{B})$ au nombre attendu pour deux variables indépendantes X et Y de mêmes cardinalités que A et B respectivement.

L'intensité d'implication est construite à partir des travaux concernant l'indice d'implication. L'indice obtenu s'exprime simplement de la manière suivante :

$$\varphi(A, B) = 1 - \Phi(q_3(A, \overline{B})) \tag{II.17}$$

ou Φ est la fonction de répartition de la loi normale centrée et réduite.

Cet indice tend rapidement vers 0 ou 1 dès que n devient assez grand (voir [Guillaume 2000, Lerman et Azé 2003]). L'indice ainsi défini ne permet donc de distinguer que deux familles de règles : celles qui sont intéressantes et celles qui ne le sont pas. Lorsque n est suffisamment grand, les règles d'une de ces familles sont toutes semblables par rapport à l'intensité d'implication.

5.10 Intensité d'implication entropique

Pour pallier un des défauts de l'intensité d'implication, [Gras et al. 2001] ont proposé une nouvelle mesure combinant l'intensité d'implication et un coefficient entropique lié à la règle étudiée. Le défaut majeur de l'intensité d'implication, à savoir tendre vers 0 ou vers 1 dès que le volume de données étudié devient trop important, est corrigé par le coefficient entropique introduit dans la mesure. Ainsi, lorsque l'intensité d'implication « classique » ne peut plus différencier deux règles R_1 et R_2, le coefficient entropique permet, si les règles sont différentiables, d'observer cette différence.

Dans les cas limites, c'est-à-dire où l'intensité d'implication est égale à 0 ou à 1, l'intensité d'implication entropique peut donc simplement se ramener au calcul du coefficient entropique.

L'intensité d'implication entropique est notée $\Psi(A, B)$ et définie par

$$\Psi(A, B) = \sqrt{\varphi(A, B) \times \tau(A, B)} \tag{II.18}$$

où $\varphi(A, B)$ représente l'intensité d'implication de la règle $A \rightarrow B$ et $\tau(A, B)$ est le coefficient entropique permettant de pondérer l'intensité d'implication.

Avant de présenter précisément le coefficient $\tau(A, B)$, nous devons introduire quelques notations liées à la notion d'information mutuelle. Notons d_{AB} la quantité

$$d_{AB} = \frac{p(A \wedge B)}{p(A) \times p(B)}$$

L'information mutuelle peut prendre l'une des trois formes suivantes :

$$\begin{aligned}
\mathcal{E} &= P(A \wedge B)\log_2(d_{AB}) + P(A \wedge \overline{B})\log_2(d_{A\overline{B}}) \\
&+ P(\overline{A} \wedge B)\log_2(d_{\overline{A}B}) + P(\overline{A} \wedge \overline{B})\log_2(d_{\overline{A}\overline{B}}) && \text{(II.19)} \\
&= E(A) - P(B)E(A|B) - P(\overline{B})E(A|\overline{B}) && \text{(II.20)} \\
&= E(B) - P(A)E(B|A) - P(\overline{A})E(B|\overline{A}) && \text{(II.21)} \\
& && \text{(II.22)}
\end{aligned}$$

Ayant défini ces quelques notations, le coefficient entropique introduit par [Gras et al. 2001] s'exprime de la manière suivante :

$$\tau(A,B) = \sqrt{G(B|A) \times G(\overline{A}|\overline{B})} \tag{II.23}$$

où $G(x|y)$ est la racine carrée positive de

$$G^2(x|y) = \begin{cases} 1 - E^2(x|y) \text{ si } P(\overline{x} \wedge y) \leq \frac{1}{2} \times P(y) \\ 0 \text{ sinon} \end{cases} \tag{II.24}$$

où $E(x)$ représente l'entropie de la distribution $(P(x), P(\overline{x}))$ et où $E(x|y)$ représente celle de la distribution conditionnelle $(P(x|y), P(\overline{x}|y))$; x et y étant deux attributs booléens.

Cet indice utilise les entropies conditionnelles $E(B|A)$ et $E(\overline{A}|\overline{B})$ qui sont, respectivement, celles des distributions à deux valeurs $(P(B|A), P(\overline{B}|A))$ et $(P(\overline{A}|\overline{B}), P(A|\overline{B}))$. La première entropie peut être récoltée comme un composant constitutif de la dernière expression (II.21) de l'information mutuelle $\mathcal{E}$ alors que la deuxième entropie est un élément composant de la précédente expression (II.20) de $\mathcal{E}$.

On notera que l'importance de l'indice $1 - E^2(B|A)$ traduit tout autant l'inclusion $\mathcal{O}(A) \subset \mathcal{O}(B)$ que celle, logiquement contraire, $\mathcal{O}(A) \subset \mathcal{O}(\overline{B})$ $(E(B|A) = E(\overline{B}|A))$. D'autre part, l'indice $1 - E^2(\overline{A}|\overline{B})$, traduit tout autant l'inclusion $\mathcal{O}(\overline{B}) \subset \mathcal{O}(\overline{A})$ que celle $\mathcal{O}(\overline{B}) \subset \mathcal{O}(A)$ $(E(\overline{A}|\overline{B}) = E(A|\overline{B}))$.

Cependant, compte tenu de la condition apparaissant dans l'équation (II.24), une valeur positive et non nulle de $\tau(A,B)$ est conditionnée par $P(B|A) > P(\overline{B}|A)$. Ainsi, l'indice d'inclusion ne prend une valeur positive et non nulle que si chacun des deux supports $P(B|A)$ et $P(\overline{A}|\overline{B})$ est supérieur à $0,5$. Or, il y a des situations, peut être non fréquentes, où $P(B|A)$ est suffisamment élevé (très supérieur à $0,5$) ; alors que $P(\overline{A}|\overline{B})$ est suffisamment bas (très inférieur à $0,5$). Il est difficile dans ce cas de rejeter toute valeur à l'implication $A \rightarrow B$. De sorte que l'indice d'inclusion possède la faiblesse de sa qualité ; à savoir tenir compte de l'implication $A \rightarrow B$ et de sa contraposée $\overline{B} \rightarrow \overline{A}$.

Ces situations sont telles que $P(B)$ est très proche de 1. Par exemple, considérons la situation présentée dans le Tableau II.5.

	A	$\overline{A}$	$\sum$
B	320	590	910
$\overline{B}$	80	10	90
$\sum$	400	600	1000

TAB. II.5 – Situation où le coefficient entropique est égal à 0.

Nous avons $P(B|A) = \frac{P(AB)}{P(A)} = \frac{320}{400} > 0,5$ et $P(\overline{A}|\overline{B}) = \frac{P(\overline{AB})}{P(\overline{B})} = \frac{10}{90} < 0,5$. Dans cette situation, $\tau(A,B) = 0$ et donc $\Psi(A,B) = 0$ alors que manifestement, la règle $A \rightarrow B$ ne peut être rejetée (support et confiance élevés).

5.11 Piatetsky-Shapiro

$$PS(A \rightarrow B) = n \times P(A) \times (P(B|A) - P(B)) \tag{II.25}$$

L'indice de Piatetsky-Shapiro [Piatetsky-Shapiro 1991] évalue l'intérêt d'une règle par rapport à son écart à l'indépendance. Cet indice est très proche de l'indice de Pearl puisqu'il s'exprime de la manière suivante : $PS(A \rightarrow B) = n \times Pearl(A \rightarrow B)$. La présence de la quantité n dans l'expression de cet indice rend celui ci moins sensible à la taille des données, contrairement à l'indice de Pearl. En effet, l'indice de Piatetsky-Shapiro peut se réécrire sous la forme suivante : $n_{AB} - \frac{n_A n_B}{n}$. Sous cette forme, il apparait plus clairement que l'indice est moins sensible à la taille des données que l'indice de Pearl. Ainsi, lorsque n augmente, la règle $A \rightarrow B$ tend à être évaluée par son indice d'implication brut.

De la même manière que l'indice de Pearl, l'indice de Piatetsky-Shapiro est symétrique.

5.12 Lœvinger

$$loevinger(A \rightarrow B) = \frac{P(B|A) - P(B)}{P(\overline{B})} \tag{II.26}$$

L'indice de Lœvinger [Loevinger 1947] est, avec le support et la confiance, l'une des plus anciennes mesures de qualité répertoriées dans le domaine de la fouille de données. Cet indice normalise la confiance centrée de la règle par rapport aux objets ne vérifiant pas la conclusion. La normalisation par rapport à la probabilité d'observer un objet ne vérifiant pas la conclusion permet de palier un des défauts de la confiance.

5.13 Moindre contradiction

$$contramin(A \rightarrow B) = \frac{P(AB) - P(A\overline{B})}{P(B)} \tag{II.27}$$

La moindre contradiction [Azé et Kodratoff 2002, Azé 2003] mesure la différence entre le nombre d'exemples et le nombre de contre-exemples de la règle, cette différence est normalisée par le nombre d'objets vérifiant la conclusion de $A \rightarrow B$.

Cette mesure permet donc de sélectionner les règles ayant plus d'exemples que de contre-exemples. De plus, si nous considérons deux règles $R_1 = A_1 \rightarrow B_1$ et $R_2 = A_2 \rightarrow B_2$, telles que $P(A_1B_1) = P(A_2B_2)$ et $P(B_1) < P(B_2)$ alors nous avons $contramin(R_1) > contramin(R_2)$.

Étant donné que nous sommes à l'origine de cette mesure, nous pouvons fournir quelques détails sur la création de celle-ci. Nous sommes partis du constat général présenté dans la section 3 qui est que le support et la confiance ne suffisent pas pour déterminer l'intérêt d'une règle. À partir de ce constat et sachant que nous sommes intéressés par la détection des règles d'association pouvant avoir de très faibles supports, nous avons étudié les propriétés de ce type de règles.

Nous reprenons l'étude présentée pour la Confiance (voir section 5.2, page 28) et nous l'adaptons aux pépites de connaissance.

Nous avons vu que l'utilisation de la mesure dite de Dépendance, aussi nommée Pearl, permet de différencier des règles qui du point de vue de la confiance sont identiques.

Les trois cas présentés sur la Figure II.11 présentent un intérêt particulier pour notre étude. Cette Figure est quasiment identique à la Figure II.9 mais, dans un effort de lisibilité, nous préférons la reproduire ici.

Dans chacun de ces cas, nous avons $P(B_i|A_j) = 1$ où $i = 1, 2$ ou 3 et $j = 1$ ou 2.

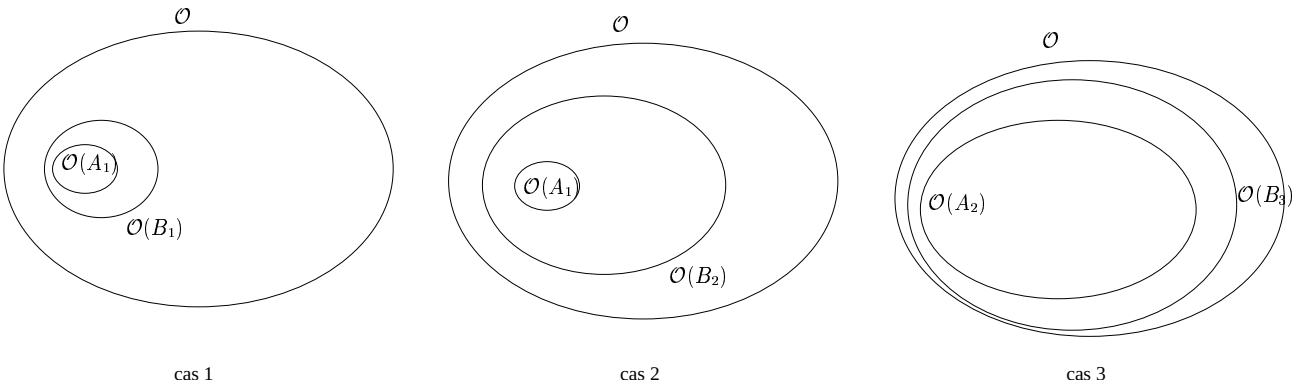


FIG. II.11 – Trois cas illustrant différents comportements de la dépendance.

cas 1
$P(B_1)$ est faible, ainsi ce cas correspond au cas d'une dépendance élevée, combinée avec un très faible support. Ce cas met en évidence des relations présentant un intérêt très élevé car les ensembles considérés sont très petits (ils peuvent donc être inconnus de l'expert), et les relations considérées ne sont jamais infirmées. Ce type de règles d'association peut être assimilé à des pépites de connaissance et seul l'avis de l'expert pourra nous indiquer si ces règles sont dues au bruit ou non.

cas 2
$\mathcal{O}(A_1)$ est « noyé » dans $\mathcal{O}(B_2)$ et, il est relativement intuitif de considérer que B_2 dépend moins de A_1, alors que B_1 dépend fortement de A_1, dans le cas 1. Dans ce cas, la dépendance $1 - P(B_2)$, est faible, comme nous nous y attendions.

Le comportement de la dépendance sur ces deux premiers cas est conforme à nos attentes. Pour ces deux cas, l'utilisation de la dépendance, comme mesure d'intérêt, peut sembler acceptable.

cas 3
L'étude du cas 3, pour lequel $P(B_3)$ et $P(A_2)$ sont toutes les deux très élevées, montre que la dépendance est faible. La règle d'assocation $A_2 \rightarrow B_3$ correspond au cas d'une relation relativement triviale et sans réel intérêt, en termes de connaissances nouvelles. Cependant, nous allons voir que la définition de la dépendance est relativement contre-intuitive dans ce cas. Nous pouvons noter une contradiction entre l'intuition classique de la dépendance et ce que nous appelerons la « dépendance discrète » (c'est-à-dire où les attributs sont de natures booléennes). Rappelons que l'intuition classique, utilisée par de nombreux auteurs, repose sur la théorie des probabilités qui précise que la probabilité conjointe de deux évenements indépendants, $P(X \wedge Y)$, est égale à $P(X) \times P(Y)$. Ainsi, selon l'intuition classique, la dépendance entre deux évenements, X et Y, est élevée lorsque la différence entre $P(X \wedge Y)$ et $P(X) \times P(Y)$ est importante.

Dans le cas discret que nous étudions ici, A_2 et B_3 sont manifestement très dépendants car $A_2 = Vrai$ nous permet de prédire $B_3 = Vrai$ avec une confiance donnée par $\frac{P(A_2 \wedge B_3)}{P(A_2)} = 1$. Cependant, nous avons bien $P(A_2) \times P(B_3) \approx 1 \approx P(A_2 \wedge B_3)$.

Notre objectif n'est pas d'étudier les détails de ce paradoxe, mais nous voulons le mettre en évidence de manière à montrer que la moindre contradiction, présente, comme la dépendance, des propriétés paradoxales, lorsque le cas 3 est pris en considération.

Cette nouvelle mesure est construite de manière à favoriser les cas où A est pratiquement to-

talement inclus dans B, c'est-à-dire où $[P(A \wedge B) - P(A \wedge \overline{B})]$ est « étonnamment » élevé (des cas où l'implication $A \rightarrow B$ est rarement infirmée). Pour prendre en considération le cas 2 présenté ci-dessus, nous proposons de normaliser notre mesure par rapport à $P(B)$.

Ainsi, la moindre contradiction a été proposée pour permettre de différencier des règles telles que $A_1 \rightarrow B_1$ et $A_1 \rightarrow B_2$. Et, comme nous l'avons indiqué dans la définition de cette mesure, nous avons bien : $contramin(A_1 \rightarrow B_1) > contramin(A_1 \rightarrow B_2)$

5.14 Mesure de Sebag-Schœnauer

$$Seb(A \rightarrow B) = \frac{P(AB)}{P(A\overline{B})} \tag{II.28}$$

La mesure de Sebag-Schœnauer [Sebag et Schoenauer 1988] prend aussi de manière explicite le nombre de contre-exemples à la règle $A \rightarrow B$. En effet, cette mesure calcule simplement le rapport entre le nombre d'exemples de la règle et son nombre de contre-exemples. Dès que la mesure est supérieure à 1 alors la règle possède plus d'exemples que de contre-exemples et inversement, dès que la mesure est inférieure à 1 alors la règle est plus souvent infirmée par les données plutôt que confirmée.

Notons que cette mesure prend ses valeurs dans l'intervalle $[0, +\infty[$.

5.15 Conviction

$$conviction(A \rightarrow B) = \frac{P(A)P(\overline{B})}{P(A\overline{B})} \tag{II.29}$$

La conviction [Brin et al. 1997a] correspond, comme nous l'avons déjà vu, à une réexpression logique de l'écart à l'indépendance de la règle étudiée. Cette mesure peut aussi s'interpréter en fonction des contre-exemples de $A \rightarrow B$. En effet, une conviction élevée indique que le nombre de contre-exemples vérifiant la règle est inférieur à celui attendu sous l'hypothèse d'indépendance entre les attributs de la règle. Donc lorsque le nombre de contre-exemples augmente, la conviction diminue pour atteindre son minimum lorsque la règle ne possède aucun exemple. La conviction représentant l'écart à l'indépendance de la règle $A \rightarrow B$ est interprétable de la manière suivante : une règle ayant une conviction égale à 2 indique qu'il est deux fois plus convaincant de considérer les attributs A et B reliés par la règle $A \rightarrow B$ que de les considérer indépendants. La conviction est très proche du lift et correspond simplement à l'étude logique de la règle $A \rightarrow B$ qui se réécrit $\overline{A \wedge \overline{B}}$. De manière similaire au lift, la conviction mesure l'écart à l'indépendance de la règle. Pour ce faire, $P(A\overline{B})$ représentant l'indice brut d'association de A et $\overline{B}$, est divisé par la valeur représentant l'indépendance des attributs A et $\overline{B}$: $P(A)P(\overline{B})$. Pour prendre en considération la négation englobant l'expression logique de la règle $A \rightarrow B$, la mesure est inversée, d'où l'expression de la conviction.

5.16 Satisfaction

$$satisfaction(A \rightarrow B) = \frac{P(\overline{B}) - P(\overline{B}|A)}{P(\overline{B})} \tag{II.30}$$

La satisfaction [Lavrac et al. 1999] étudie la $A \rightarrow \overline{B}$. Cette mesure peut se réécrire simplement en $1 - lift(A \rightarrow \overline{B})$. Ainsi, si le lift de la règle $A \rightarrow \overline{B}$ est élevé alors la satisfaction de $A \rightarrow B$ sera faible. La satisfaction permet donc d'apprécier si la règle $A \rightarrow B$ est plus intéressante que la règle $A \rightarrow \overline{B}$. Lorsque le nombre de contre-exemples de la règle $A \rightarrow B$ augmente alors le nombre de contre-exemples de la règle $A \rightarrow \overline{B}$ augmente et la satisfaction de $A \rightarrow B$ diminue.

5.17 J-mesure

$$J - mesure(A \rightarrow B) = P(AB) log \frac{P(AB)}{P(A)P(B)} + P(A\overline{B}) log \frac{P(A\overline{B})}{P(A)P(\overline{B})} \quad \text{(II.31)}$$

La présence du nombre de contre-exemples dans l'expression de la J-mesure [Goodman et Smyth 1988] permet d'estimer la quantité d'information apportée par l'étude de la règle $A \rightarrow B$. Rappelons que la J-mesure évalue de manière identique les règles $A \rightarrow B$ et $A \rightarrow \overline{B}$.

5.18 Spécificité

$$specificite(A \rightarrow B) = P(\overline{A}|\overline{B}) \quad \text{(II.32)}$$

La spécificité [Lavrac et al. 1999] représente la confiance de la règle $\overline{B} \rightarrow \overline{A}$. Cette règle possède les mêmes contre-exemples que la règle $A \rightarrow B$. Ainsi l'étude de la règle $\overline{B} \rightarrow \overline{A}$ permet d'apprécier l'intérêt de la règle $A \rightarrow B$. En effet, une règle $A \rightarrow B$ ayant une spécificité élevée indique que la probabilité d'observer l'attribut $\overline{A}$ est élevée sachant que l'on a observé l'attribut $\overline{B}$. Donc, le nombre d'objets vérifiant $\overline{A}$ et $\overline{B}$ est faible ce qui confirme l'intérêt pour la règle $A \rightarrow B$ qui possède alors elle aussi peu de contre-exemples.

5.19 Fiabilité négative

La fiabilité négative introduite dans [Lavrac et al. 1999] permet de mesurer la capacité d'une règle à sélectionner les individus ne vérifiant vraiment pas la règle. Cette mesure est similaire à la confiance qui permet d'évaluer la capacité d'une règle à sélectionner les individus vérifiant vraiment la règle.

La définition de cette mesure est la suivante

$$fiabilite - negative(A \rightarrow B) = P(\overline{B}|\overline{A}) \quad \text{(II.33)}$$

Ainsi, si nous considérons que, étant donnés A et B, les individus sont répartis en deux groupes : les individus « positifs » et les individus « négatifs ». Considérons que les individus positifs (resp. négatifs) sont ceux qui possèdent l'attribut B à $Vrai$ (resp. $Faux$). Parmi les positifs (resp. négatifs), certains individus sont de vrais positifs (resp. négatifs) et possèdent l'attribut A à $Vrai$ (resp. $Faux$).

Étant données ces définitions proches de celles utilisées en apprentissage supervisé, la fiabilité négative permet donc d'évaluer la capacité d'une règle à détecter les vrais négatifs parmi les individus négatifs.

Il est évident que cette mesure possède les mêmes qualités et défauts que ceux déjà mis en évidence pour la confiance.

mesure	expression	référence
Support	$P(AB)$	[Agrawal et al. 1993]
Confiance	$P(B\|A)$	[Agrawal et al. 1993]
Rappel	$P(A\|B)$	[Lavrac et al. 1999]
confiance centrée	$P(B\|A) - P(B)$	[Lallich et Teytaud 2004]
Lift	$\frac{P(AB)}{P(A)P(B)}$	[IBM 1996]
Moindre contradiction	$\frac{P(AB)-P(A\overline{B})}{P(B)}$	[Azé et Kodratoff 2002]
Piatetsky-Shapiro	$nP(A)(P(B\|A) - P(B))$	[Piatetsky-Shapiro 1991]
Lœvinger	$\frac{P(B\|A)-P(B)}{P(\overline{B})}$	[Loevinger 1947]
Corrélation	$\frac{P(AB)-P(A)P(B)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}}$	[Lerman 1981]
Indice d'implication	$q_3(A,\overline{B}) = \sqrt{n}\frac{P(A\overline{B})-P(A)P(\overline{B})}{\sqrt{P(A)P(\overline{B})}}$	[Lerman et al. 1981]
Pearl	$P(A) * \|P(B\|A) - P(B)\|$	[Pearl 1988]
Nouveauté	$P(AB) - P(A)P(B)$	[Lavrac et al. 1999]
Satisfaction	$\frac{P(\overline{B})-P(\overline{B}\|A)}{P(\overline{B})}$	[Lavrac et al. 1999]
Spécificité	$P(\overline{A}\|\overline{B})$	[Lavrac et al. 1999]
Fiabilité négative	$P(\overline{B}\|\overline{A})$	[Lavrac et al. 1999]
J-mesure	$P(AB)log\frac{P(AB)}{P(A)P(B)} + P(A\overline{B})log\frac{P(A\overline{B})}{P(A)P(\overline{B})}$	[Goodman et Smyth 1988]
Sebag-Schœnauer	$\frac{P(AB)}{P(A\overline{B})}$	[Sebag et Schoenauer 1988]
Conviction	$\frac{P(A)P(\overline{B})}{P(A\overline{B})}$	[Brin et al. 1997a]
Intensité d'implication	$\varphi(A,B) = 1 - \Phi(q_3(A,\overline{B}))$	[Gras 1979]
Intensité d'implication entropique	$\Psi(A,B) = \sqrt{\varphi(A,B) \times \tau(A,B)}$	[Gras et al. 2001]

TAB. II.6 – Mesures de qualité

5.20 Étude de situations caractéristiques pour $A \rightarrow B$

Nous nous intéressons au comportement des mesures de qualité présentées dans le Tableau II.6 dans chacun des cas suivants :

a. La règle ne possède aucun exemple. A et B sont totalement incompatibles, $P(AB) = 0$.

b. A et B sont indépendants, $P(AB) = P(A)P(B)$.

c. $A \rightarrow B$ est une règle logique, $P(AB) = P(A)$ et $A \subset B$.

La Figure II.12 illustre les cas a, b et c pour une règle $A \rightarrow B$.

Ces trois situations sont intéressantes car elles correspondent à des cas simples pour lesquels, l'intérêt d'une règle est facilement évaluable. En effet, dans le cas a, la règle $A \rightarrow B$ n'a aucun lieu d'être et son intérêt est donc nul. L'évaluation d'une telle règle par une mesure de qualité m doit, soit conduire à une valeur nulle pour la mesure, soit fournir la valeur minimale associée à la mesure. L'analyse du Tableau II.7 montre que, pour la majorité des mesures, l'évaluation de l'intérêt d'une telle règle conduit bien à la valeur minimale (ou nulle) associée à la mesure. En revanche, et ce de manière contre-intuitive pour l'expert, la J-mesure, la spécificité et la conviction n'associent pas leur valeur minimale à cette situation.

Le second cas, correspondant à une situation d'indépendance entre les attributs A et B, est caractérisé pour la plupart des mesures, par une évaluation égale à la valeur nulle.

Enfin, le troisième cas correspond aux fameuses pépites de connaissance que nous recherchons. Il est intéressant de constater que de nombreuses mesures classiquement utilisées dans la recherche des règles d'association évaluent toutes ce type de règle de la même manière. Ainsi, la confiance, la mesure de Lœvinger, de Sebag-Schœnauer, la spécifité la conviction évaluent les règles logiques à 1.

Et ce faisant, elles sont incapables de distinguer les cas déjà présentés dans la figure II.9. Nous noterons, que parmi les mesures non symétriques et donc capable de différencier les règles $A \rightarrow B$ des règles $B \rightarrow A$, seules, le rappel, la J-mesure, la moindre contradiction, la mesure de Pearl et la satisfaction gardent un pouvoir discriminant pour ce type de règles.

Le Tableau II.7 présente les valeurs prises par les mesures de qualité dans chacun de ces cas.

L'intensité d'implication classique est une mesure que nous allons étudier dans la suite de cette thèse et particulièrement dans le chapitre III où nous proposons une « amélioration » de l'intensité d'implication classique pour corriger un de ces défauts. Nous essayerons donc, dans la mesure du possible, d'étudier les cas limites et intéressants pour cette mesure. Ainsi, dans le cas où $P(AB) = 0$, il n'est pas évident que l'intensité d'implication soit égale à 0. En effet, en utilisant les tables de la loi normale, dès que l'indice d'implication $q_3(A, \overline{B})$ devient supérieur ou égal à la valeur 3, l'intensité d'implication associée est considérée comme nulle. Dans le cas d'incompatibilité entre A et B associé à $P(AB) = 0$, il suffit donc que

$$P(B)\sqrt{\frac{nP(A)}{P(\overline{B})}} \geq 3 \tag{II.34}$$

pour pouvoir conclure que l'intensité d'implication est nulle. Or, il est aisé de montrer qu'il existe des configurations de n, $P(A)$ et $P(B)$ pour lesquelles l'intensité d'implication n'est pas nulle.

Si nous considérons que les observations $P(A)$ et $P(B)$ sont fixées, il suffit de trouver la valeur de n garantissant l'inégalité II.34, c'est-à-dire trouver n vérifiant l'inégalité II.35.

$$n \geq 9\frac{1 - P(B)}{P(A)P(B)^2} \tag{II.35}$$

Le Tableau II.8 présente pour quelques valeurs de $P(A)$ et $P(B)$ la valeur minimale de n permettant de conclure que l'intensité d'implication d'une règle $A \rightarrow B$ telle que A et B sont incompatibles, est nulle.

mesure	**incompatibilité**	**indépendance**	**règle logique**
support	0	$P(A)P(B)$	$P(A)$
confiance	0	$P(B)$	1
rappel	0	$P(A)$	$\frac{1}{P(B)}$
confiance centrée	$-P(B)$	0	$P(\overline{B})$
Lift	0	1	$\frac{1}{P(B)}$
Moindre contradiction	$\frac{-P(A)}{P(B)}$	$\frac{P(A)(2P(B)-1)}{P(B)}$	$\frac{P(A)}{P(B)}$
Piatetsky-Shapiro	$-nP(A)P(B)$	0	$nP(A)P(\overline{B})$
Lœvinger	$-\frac{P(B)}{P(\overline{B})}$	0	1
Corrélation	$-\frac{\sqrt{P(A)P(B)}}{\sqrt{P(\overline{A})P(\overline{B})}}$	0	$\frac{\sqrt{P(A)P(\overline{B})}}{\sqrt{P(\overline{A})P(B)}}$
Indice d'implication	$P(B)\sqrt{\frac{nP(A)}{P(\overline{B})}}$	0	$-\sqrt{nP(A)P(\overline{B})}$
Pearl	$P(A)P(B)$	0	$P(A)P(\overline{B})$
Novelty	$-P(A)P(B)$	0	$P(A)P(\overline{B})$
Satisfaction	$-\frac{P(A)}{P(\overline{B})}$	0	$\frac{P(A)}{P(\overline{B})}$
Spécificité	$\frac{P(\overline{A})-P(B)}{P(\overline{B})}$	$P(\overline{A})$	1
J-mesure	$P(A)log\left(\frac{1}{P(B)}\right)$	0	$P(A)log\left(\frac{1}{P(B)}\right)$
Sebag-Schœnauer	0	$\frac{P(B)}{P(\overline{B})}$	∞
Conviction	$\frac{P(\overline{B})-1}{P(\overline{B})}$	0	1
Intensité d'implication	variable (valeur minimale : 0)	$\approx 0,5$	variable (valeur maximale : 1)

TAB. II.7 – Comportements des mesures de qualité dans les situations remarquables

$P(A)$	$0,2$	$0,2$	$0,1$	$0,05$	$0,03$
$P(B)$	$0,2$	$0,15$	$0,1$	$0,05$	$0,03$
n	900	1700	8100	68400	323334

TAB. II.8 – Valeurs minimales de n nécessaires pour annuler l'intensité d'implication de $A \rightarrow B$ en cas d'incompatibilité entre A et B

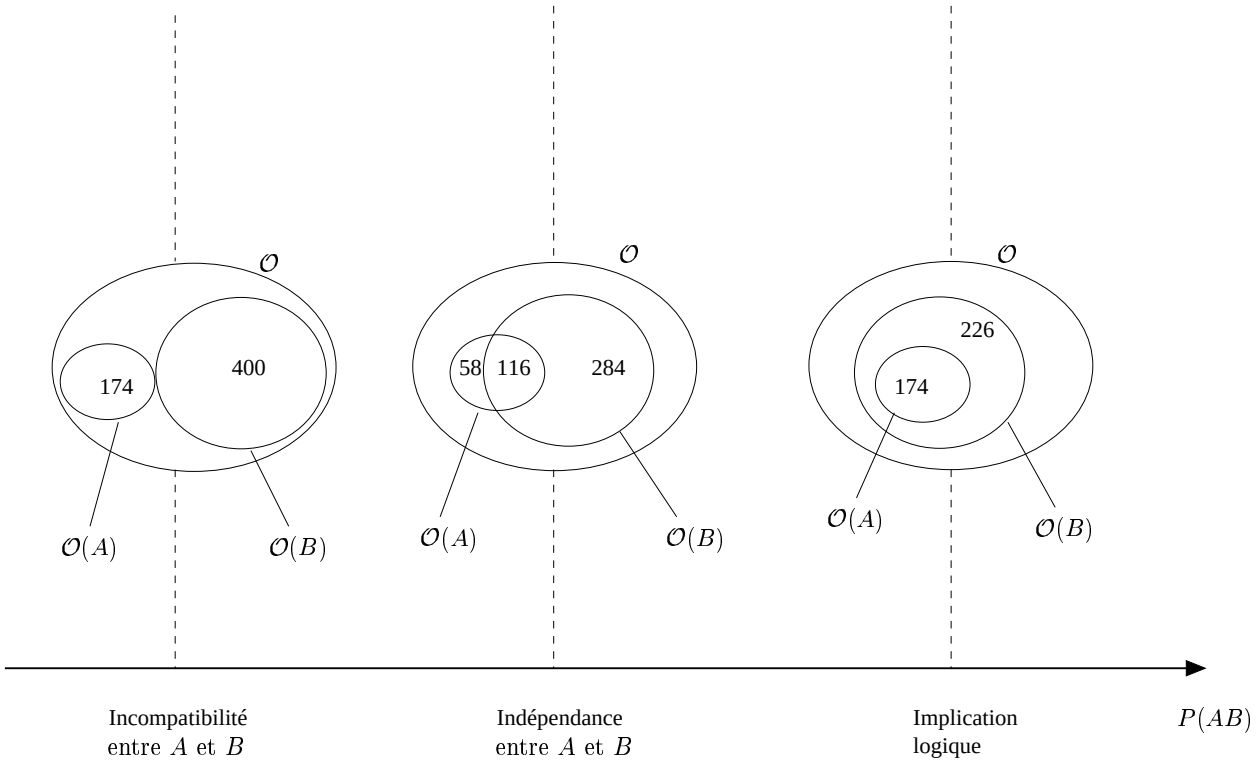


FIG. II.12 – Illustration des cas a., b. et c.

Bien évidemment, l'utilisation d'autres mesures de qualité permet de lever ce problème. De plus, ces cas ne correspondent pas aux cas intéressant pour les pépites de connaissance car leur confiance est nulle. Il est cependant intéressant de noter le comportement contre intuitif de la mesure dans cette situation.

Nous pouvons constater que lorsque les items A et B sont indépendants, la plupart des mesures prennent des valeurs constantes et donc indépendantes de la taille de A et B. Les quelques exceptions sont le support, la confiance, le rappel, la moindre contradiction, la spécificité et la mesure de Sebag-Schœnauer.

De même, lorsque $A \rightarrow B$ est une règle logique, plusieurs mesures prennent des valeurs constantes évaluant ainsi de manière identique les cas 1, 2 et 3 présentés sur la Figure II.13.

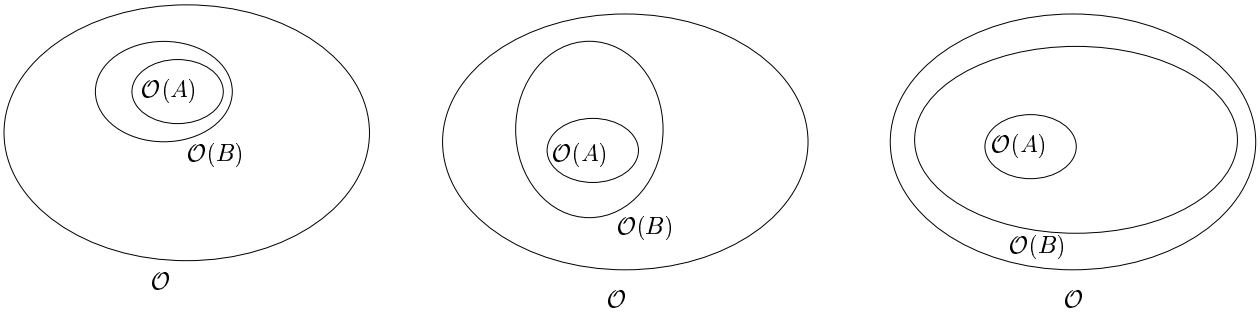


FIG. II.13 – Différents cas de règles logiques.

6 Travaux unificateurs sur les mesures de qualité

Nous présentons dans cette section plusieurs travaux apportant un éclairage nouveau et intéressant sur les mesures de qualité existantes. Ainsi, nous présentons en 6.1 une partie des travaux de S. Lallich et O. Teytaud [Lallich et Teytaud 2004]. Ces travaux concernent l'analyse d'une partie des mesures de qualité présentées dans le Tableau II.6. L'analyse réalisée permet de mieux comprendre le lien entre diverses mesures pouvant être vues comme des fonctions affines de la confiance.

La section 6.2 présente les travaux de Lavrac, Flach et Zupan [Lavrac et al. 1999]. Ces travaux sont centrés sur la nouveauté et les auteurs montrent que plusieurs mesures, après de simples normalisations, sont égales à la nouveauté.

6.1 Transformations affines de la confiance

[Lallich et Teytaud 2004] ont étudié les mesures de qualité sous un angle nouveau et qui apporte un éclairage intéressant sur les différentes mesures étudiées. La plupart des mesures de qualité peuvent s'exprimer comme des transformations affines de la confiance (voir définition 8).

Définition 8 *Une mesure de qualité m est une transformation affine de la confiance si m peut s'exprimer sous la forme suivante*

$$m(A \rightarrow B) = \theta_1(P(B|A) - \theta_0)$$

avec θ_0 et θ_1 ne dépendant que des marges relatives de la table qui croise A et B et éventuellement de n (voir Tableau II.1).

Ces mesures peuvent alors être simplement interprétées comme un centrage-réduction de la confiance par rapport aux paramètres θ_0 et θ_1. La plupart des mesures sont centrées sur $P(B)$ à l'exception de la moindre contradiction (centrée sur $0,5$), du lift (centré sur 0), de la nouveauté (centrée sur $P(A) \times P(B)$) et de la satisfaction (centrée sur $1 + \frac{P(\overline{A})P(B)}{P(A)}$).

Le centrage de la confiance par rapport à $P(B)$ permet de corriger un des défauts de la confiance, en comparant la valeur observée à celle attendue sous l'hypothèse d'indépendance. Le changement d'échelle (la réduction) diffère suivant les mesures et suivant le but recherché. De plus, ce changement d'échelle permet de différencier deux mesures centrées sur $P(B)$. Le Tableau II.9 rappelle pour chaque mesure son expression analytique et présente les deux paramètres θ_0 et θ_1.

Comme le souligne [Lallich et Teytaud 2004], ces mesures répondent au principal défaut de la confiance mais étant issues de la confiance, elles héritent toutes de certains de ses défauts majeurs. Ainsi, à marges fixées, ces mesures sont une fonction affine du nombre de contre-exemples et varient donc linéairement par rapport à cette valeur. D'autre part, toutes ces mesures, à l'exception de la mesure de Piatesky-Shapiro et de l'indice d'implication, ne dépendent par de n et elles sont donc invariantes par variation de la taille des données.

L'étude des mesures de qualité avec ce nouvel éclairage permet de mettre plus facilement en évidence les différents défauts et qualités détaillés lors de la présentation des mesures.

6.2 Mesure de qualité unique

Lavrac, Flach et Zupan [Lavrac et al. 1999] présentent une étude de la nouveauté.

Dans un premier temps, ils introduisent les mesures de qualité suivantes : confiance, fiabilité négative, sensibilité, spécificité, nouveauté et satisfaction.

mesure	expression	θ_0	θ_1
confiance centrée	$P(B\|A) - P(B)$	$P(B)$	1
Lift	$\frac{P(AB)}{P(A)P(B)}$	0	$\frac{1}{P(B)}$
Moindre contradiction	$\frac{P(AB)-P(A\overline{B})}{P(B)}$	0.5	$2 \times \frac{P(A)}{P(B)}$
Piatetsky-Shapiro	$NP(A)(P(B\|A) - P(B))$	$P(B)$	$N \times P(A)$
Lœvinger	$\frac{P(B\|A)-P(B)}{P(\overline{B})}$	$P(B)$	$\frac{1}{P(\overline{B})}$
Corrélation	$\frac{P(AB)-P(A)P(B)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}}$	$P(B)$	$\frac{\sqrt{P(A)}}{\sqrt{P(\overline{A})P(B)P(\overline{B})}}$
Indice d'implication	$\sqrt{n}\frac{P(A\overline{B})-P(A)P(\overline{B})}{\sqrt{P(A)P(\overline{B})}}$	$P(B)$	$-\sqrt{n}\sqrt{\frac{P(A)}{P(\overline{B})}}$
Pearl	$P(A) \times \|P(B\|A) - P(B)\|$	$P(B)$	$P(A)$
Nouveauté	$P(AB) - P(A)P(B)$	$P(A)P(B)$	0
Satisfaction	$\frac{P(\overline{A})-P(\overline{A}\|B)}{P(\overline{B})}$	$1 + \frac{P(\overline{A})P(B)}{P(A)}$	$\frac{P(A)}{P(B)P(\overline{B})}$

TAB. II.9 – Transformations affines de la confiance.

Les auteurs ont placé la nouveauté au centre de leur étude car ils considèrent la nouveauté comme une mesure relative en ce sens qu'elle permet de comparer le support d'une règle $A \rightarrow B$ avec la valeur attendue du support sous l'hypothèse d'indépendance des attributs A et B.

Dans un deuxième temps, ils proposent de redéfinir la confiance, la sensibilité, la spécificité et la fiabilité négative de manière relative en les normalisant par rapport aux valeurs attendues sous l'hypothèse d'indépendance (voir Tableau II.10).

mesure	expression
confiance relative	$P(B\|A) - P(B)$
fiabilité négative relative	$P(\overline{B}\|\overline{A}) - P(\overline{B})$
sensibilité relative	$P(A\|B) - P(A)$
spécificité relative	$P(\overline{A}\|\overline{B}) - P(\overline{A})$

TAB. II.10 – Mesures de qualité relatives.

Ensuite, à partir de ces mesures, une deuxième normalisation est effectuée. Cette normalisation a été initialement conçue pour pallier à un des défauts de la confiance relative, à savoir trouver des règles ayant un faible support et une confiance relative très élevée. La nouvelle mesure obtenue est appelée la *confiance relative pondérée* et s'exprime de la manière suivante :

$$P(A)(P(B|A) - P(B))$$

L'introduction de $P(A)$ dans la mesure permet de pondérer la confiance relative par le support de la prémisse de la règle étudiée, rendant ainsi difficile l'apparition de règle ayant un faible support et une confiance relative élevée. Nous pouvons constater que la confiance relative pondérée est égale à la nouveauté.

Les auteurs proposent de pondérer les autres mesures relatives de la même manière. Le Tableau II.11 présentent les nouvelles mesures obtenues.

Ayant défini ces quatre nouvelles mesures, il est aisé de montrer que ces mesures sont toutes égales à la nouveauté. Ainsi, bien que sémantiquement différentes et évaluant, en apparence, l'intérêt de la règle $A \rightarrow B$ sous différents aspects, ces mesures ne font qu'évaluer la nouveauté de la règle.

Selon, les auteurs, la nouveauté est donc une mesure centrale et suffisante pour évaluer l'intérêt d'une règle. Cependant, comme nous l'avons montré lors de la présentation de la nouveauté, cette

mesure	expression
confiance relative pondérée	$P(A)(P(B\|A) - P(B))$
fiabilité négative relative pondérée	$P(\overline{A})P(\overline{B}\|\overline{A}) - P(\overline{B})$
sensibilité relative pondérée	$P(B)(P(A\|B) - P(A))$
spécificité relative pondérée	$P(\overline{B})(P(\overline{A}\|\overline{B}) - P(\overline{A}))$

TAB. II.11 – Mesures relatives pondérées.

mesure possède certains défauts lorsque le problème de la recherche de pépites de connaissance est abordé. Bien qu'elle regroupe de nombreux aspects de l'intérêt d'une règle nous pensons qu'il est difficile de se contenter de cette unique mesure pour évaluer l'intérêt d'une pépite de connaissance.

7 Travaux de Hamilton et Hilderman

Hilderman et Hamilton [Hilderman et Hamilton 1999] présentent une étude des mesures d'intérêt utilisées pour la découverte de connaissances. L'étude qu'ils ont réalisé est très générale et concerne différents aspects de la découverte de connaissances dont les règles d'association. Parmi les nombreuses mesures étudiées, cinq mesures objectives et une mesure subjective peuvent être utilisées dans le cadre de la recherche de règles d'association. Outre la confiance, les nouvelles mesures étudiées sont la surprise [Freitas 1998], l'intérêt [Gray et Orlowska 1998, Dong et Li 1998] et les patrons de règles [Klemettinen et al. 1994].

7.1 évaluation objective de la surprise d'une règle

A. Freitas [Freitas 1998] propose trois méthodes objectives pour mesurer la surprise d'une règle.

7.1.1 étude des « petites » règles

[Freitas 1998] propose de se focaliser sur les « petites » règles ("small disjuncts" [Provost et Aronis 1996]), c'est-à-dire des règles ayant un faible support. Une telle règle est considérée comme surprenante si la généralisation minimale de la règle conduit à une règle ayant une conclusion différente de celle « prédite » par la règle dont elle est issue. Soit $R = A_1, A_2, ..., A_m \rightarrow B$ une règle et R_i la règle obtenue par généralisation de R en enlevant l'attribut A_i de la prémisse. Pour évaluer la surprise d'une règle, il suffit de mesurer le nombre de fois où la conclusion de la règle R diffère de celle de la règle R_i, $1 \leq i \leq m$.

$$DisjSurp_{norm}(R) = \frac{\sum_i DiffCcl_i(R)}{m}$$

$$avecDiffCcl_i(R) = \begin{cases} 1 \text{ si } ccl(R_i) \neq ccl(R) \\ 0 \text{ sinon} \end{cases}$$

Plus la valeur de $DisjSurp_{norm}(R)$ est élevée et plus la règle R est surprenante.

7.1.2 étude du gain d'une règle

Dans cette approche, la prémisse des règles n'est plus considérée comme un tout (comme dans [Piatetsky-Shapiro 1991]), mais comme un ensemble d'attributs, chaque attribut étant considéré individuellement.

Une mesure permettant de mesurer le gain informationel de la prémisse d'une règle est proposée. Cette mesure permet de comparer deux règles R_1 et R_2 qui sont équivalentes du point de vue des autres mesures de qualité utilisées (même support et confiance). Ainsi, R_1 est considérée comme plus suprenante que R_2 si la prémisse de R_1 contient des attributs ayant un plus faible gain informationel que ceux de R_2.

Le gain informationel d'un attribut A_i est défini de la manière suivante :

$$InfoGain(A_i) = Info(G) - Info(G|A_i)$$

avec

$$Info(G) = -\sum_{j=1}^{n} Pr(G_j) log_2 Pr(G_j)$$

$$InfoGain(G|A_i) = \sum_{k=1}^{p} Pr(A_i^k) \left(-\sum_{j=1}^{n} Pr(G_j|A_i^k) log_2 Pr(G_j|A_i^k) \right)$$

où G représente la conclusion des règles et G_j représente une valeur possible de la conclusion. A_i^k représente la k^{eme} valeur possible de l'attribut A_i.

Dans le cadre de la recherche de règles d'association, tous les attributs peuvent être en prémisse et en conclusion. Ainsi, G peut être égal à un quelconque attribut A_i, $1 \leq i \leq n$ (n le nombre d'attributs de la base de données).

La surprise d'une règle peut alors être évaluée avec la formule :

$$GainSurprise(R) = \frac{1}{\sum_{i=1}^{m} \frac{InfoGain(A_i)}{m}}$$

où m représente le nombre d'attributs en prémisse de la règle R.

Plus cette valeur est élevée et plus la règle est surprenante.

7.1.3 Détection des paradoxes de Simpson

Un paradoxe de Simpson (voir [Pearl 1999] extrait du chapitre 6 de [Pearl 2000]) correspond à un événement C qui

- augmente la probabilité d'occurence d'un événement E dans une population p
- et diminue la probabilité de l'événement E dans toutes les sous-populations de p

Ce paradoxe peut se formuler de la manière suivante : Soient F et $\neg F$, deux propriétés complémentaires décrivant deux populations issues d'une population p. Si C correspond à un paradoxe de Simpson, alors les inégalités suivantes sont vérifiées

$$P(E|C) > P(E|\neg C)$$

$$P(E|C,F) < P(E|\neg C,F)$$

$$P(E|C,\neg F) < P(E|\neg C,\neg F)$$

Voici un exemple permettant de mieux comprendre ce phénomène.

Exemple : Considérons une population p d'individus constituées d'hommes et de femmes. Posons F, la propriété d'être une femme et $\neg F$ la propriété d'être un homme. Soit C l'événement associé à la prise d'un médicament par les individus de la population. Soit E l'effet du médicament sur la population.

p	E	$\neg E$	$P(E\|C)$
C	20	20	50%
$\neg C$	16	24	40%

$\neg F$	E	$\neg E$	$P(E\|C)$
C	18	12	60%
$\neg C$	7	3	70%

F	E	$\neg E$	$P(E\|C)$
C	2	8	20%
$\neg C$	9	21	30%

TAB. II.12 – Illustration d'un paradoxe de Simpson.

Un paradoxe de Simpson est observable si l'effet du médicament sur toute la population est significatif alors que sur chacune des sous-populations (hommes/femmes), l'effet inverse est observable.

Le paradoxe est dû au fait que C est considéré comme causal pour E alors que C n'est en fait qu'une évidence positive pour E. En effet, sur cet exemple, il est aisé de constater que la proportion d'hommes à consommer le médicament est beaucoup plus élevée que la proportion de femmes. Ainsi, lorsqu'un patient consomme le médicament, la « bonne » conclusion serait de dire que ce patient est un homme. Pour réaliser correctement l'étude, il faut isoler le facteur pertinent, ici F, et effectuer les observations pour F et $\neg F$. Ainsi, le premier tableau représente l'impact du médicament sur la population sans information sur le sexe des patients, alors que connaissant le sexe les tableaux liés à F et $\neg F$ représentent bien l'efficacité du médicament dans les sous-populations.

Ayant posé le problème de la sorte, le paradoxe disparait.

Ceci étant dit, A. Freitas propose l'algorithme 2 pour détecter les paradoxes de Simpson.

Cet algorithme est très général et peu être adapté à différents cas (comme le souligne l'auteur). Ainsi, dans le cadre de la recherche de règles d'association sur des attributs binaires, la liste L_2 est égale à la liste L_1 et $m = 2$. Connaissant les paradoxes de Simpson contenu dans ses données, l'utilisateur peut les analyser un par un ou écarter les attributs présents en conclusion des paradoxes de la liste des attributs qui l'intéresse.

Bien que ces paradoxes soient très intéressants, leur détection est très coûteuse.

7.2 L'intérêt selon G. Dong et J. Li

G. Dong et J. Li [Dong et Li 1998] proposent d'évaluer l'intérêt d'une règle en considérant son caractère inattendu par rapport aux règles d'association de son voisinage. Le voisinage d'une règle est défini par l'ensemble des règles d'association situées à une distance D de la règle étudiée.

La distance entre deux règles R_1 et R_2 est définie de la manière suivante :

$$D(R_1, R_2) = \delta_1|(A_1 \cup B_1)\Theta(A_2 \cup B_2)| + \delta_2|A_1\Theta A_2| + \delta_3|B_1\Theta B_2|$$

$$\text{où} \begin{cases} R_i = A_i \rightarrow B_i \ (i = 1, 2) \\ \delta_i \text{ sont des paramètres permettant de pondérer l'importance de chacun des trois termes} \\ \Theta \text{ est un opérateur de différence ensembliste symétrique, c'est-à-dire } X\Theta Y = (X - Y) \cup (Y - X) \end{cases}$$

Le $r - voisinage$ d'une règle R_0 est défini par

$$N(R_0, r) = \{R | D(R, R_0) \geq r, R = A \rightarrow B\}$$

Ayant défini le $r - voisinage$ d'une règle, deux types de règles peuvent être isolées :
- les règles ayant une confiance inattendue
- les règles isolées

Algorithme 2 Recherche des paradoxes de Simpson.

Entrée :

L_G : liste des attributs pouvant être en conclusion

Sortie :

paradoxes : liste des paradoxes de Simpson

Début

$paradoxes \leftarrow \emptyset$

L_1 : liste des attributs binaires pouvant être utilisés en prémisse

L_2 : liste des attributs discrets pouvant être utilisés en prémisse

Pour tout $(G \in L_G)$ **Faire**

Pour tout $(A_1 \in L_1)$ **Faire**

répartir les individus en deux populations P_1 et P_2 selon les valeurs de A_1

$Pr(G_1) \leftarrow Pr(G = "yes" | A_1 = 1)$

$Pr(G_2) \leftarrow Pr(G = "yes" | A_1 = 2)$

Pour tout $(A_2 \in L_2$ tel que $A_2 \neq A_1)$ **Faire**

Pour $(i \leftarrow 1$ to $2)$ **Faire**

répartir les individus de P_i en deux populations P_{i1} et P_{im} selon les valeurs de A_2

Pour $(j \leftarrow 1$ to $m)$ **Faire**

$Pr(G_{ij}) \leftarrow Pr(G = "yes" | A1 = i, A_2 = j)$

Fin Pour

Fin Pour

Si $(Pr(G_1) > Pr(G_2)$ et $Pr(G_{1j}) \leq Pr(G_{2j}), j = 1, ..., m)$ ou $(Pr(G_1) < Pr(G_2)$ et $Pr(G_{1j}) \geq Pr(G_{2j}), j = 1, ..., m)$ **Alors**

$paradoxes \leftarrow paradoxes \cup \{A_1, A_2 = j \rightarrow G\}$

Fin Si

Fin Pour

Fin Pour

Fin Pour

Retourner *paradoxes*

Fin

Posons E l'ensemble des règles d'association vérifiant les contraintes de support et de confiance imposées par l'utilisateur.

les règles ayant une confiance inattendue

Une règle R_0 est considérée comme ayant une confiance inattendue si

$$||conf(R_0) - \mu(R_0, r)| - \sigma(R_0, r)| > t_1$$

avec $conf(R_0)$, la confiance de la règle R_0, $\mu(R_0, r)$ et $\sigma(R_0, r)$ représentant respectivement la moyenne et l'écart-type de la confiance des règles situées dans $E \cap N(R_0, r) - \{R_0\}$. Et t_1 est un seuil déterminé par l'utilisateur.

les règles isolées

Une règle R_0 est considérée comme isolée si

$$card(N(R_0, r)) - card(E \cup N(R_0, r)) > t_2$$

où $card(X)$ est égal au nombre d'éléments de l'ensemble X et t_2 est un seuil déterminé par l'utilisateur.

Cette approche est intéressante mais repose sur la confiance et elle hérite donc des défauts déjà évoqués de la confiance. De plus, E ne contenant que les règles vérifiant les contraintes imposées par l'utilisateur, celui-ci risque de rater les règles les plus surprenantes contenues dans ses données.

7.3 L'intérêt selong B. Gray et M.E. Orlowska

B. Gray et M.E. Orlowska [Gray et Orlowska 1998] proposent d'évaluer l'intérêt d'une règle d'association $A \rightarrow B$ en estimant l'indépendance entre la conclusion de la règle et sa prémisse. Ils proposent la mesure suivante pour estimer cette indépendance :

$$I(A \rightarrow B) = \left(\left(\frac{P(AB)}{P(A) \times P(B)} \right)^k - 1 \right) \times (P(A) \times P(B))^m$$

où k et m sont des paramètres permettant de pondérer respectivement l'influence du lift et des supports.

7.4 Les patrons de règles

Les patrons de règles [Klemettinen et al. 1994] permettent à l'utilisateur de spécifier les règles qui l'intéressent et celles qui ne l'intéressent pas parmi un ensemble de règles trouvées par un algorithme tel qu'Apriori. Un patron peut prendre la forme suivante : $P_1, \ldots, P_k \rightarrow P_{k+1}$ où chaque $P_i, 1 \leq i \leq k+1$ est soit le nom d'un attribut, soit le nom d'une classe (en apprentissage supervisé), soit une expression de la forme C^+ ou C^* (avec C le nom d'une classe).

L'expression C^* signifie que les règles (vérifiant le patron) peuvent contenir zéro ou plusieures instances de la classe C.

L'expression C^+ signifie que les règles (vérifiant le patron) doivent contenir au moins une instance de la classe C.

En apprentissage non supervisé, nous considérons que n'importe quel attribut peut jouer le rôle de la classe.

L'exemple suivant, extrait de [Klemettinen et al. 1994], présente un patron de règles utilisé sur des données universitaires.

Exemple : Considérons une base de données contenant des renseignements sur les cursus universitaires. Un utilisateur peut être intéressé par le cours « conception et analyse des algorithmes ». Plus précisément, il peut rechercher les associations contenant ce cours en conclusion des règles et un autre cours dans la prémisse. Le patron « cours$^+$ $\rightarrow$ conception et analyse des algorithmes » permet de ne retenir que les règles vérifiant l'expression régulière placée en prémisse et contenant le cours « conception et analyse des algorithmes » en conclusion. L'expression régulière de la prémisse « cours$^+$ » doit être interprétée de la manière suivante : au moins un cours en prémisse.

De la même manière, l'utilisateur peut définir des patrons pour exclure les règles qui ne l'intéressent pas.

Cette approche est utile pour filtrer un ensemble de règles obtenues par un algorithme d'extraction de règles d'association. L'utilisation de cette approche pour extraire les règles suppose que l'utilisateur soit capable de définir les patrons avant la phase de recherche des règles. Or, bien souvent, l'utilisateur rencontre des difficultés pour définir précisément ce qu'il recherche dans les données.

8 Conclusion

Les différentes mesures de qualité que nous avons présentées dans ce chapitre sont :

- soit intégrées dans un algorithme d'extraction de règles d'association qui, dans la plupart des cas, exploite la propriété d'anti-monotonie du support pour filtrer les itemsets en cours de construction.
- soit utilisées dans des systèmes de filtrage de règles obtenues avec des algorithmes du type Apriori.

Dans le deuxième cas, le choix de la mesure à utiliser pour filtrer les très nombreuses règles est souvent difficile pour l'expert. Les travaux de [Lenca et al. 2003a, Lenca et al. 2003b] présentent une approche d'aide multicritères à la décision permettant d'aider l'expert à choisir la ou les mesures de qualité les mieux adaptées à ses critères.

Le point de départ de la plupart des techniques de filtrage existantes est un ensemble de règles (obtenu avec un algorithme tel qu'Apriori) et présentant le défaut majeur de contenir excessivement trop de règles, d'où la nécessité d'une étape de filtrage pour ne retenir que les règles les plus pertinentes.

Nous pensons que nous pouvons utiliser les mesures de qualité directement lors de la phase d'extraction des règles, sans pour autant avoir besoin de définir un support minimal pour les règles recherchées et prendre alors le risque de rater des pépites de connaissance.

Nous proposons, dans le chapitre III, un algorithme d'extraction de pépites de connaissance, à partir de données booléennes. Cet algorithme fonctionne généralement sans utiliser de support minimal et permet d'obtenir un ensemble de taille nettement inférieure à l'ensemble de règles obtenu par Apriori lorsque nous lui retirons la contrainte du support minimal.

Notre algorithme peut aussi utiliser un support minimal pour filtrer les règles qui sont *a priori* trop sensibles au bruit pour pouvoir être considérées comme des pépites de connaissance. Cet aspect de notre approche est présenté dans le chapitre IV.

CHAPITRE III

Extraction de pépites de connaissance

1 Introduction

Nous avons vu dans le chapitre précédent diverses mesures de qualité permettant d'extraire les règles d'association vérifiant une partie des critères de qualité présentés dans ce même chapitre. Comme nous l'avons vu, la plupart des approches sont fondées sur l'algorithme APRIORI et utilisent donc la propriété du support pour filtrer les règles candidates. Puis, à partir des règles obtenues, une ou plusieurs mesures de qualité sont appliquées pour sélectionner les meilleures règles.

La définition d'un support minimal, bien que très utile pour parcourir l'espace de recherche, est aussi très contraignante pour l'expert qui, sans connaissance *a priori* sur les règles recherchées, est souvent incapable de définir cette valeur minimale pour le support.

Nous proposons donc dans ce chapitre une approche permettant d'extraire un ensemble de règles intéressantes ayant la propriété d'être peu contredites par les données. Cette nouvelle approche ne nécessite pas la définition d'un support minimal. Elle est fondée sur l'utilisation de la mesure de qualité contramin (aussi appelée moindre contradiction et présentée dans le chapitre précédent) et utilise des contraintes proches de celles proposées par S. Sahar [Sahar 1999].

Nous commencerons par présenter les travaux de S. Sahar. Puis nous présenterons notre approche (section 3), ainsi que les résultats des expérimentations réalisées sur des données de l'UCI [Blake et Merz 1998] pour valider notre approche. Enfin, les travaux réalisés avec I.C. Lerman [Lerman et Azé 2003] concernant l'extraction des règles d'association les plus surprenantes, étant donné un ensemble de règles, seront présentés dans la section 5.

2 Sélection interactive des règles intéressantes

[Sahar 1999] présente une méthode de filtrage de règles d'association fondée sur une interaction avec l'utilisateur. Cette méthode utilise un algorithme d'extraction de règles d'association classique tel que APRIORI pour obtenir l'ensemble $\mathcal{E}$ des règles d'association vérifiant les contraintes de support et de confiance (ou autre mesure de qualité). Cet ensemble est souvent très volumineux et ne peut donc pas être proposé intégralement à l'expert. S. Sahar propose de présenter les règles une par une à l'expert selon un ordre bien particulier (voir section 2.3). Le travail de l'expert consiste alors à

analyser la règle et à la valider selon les critères présentés dans la section suivante. Les informations obtenues sur la règle analysée permettent d'élaguer efficacement l'ensemble des règles restantes.

2.1 Critères de validation

Lorsqu'une règle est proposée à l'expert, deux informations lui sont demandées :
- la règle est-elle vraie (V) ou fausse (F) ?
- la règle est-elle intéressante (I) ou non intéressante (NI) ?

Étant données ces deux informations, une règle appartient à l'une des quatre catégories suivantes :

VI : règle considérée comme vraie et intéressante par l'expert. Ces règles ne permettent pas d'élaguer l'ensemble des règles.

FI : règle considérée comme fausse et intéressante par l'expert. Par exemple, la règle « homme $\rightarrow$ marié »[1] est généralement fausse mais l'expert peut être intéressé par une spécialisation de la prémisse de la règle telle que « homme $\wedge$ possède un Mini Van $\rightarrow$ marié ». Ces règles, étiquetées FI, permettent d'élaguer la partie de $\mathcal{E}$ qui contient les règles de la forme « homme $\rightarrow$ marié$\wedge X$ ».

VNI : règle considérée comme vraie et non intéressante par l'expert. Par exemple, la règle « mari $\rightarrow$ marié » est vraie mais elle ne représente pas une connaissance intéressante pour l'expert. Ce type de règle permet d'élaguer efficacement l'ensemble $\mathcal{E}$. Sur cet exemple, toutes les règles contenant l'attribut « mari » en prémisse et l'attribut « marié » en conclusion sont élaguées.

FNI : règle considérée comme fausse et non intéressante par l'expert. Les règles appartenant à cette catégorie sont proches de celles appartenant à la catégorie FI. La différence majeure entre ces deux catégories est liée à la définition de l'intérêt de l'expert. Par exemple, la règle « homme $\rightarrow$ marié » précédemment associée à la catégorie FI peut être étiquetée FNI par un expert qui est intéressé par les règles concluant sur des revenus annuels supérieurs à 50000$ et non par des règles caractérisant les relations matrimoniales. L'élagage réalisé grâce à ce type de règles est identique à celui obtenu pour la catégorie VNI.

Lorsqu'une règle est associée à l'une de ces quatre catégories, le système proposé par S. Sahar réalise deux opérations :
- construction d'une base de connaissances.
  Toutes les règles considérées comme vraies par l'expert sont ajoutées dans la base de connaissances. Nous ne détaillerons pas cet aspect du système dans cette thèse.
- filtrage de l'ensemble de règles.

2.2 Élagage (ou filtrage) des règles

Les règles appartenant aux catégories VNI, FI et FNI participent à l'élagage de l'ensemble des règles.

Étant donné une règle, $R = a \rightarrow b$, associée aux catégories VNI ou FNI, l'élagage consiste à éliminer de $\mathcal{E}$ toutes les règles appartenant à $famille(R|\mathcal{E})$ avec

$$famille(R|\mathcal{E}) = \{A \rightarrow B | A \rightarrow B \in \mathcal{E}, a \in A, b \in B\}$$

L'élagage réalisé pour les règles étiquetées FI est légèrement différent du précédent. Étant donné une règle $R = a \rightarrow b$ étiquetée FI, seules les règles $R' = a \rightarrow B$ avec $b \in B$ peuvent être supprimées

[1] obtenue à partir de la base de données « Adult » de l'UCI [Blake et Merz 1998]

car ces règles seront toutes fausses puisqu'elles possèdent exactement la même prémisse que la règle étiquetée FI. Par contre, les règles appartenant à $\{A \rightarrow B | a \in A, b \in B, A - \{a\} \neq \emptyset, B - \{b\} \neq \emptyset\}$ sont potentiellement intéressantes.

2.3 Sélection des règles proposées à l'expert

L'intérêt de la méthode de S. Sahar réside dans l'interaction mise en place avec l'expert pour élaguer un ensemble de règles d'association. Le temps de l'expert étant précieux, il convient d'optimiser l'interaction mise en place. Pour atteindre cet objectif, S. Sahar propose de sélectionner les règles pour lesquelles l'élagage sera maximal si l'expert les classes dans les catégories adaptées.

Ainsi, pour chaque règle $R = a \rightarrow b$, le système calcule $s_R = card(famille(R|\mathcal{E}))$ et les règles sont proposées à l'utilisateur par valeurs décroissantes de s_R. À chaque itération de l'algorithme, ces valeurs sont remises à jour en fonction de l'élagage effectué. Lorsque les valeurs de s_R deviennent inférieures à un seuil prédéfini pendant deux itérations consécutives, l'expert est informé qu'il peut arrêter l'évaluation des règles s'il le désire car le gain potentiel (en termes de règles élaguées) devient minime.

2.4 Avantages et inconvénients

[Sahar 1999] montre qu'il suffit, en moyenne, de cinq interventions de l'expert pour diviser par deux la taille de l'ensemble $\mathcal{E}$. Ces résultats ont été obtenus sur quelques bases de données de l'UCI.

Cette approche permet donc de rapidement diminuer le volume de règles à analyser et ceci en solicitant relativement peu l'expert.

Cependant, un des inconvénients majeurs, selon nous, est lié au calcul de l'ensemble $\mathcal{E}$. L'ensemble de règles candidates $\mathcal{E}$ est obtenu avec un algorithme du type Apriori. Des contraintes sont donc imposées sur le support minimal des règles recherchées et il existe alors un risque de rater les règles vraiment intéressantes pour l'utilisateur. Si la contrainte liée au support est relâchée, alors le nombre de règles obtenues est en général beaucoup trop élevé pour pouvoir appliquer des méthodes d'analyse ou de filtrage de règles.

Cette méthode reste malgré tout intéressante et nous avons utilisé quelques un de ses principes pour concevoir celle que nous présentons dans la section suivante.

3 Recherche des pépites de connaissance

3.1 Introduction

Comme nous l'avons déjà vu, l'une des difficultés principales des systèmes d'extraction de règles d'association dans des données est de proposer à l'expert un ensemble de règles intéressantes et utiles.

L'objectif principal de cette thèse est de découvrir des « pépites » de connaissance dans les données. Une pépite de connaissance est caractérisée par rapport à un ensemble de règles et est définie simplement comme étant *potentiellement utile et nouvelle pour l'expert.*

Nous avons choisi de minimiser l'interaction avec l'expert dans le processus d'extraction des pépites de connaissance. Ce choix est motivé par le fait que bien souvent l'expert (pas forcément familier des systèmes d'extraction de connaissances) peut difficilement fournir au système les informations nécessaires pour effectuer l'extraction des règles. Ces informations sont souvent réduites à la définition de bornes inférieures (ou supérieures) pour le support, la confiance et autres mesures de qualité.

Malheureusement, bien que ces informations soient *a priori* simples à comprendre pour l'expert, celui-ci possède rarement le recul suffisant sur ses données pour définir de manière sûre les bornes demandées. C'est-à-dire, définir des bornes telles qu'aucune connaissance intéressante ne possède un support et une confiance (par exemple) inférieurs aux seuils définis.

Nous proposons donc une méthode permettant de ne plus utiliser ces contraintes pour extraire les pépites de connaissance tant recherchées.

3.2 Principe de la méthode

Puisque nous avons supprimé la contrainte du support minimal, nous ne pouvons plus utiliser les techniques d'élagage employées dans les algorithmes classiques d'extraction de règles d'association.

Nous devons donc utiliser d'autres propriétés pour limiter le nombre de règles obtenues qui augmente souvent de manière « dramatique » dès que le support minimal diminue.

Sachant que nous voulons que l'expert puisse interpréter les règles que nous lui proposerons, nous avons choisi d'utiliser des mesures de qualité relativement intuitives. Parmi les différentes mesures présentées dans le chapitre II, nous avons utilisé la moindre contradiction [Azé et Kodratoff 2002] pour évaluer l'intérêt de nos règles et extraire les pépites de connaissance. Ce choix est justifié d'une part par la simplicité de la mesure et d'autre part par les performances de cette mesure en terme de résistance au bruit (voir chapitre IV).

Nous tenons à préciser que l'algorithme proposé, bien que conçu pour la moindre contradiction, est indépendant de la mesure de qualité utilisée, pourvu que cette dernière possède la propriété d'être bornée. Lorsque nous avons testé la résistance au bruit de notre algorithme (voir chapitre IV), nous avons comparé plusieurs mesures ayant cette propriété et la moindre contradiction est celle qui présente le meilleur comportement, d'où notre choix.

La moindre contradiction est, rappelons le, une transformation affine de la confiance (voir chapitre II, section 6.1) et elle ne possède pas de propriété de monotonie ou d'anti-monotonie (voir figure III.1). Il est donc difficile (voire impossible) de rechercher de manière exhaustive l'intégralité des règles d'association les moins-contradictoires.

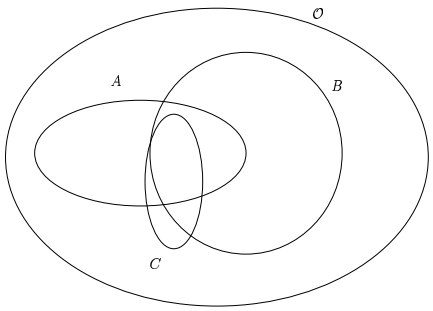


FIG. III.1 – contramin$(A \rightarrow B) < 0$ et contramin$(AC \rightarrow B) > 0$ donc la moindre contradiction n'est pas une propriété anti-monotone.

Cette contrainte nous a amené à réduire l'ensemble des règles recherchées en nous limitant aux règles d'association ayant les propriétés suivantes :

(i) être les règles se distinguant le plus des autres règles

(ii) être telles que les prémisses des règles ne contiennent pas plus de K_{max} attributs et telles que la conclusion de celles-ci soit réduite à un seul attribut.

Ainsi, contrairement aux méthodes classiques d'extraction de règles d'association qui permettent d'obtenir la totalité des règles d'association vérifiant les contraintes de l'utilisateur, nous n'obtiendrons pas la totalité des règles les moins contredites mais seulement les plus significatives d'entre elles.

Ceci nous permettra donc de réduire significativement la taille de l'ensemble des règles obtenu.

Nous présentons, dans la section suivante, un algorithme permettant d'extraire les règles d'association les moins-contredites satisfaisant les points (i) et (ii).

3.3 Algorithme d'extraction des pépites de connaissance

Pour satisfaire la première condition, (i), nous proposons d'utiliser une approche classique en Analyse de Données [Lerman 1984, Daudé 1992], à savoir normaliser les valeurs de chaque mesure par rapport à l'ensemble des mesures observées. Cette normalisation est effectuée de la manière suivante : (1) la valeur de la mesure est centrée par rapport à la valeur moyenne des mesures observées, puis (2) cette valeur est réduite par rapport à l'écart-type observé pour ces mêmes règles.

La deuxième condition, (ii), est justifiée par le fait que nous considérons que des règles d'association ayant un grand nombre d'attributs en prémisse et plus d'un attribut en conclusion, sont difficilement interprétables. Même pour un expert du domaine, de trop longues règles sont difficilement compréhensibles, c'est pourquoi, nous avons choisi de limiter le nombre d'attributs en prémisse et en conclusion des règles recherchées.

Dans une version précédente de notre algorithme [Azé 2003] les pépites de connaissance étaient extraites par niveau dans le treillis. Dans cette première version, dont l'algorithme est très proche de celui présenté ici, toutes les règles ayant exactement un attribut en prémisse étaient engendrées, puis les règles les plus significatives (c'est-à-dire dont la moindre contradiction, après normalisation, était supérieure à 1) étaient retenues comme pépites de connaissance pour être proposées à l'expert. Le principe de l'algorithme restant le même concernant la non spécialisation des pépites de connaissance trouvées.

Les pépites de connaissance obtenues, ayant K attributs en prémisse, étaient donc les règles les moins contredites parmi les règles ayant K attributs en prémisse. L'inconvénient de cette méthode est que certains attributs ne figuraient jamais en conclusion des règles car les règles les contenant étaient « écrasées » par les autres règles.

Nous avons donc modifié l'algorithme, en y ajoutant simplement une boucle pour faire figurer tous les attributs (un à un) en conclusion des règles. Les pépites de connaissance obtenues sont donc relatives à un nombre K d'attributs en prémisse et à un attribut donné en conclusion.

De plus, toutes les pépites trouvées sont relatives à un attribut donné (celui placé en conclusion des règles).

Le Tableau III.1 présente les notations utilisées dans l'algorithme 3.

Détails de l'étape itérative : $K \geq 1$ (c'est-à-dire, A contient K attributs)
Les règles appartenant à $\mathcal{E}'_K$ sont conservées sans être modifiées car, à l'étape $K+1$, les nouvelles règles obtenues ne doivent pas être des spécialisations de règles déjà obtenues à l'étape K. Cette heuristique est définie de manière à ne pas surcharger l'expert. Notre mesure ne possédant pas de propriété de monotonie, nous ne pouvons pas garantir qu'il n'existe pas de règle $A \wedge X \rightarrow B$ plus spécifique que $A \rightarrow B$ et qui soit moins contredite. Cependant, si nous choisissons de spécialiser l'intégralité des règles les moins contredites, nous obtenons alors un ensemble de règles trop volumineux pour pouvoir être soumis à l'expert. Nous pensons qu'il est plus aisé pour l'expert d'examiner un ensemble de règles générales et d'étudier, si besoin, un sous-ensemble de règles issu de la spécialisation d'une règle qu'il aura explicitement choisie.

Algorithme 3 ExtrairePépites($\mathcal{D}_n^p, K_{max}, min_{sup}$)

Entrée :

$\mathcal{D}_n^p$: la base de données étudiée

K_{max} : nombre maximal d'attributs en prémisse, défini par l'utilisateur

min_{sup} : support minimal pour les pépites de connaissance

Sortie :

$\mathcal{E}$: ensemble des règles d'association les moins-contredites par les données contenant au plus K_{max} attributs en prémisse et telles que $\{confiance(R) > 0,5 \ \forall R \in \mathcal{E}\}$.

Début

$K = 1$

$T_K \leftarrow \mu_0 \leftarrow \sigma_0 \leftarrow 0$

$\mathcal{E} \leftarrow \emptyset$

$\mathcal{E}_1 \leftarrow \{$ *l'ensemble de tous les attributs de la base étudiée* $\}$

Pour tout $(X_j \in \mathcal{E}_1)$ **Faire**

Tant que $(\mathcal{E}_K \neq \emptyset)$ and $(T_K < 1)$ and $(K \leq K_{max})$ **Faire**

-- Génération de $\mathcal{C}_K$ à partir de $\mathcal{E}_K$

Si $(K = 1)$ **Alors**

-- Cas particulier où $\mathcal{E}_K$ ne contient que des attributs

$\mathcal{C}_K \leftarrow \{X_i \rightarrow X_j / X_i \in \mathcal{E}_1 \ , \ X_j \in \mathcal{E}_1 \ , \ i \neq j \ , \ support(X_i \rightarrow X_j) > min_{sup}\}$

Sinon

-- Cas général où $\mathcal{E}_K$ contient des règles

$\mathcal{C}_K \leftarrow \{X_i, X_l \rightarrow X_j / X_i \rightarrow X_j \in \mathcal{E}_K \ , \ X_l \rightarrow X_j \in \mathcal{E}_K \ , \ i \neq l \ , \ support(X_i, X_l \rightarrow X_j) > min_{sup}\}$

Fin Si

$\mathcal{E}_K^+ \leftarrow \{R \in \mathcal{C}_K / contramin(R) > T_K\}$

$\mathcal{E}_K^- \leftarrow \{R \in \mathcal{C}_K / contramin(R) \leq T_K\}$

Si $(\mathcal{E}_K^+ = \emptyset)$ **Alors**

$T_{K+1} \leftarrow T_K$ *-- Dans ce cas, μ_K et σ_K ne sont pas calculables*

$\mathcal{E}_{K+1} \leftarrow \mathcal{E}_K^-$

Sinon

Calcul de μ_K et σ_K

$\mathcal{E}'_K \leftarrow \{R \in \mathcal{E}_K^+ / \frac{contramin(R) - \mu_K}{\sigma_K} > 1\}$

$\mathcal{E} \leftarrow \mathcal{E} \cup \mathcal{E}'_K$

-- Préparation du niveau suivant

$\mathcal{E}_{K+1} \leftarrow \mathcal{E}_K^- \cup (\mathcal{E}_K^+ - \mathcal{E}'_K)$ *-- {règles non proposées à l'expert}*

$T_{K+1} \leftarrow \mu_K + \sigma_K$

Fin Si

$K \leftarrow K + 1$

Fin Tant que

Fin Pour

Retourner $\mathcal{E}$

Fin

symbole	signification
$\mathcal{D}_n^p$	base de données contenant n individus décrits par p attributs
K	nombre d'attributs dans la prémisse d'une règle d'association $A \rightarrow B$
$\mathcal{E}_K$	ensemble de toutes les règles d'association utilisé comme ensemble générateur pour $\mathcal{C}_K$
$\mathcal{C}_K$	ensemble des règles d'association « candidates » obtenues à partir de $\mathcal{E}_K$ contenant les règles d'association les moins-contradictoires
T_K	seuil d'élagage pour les règles d'association les moins-contradictoires
$\mathcal{E}_K^+$	sous-ensemble des règles de $\mathcal{C}_K$ dont la moindre contradiction se situe au delà du seuil T_K
$\mathcal{E}_K^-$	sous-ensemble des règles de $\mathcal{C}_K$ dont la moindre contradiction se situe en dessous du seuil T_K
μ_K	moyenne des règles d'association les moins-contradictoires appartenant à $\mathcal{E}_K^+$
σ_K	écart-type des règles d'association les moins-contradictoires appartenant à $\mathcal{E}_K^+$
$\mathcal{E}'_K$	ensemble des règles d'association les moins-contradictoires parmi $\mathcal{E}_K^+$
$\mathcal{E}$	ensemble des règles d'association telles que la moindre-contradiction soit la plus élevée

TAB. III.1 – Notations utilisées dans l'algorithme EXTRAIREPÉPITES.

Notons que la non spécialisation des règles proposées à l'expert peut être comparée à l'approche retenue par [Sahar 1999] pour filtrer un ensemble de règles. En effet, comme nous l'avons vu précédemment, S. Sahar utilise l'expert pour réduire de manière interactive l'ensemble des règles d'association. Dans notre approche, la réduction de l'ensemble des règles d'association est réalisée automatiquement.

De plus, et contrairement à la méthode proposée par S. Sahar, nous pouvons proposer à l'expert certaines pépites de connaissance que le système de S. Sahar ne peut détecter.

En effet, si nous considérons la règle générale *tous les humains aiment un autre humain* et que nous considérons que l'expert utilisant le système de S. Sahar étiquette cette règle FNI. Dans ce cas, toutes les spécialisations de cette règle seront supprimées, y compris la règle *Pierre Curie aimait Marie* qui peut intéresser l'expert.

Notre système, s'il ne propose pas la règle générale à l'expert, va malgré tout engendrer et analyser toutes les spécialisations de celle-ci.

Et, dans les deux systèmes, celui de S. Sahar et le nôtre, il est aisé de proposer à l'expert la possibilité de spécialiser une règle qu'il aura jugé intéressante ou que le système lui aura proposé.

Comme le nombre de règles situées au dessus d'un seuil de moindre contradiction prédéfini augmente très rapidement avec K, et comme nous voulons éviter d'engendrer trop de règles, nous devons définir une nouvelle heuristique permettant de faire en sorte que le seuil d'élagage augmente avec K. Nous nous focalisons alors sur la spécialisation des règles appartenant à $\mathcal{E}_{K+1} = \mathcal{E}_K^- \cup (\mathcal{E}_K^+ - \mathcal{E}'_K)$ et nous initialisons le nouveau seuil d'élagage, pour les règles les moins-contradictoires, avec $\mu_K + \sigma_K$, de manière à ne retenir que les « meilleures » règles les moins-contradictoires.

3.4 Complexité de l'algorithme

Nous proposons dans cette section d'évaluer la complexité de notre algorithme en comparant le coût de notre approche à celui de l'algorithme APRIORI. Nous allons étudier deux situations : le cas le pire et le cas général.

3.4.1 Cas le pire

Dans le pire des cas, aucune pépite de connaissance n'est trouvée par notre algorithme et le coût de notre approche est en $O(p^{K_{max}+1})$.

Donc dans le pire des cas, la complexité de l'algorithme est identique à celle de l'algorithme APRIORI (lorsqu'aucun item n'est élagué grace à la propriété du support).

PREUVE :

Si aucun élagage n'est effectué par l'algorithme alors, pour un attribut placé en conclusion, tous les itemsets de taille 1 à K_{max} seront engendrés comme prémisse. Pour une conclusion donnée, le nombre de premisses possibles est égal à $\sum_{i=1}^{K_{max}} C_{p-1}^{i}$. Sachant que les p attributs peuvent être en conclusion, le nombre de règles est égale à $p \times \sum_{i=1}^{K_{max}} C_{p-1}^{i}$.

On a

$$\begin{aligned}
C_{p-1}^{i} &= \frac{(p-1)!}{i! \times (p-1-i)!} \\
p \times \sum_{i=1}^{K_{max}} C_{p-1}^{i} &= p \times \sum_{i=1}^{K_{max}} \prod_{j=0}^{i-1} \frac{(p-1-j)}{i!} \\
&= p \times \sum_{i=1}^{K_{max}} \frac{1}{i!} \prod_{j=0}^{i-1} (p-1-j) \\
&\leq p \times \sum_{i=1}^{K_{max}} \frac{(p-1)^{K_{max}}}{i!} \\
&\leq p \times (p-1)^{K_{max}} \times \sum_{i=1}^{K_{max}} \frac{1}{i!} \\
&< p^{K_{max}+1} \times \sum_{i=1}^{K_{max}} \frac{1}{i!}
\end{aligned} \tag{III.1}$$

D'où le coût de l'approche en $O(p^{K_{max}+1})$.

♦

3.4.2 Cas général

Dans le cas général, des pépites de connaissances sont trouvées et, sachant que notre algorithme inclut un procédé permettant de ne pas spécialiser les règles les moins contredites, le coût de notre approche devient inférieur à celui de l'algorithme APRIORI. L'élagage réalisé par l'algorithme d'extraction des pépites de connaissance permet de réduire le nombre de règles à engendrer et à analyser.

Apport de l'élagage dans l'algorithme

Dans notre approche, aucune des règles les moins contredites (pépites de connaissance) n'est spécialisée, ce qui représente donc un gain en terme de temps de calcul et d'espace mémoire nécessaire pour stocker les règles calculées.

Nous proposons d'évaluer le nombre de règles que nous ne spécialisons pas et qui représente donc un gain par rapport à l'approche APRIORI. Si on considère que, étant donné un ensemble de règles $\mathcal{E}'_K$, toutes les spécialisations de ces règles ont un support non nul et une confiance strictement supérieure à $0,5$, alors l'algorithme APRIORI ne réalisera aucun élagage.

L'algorithme d'extraction des pépites de connaissance ne spécialise aucune des règles proposées à l'expert.

Pour évaluer l'apport de l'élagage dans notre algorithme, nous devons considérer le premier ensemble de pépites de connaissance trouvé. Cet ensemble de règles est noté $\mathcal{E}'_K$ dans notre algorithme, notons $n'_K = card(\mathcal{E}'_K)$.

Notons $n'_K(j)$ le nombre de pépites de connaissance ayant l'attribut X_j en conclusion.

Le calcul du nombre de règles non spécialisées dépend de trois paramètres :

- p : le nombre d'attributs de la base de données
- K : le nombre d'attributs en prémisse des pépites de connaissance trouvées
- $n'_K(j)$: le nombre de pépites trouvées ayant un attribut donné en conclusion.

élagage dès la première itération de l'algorithme

Le cas le plus simple et le plus favorable (c'est-à-dire maximisant l'élagage) est celui où des pépites de connaissance sont trouvées dès la première itération de l'algorithme (pour $K = 1$).

Appelons $elagage(p, K+1, n'_K)$ le nombre de règles non spécialisées au niveau $K+1$, étant données n'_K pépites de connaissance.

Propriété 1 *Lorsque les pépites proposées à l'expert (et concluant sur un attribut $X_j, 1 \leq j \leq p$) ne contiennent qu'un seul attribut en prémisse, le nombre de règles non spécialisées est égal à*

$$elagage(p, 2, n'_1(j)) = n'_1(j) \times \left((p-2) - \frac{n'_1(j) - 1}{2} \right)$$

La preuve de cette propriété est la suivante :

PREUVE :

Sachant que les $n'_1(j)$ pépites ont toutes la même conclusion, X_j, seule la spécialisation des prémisses nous intéresse. Ces prémisses ne contiennent qu'un seul attribut choisi parmi les $p-1$ attributs de la base différents de X_j. Donc, pour une prémisse donnée, nous pouvons obtenir $p-2$ spécialisations par simple ajout d'un attribut dans la prémisse. Ainsi, les $n'_1(j)$ pépites engendrent $n'_1(j) \times (p-2)$ nouvelles règles obtenues à partir des spécialisations de leurs prémisses.

Deux prémisses différentes (réduites à un seul attribut), X_1 et X_2, engendrent chacune $p-2$ nouvelles prémisses correspondant à leurs spécialisations. Parmi les prémisses engendrées, il existe une prémisse commune qui est $X_1 \wedge X_2$ obtenue à partir de la spécialisation de X_1 et $X_2 \wedge X_1$ obtenue à partir de la spécialisation de X_2.

Nous devons prendre ces recouvrements en considération pour calculer le nombre de spécialisations obtenues à partir de $n'_1(j)$ prémisses.

Le nombre exact de recouvrements est égal à $C^2_{n'_1(j)} = \frac{n'_1(j) \times (n'_1(j)-1)}{2}$. Cette valeur correspond aux nombres de prémisses ayant exactement deux attributs, engendrées à partir de $n'_1(j)$ prémisses (réduites à un unique attribut).

Le nombre de règles non spécialisées, étant donné $n'_1(j)$ règles concluant sur l'attribut X_j, est donc bien égal à $n'_1(j) \times \left((p-2) - \frac{n'_1(j)-1}{2} \right)$.

♦

Le gain de notre approche par rapport à l'algorithme APRIORI, dans ce cas, s'exprime donc de la manière suivante

$$gain(ExtrairePepites/Apriori|K=1,X_j) = \frac{elagage(p,2,n'_1(j))}{C^2_{p-1}}$$

où C^2_{p-1} est le nombre de règles ayant deux attributs en prémisse obtenues à partir des $p-1$ règles ayant un attribut en prémisse.

Après simplification des calculs, le gain de notre approche par rapport à l'algorithme APRIORI, pour les règles de la forme $X_i \rightarrow X_j, i \neq j$, est égal à

$$gain(ExtrairePepites/Apriori|K=1,X_j) = \frac{n'_1(j)(2p-n'_1(j)-3)}{(p-1)(p-2)}$$

Élagage pour $K \geq 2$

Dans le cas plus général où la prémisse des pépites contient plus d'un attribut, c'est-à-dire lorsque le premier élagage intervient pour une valeur de K supérieure ou égale à 2, le calcul exact du nombre de règles non spécialisées est plus difficile (voir impossible).

Ce problème est dû au nombre de recouvrement entre les spécialisations des règles supprimées lors de la $K^{ième}$ itération de l'algorithme qui n'est plus aussi simple à calculer que pour le cas précédent.

La Figure III.2 présente l'élagage des spécialisations des règles $\{AB \rightarrow C, AF \rightarrow C, BE \rightarrow C\}$.

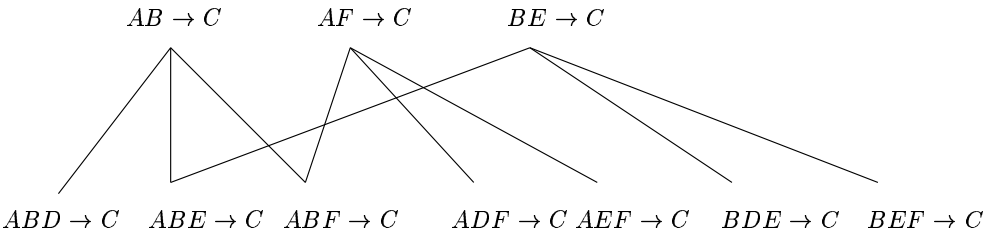


FIG. III.2 – Spécialisation des règles $\{AB \rightarrow C, AF \rightarrow C, BE \rightarrow C\}$ étant donnés les attributs A, B, C, D, E et F.

Nous pouvons voir que la règle $ABE \rightarrow C$ est une spécialisation des règles $AB \rightarrow C$ et $BE \rightarrow C$ alors que la règle $BDE \rightarrow C$ n'est engendrée que par $BE \rightarrow C$. Le calcul exact de ces recouvrements, sans engendrer les règles spécialisées, est un problème difficile.

Nous proposons donc d'estimer le nombre de règles élaguées pour le niveau $K+1$ lorsque les n'_K pépites trouvées au niveau K sont les premières pépites trouvées par l'algorithme. Cette valeur peut être bornée par $binf(p, K+1, n'_K)$ et $bsup(p, K+1, n'_K)$ (voir la propriété 2).

Ainsi, seules les bornes inférieure et supérieure, $binf(p, K+1, n'_K)$ et $bsup(p, K+1, n'_K)$, sont calculables.

Les bornes inférieures et supérieures correspondent aux situations extrêmes pouvant être observées lors de l'élagage de n'_K règles de taille K. L'estimation du nombre de spécialisations non engendrées est liée au nombre de recouvrements possibles entre les spécialisations des n'_K règles élaguées. Ainsi, dans le meilleur des cas, les spécialisations des n'_K règles ne se recouvrent jamais

(d'où l'expression de la borne supérieure) et dans le pire des cas, le recouvrement des spécialisations est maximal (d'où l'expression de la borne inférieure).

Propriété 2 *L'élagage réalisé par notre algorithme est égal au nombre de règles non spécialisées parmi les n'_K pépites de connaissance obtenues lors de la $K^{ème}$ itération.*

Si on suppose qu'aucun élagage n'a été réalisé lors des itérations précédentes, alors le nombre de règles élaguées noté $\boldsymbol{elagage(p, K+1, n'_K)}$ *est tel que :*

$$binf(p, K+1, n'_K) \leq binf_{precise}(p, K+1, n'_K) \leq elagage(p, K+1, n'_K) \leq bsup(p, K+1, n'_K)$$

Les propriétés 3 et 4 définissent respectivement la borne supérieure et la borne inférieure. La borne $binf_{precise}(p, K+1, n'_K)$ représente une borne inférieure plus précise que $binf(p, K+1, n'_K)$ et son expression est donnée dans la propriété 5.

Propriété 3 *Il existe une borne supérieure au nombre de règles contenant $K+1$ attributs en prémisse engendrés à partir de $n'_K(j)$ règles contenant K attributs en prémisse et ayant l'attribut X_j en conclusion. Cette valeur, que nous nommerons* $\boldsymbol{bsup(p, K+1, n'_K(j))}$*, est égale à :*

$$bsup(p, K+1, n'_K(j)) = n'_K \times (p - 1 - K)$$

PREUVE :

La démonstration de l'existence de la borne supérieure est fondée sur l'observation suivante : dans le meilleur des cas, les $n'_K(j)$ règles supprimées sur le niveau K du treillis sont toutes deux à deux séparées par strictement plus d'un attribut (voir Figure III.3 sur laquelle nous n'avons fait figurer que les prémisses des règles concluant toutes sur l'attribut X_j). Rappelons que les K attributs des prémisses de ces règles sont choisis parmis $p-1$ attributs différents de X_j.

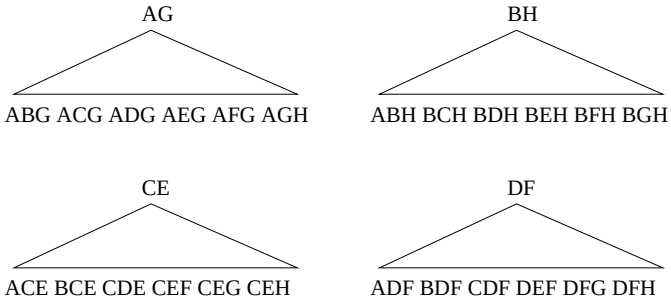


FIG. III.3 – Aucun recouvrement entre les prémisses des spécialisations des règles $\{AG \rightarrow X_j, BH \rightarrow X_j, CE \rightarrow X_j, DF \rightarrow X_j\}$.

Ainsi chaque règle supprimée engendre $(p-1-K)$ règles contenant $K+1$ attributs en prémisse. Ces règles sont obtenues par ajout d'un attribut dans la prémisse de la règle courante, l'attribut ajouté étant choisi parmi les attributs disponibles non présents dans la prémisse de la règle courante. Si le nombre d'attribut de la prémisse est égal à K et le nombre total d'attribut égal à p, alors chaque règle dispose de $p-1-K$ attributs non utilisés pour construire des règles contenant $K+1$ attributs en prémisse.

D'où l'existence de la borne supérieure : $n'_K(j) \times (p-1-K)$.

♦

Propriété 4 *Il existe une borne inférieure au nombre de règles contenant $K+1$ attributs en prémisse engendrées à partir de $n'_K(j)$ règles contenant K attributs en prémisse et ayant l'attribut X_j en conclusion.*

Cette valeur, que nous nommerons $\boldsymbol{binf(p, K+1, n'_K(j))}$*, est égale à :*

$$binf(p, K+1, n'_K(j)) = C^{K+1}_{p_{min}(n'_K(j),K)}$$

où

$$C^{K}_{p_{min}(n'_K(j),K)} = n'_K(j)$$

PREUVE :

Pour prouver l'existence de la borne inférieure et pour calculer sa valeur, plaçons-nous dans le pire des cas : toutes les règles supprimées sont deux à deux séparées par exactement un attribut (voir Figure III.4 où seules les prémisses des règles ayant la même conclusion sont indiquées).

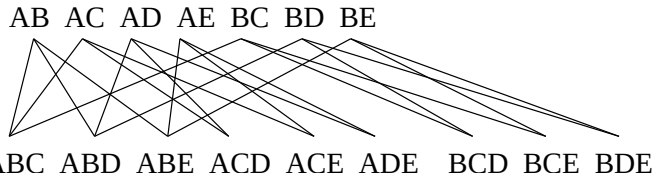


FIG. III.4 – Recouvrement maximal dans les règles spécialisées.

Ainsi, le nombre de règles contenant $K+1$ attributs en prémisse obtenues à partir des $n'_K(j)$ règles supprimées est minimal.

Pour déterminer la borne inférieure, commencons par calculer le nombre minimal d'attributs nécessaires pour obtenir n'_K règles contenant K attributs en prémisse. Cette valeur, que nous nommerons $p_{min}(n'_K(j), K)$, ne peut être déterminée de manière exacte. En effet, cette valeur est telle que $C^{K}_{p_{min}(n'_K(j),K)} = n'_K(j)$ et il n'est pas possible d'obtenir la valeur de $p_{min}(n'_K(j), K)$ autrement que par approximation.

Si nous supposons cette valeur connue, nous pouvons alors déterminer le nombre de règles ayant $K+1$ attributs en prémisse obtenues à partir des $n'_K(j)$ règles étudiées. Cette valeur est simplement égale à $C^{K+1}_{p_{min}(n'_K(j),K)}$, c'est-à-dire le nombre de règles contenant $K+1$ attributs en prémisse, les attributs étant choisi parmi un ensemble de $p_{min}(n'_K(j), K)$ attributs.

♦

La borne inférieure que nous avons obtenue et qui permet d'apprécier le gain réalisé par notre algorithme par rapport à l'algorithme APRIORI sous-estime très souvent le nombre réel de règles élaguées par notre approche.

Nous pouvons calculer une borne inférieure beaucoup plus précise que $binf(p, K+1, n'_K(j))$ en utilisant le graphe des $n'_K(j)$ règles supprimées. Dans ce graphe, deux règles sont reliées par un arc si et seulement si les deux règles diffèrent exactement d'un attribut.

La Figure III.5 présente un tel graphe.

Un simple parcours de ce graphe permet d'obtenir une borne inférieure très précise pour le nombre de règles ayant $K+1$ attributs en prémisse et pouvant être obtenus à partir de ce graphe.

Propriété 5 *Considérons un ensemble de $n'_K(j)$ prémisses de taille K et un ensemble de p attributs.*

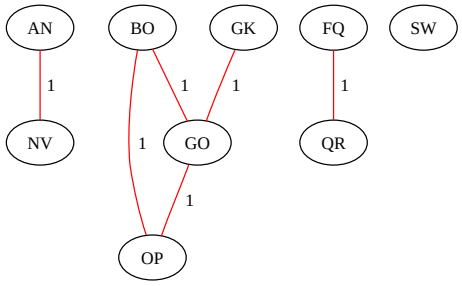


FIG. III.5 – Graphe de 8 règles (de la forme $X_i \rightarrow Y$) dont les prémisses contiennent 2 attributs. (seules les prémisses (X_i) sont présentées).

Soit R_1, ..., $R_{n'_K(j)}$ les $n'_K(j)$ prémisses du graphe et $\Pi_1^{n'_K(j)}$ une permutation des nœuds correspondant à l'ordre des nœuds rencontrés lors d'un parcours du graphe en profondeur (ou en largeur).

Lors de ce parcours, où chaque nœud n'est visité qu'une et une seule fois, il suffit de maintenir la liste des nœuds visités pour pouvoir calculer la borne inférieure recherchée.

Soit $voisins_{visite}(A_i)$ le nombre de voisins du nœud A_i déjà visités lorsque le nœud A_i est visité.

Le nombre de prémisses différentes de taille $K+1$ engendrées par les n'_K règles est inférieur ou égal à

$$binf_{precise}(p, K+1, n'_K(j)) = bsup(p, K+1, n'_K(j)) - \sum_{i=\Pi_1^{n'_K(j)}(1)}^{\Pi_1^{n'_K(j)}(n'_K(j))} voisins_{visite}(A_i)$$

PREUVE :

Les n'_K prémisses peuvent être représentées par un ensemble de graphes connexes où deux prémisses sont reliées par un arc si elles diffèrent exactement d'un attribut (voir Figure III.5).

Étant donné une prémisse A de taille K, le remplacement de chacun des K attributs de A par un des $(p-1-K)$ attributs non présents dans la prémisse engendre une prémisse de même taille et n'ayant qu'un attribut de différence avec A. Toutes ces prémisses représentent des voisins de A dans le graphe étudié. Le nombre de voisins de A est donc au plus égal à $K \times (p-1-K)$.

Lorsque le graphe est connexe, il existe un chemin permettant de visiter tout les nœuds du graphe, une et une seule fois, et sur ce chemin, deux nœuds consécutifs ne diffèrent que d'un attribut. Considérons un tel chemin : A_1, A_2, ..., $A_{p_{n'_K(j)}}$, la prémisse A_1 engendre $C^1_{p-K} = p-K$ prémisses de taille $K+1$ par ajout d'un attribut choisi parmi les $(p-K-1)$ attributs non présents dans A_1.

La prémisse A_2 ayant un attribut commun avec A_1, engendre elle aussi $(p-1-K)$ prémisses de taille $K+1$ dont une prémisse (celle contenant l'attribut partagé avec A_1) a déjà étant engendrée. Le nombre de nouvelles prémisses de taille $K+1$ engendrées par A_2 est donc égal à $(p-K-2)$.

La prémisse A_i engendre $(p-K-i)$ nouvelles prémisses de taille $K+1$ et non engendrées par les $i-1$ premières prémisses rencontrées sur le chemin.

En effet, si l'on considère une prémisse de taille K appartenant au graphe et ayant i voisins, cette prémisse engendrera $(p-1-K)$ prémisses dont i seront aussi engendrées par ces voisins. Ainsi, si les prémisses sont ordonnées selon l'ordre dans lequel elles sont rencontrées dans

le graphe lors d'un parcours en profondeur (ou en largeur) alors la $j^{ème}$ prémisse engendre $p-1-K-voisins_{visite}(A_j)$ nouvelles prémisses. Seuls les voisins de la prémisse A_j déjà visités lors du parcours ont engendré chacun une prémisse qui est aussi engendrée par A_j.

Notons $\Pi_1^{n'_K}$ la permutation des prémisses de l'ensemble des graphes correspondant au parcours effectué. Avec cette notation, $\Pi_1^{n'_K}(i)$ correspond à la $i^{ème}$ prémisse rencontrée lors du parcours. Le nombre de prémisses engendré par spécialisation des n'_K prémisses de taille K est donc égal à $bsup(p, K+1, n'_K) - \sum_{i=\Pi_1^{n'_K}(1)}^{\Pi_1^{n'_K}(n'_K)} voisins_{visite}(A_i)$.

♦

Cette borne, bien que très précise, nécessite la construction et le parcours du graphe des n'_K prémisses.

Ces premiers travaux concernant l'estimation de la complexité de notre algorithme nous ont permis de déterminer cette borne qui s'avère expérimentalement très précise. Cependant, la nécessité de construire le graphe des prémisses minimise l'intérêt et l'utilisation de cette borne inférieure. De nouveaux travaux sont donc nécessaires pour approfondir l'étude de la complexité. La connaissance ou l'encadrement précis du nombre de règles élaguées pourrait permettre une interaction plus souple avec l'utilisateur en lui indiquant que la recherche de certaines pépites de connaissance peuvent s'avérer plus rapide que d'autres.

4 Validation de l'approche

Pour valider notre algorithme, nous avons étudié deux types de données différentes : des données « académiques » et des données ancrées dans la vie réelle.

Dans la suite de cette thèse, nous considérons que des **données ancrées** (dans la vie réelle) sont des données issues de données réelles et qui intéressent un ou plusieurs experts. Par opposition, **des données non ancrées** sont des données, soit totalement artificielles, soit n'intéressant pas les experts.

Les données académiques présentent l'avantage majeur d'avoir été déjà très étudiées et nous pouvons donc disposer d'informations les concernant. Ces données sont disponibles, il est aisé de reproduire les expérimentations que nous avons réalisées. Le désavantage évident de ces données est de ne pas être ancrées dans la vie réelle et il est donc difficile de trouver des experts s'y intéressant vraiment. Ces données ne nous ont permis, dans un premier temps, que de tester et de valider notre approche d'un point de vue fonctionnel.

Pour obtenir une validation plus significative de nos résultats, nous devons avoir recours à un expert qui analysera les connaissances obtenues par notre système. Il est difficile de trouver un expert capable de bien évaluer l'intérêt des connaissances obtenues à partir de données non ancrées. Nous avons donc, en collaboration avec un expert, étudié des données ancrées, issues de textes et qui présentent le très net avantage d'intéresser notre expert. Ces données, ainsi que les résultats obtenus, sont présentées dans le chapitre V.

Nous ne présenterons donc dans ce chapitre que les résultats obtenus pour les données non ancrées.

4.1 Présentation des données

Avant d'utiliser notre méthode sur des bases de données réelles, nous avons choisi de la tester sur huit bases de données de l'UCI [Blake et Merz 1998]. Nous considérons que ces bases de données

Base	# d'attributs discrets	# d'attributs booléens	# d'enregistrements	Valeurs manquantes ?
car	7	25	1728	non
monks-1	7	19	432	non
monks-2	7	19	432	non
monks-3	7	19	432	non
mushrooms	23	125	8124	oui
nursery	9	32	12960	non
tic-tac-toe	10	29	958	non
votes	17	49	435	oui

TAB. III.2 – Quelques bases de données étudiées.

sont « artificielles » car bien que très utilisées pour évaluer de nombreux systèmes d'extraction de connaissances, elles n'en sont pas moins relativement éloignées des centres d'intérêts de la plupart des experts.

Les bases de l'UCI sont initialement prévues pour évaluer des systèmes d'apprentissage supervisés. Notre approche étant non supervisée, nous assimilons les classes contenues dans ces bases à de simples attributs. Ainsi, les conclusions des règles d'association obtenues ne sont pas limitées aux classes définies dans les données.

Dans cette série d'expériences, nous n'avons retenu que des bases de données discrètes, puis nous avons transformé toutes ces bases en bases de données booléennes. La transformation des attributs discrets en attributs booléens consiste simplement à créer pour chaque valeur discrète d'un attribut un nouvel attribut booléen.

Le Tableau III.2 présente le détail des huit bases étudiées.

Nous pouvons voir que la plupart des bases n'ont aucune valeur manquante, à l'exception des bases « mushrooms » et « votes ».

Dans la base « mushrooms » (qui est une base non bruitée), les valeurs manquantes correspondent à une mesure non effectuée. Nous les remplaçons par la valeur $Faux$, rendant ainsi compte de l'absence totale d'information pour l'individu concerné. Ce choix modifie légèrement les données et présente le seul défaut d'augmenter artificiellement la contradiction des règles.

Dans la base « votes » (qui représente les votes de la chambre des congrès des États-Unis d'Amérique pour l'année 1984), les valeurs manquantes correspondent à un vote non exprimé. Ces valeurs ont été remplacées par la valeur « non exprimé » qui s'ajoute aux valeurs discrètes possibles.

4.2 Résultats obtenus

Lors de la validation de notre approche, nous avons voulu mettre en évidence les deux points suivants :

1. montrer que le volume de règles engendrées par notre approche est inférieur au volume de règles engendrées par APRIORI, avec les mêmes contraintes initiales : $support > 0$ et $confiance > 0,5$;
2. montrer que l'ensemble des règles obtenues par notre approche représente une source de connaissances intéressantes pour l'expert du domaine ;

Lors de la première phase de la validation, le volume de règles extraites par notre algorithme est comparé avec les règles extraites par l'algorithme APRIORI. Les contraintes suivantes sont utilisées pour effectuer l'extraction :

– $support_{mini} = 0$

Base	APRIORI	notre approche
car	1617	5
monks-1	1935	17
monks-2	1796	20
monks-3	1987	17
mushrooms	680088	386
nursery	3319	7
tic-tac-toe	8094	32
votes	17681	232

TAB. III.3 – Taille des ensembles de règles obtenues.

Bases	APRIORI			notre approche		
	$K=1$	$K=2$	$K=3$	$K=1$	$K=2$	$K=3$
car	30	281	1306	2	3	0
monks-1	46	366	1523	9	8	0
monks-2	50	343	1403	10	10	0
monks-3	46	372	1569	9	8	0
mushrooms	1727	48316	630045	265	121	0
nursery	22	378	2919	4	3	0
tic-tac-toe	28	802	7264	0	32	0
votes	628	14338	161925	114	118	0

TAB. III.4 – Détail des ensembles de règles obtenus.

- $confiance_{mini} = 0,5$
- $K_{max} = 3$

Les résultats obtenus sont présentés dans le tableau III.3.

Nous pouvons voir que le volume de règles engendré par notre approche est nettement inférieur à celui obtenu par une approche de recherche exhaustive des règles d'association. L'application d'un post-traitement, reproduisant le comportement de notre algorithme, sur l'ensemble des règles engendrées par APRIORI, en appliquant les contraintes suivantes : $support > 0$ et $confiance > 0.5$, doit produire les mêmes règles que celle obtenues par notre approche. L'un des avantages de notre approche réside dans le fait que nous ne sommes pas contraints de produire l'intégralité des règles pour ensuite devoir les élaguer à l'aide d'heuristiques. Ainsi, les spécialisations des règles les « moins contradictoires » ne sont pas engendrées.

Les résultats obtenus sont détaillés dans le tableau III.4.

La deuxième phase de la validation nécessite l'expertise d'un spécialiste du domaine. Il est indispensable de faire valider les ensembles de règles obtenus par un expert, de manière à valider l'approche globale. Le faible nombre de règles obtenues n'est pas garant de la qualité de celles-ci, seul un expert peu apprécier la qualité des règles.

Malheureusement, les experts sont rares et nous ne disposons pas de spécialistes pour les bases de données étudiées. Cependant, en utilisant nos connaissances généralistes, nous avons pu valider les règles obtenues à partir des bases *car*, *nursery* et *tic-tac-toe*. Ces règles, peu nombreuses et facilement interprétables pour un néophyte, représentent des connaissances valides (pour *car* et *nursery*) mais non nouvelles, compte tenu de nos connaissances. Notre manque d'expertise sur ces données ne nous

permet pas de conclure sur la qualité des règles obtenues par rapport aux connaissances « attendues » de ces bases. Les règles obtenues à partir de *tic-tac-toe* ne sont pas très informatives.

Ces résultats ne permettent pas d'évaluer la capacité de notre algorithme à extraire les pépites de connaissances. Seuls des résultats validés par des experts peuvent être considérés comme une validation positive pour notre approche.

Nous verrons dans le chapitre V que nous avons obtenu une telle validation sur des données issues de textes.

5 Extraction des règles les plus surprenantes

Dans cette section, nous proposons une solution permettant d'extraire les règles d'association les plus surprenantes, étant donné un ensemble de règles. Cette solution présentée dans [Lerman et Azé 2003] avait déjà été exprimée dans [Lerman et al. 1981] mais sous une forme sensiblement moins élaborée et moins justifiée.

Pour obtenir les règles les plus surprenantes, nous proposons d'utiliser une mesure probabiliste fondée sur l'Intensité d'Implication Classique présentée dans le chapitre II.

Rappelons que l'intensité d'implication possède le défaut de tendre rapidement vers 0 ou vers 1 dès que la taille des données devient trop importante.

[Lerman et Azé 2003] présentent une solution permettant d'arriver à un indice probabiliste d'implication discriminant quelle que soit la taille des données (n). Cette solution est fondée sur la réduction globale des similarités implicatives de la forme $q_3(A, \overline{B})$ (equation II.14, page 32).

Nous présentons le principe de la méthode dans la section suivante, puis nous montrons une évaluation du comportement du nouvel indice proposé par rapport à la taille de n. Les expérimentations réalisées, ainsi que les résulats obtenus, sont présentés dans la section 5.2.

5.1 Similarité implicative de vraisemblance du lien dans le contexte

Étant donné un couple d'attributs (A, B), nous proposons d'évaluer l'intérêt de la règle d'association $A \rightarrow B$ par rapport à un ensemble de règles vérifiant les mêmes propriétés que la règle $A \rightarrow B$. Cette méthode, proche de celle utilisée pour extraire les règles d'association, permet étant donné un ensemble de règles d'association homogène (c'est-à-dire que les règles de l'ensemble vérifient toutes les mêmes propriétés) d'évaluer les règles d'association les plus intéressantes.

Un des problèmes posés consiste dans le choix de la base formée par les couples d'attributs qui va servir à la normalisation globale. Celui qui a été retenu dans l'analyse expérimentale qui suit a consisté, étant donnée une base de données de n individus décrits par un ensemble $\mathcal{A}$ d'attributs, à considérer tous les couples distincts d'attributs (A, B) avec $A \in \mathcal{A}$ et $B \in \mathcal{A}$.

Dans [Lerman et al. 1981] un choix plus sélectif a été préconisé. La contrainte $n(A) < n(B)$ a été imposée sur les couples d'attributs retenus de sorte que l'absortion totale de $\mathcal{C}(A)$ par $\mathcal{C}(B)$ puisse se réaliser.

L'usage veut qu'on ne puisse avoir à considérer une implication de la forme $A \rightarrow B$, que si le support et la confiance de la règle $A \rightarrow B$ sont respectivement supérieurs à des seuils s_0 et c_0 que l'expert peut fixer [Agrawal et al. 1993, Guillaume 2000, Bastide et al. 2002].

Ainsi, nous ne retenons que les couples d'attributs (A, B) vérifiant les contraintes suivantes :
- $n(A) < n(B)$
- $support(A \rightarrow B) \geq s_0$
- $confiance(A \rightarrow B) \geq c_0$

Notons $\mathcal{G}$ l'ensemble des couples d'attributs vérifiant les contraintes ci-dessus.

$$\mathcal{G} = \{(A,B) | A \neq B, n(A) < n(B), support(A \rightarrow B) \geq s_0, confiance(A \rightarrow B) \geq c_0\}$$

Nous proposons donc de calculer, pour chacun des couples d'attributs de $\mathcal{G}$, l'indice d'implication $q_3(A, \overline{B})$, puis de normaliser les valeurs obtenues par rapport à la valeur moyenne et à l'écart type des indices obtenus.

Le nouvel indice proposé et noté « indice d'implication global », par opposition à l'indice d'implication q_3 qui est « local » aux règles, s'exprime donc de la manière suivante :

$$q_3^g(A, \overline{B}) = \frac{q_3(A, \overline{B}) - moy_e\{q_3|\mathcal{G}\}}{\sqrt{var_e\{q_3|\mathcal{G}\}}} \tag{III.2}$$

où $moy_e\{q_3|\mathcal{G}\}$ et $var_e\{q_3|\mathcal{G}\}$ désignent respectivement la moyenne et la variance empirique de q_3 sur $\mathcal{G}$.

Ainsi, l'indice probabiliste discriminant (de vraisemblance du lien) d'une règle d'association $(A \rightarrow B)$ issue de l'ensemble $\mathcal{G}$, s'exprime par

$$\mathcal{J}_n(A,B) = 1 - \Phi[q_3^g(A, \overline{B})] \tag{III.3}$$

où Φ désigne, comme pour l'intensité d'implication, la fonction de répartition de la loi normale centrée et réduite.

Nous appelerons ce nouvel indice, $\mathcal{J}_n$, **l'Intensite d'Implication Normalisée**.

5.2 Résultats expérimentaux

5.2.1 Le protocole expérimental

Nous avons réalisé deux séries d'expériences sur la base de données « mushrooms ».

Le choix de cette base est lié au « désir » d'observer le comportement des indices testés sur des données « réelles » et non bruitées.

Le protocole expérimental utilisé est le suivant :

1. Isoler les couples (A, B) tel que $n(A) < n(B)$
2. Pour chacun de ces couples, calculer $q_3(A, \overline{B})$ et $\varphi(A, B)$
3. Ensuite, calculer la moyenne et l'écart-type de q_3 pour l'ensemble des couples (A, B) retenus et normaliser l'indice q_3 avant de faire appel à la loi normale.

Lors de la première série d'expériences, la sélection des couples (A, B) n'est conditionnée que par la contrainte suivante : $n_A \neq 0$, $n_B \neq 0$ et $n_{AB} \neq 0$.

Dans la deuxième série d'expériences, les contraintes appliquées aux couples sélectionnés sont liées au support et à la confiance de chacun des couples considéré. Ainsi, les couples (A, B) retenus, dans la deuxième série d'expériences, sont tels que $support(A \rightarrow B) \geq s_0$ et $confiance(A \rightarrow B) \geq c_0$.

Cette seconde série d'expériences permet d'observer le comportement des différents indices étudiés sur des couples ayant de meilleures propriétés (au sens de la fouille de données classique).

L'objectif principal de ces expériences est d'observer le comportement des indices lorsque la taille de $\mathcal{O}$ augmente sans que les cardinaux de $\mathcal{O}(A)$, $\mathcal{O}(B)$ et $\mathcal{O}(A) \cap \mathcal{O}(B)$ soient modifiés. Ainsi,

nous choisissons un couple (A, B) satisfaisant les contraintes du protocole et nous augmentons la valeur de n, sans modifier les valeurs n_A, n_B et n_{AB}. Tout se passe comme si on ajoutait des objets fictifs où tous les attributs sont à $Faux$.

L'algorithme 4 permet de réaliser ces expériences.

Algorithme 4 Évaluation du comportement de l'IIC et de l'IIN par rapport à la taille des données.

Entrée :

$\mathcal{G}$: ensemble des couples (A, B) satisfaisant les contraintes de l'expérience

$seuil_n$: valeur maximale pour n

Début

Tant que $(n < seuil_n)$ **Faire**

Pour tout $(A, B) \in E$ **Faire**

calcul de $q_3(A, \overline{B})$

Fin Pour

Calcul de la moyenne et de l'écart-type des valeurs de l'indice q_3 obtenues

Pour tout $(A, B) \in E$ **Faire**

Calcul de l'intensité d'implication « classique » (IIC)

Calcul de l'intensité d'implication normalisée (IIN)

Fin Pour

$n \leftarrow n + constante$

Fin Tant que

Fin

Les résultats obtenus pour les deux séries d'expériences réalisées sont détaillés dans la section suivante.

5.2.2 Première série d'expériences

La première série d'expériences (aucune contrainte sur les couples (A, B)) a permis de mettre en évidence plusieurs propriétés intéressantes de l'indice normalisé.

- cet indice se montre discriminant quel que soit la valeur de n
- de plus les résultats montrent que le comportement de l'indice est plus soutenu relativement aux implications $A \rightarrow B$ lorsque $n_A < n_B$

Nous pouvons voir, sur les Figures III.6 et III.7, que l'indice normalisé présente un comportement discriminant, étant donnés les caractéristiques cardinales des attributs A et B mis en jeu, contrairement à l'indice classique (toujours égal à 1).

Enfin, comme le montre la Figure III.8, l'indice normalisé tend vers des valeurs relativement modérées, contrairement à l'indice classique (qui tend toujours vers 1), même lorsque la valeur de n_{AB} est relativement faible par rapport à n_A (voir la Figure III.9 associée au résultat III.8).

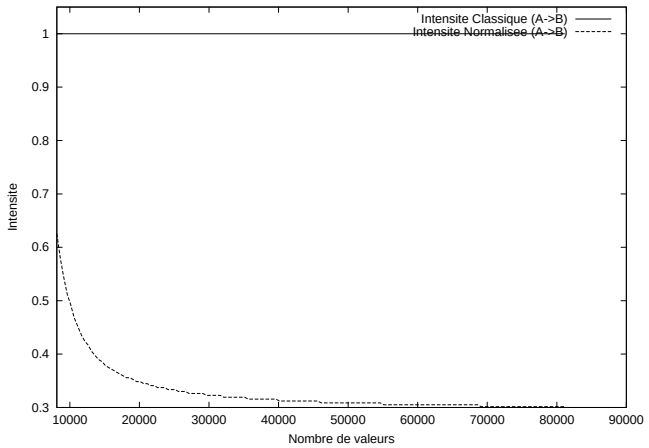


FIG. III.6 – Cas où $n_A = 192$, $n_B = 1202$, $n_{AB} = 96$, $s_0 = 0$, $c_0 = 0$.

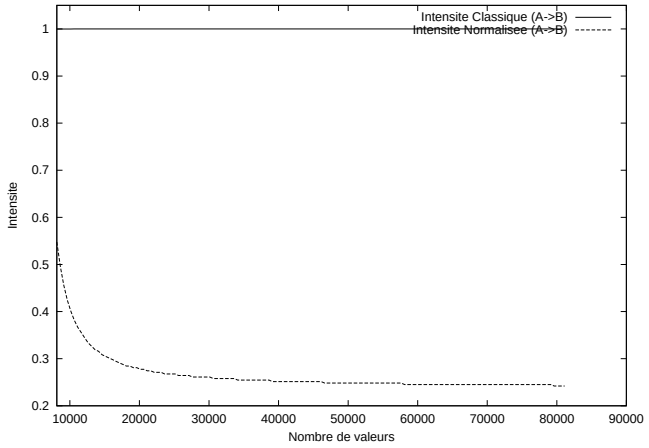


FIG. III.7 – Cas où $n_A = 192$, $n_B = 828$, $n(A \wedge b) = 64$, $s_0 = 0$, $c_0 = 0$.

5.2.3 Deuxième série d'expériences

La deuxième série d'expériences permet de se focaliser sur des couples (A, B) ayant des propriétés de support et de confiance définies par l'utilisateur. Pour l'ensemble des expériences réalisées, nous avons fixé $s_0 = 0,1$ et $c_0 = 0,9$. Ainsi, nous ne retiendrons que les couples (A, B) tels que $support(A \rightarrow B) \geq 0,1$ et $confiance(A \rightarrow B) \geq 0,9$.

Ces seuils correspondent à des seuil relativement usuels en extraction de règles d'association et des valeurs similaires ont été souvent utilisées pour la base de données « mushrooms » ([Lehn 2000, Bastide et al. 2002].

Les résultats obtenus montrent que sur cet ensemble réduit de couples (A, B), l'indice normalisé se montre toujours plus discrimant que l'indice classique. De plus, comme le montre la Figure III.11, l'indice normalisé se montre plus discriminant dans ce cadre expérimental que dans la première série d'expériences (Figure III.12). Ces courbes correspondent au cas présenté dans la Figure III.10.
Dans la deuxième série d'expériences, les couples retenus correspondent à des relations « fortes » et la relation présentée sur la Figure III.10 devient de moins en moins « intense » lorsque n augmente, comparativement aux autres relations présentes. Ainsi, l'opération de filtrage (basée sur l'utilisation

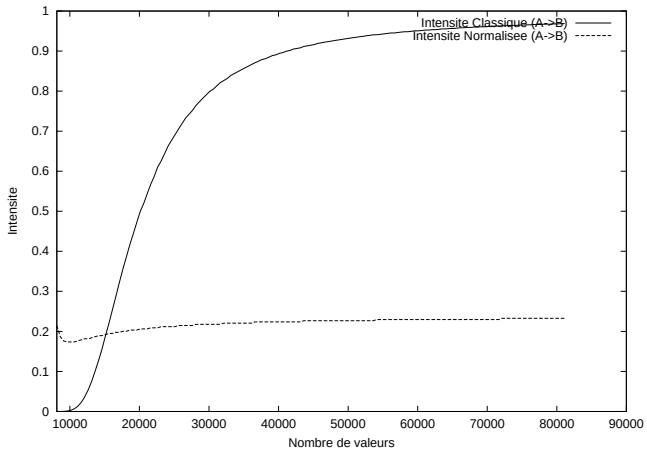


FIG. III.8 – Cas où $n_A = 452$, $n_B = 2320$, $n_{AB} = 52$, $s_0 = 0$, $c_0 = 0$.

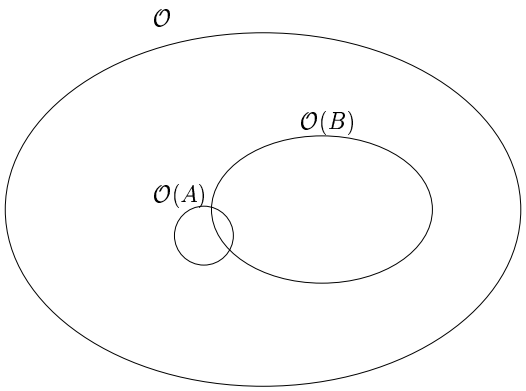


FIG. III.9 – Cas où $n_A = 452$, $n_B = 2320$, $n_{AB} = 52$.

des seuils s_0 pour le support et c_0 pour la confiance) permet de retenir un ensemble de relations ayant des fortes valeurs pour les indices étudiées. L'association que nous étudions, Figure III.10, devient donc de moins en moins pertinente, relativement aux autres relations respectant les conditions s_0 et c_0 imposées, lorsque la valeur de n augmente significativement.

Nous allons étudier relativement au protocole expérimental ci-dessus, la situation d'inclusion totale définie par $n_A = 1296$, $n_B = 2304$ et $n_{AB} = 1296$, représentée par la Figure III.10. La Figure III.11 montre l'évolution de l'indice normalisé, calculé dans le contexte, lorsque le nombre d'objets augmente à partir de $n = 8124$ par adjonction d'objets où tous les attributs sont à $Faux$. De toute façon, la valeur de l'indice se trouve discriminée. Elle demeure forte, supérieur à $0,98$, pour n croissant jusqu'à 80000. Néanmoins, la valeur de l'indice va en décroissant vers un palier pour n augmentant. C'est que, pour n devenant grand, l'implication ci-dessus devient moins saisissante relativement à d'autres implications. Ces dernières peuvent ne pas correspondre à des inclusions totales ; mais elles doivent concerner des situations où n_A et n_B sont sensiblement plus gros et proches.

Relativement à la dernière série d'expériences et à la Figure III.12, il y a lieu d'ajouter qu'il ne faut pas s'étonner que la valeur de l'indice normalisé puisse tomber assez bas lorsque n augmente. On se trouve en effet dans le contexte du graphe $\mathcal{G}$ des couples d'attributs (A, B) où la relation

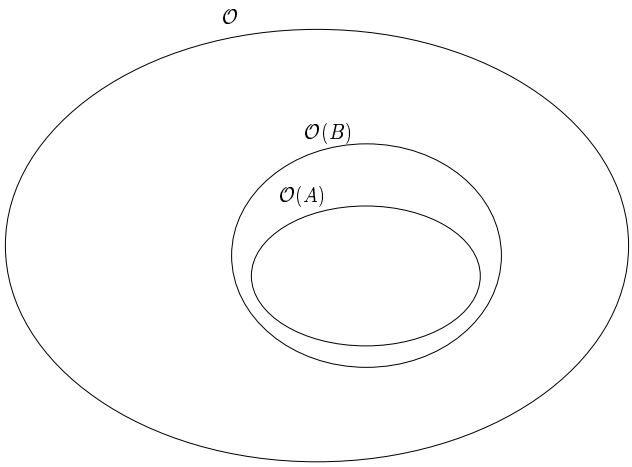


FIG. III.10 – cas où $n_A = 1296$, $n_B = 2304$, $n_{AB} = 1296$.

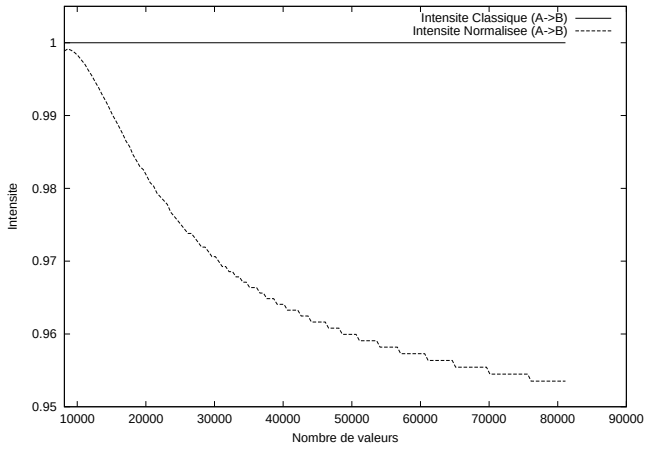


FIG. III.11 – cas où $n_A = 1296$, $n_B = 2304$, $n_{AB} = 1296$, $s_0 = 0$, $c_0 = 0$.

d'implication est très forte. Ainsi, en prenant $n = 80000$ (par l'adjonction d'objets, où tous les attributs sont à $Faux$), on a pour chaque (A, B) retenu, $n_{AB} \geq 8000$ et $\frac{n_{AB}}{n_A} \geq 0,9$. Les indices se calculent sur la base des 8124 objets initiaux. Ils correspondent à des quasi inclusions d'un très gros $\mathcal{O}(A)$ dans un à peine moins gros $\mathcal{O}(B)$.

Lorsque le seuil minimal de confiance est abaissé à $0,5$, nous pouvons constater sur la Figure III.13 que la relation étudiée devient plus significative par rapport à l'ensemble de relations considéré. Ce comportement est attendu et lié au fait que l'ensemble de relations étudié contient des relations beaucoup moins fortes que dans le cas de la Figure III.12. Ce nouvel indice permet donc de mieux rendre compte du degré d'étonnement d'une règle par rapport aux autres règles étudiées.

5.3 Validation comparative du nouvel indice

Nous avons aussi comparé ce nouvel indice avec la moindre contradiction, ceci dans le but d'évaluer la capacité de cet indice à extraire des règles intéressantes. Ces deux méthodes sont conçues pour éviter le piège des règles triviales presque toujours vérifiées sur la majorité des données. Nous avons donc voulu comparer les règles fournies par ces deux méthodes sur la base de données

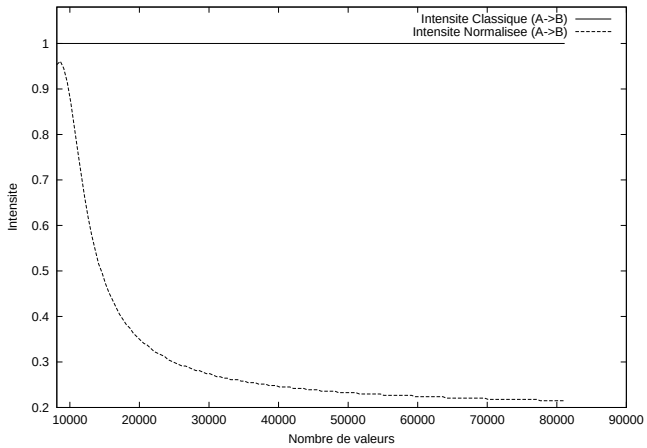


FIG. III.12 – cas où $n_A = 1296$, $n_B = 2304$, $n_{AB} = 1296$, $s_0 = 0,1$, $c_0 = 0,9$.

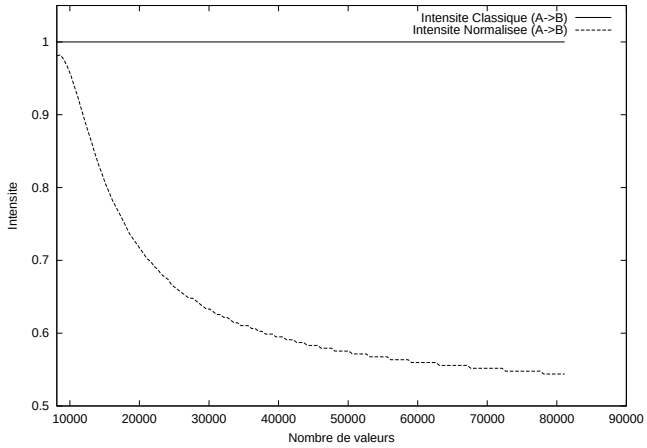


FIG. III.13 – cas où $n_A = 1296$, $n_B = 2304$, $n_{AB} = 1296$, $s_0 = 0,1$, $c_0 = 0,5$.

« mushrooms » pour voir si elles se comportaient de manière similaires ou non.

Nous avons extrait de cette base les règles d'association de la forme $A \rightarrow B$, c'est-à-dire contenant exactement un attribut en prémisse et un en conclusion.

Toutes les règles $A \rightarrow B$ ayant un support non nul et une confiance strictement supérieure à 50% ont été engendrées. Notons $\mathcal{E}_R$ cet ensemble de règles. Puis, la moindre contradiction et l'Intensité d'Implication Normalisée ont été utilisées pour filtrer l'ensemble $\mathcal{E}_R$. Seules les règles dont la mesure, après normalisation par centrage-réduction, était supérieure à 1 ont été conservées.

Étant donnée cette contrainte (déjà introduite et utilisée dans l'algorithme EXTRAIREPÉPITES), la moindre contradiction trouve 224 règles et l'Intensité d'Implication Normalisée 363.

Parmi celles-ci, 108 règles sont communes aux deux mesures, ce qui confirme bien notre hypothèse qu'elles se comportent de façon comparable.

Parmi ces règles communes, le cas général est celui d'une règle avec un support compris entre $0,2$ et $0,5$ et une confiance comprise entre $0,7$ et 1. Notons que $0,2$ de support signifie que tout de même plus de 1600 exemples sur les 8124 de la base vérifient cette règle.

Quelques règles ont une moindre contradiction très faible (voire nulle) et un support « minuscule » de $0,024$ ce qui correspond à 192 instances dans la base. La plus contredite d'entre elles

est telle que $A \rightarrow B$ est soutenue par 256 exemples et contredite par 36, ceux-ci soutenant donc $A \rightarrow \overline{B}$. De telles règles avec un si petit support sont peut-être bien dues au bruit comme la plus contredite d'entre elles, mais il faut avouer que les règles affirmées par 192 cas et contredites par aucun méritent au moins un examen sérieux par un expert, dans le cas d'une base ancrée.

Nous ne pouvons donc que nous féliciter du fait que les deux mesures étudiées ont été capables de détecter de telles règles représentant de possibles pépites de connaissance.

Inversement, la moindre contradiction ne trouve qu'une seule règle un peu triviale, avec un support de $0,97$. Cette règle triviale présente une Intensité d'Implication Normalisée de $0,758$, ce qui en fait la règle acceptée par l'Intensité d'Implication Normalisée ayant la plus faible des valeurs observées. Ceci confirme bien que l'Intensité d'Implication Normalisée tend à éliminer ce type de règles en général.

Quant aux règles qui ne sont détectées que par l'une des deux méthodes, elles présentent une tendance générale à avoir un support très faible dans le cas de l'Intensité d'Implication Normalisée. Dans ce cas, sur les 255 règles concernées, 69 ont un support inférieur à $0,1$; les autres règles ont un support compris entre $0,1$ et $0,31$.

Inversement, les règles détectées par la moindre contradiction ont tendance à avoir un support un peu plus élevé. Sur les 116 règles concernées, 32 ont un support inférieur à $0,1$ (c'est-à-dire une proportion comparable à l'Intensité d'Implication Normalisée), mais 71 ont un support supérieur à $0,4$.

Ceci confirme donc le rôle joué par chacune de ces deux mesures. L'Intensité d'Implication Normalisée va détecter des règles à très faible support, mais peut accepter des règles assez contredites (nous en verrons un exemple sur les données ancrées, chapitre V, section 7). La moindre contradiction aura tendance à accepter des règles de plus fort support et à éliminer celles de très faible support mais, bien entendu, sélectionne des règles très peu contredites dans les données.

De cette expérimentation sur des données non ancrées, il ressort que ces deux mesures sont capables de détecter des pépites de connaissance de nature légèrement différentes. Un premier passage devrait se focaliser sur les règles considérées comme les meilleures par les deux méthodes, mais on peut envisager de les utiliser en séquence pour ne pas risquer d'oublier une pépite de connaissance.

6 Conclusion

Nous avons vu dans ce chapitre un algorithme permettant d'extraire, à partir de données booléennes, des règles d'association pouvant avoir de très faibles supports. La méthode proposée est fondée sur l'utilisation de la mesure de qualité nommée la « moindre contradiction » et n'utilise pas, contrairement à la plupart des approches classiques, un support minimal associé aux règles recherchées. Nous avons choisi de ne pas utiliser cette contrainte car nous pensons qu'il n'est pas toujours aisé pour l'expert de définir un tel support. De plus, certaines règles d'association peuvent avoir un faible support et être très intéressantes pour l'expert.

Ces règles, que nous avons appelées des pépites de connaissance, sont souvent inconnues de l'expert et sont, soit indétectables par les systèmes classiques, soit difficilement détectables dans l'ensemble des innombrables règles souvent trouvées par les systèmes classiques. Notre algorithme garantit par construction de fournir relativement peu de règles car seules les règles d'association ayant une moindre contradiction suffisament élevée par rapport aux autres règles sont retenues et proposées à l'expert.

Les expérimentations que nous avons réalisées montrent que les pépites de connaissance obtenues sont effectivement peu nombreuses et qu'elles représentent relativement souvent des connaissances intéressantes.

Ce chapitre présente aussi une solution permettant de corriger l'un des défauts de l'Intensité d'Implication. Rappelons que cette mesure de qualité permet de mesurer l'étonnement statistique d'observer peu de contre exemples pour une règle $A \rightarrow B$ par rapport à la valeur attendue sous l'hypothèse d'indépendance entre A et B. Le défaut majeur de cette mesure est de converger rapidement vers ses valeurs extrêmes (0 ou 1) dès que les données étudiées deviennent trop volumineuses. Cependant, les fondements statistiques de l'Intensité d'Implication font de cet indice une mesure particulièrement intéressante pour évaluer l'intérêt des connaissances obtenues. La solution que nous avons proposée dans ce chapitre pour diminuer l'impact de la taille des données sur cet indice permet de conserver les propriétés statistiques de ce dernier. Nous pouvons alors utiliser cette nouvelle mesure, nommée l'Intensité d'Implication Normalisée, pour extraire des connaissances et les comparer entre elles quelque soit la taille des données.

Nous avons réalisé des expériences comparatives entre l'Intensité d'Implication Normalisée et la moindre contradiction dans le but d'évaluer la capacité de ces deux mesures à extraire des pépites de connaissance. Les premiers résultats obtenus, et confirmés par ceux présentés dans le chapitre V, montrent que la moindre contradiction se comporte bien par rapport à l'Intensité d'Implication Normalisée. La simplicité de la moindre contradiction par rapport à l'aspect statistique de l'Intensité d'Implication Normalisée n'altère en rien ses capacités à extraire des connaissances fiables et utiles.

Notons toutefois que nous étant débarrassés du seuil de support minimal permettant de contrôler les règles d'association détectées, nous prenons le risque réel de détecter des connaissances ayant un faible support, une forte valeur pour la mesure de qualité utilisée et étant pourtant issues d'un phénomène aléatoire tel que du bruit. Il devient alors nécessaire d'étudier l'impact des données bruitées sur les connaissances obtenues, ceci pour envisager des solutions permettant de détecter les règles potentiellement bruitées et de prévenir l'expert lors de l'analyse de celles-ci.

Le chapitre suivant présente une étude du bruit et de son impact sur les connaissances extraites à partir des données. Plusieurs solutions permettant d'atténuer les effets des données bruitées sur les connaissances obtenues y sont proposées et étudiées.

Chapitre IV

Étude de l'influence du bruit sur les connaissances

1 Nature du bruit

L'analyse et l'étude de données issues de mondes réels implique de prendre en considération la nature des processus ayant engendré les données. Dans le monde académique, les données sont supposées parfaites et donc le problème du bruit n'est quasiment jamais abordé. Lorsqu'un système de fouille de données est destiné à travailler avec des données réelles, il devient très important d'étudier le comportement du système en présence de données bruitées. En effet, les données issues du monde réel, tels que des capteurs de systèmes de productions industriel, des informations relatives à des patients dans un milieu hospitalier, etc. sont très rarement parfaites.

Les imperfections liées à ces données peuvent être de plusieurs natures :

- données manquantes : capteur défecteux de temps en temps
- absence des données utiles : examen pertinent non réalisé pour un patient
- données erronées : capteur déreglé, mauvaise lecture des instruments
- *etc.*.

Dans le cadre de la fouille de données supervisée, de nombreuses études sur les données bruitées et l'impact de ces données ont été réalisées.

Par contre, dans le cadre de la fouille de données non supervisée, très peu d'études ont été réalisées. Ceci est probablement lié au fait qu'il est très difficile de détecter du bruit dans ce cadre de travail.

Certaines formes de bruit ne peuvent être traitées de manière automatique. Si l'on suppose l'existence d'un système capable de détecter l'absence des données utiles, le système pourra alors indiquer à l'utilisateur que ses données ne contiennent probablement pas les informations utiles. Mais en aucun cas, le système ne sera capable d'indiquer à l'utilisateur quelles sont les « bonnes » informations à ajouter. Cette forme de bruit rend les données difficilement exploitables par un système de fouille de données. Nous ne l'étudierons donc pas dans la suite de cette thèse.

Pour traiter le cas des données manquantes, nous proposons classiquement soit de remplacer les données manquantes par la valeur moyenne observée sur l'ensemble des données pour l'attribut affecté, soit de créer un nouvel attribut pour ces informations manquantes.

Nous nous focaliserons, dans la suite de cette thèse, sur une seule forme de bruit : *les données erronées*.

2 Étude de l'impact du bruit lié aux données erronées

Lorsque les données étudiées sont bruitées et plus précisément, lorsque les valeurs de certains attributs sont erronées, la recherche de règles d'association dans de telles données risque de fournir des règles incorrectes.

Considérons une base de données $\mathcal{B}$ non bruitée et l'ensemble des règles d'association trouvées à partir de $\mathcal{B}$ en utilisant une mesure de qualité m : $E^m_{\mathcal{B}}$.

Si l'on suppose qu'un phénomène extérieur altère $\mathcal{B}$, nous obtenons alors une nouvelle base $\mathcal{B}'$ qui correspond à une version bruitée de $\mathcal{B}$. En appliquant sur $\mathcal{B}'$ les mêmes techniques de recherche de règles d'association que celles appliquées sur $\mathcal{B}$, nous obtenons un nouvel ensemble de règles d'association : $E^m_{\mathcal{B}'}$.

L'impact du bruit ayant altéré $\mathcal{B}$ peut se mesurer soit directement sur les données, soit sur les connaissances obtenues à partir des données. Dans le premier cas, il suffit de mesurer la différence entre $\mathcal{B}$ et $\mathcal{B}'$ pour estimer le taux de données erronées.

Dans le second cas, il suffit d'observer les différences entre les ensembles $E^m_{\mathcal{B}}$ et $E^m_{\mathcal{B}'}$ pour mesurer l'effet du bruit sur les règles d'association détectées avec la mesure m.

Lorsque nous travaillons avec des données réelles, nous ne savons pas si les données sont bruitées ou non. Il est donc impossible de calculer l'impact réel du bruit.

Par contre, nous pouvons estimer l'impact du bruit sur les connaissances extraites en supposant les données correctes et en étudiant les effets d'un bruit introduit artificiellement dans ces données.

Nous sommes alors ramené à étudier la différence entre $E^m_{\mathcal{B}}$ et $E^m_{\mathcal{B}'}$.

Si on suppose que $E^m_{\mathcal{B}}$ représente les connaissances non bruitées, alors

- les règles contenues dans $E^m_{\mathcal{B}} \cap E^m_{\mathcal{B}'}$ sont les règles correctement retrouvées dans $\mathcal{B}'$
- les règles contenues dans $E^m_{\mathcal{B}} - E^m_{\mathcal{B}'}$ sont les règles correctes et non trouvées dans $\mathcal{B}'$
- et enfin, les règles contenues dans $E^m_{\mathcal{B}'} - E^m_{\mathcal{B}}$ sont les règles erronées trouvées dans $\mathcal{B}'$ et absentes de $\mathcal{B}$

La Figure IV.1 présente les trois groupes de règles obtenues en étudiant les données « saines » et les données bruitées.

L'impact du bruit sur un système d'extraction de règles d'association peut donc se mesurer en calculant le pourcentage de règles correctes perdues et le pourcentage de règles erronées trouvées.

Nous présentons dans la section suivante plusieurs méthodes permettant de bruiter les données. Ces méthodes correspondent à des formes de bruit légèrement différentes.

3 Différentes formes de bruit étudiées

Comme nous étudions le problème de la détection de règles d'association, tous les attributs peuvent se trouver soit dans la prémisse, soit dans la conclusion d'une règle. Ainsi, l'ensemble des attributs peut être affecté par le bruit, mais un attribut bruité ne peut modifier que les règles contenant cet attribut donné. C'est pourquoi nous nous sommes focalisés sur l'étude de trois différents types de bruit.

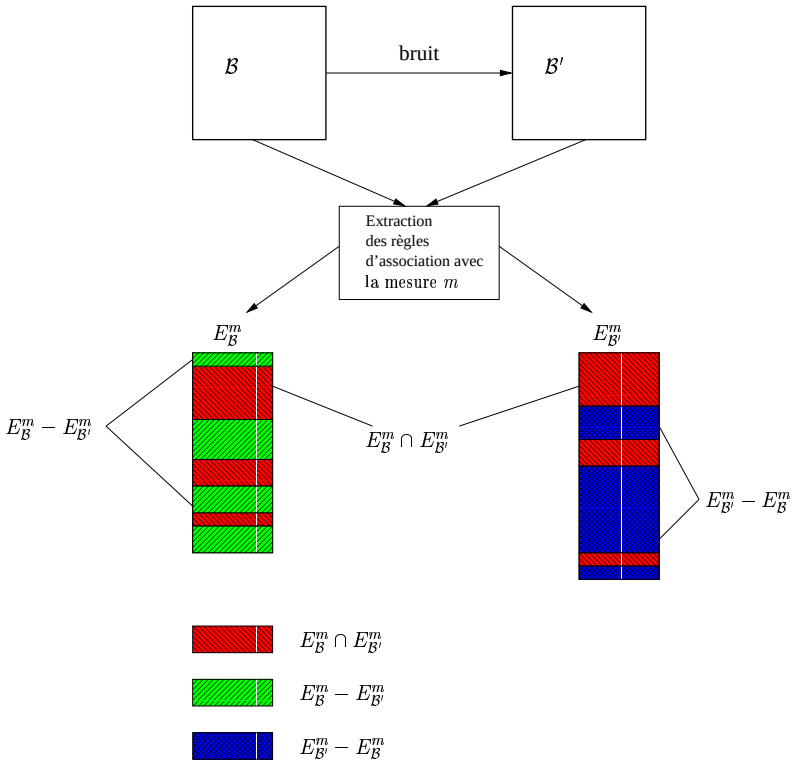


FIG. IV.1 – Étude de l'impact du bruit sur les règles d'association.

Quelque soit la forme de bruit étudiée, le choix des couples (individu, attribut) dont la valeur est modifiée (pour simuler l'apparition de bruit dans les données) est réalisé par un tirage uniforme et sans remise.

3.1 Un seul descripteur bruité

La première méthode consiste à introduire du bruit dans un attribut X, par exemple, et à étudier les effets du bruit sur les règles contenant cet attribut donné. L'ensemble des règles ne contenant pas l'attribut X n'est pas modifié et l'effet du bruit est lié au nombre de règles contenant l'attribut X. De manière à préciser notre approche, considérons le cas d'une règle liant deux attributs binaires, par exemple $X = Vrai$ et $Y = Vrai$ (ce type de règle sera celui que nous étudierons dans la suite de cette section). Supposons que cet attribut X est bruité, avec 5% de bruit. Nous introduisons le bruit en renversant 5% des valeurs de l'attribut X, c'est-à-dire en changeant les valeurs $Vrai$ par $Faux$ et inversement. La quantité de règles telles que $X = Vrai$ et $Y = Vrai$ est alors modifiée.

En fait, nous observons deux changements dûs à ce type de bruit :

- Quand une valeur $Vrai$ devient égale à $Faux$, des règles contenant $X = Vrai$ et $Y = Vrai$ vont disparaître.
- Quand une valeur $Faux$ devient égale à $Vrai$, des règles contenant $X = Vrai$ et $Y = Vrai$ vont apparaître.

Avec cette première méthode, nous pouvons isoler le bruit et comprendre ses effets. La proportion de règles qui disparaissent est liée à la présence de règles contenant $X = Vrai$ et $Y = Vrai$ dans les données ; inversement, la proportion de règles créées par le bruit est liée à la présence de règles contenant $X = Faux$ et $Y = Vrai$ dans les données.

L'algorithme 5 permet d'introduire ce type de bruit dans une base de données.

Algorithme 5 *IntroduireBruit-a*$(\mathcal{B}_n^p, j, \alpha_{noise})$

Entrée :

$\mathcal{B}_n^p$: base de données à bruiter contenant n individus décrits par p attributs

j : indice de l'attribut bruité (entre 0 et 1)

α_{noise} : bruit introduit dans les données

Sortie :

$\mathcal{B}'$: base de données bruitée

Début

$nb\_objets\_a\_bruiter \leftarrow int(\alpha_{noise} * p)$ *-- int(x) : partie entière de (x)*

$\mathcal{B}' \leftarrow \mathcal{B}$

Pour $(i \leftarrow 1\,;\, i \leq n\,;\, i++)$ **Faire**

$nonBruite[i] \leftarrow 1$

Fin Pour

Pour $(k \leftarrow 1\,;\, k \leq nb\_objets\_a\_bruiter\,;\, k++)$ **Faire**

Répéter

$i \leftarrow$ entier choisi aléatoirement entre 1 et n

Jusqu'à ce que $(nonBruite[i] = 1)$

$nonBruite[i] \leftarrow 0$

$\mathcal{B}'(\alpha_i^j) \leftarrow \overline{\mathcal{B}(\alpha_i^j)}$ *-- introduction du bruit en inversant la valeur de* $\mathcal{B}(\alpha_i^j)$

Fin Pour

Retourner $\mathcal{B}'$

Fin

3.2 Répartition aléatoire du bruit dans les données

La seconde méthode consiste à introduire le bruit de manière aléatoire dans la base de données. Cette fois, tous les couples (individus, attributs) peuvent donc être bruités. Le but de cette expérience est de montrer comment les différentes mesures réagissent lorsque les attributs sont bruités de manière aléatoire avec un faible pourcentage de bruit.

L'algorithme 6 permet d'introduire ce type de bruit dans les données.

3.3 Quelques attributs bruités

La troisième méthode consiste à introduire différents niveaux de bruit sur quelques attributs de la base. Le but de cette expérience est d'étudier la sensibilité au bruit des mesures lorsque les données sont globalement peu bruitées (1%) mais que ce bruit est lié à quelques attributs seulement. Nous avons l'impression que ce type de bruit reflète mieux les situations réelles auxquelles nous sommes confrontés. En effet, il est peu probable que toutes les données soient bruitées, c'est-à-dire

Algorithme 6 *IntroduireBruit-b*$(\mathcal{B}_n^p, \alpha_{noise})$

Entrée :

$\mathcal{B}_n^p$: base de données à bruiter

α_{noise} : bruit introduit dans les données (entre 0 et 1)

Sortie :

$\mathcal{B}'$: base de données bruitée

Début

$nb\_objets\_a\_bruiter \leftarrow int(\alpha_{noise} * p * n)$ $\quad -- int(x)$ *: partie entière de (x)*

$\mathcal{B}' \leftarrow \mathcal{B}$

Pour $(i \leftarrow 1\,;\, i \leq n\,;\, i++)$ **Faire**

Pour $(j \leftarrow 1\,;\, j \leq p\,;\, j++)$ **Faire**

$nonBruite[i][j] \leftarrow 1$

Fin Pour

Fin Pour

Pour $(k \leftarrow 1\,;\, k \leq nb\_objets\_a\_bruiter\,;\, k++)$ **Faire**

Répéter

$i \leftarrow$ entier choisi aléatoirement entre 1 et n

$j \leftarrow$ entier choisi aléatoirement entre 1 et p

Jusqu'à ce que $(nonBruite[i][j] = 1)$

$nonBruite[i][j] \leftarrow 0$

$\mathcal{B}'(\alpha_i^j) \leftarrow \overline{\mathcal{B}(\alpha_i^j)}$ $\quad -- $ *introduction du bruit en inversant la valeur de* $\mathcal{B}(\alpha_i^j)$

Fin Pour

Retourner $\mathcal{B}'$

Fin

incorrectes. Si tel était le cas, *a priori* aucune méthode ne pourrait extraire des connaissances valides à partir de ces données. Par contre, il est relativement raisonnable de considérer que la majorité des attributs de la base sont fiables, et que seulement quelques attributs sont bruités.

Par exemple, considérons une base d'individus pour lesquels, les informations suivantes sont renseignées : âge, taille, poids, sexe, nationalité. Cette base de données est remplie par un employé de mairie disposant de la carte d'identité de chaque individu. La probabilité pour que les informations : taille, sexe, âge et nationalité soient incorrectes est très faible car ces informations sont présentes sur une carte d'identité. Par contre, le poids n'y figure pas, la probabilité d'erreur est donc beaucoup plus élevée pour cet attribut que pour les quatre précédents.

Cette troisième méthode essaye donc de rendre compte de ce phénomène en introduisant différents niveaux de bruit, sur quelques attributs de la base. Le nombre d'attributs à bruiter est un paramètre contrôlé par l'expert du domaine étudié. En fonction du type de données manipulées et des conditions dans lesquelles ces données ont été collectées, le nombre d'attributs bruités peut varier significativement.

De manière à mieux comprendre l'impact de ce bruit sur les règles obtenues, nous devons introduire du bruit sur quelques attributs, par exemple 3, et ce pour différents niveaux de bruit, par exemple 1%, 5% et 10%. Pour chaque triplet de couple $(Attribut, bruit)$, nous observons un certain bruit résultant dans les règles obtenues. La moyenne de ces différentes valeurs de bruit nous permet d'apprécier l'impact moyen de ce bruit sur notre approche. La combinatoire de cette méthode est élevée et l'évaluation de ce bruit sur une base de données telle que « mushrooms » est très coûteuse en temps de calcul.

Algorithme 7 *IntroduireBruit-c*$(\mathcal{B}_n^p, \mathcal{L})$

Entrée :

$\mathcal{B}_n^p$: base de données à bruiter

$\mathcal{L}$: liste de couples $(attribut, bruit)$

Sortie :

$\mathcal{B}'$: base de données bruitée

Début

$\mathcal{B}' \leftarrow \mathcal{B}$

Pour tout $((att_i, bruit_i) \in \mathcal{L})$ **Faire**

$\mathcal{B}' \leftarrow IntroduireBruit - a(\mathcal{B}', att_i, bruit_i)$

Fin Pour

Retourner $\mathcal{B}'$

Fin

L'algorithme 7 permet d'introduire cette forme de bruit dans les données.

4 Analyse de l'impact du bruit sur des données artificielles

Nous avons établi un protocole expérimental permettant d'étudier l'impact du bruit sur des données artificielles. L'objectif de l'étude réalisée est d'une part de déterminer s'il est possible d'extraire des connaissances « fiables » à partir de données bruitées et d'autre part d'analyser le

comportement de notre algorithme de recherche de pépites de connaissance en présence de données bruitées.

Le protocole mis en place est le suivant :

- création d'une base de données artificielle $\mathcal{B}_n^p$
- extraction des pépites de connaissance à partir de $\mathcal{B}_n^p$ en utilisant une des mesures de qualité suivante : Confiance, Lœvinger, Nouveauté, Moindre-contradiction, Intensité d'implication classique, Intensité d'implication entropique $\Rightarrow E^m_{\mathcal{B}_n^p}$
- introduction de bruit de type b dans les données $\mathcal{B}'^p_n$
- extraction des pépites de connaissance à partir de $\mathcal{B}'^p_n$ en utilisant la même mesure que celle ayant permis d'obtenir $E^m_{\mathcal{B}_n^p} \Rightarrow E^m_{\mathcal{B}'^p_n}$
- comparaison de $E^m_{\mathcal{B}_n^p}$ et de $E^m_{\mathcal{B}'^p_n}$
- calcul du support moyen des règles apparaissant dans $E^m_{\mathcal{B}'^p_n}$ et de celles disparaissant de $E^m_{\mathcal{B}_n^p}$.

Nous avons appliqué ce protocole sur des données contenant entre $n = 10000$ et $n = 50000$ (par pas de 1000) individus, et pour lesquelles le nombre d'attributs varie entre $p = 10$ et $p = 100$ (par pas de 5). Pour chaque base de données obtenue, l'introduction du bruit et la comparaison des règles bruitées avec les règles saines a été realisée dix fois. Pour chaque couple de paramètres (n, p), nous avons réalisé dix séries d'expériences.

Les données artificielles ont été engendrées avec le générateur aléatoire IBMdataGen mis à disposition par M. Zaki[1].

Le bruit étudié dans ces expériences est un bruit réparti de manière aléatoire sur les données (voir l'Algorithme 6). Compte tenu de la nature des données, seule l'étude de l'impact de cette forme de bruit est significative.

Nous avons observé, sur les diverses expériences réalisées, que le nombre de transactions n'influe pas sur le support moyen des règles disparaissant ou apparaissant. Par contre, le nombre d'attributs est un facteur important et son influence sur le support moyen des règles altérées par le bruit est significative.

Les résultats obtenus et présentés sur les Figures IV.2 à IV.7 correspondent donc aux courbes obtenues en moyennant les résultats pour les diverses valeurs du nombre de transactions. Nous n'avons retenu que les résultats pour lesquels au moins 20% des expérimentations sont « réussies » Nous considérons qu'une expérience est réussie si elle est interprétable, c'est-à-dire si les données engendrées aléatoirement sont telles qu'on puisse en extraire des connaissances. Il existe en effet certaines bases de données aléatoires dont les attributs sont si décorrélés qu'il n'existe pas de règles d'association entre attributs. Nous avons donc éliminé les résultats associés au nombre d'attribut pour lesquels les expérimentations ne réussissent que dans moins de 20% des expérimentations. Ce choix est lié à la volonté de ne pas biaiser l'interprétation des courbes. Un biais évident étant le nombre d'expériences utilisées pour calculer le support moyen. Si pour une valeur d'un attribut, nous avons 30 expériences différentes qui sont utilisées pour calculer le support moyen des règles disparaissant et pour un autre attribut, nous n'avons que 2 expériences différentes ayant « réussi », il nous semble difficile de comparer les supports moyens entre eux.

Sur ces courbes, le nombre d'attributs est représenté sur l'axe des abscisses et le support moyen des règles disparaissant (ou apparaissant) sur l'axe des ordonnées. Ces résultats ont été obtenus avec l'introduction de 5% de bruit de la forme b (c'est-à-dire bruit aléatoire réparti sur toutes les données).

Pour la plupart des mesures, dès que le nombre d'attributs est supérieur à 40, le support moyen se stabilise autour d'une valeur dépendant de la mesure étudiée. Les résultats obtenus pour les

[1] http ://www.cs.rpi.edu/∼zaki/software/

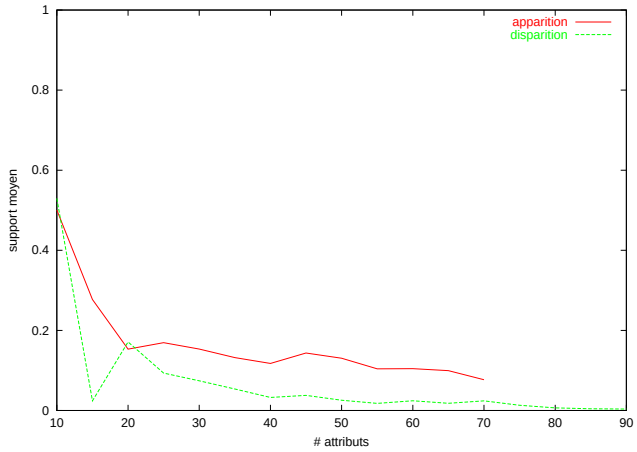


FIG. IV.2 – Support moyen des règles bruitées obtenues avec la Confiance.

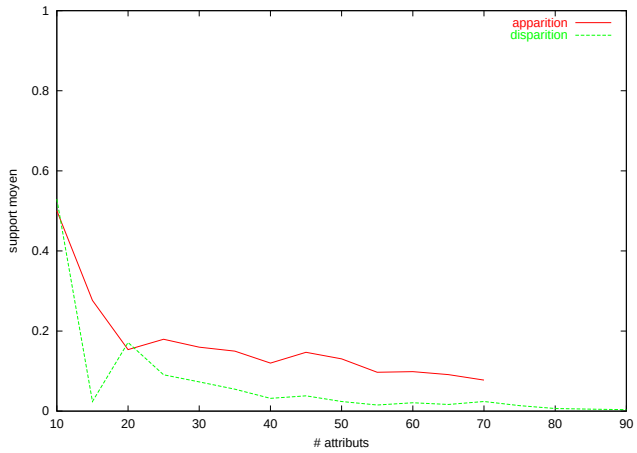


FIG. IV.3 – Support moyen des règles bruitées obtenues avec l'indice de Loevinger.

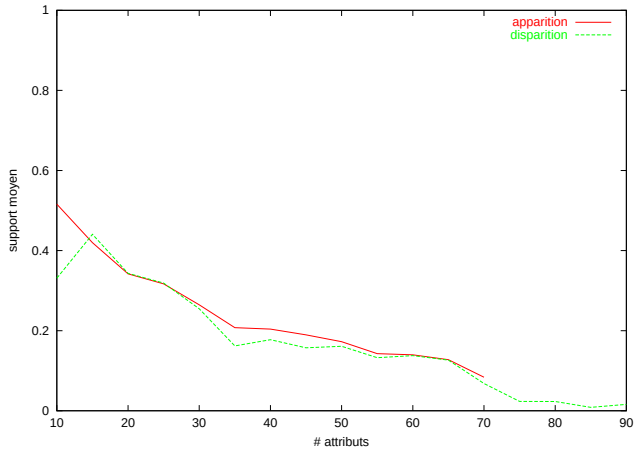


FIG. IV.4 – Support moyen des règles bruitées obtenues avec la Nouveauté.

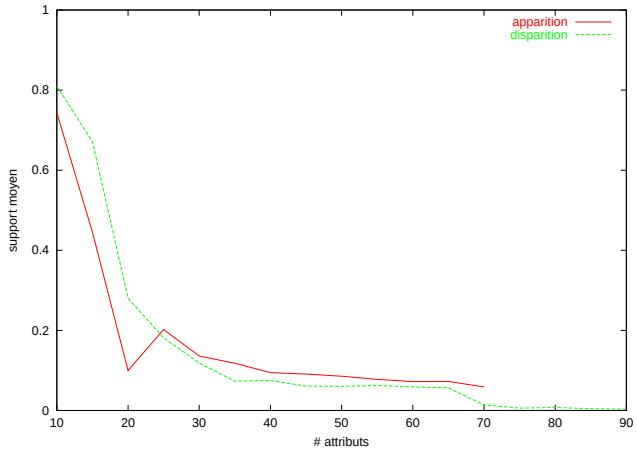


FIG. IV.5 – Support moyen des règles bruitées obtenues avec l'Intensité d'Implication Classique.

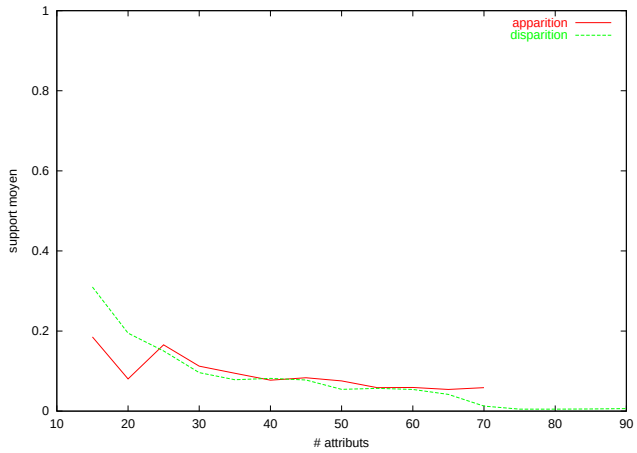


FIG. IV.6 – Support moyen des règles bruitées obtenues avec l'Intensité d'Implication Entropique.

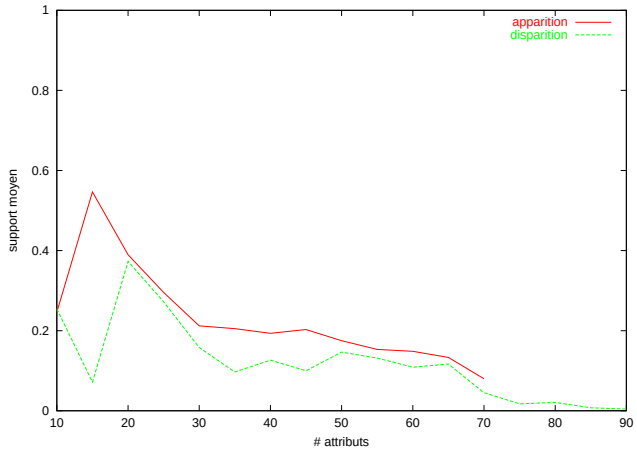


FIG. IV.7 – Support moyen des règles bruitées obtenues avec la Moindre contradiction.

bases de données contenant entre 10 et 40 attributs (voir les Tableaux IV.1 et IV.2) montrent qu'il est difficile d'obtenir des pépites de connaissance « fiables » sur de telles données. En effet, pour la majorité des mesures étudiées, le support moyen des règles apparaissant ou disparaissant est si élevé qu'il est d'une part difficile d'utiliser des méthodes d'extraction de connaissance utilisant le support comme seuil d'élagage. Et d'autre part, qu'il est difficile de proposer une méthode permettant de minimiser l'impact du bruit sur les connaissances extraites.

En contre-partie, la quasi stabilité du support moyen des règles apparues ou perdues avec le bruit nous permet d'utiliser soit des techniques classiques limitant le support des règles à extraire, soit une méthode plus élaborée essayant de déterminer, à partir des données, le support minimal tel que l'impact du bruit soit peu significatif.

Les Tableaux IV.1 et IV.2 présentent un résumé des Figures IV.2 à IV.7. Les différents points ou zones caractéristiques observés lors des expériences sont répertoriés dans ces tableaux. Nous y avons fait figurer les valeurs maximales du support observé pour chaque mesure ainsi que les deux zones les plus larges où le support est stationnaire. Une telle zone est appelée un **plateau**. Le support est considéré comme stationnaire si les valeurs du support appartenant au plateau sont toutes comprises entre $\mu_{plateau} \times (1-\delta)$ et $\mu_{plateau} \times (1+\delta)$ où $\mu_{plateau}$ est la valeur moyenne du support sur le plateau et δ est un paramètre permettant de contrôler la sensibilité du plateau.

Les Tableaux IV.1 et IV.2 présentent les plateaux obtenus lorsque $\delta = 0.5$. Les plateaux sont indiqués par largeur décroissante.

Rappelons que le pas utilisé pour le nombre d'attributs est égal à 5, ainsi un plateau entre p_i et $p_i + 5$ correspond à une zone pour laquelle nous n'avons que deux séries de mesures, une pour $p = p_i$ et l'autre pour $p = p_i + 5$. Un tel plateau n'est donc pas très informatif.

Mesure	1^{er} maximum		$2^{ème}$ maximum		1^{er} plateau		$2^{ème}$ plateau	
	# att	support	# att	support	zone	support	zone	support
IIC	10	0,744	15	0,445	40 – 65	0,081	25 – 30	0,169
Nouveaute	10	0,516	15	0,420	35 – 65	0,168	10 – 20	0,425
Confiance	10	0,501	15	0,277	20 – 55	0,137	60 – 70	0,093
IIE	15	0,185	25	0,165	40 – 65	0,067	25 – 30	0,138
Lœvinger	10	0,501	15	0,276	20 – 50	0,148	55 – 70	0,090
Moindre Contradiction	15	0,547	20	0,390	30 – 65	0,177	20 – 25	0,342

TAB. IV.1 – Résumé des expériences réalisées lorsque 5% de bruit est introduit dans les données (bruit de la forme *b*). Ces informations sont liées aux règles qui apparaissent.

Mesure	1^{er} maximum		$2^{ème}$ maximum		1^{er} plateau		$2^{ème}$ plateau	
	# att	support	# att	support	zone	support	zone	support
IIC	10	0,810	15	0,671	35 – 65	0,063	75 – 80	0,006
Nouveaute	15	0,441	20	0,343	35 – 65	0,150	10 – 25	0,358
Confiance	10	0,530	20	0,171	50 – 70	0,021	80 – 85	0,005
IIE	15	0,310	20	0,195	35 – 60	0,066	75 – 90	0,005
Lœvinger	10	0,530	20	0,171	55 – 65	0,017	25 – 35	0,072
Moindre Contradiction	20	0,373	25	0,272	35 – 65	0,117	75 – 80	0,018

TAB. IV.2 – Résumé des expériences réalisées lorsque 5% de bruit est introduit dans les données (bruit de la forme *b*). Ces informations sont liées aux règles qui disparaissent.

Cette représentation condensée des résultats nous permet de voir que :

- comme le montrent les courbes, toutes les mesures présentent une grande sensibilité au bruit (valeur très élevée pour le support des règles apparaissant ou disparaissant) lorsque les données contiennent très peu d'attributs
- la différence entre les supports associés aux deux premiers extrema est généralement très élevée alors que ces valeurs sont liées à des données décrites par un nombre proche d'attributs
- les plateaux permettent de visualiser des zones pour lesquelles le support moyen des règles disparaissant (ou apparaissant) reste relativement stable. L'utilisation de ces zones n'a pas encore été étudiée de manière approfondie mais nous pensons qu'elles pourraient permettre de dresser des cartes de compétences des différentes mesures de qualité. Ces cartes pouvant ensuite être utilisées par les experts pour choisir la ou les mesures de qualité qui semblent les mieux adaptées à leurs données (c'est-à-dire présentant un plateau contenant le nombre d'attributs des données à étudier).

Nous présentons dans la section suivante, un algorithme permettant de déterminer le support minimal tel que les règles ayant une forte valeur pour la mesure de qualité utilisée ne disparaissent plus lorsque les données sont bruitées.

4.1 Une solution pour réduire l'impact du bruit

Nous avons vu, sur des données artificielles, que quelque soit la mesure de qualité utilisée pour extraire les pépites de connaissance, il existe un support maximal associé aux règles altérées par le bruit. Ce support peut permettre d'élaguer les règles proposées à l'expert et donc de réduire la probabilité de proposer à l'expert des règles dues au bruit.

Lorsque nous étudions le problème du bruit, nous sommes confrontés à trois catégories de règles (voir Figure IV.1).

1. Les règles stables en présence de bruit
2. Les règles qui disparaissent lorsque les données sont bruitées
3. Les règles qui apparaissent lorsque les données sont bruitées

Sachant que les règles doivent être évaluées par un expert humain, il est important de minimiser la quantité de règles purement dues au bruit (et qui sont donc par définition fausses). Idéalement, il ne faudrait proposer à l'expert que les règles « stables ».

Pour atteindre cet objectif, nous proposons dans un premier temps, de minimiser le nombre de règles ayant une forte valeur pour la mesure de qualité utilisée et qui disparaissent dès que les données sont légèrement bruitées.

Par exemple, pour la moindre-contradiction, certaines règles ayant un faible support sont très peu contredites et ont donc une forte valeur pour la moindre-contradiction. Et dès que nous introduisons une faible quantité de bruit, ce type de règles d'association disparaît de l'ensemble des pépites de connaissance, ceci étant dû à l'apparition de contradictions liées au bruit. Le support de ces règles instables peut être utilisé pour élaguer l'ensemble des pépites recherchées.

Nous nous proposons donc de déterminer de manière automatique ce support minimal.

Weiss et Hirsh ont montré dans [Weiss et Hirsh 1998] que les « small disjuncts » sont très sensibles au bruit et qu'ils dégradent significativement les résultats du processus d'apprentissage. Ces observations, effectuées dans le cadre de l'apprentissage supervisé, peuvent être réutilisées dans le cadre de la découverte non supervisée de règles d'association, où nous considérons que les « small disjuncts » sont équivalents aux règles d'association ayant un très faible support.
Pour prendre ce problème en considération, nous introduisons le concept de « support minimum pour la résistance à un α-bruit ». Ce concept est évalué comme étant le support stable pour lequel

les règles d'association, ayant une valeur élevée pour la mesure de qualité étudiée, ne disparaissent plus lorsque nous introduisons N fois $\alpha_{noise}\%$ de bruit dans les données. Le principe de l'algorithme est le suivant :

- extraire les règles d'association
- introduire $\alpha_{noise}\%$ de bruit dans les données
- extraire les nouvelles règles d'association
- déterminer l'ensemble des règles disparues à cause du bruit
- puis dans cet ensemble, isoler le sous-ensemble de règles ayant une forte valeur pour la mesure étudiée (c'est-à-dire proche de la valeur maximale trouvée) $\Rightarrow \mathcal{E}^{-}_{noise}$
- calculer le support maximal des règles appartenant à $\mathcal{E}^{-}_{noise}$
- utiliser ce support comme seuil d'élagage

Nous obtenons ainsi un nouvel algorithme qui est très semblable à la méthode classique permettant de rendre les réseaux de neurones résistants au bruit (voir, par exemple, [Haykin 1998]).

Nommons $S_{\alpha_{noise}}$ le « support minimal pour la résistance à un α-bruit ».

Pour l'instant, l'implantation actuelle ne calcule le support minimum que pour les règles ayant exactement un seul attribut en prémisse.

Les résultats expérimentaux obtenus sur des données artificielles montrent que l'utilisation de cet algorithme permet de réduire significativement, et ce pour la majorité des mesures étudiées, le taux de règles disparaissant lorsque les données sont bruitées. Par contre, le revers de l'approche est lié aux règles qui apparaissent lorsque les données sont bruitées. Les tableaux IV.3 à IV.7 présentent les résultats obtenus pour les différentes mesures étudiées lorsque la détection de $S_{\alpha_{noise}}$ est réalisée. Ces résultats ont été obtenus sur des bases de données contenant 30000 transactions décrites par 40 attributs.

bruit introduit	bruit observé	
	-	+
1,00	15,98	6,52
2,00	22,96	8,55
5,00	31,57	14,20
10,00	42,55	22,90

sans la détection du support $S_{\alpha_{noise}}$

bruit introduit	bruit observé	
	-	+
1,00	9,94	13,72
2,00	21,25	32,57
5,00	35,63	46,47
10,00	44,81	68,70

avec la détection du support $S_{\alpha_{noise}}$

TAB. IV.3 – Impact de la détection de $S_{\alpha_{noise}}$ pour la Confiance.

bruit introduit	bruit observé	
	-	+
1,00	12,96	8,30
2,00	19,30	10,47
5,00	27,95	16,83
10,00	40,33	29,76

sans la détection du support $S_{\alpha_{noise}}$

bruit introduit	bruit observé	
	-	+
5,00	11,54	35,38
10,00	28,00	50,99

avec la détection du support $S_{\alpha_{noise}}$

TAB. IV.4 – Impact de la détection de $S_{\alpha_{noise}}$ pour l'indice de Lœvinger.

Les résultats obtenus montrent qu'il semble difficile de réduire les effets du bruit. En minimisant l'impact du bruit sur les règles intéressantes et disparaissant en présence de peu de bruit, notre approche provoque une augmentation dramatique du nombre de règles purement dues au bruit. Ces règles sont difficiles à détecter et imposent une surcharge de travail inutile à l'utilisateur.

Algorithme 8 algorithme permettant de déterminer le support minimal garantissant la stabilité des règles les moins-contradictoires en présence d'un bruit α_{noise}

Entrée :

$\mathcal{B}_n^p$: la base de données étudiée

α_{noise} : bruit introduit dans les données

N : nombre d'itérations pour obtenir le support minimal $S_{\alpha_{noise}}$

Sortie :

$S_{\alpha_{noise}}$: support minimal garantissant la stabilité des règles en présence d'un bruit α_{noise}

Début

$S_{\alpha_{noise}} \leftarrow 0$; $\mathcal{E}_{noise} \leftarrow \emptyset$; $S_{noise}^{prec} \leftarrow 0$; $N \leftarrow 1$

Tant que $(S_{\alpha_{noise}} \neq S_{noise}^{prec})$ **et** $(N \leq 100)$ **Faire**

$S_{noise}^{prec} \leftarrow S_{\alpha_{noise}}$

Si $(\mathcal{E}_{noise} = \emptyset)$ **Alors**

$\mathcal{E}'_1 \leftarrow ExtrairePepites(\mathcal{B}, 1, S_{\alpha_{noise}})$

$m_{max} \leftarrow max(m(R), R \in \mathcal{E}'_1)$

$\mathcal{E}_{noise} \leftarrow \mathcal{E}'_1$

Sinon

$\mathcal{E}'_1 \leftarrow ExtrairePepites(\mathcal{B}', 1, S_{\alpha_{noise}})$

$\mathcal{E}_{noise} \leftarrow \mathcal{E}_{noise} \cap \mathcal{E}'_1$ *-- règles stables*

$\mathcal{E}_{noise}^{-} \leftarrow \mathcal{E}_{noise} - \mathcal{E}'_1$ *-- règles disparues à cause du bruit*

$\mathcal{E}_{noise}^{+} \leftarrow \mathcal{E}'_1 - \mathcal{E}_{noise}$ *-- règles apparues à cause du bruit*

-- détection des « meilleures » règles ayant disparues

ne conserver dans $\mathcal{E}_{noise}^{-}$ que les règles R telle que $m(R) \approx m_{max}$

$S_{max}^{-} \leftarrow support(R)\ tq\ R \in \mathcal{E}_{noise}^{-} \wedge \forall R' \in \mathcal{E}_{noise}^{-},\ support(R) \geq support(R')$

$S_{\alpha_{noise}} \leftarrow max(S_{\alpha_{noise}}, S_{max}^{-})$

Fin Si

Si $(S_{\alpha_{noise}} \neq S_{noise}^{prec})$ **Alors**

$N \leftarrow 1$ *-- Supports différents donc redémarrage d'une nouvelle boucle de 100 itérations*

Sinon

$N \leftarrow N + 1$ *-- Supports identiques donc poursuite de la boucle courante de bruitage*

Fin Si

$\mathcal{B}' \leftarrow IntroduireBruit - b(\mathcal{B}, \alpha_{noise})$

Fin Tant que

Retourner $S_{\alpha_{noise}}$

Fin

bruit introduit	bruit observé	
	-	+
1,00	9,48	4,92
2,00	13,05	6,99
5,00	21,75	11,77
10,00	31,88	18,65

sans la détection du support $S_{\alpha_{noise}}$

bruit introduit	bruit observé	
	-	+
1,00	32,08	21,67
2,00	39,33	23,00
5,00	71,48	40,89
10,00	72,35	18,56

avec la détection du support $S_{\alpha_{noise}}$

TAB. IV.5 – Impact de la détection de $S_{\alpha_{noise}}$ pour la Nouveauté.

bruit introduit	bruit observé	
	-	+
1,00	10,22	17,64
2,00	13,43	24,19
5,00	19,73	35,54
10,00	26,10	45,07

sans la détection du support $S_{\alpha_{noise}}$

bruit introduit	bruit observé	
	-	+
2,00	12,38	26,19
5,00	19,30	35,44
10,00	24,19	45,82

avec la détection du support $S_{\alpha_{noise}}$

TAB. IV.6 – Impact de la détection de $S_{\alpha_{noise}}$ pour l'Intensité d'Implication Classique.

bruit introduit	bruit observé	
	-	+
1,00	11,51	12,19
2,00	13,85	17,23
5,00	19,20	25,15
10,00	30,35	33,52

sans la détection du support $S_{\alpha_{noise}}$

bruit introduit	bruit observé	
	-	+
1,00	15,42	19,58
2,00	12,49	27,52
5,00	17,02	32,90
10,00	28,46	54,33

avec la détection du support $S_{\alpha_{noise}}$

TAB. IV.7 – Impact de la détection de $S_{\alpha_{noise}}$ pour l'Intensité d'Implication Entropique.

4.2 Une alternative au bruit

N'oublions pas qu'un de nos objectifs majeurs est de minimiser le travail de l'utilisateur. Il est donc important de lui proposer un ensemble de règles qui vérifie au mieux les critères de qualité imposés par l'utilisateur (voir chapitre II) et qui soit le plus résistant possible au bruit.

Nous proposons donc une approche visant, non pas à réduire l'impact du bruit dans les règles extraites, mais permettant d'associer à chaque règle proposée à l'expert une estimation de sa résistance au bruit.

Considérons une base de données $\mathcal{B}$ à partir de laquelle un ensemble de règles $\mathcal{E}$ est extrait (en utilisant la mesure de qualité m). Nous proposons d'introduire du bruit dans $\mathcal{B}$, puis à partir de la nouvelle base bruitée $\mathcal{B}'$, d'extraire le nouvel ensemble de règles $\mathcal{E}'$. Cette première étape, en tout point identique à celle déjà présentée dans les sections précédentes, est suivie d'une phase de comparaison des règles de $\mathcal{E}$ avec celles de $\mathcal{E}'$. L'objectif de cette comparaison est de vérifier si les règles trouvées à partir de $\mathcal{B}$ sont présentes dans $\mathcal{E}'$. Si tel est le cas, alors ces règles sont considérées comme fiables.

La répétition de cette opération permet d'obtenir pour chaque règle de $\mathcal{E}$ un pourcentage de « fiabilité » (par rapport au bruit).

L'algorithme 9 correspond à cette nouvelle approche.

Nous avons validé cette nouvelle approche sur le même type de données artificielles que celles utilisées précédemment. Nous nous sommes focalisés sur les pépites de connaissance contenant exactement un attribut en prémisse et un en conclusion. Les résultats obtenus permettent d'établir un classement entre les différentes mesures de qualité étudiées. Le classement obtenu est relatif à la fiabilité moyenne observée sur $N = 20$ itérations, et pour 10 bases de données différentes contenant 30000 transactions décrite par 50 attributs. Pour obtenir ce classement, nous avons réalisé deux séries d'expériences, une première en introduisant 1% de bruit dans les données et la seconde en introduisant 5% de bruit (le bruit est toujours de la forme b).

La figure IV.8 présente l'évolution de la fiabilité, pour les six mesures étudiées, en fonction du bruit introduit dans les données.

Toutes les mesures ont un comportement relativement proche lorsque les données sont bruitées. Cependant parmi ces six mesures testées, la Moindre Contradiction est la mesure qui semble la plus fiable.

Pour compléter ces résultats et valider l'approche retenue, nous avons aussi mesuré le pourcentage de règles qui, pour chaque mesure, sont présentes dans $\mathcal{E}$ et ne sont jamais retrouvées dans $\mathcal{E}'$ pour l'ensemble des 20 expérimentations réalisées. La Moindre Contradiction et la Nouveauté sont les deux mesures présentant les plus faibles pourcentages de règles qui ne sont jamais retrouvées.

Les Tableaux IV.8 et IV.9 présentent le pourcentage moyen de règles qui ne sont jamais retrouvées lorsque les données sont bruitées (et ce pour les différentes valeurs de bruit introduit dans les données).

4.3 Conclusion sur les données aléatoires

Nous avons montré, à travers les diverses approches étudiées et les expérimentations associées, que la prise en considération du bruit pouvant exister dans les données ou pouvant les altérer représente un aspect important de la qualité des connaissances extraites à partir des données.

Bien que nous ayons réduit notre étude à une forme de bruit relativement simpliste, nous avons constaté qu'il est difficile de proposer des solutions permettant de réduire l'impact du bruit sur les connaissances obtenues.

Algorithme 9 calcul de la fiabilité des règles.

Entrée :

$\mathcal{B}_n^p$: la base de données étudiée

α_{noise} : bruit introduit dans les données

N : nombre d'itérations pour calculer la fiabilité moyenne associée à chaque règle

Sortie :

$\mathcal{E}$: ensemble de règles

Début

$\mathcal{E}_{noise} \leftarrow \emptyset$; $N \leftarrow 1$

$\mathcal{E} \leftarrow ExtrairePepites(\mathcal{B}, 1, 0)$

−− on associe un compteur aux règles pour déterminer leur fiabilité

Pour tout $(R \in \mathcal{E})$ **Faire**

$R.cpt \leftarrow 0$

Fin Pour

Pour $(i \leftarrow 1 ; i \leq N ; i++)$ **Faire**

$\mathcal{B}' \leftarrow IntroduireBruit - b(\mathcal{B}, \alpha_{noise})$

$\mathcal{E}' \leftarrow ExtrairePepites(\mathcal{B}', 1, 0)$

$\mathcal{E}_{stable} \leftarrow \mathcal{E} \cap \mathcal{E}'$ *−− règles stables*

−− on incrémente le compteur associé aux règles

Pour tout $(R \in \mathcal{E}_{stable})$ **Faire**

$R.cpt \leftarrow R.cpt + 1$

Fin Pour

Fin Pour

−− on normalise le compteur associé aux règles

Pour tout $(R \in \mathcal{E})$ **Faire**

$R.cpt \leftarrow \frac{R.cpt}{N}$

Fin Pour

Retourner $\mathcal{E}$

Fin

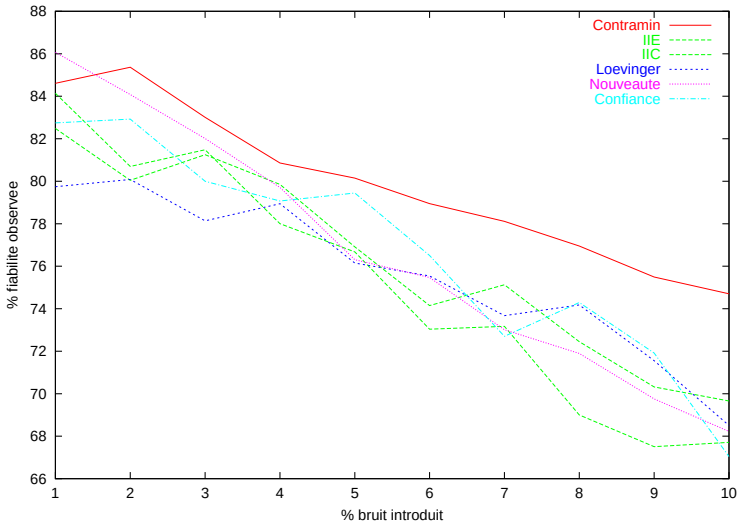


FIG. IV.8 – Évolution de la fiabilité en fonction du bruit introduit.

Mesure	Bruit introduit				
	1	2	3	4	5
Intensite d'Implication Entropique	0,988	3,514	8,452	8,577	10,308
Intensite d'Implication Classique	0	0,854	3,076	0	3,076
Loevinger	5,106	7,922	9,624	16,15	13,495
Nouveaute	0	0,891	0,463	0	0
Confiance	8,457	9,33	9,671	14	17,602
Moindre Contradiction	0,769	0	0,346	0,716	0,346

TAB. IV.8 – Pourcentage de règles qui ne sont jamais retrouvées lorsque les données sont bruitées (entre 1% et 5% de bruit).

Mesure	Bruit introduit				
	6	7	8	9	10
Intensite d'Implication Entropique	8,448	13,265	12,136	13,266	14,778
Intensite d'Implication Classique	0	3,076	2,82	5,7	5,113
Loevinger	15,903	19,027	21,088	24,425	23,602
Nouveaute	0,427	0,855	2,396	1,316	1,42
Confiance	15,632	17,041	21,097	24,263	19,691
Moindre Contradiction	0,396	1,154	2,342	0,396	1,642

TAB. IV.9 – Pourcentage de règles qui ne sont jamais retrouvées lorsque les données sont bruitées (entre 6% et 10% de bruit).

Les cartes de compétence présentées (Figures IV.2 à IV.7) permettent de visualiser le support moyen des règles affectées par le bruit pour chacune des mesures de qualité testées. Ces cartes de compétence montrent qu'une approche basée sur l'utilisation du support comme critère d'élagage pour réduire l'impact du bruit sur les données est difficilement envisageable compte tenu du fait que le support moyen des règles bruitées est relativement élevé pour la plupart des mesures testées.

L'approche que nous avons proposée, fondée sur la détection automatique du support minimal permettant de ne plus perdre les meilleures règles lorsque les données sont bruitées, a permis de réduire significativement le pourcentage de règles disparaissant lorsque les données sont bruitées. Malheureusement, le revers de la médaille est pire que cet aspect positif. En effet, cette première approche a entraîné une augmentation dramatique du pourcentage de règle apparaissant lorsque les données sont bruitées et ce faisant augmente significativement le travail de l'expert pour trier les bonnes règles des règles erronées.

Nous avons donc opté pour une nouvelle approche dont l'objectif majeur est d'assister l'expert dans le choix et l'analyse des règles que nous lui proposons plutôt que de concevoir un filtre visant à sélectionner automatiquement les meilleures règles. Cette nouvelle approche permet d'associer à chaque règle détectée par l'algorithme d'extraction des pépites de connaissance un pourcentage de fiabilité correspondant à la résistance de la règle lorsque les données contiennent $\alpha_{noise}\%$ de bruit réparti de manière aléatoire. Les résultats obtenus sur des données aléatoires ont montré que les différentes mesures testées ont un comportement globalement très similaire. Notons toutefois que la Moindre Contradiction est l'une des mesures (avec la nouveauté) qui se comporte « le mieux » (c'est-à-dire pourcentage moyen de résistance au bruit le plus élevé) lorsque les données sont bruitées.

5 Analyse de l'impact du bruit sur des données réelles

L'étude de l'impact du bruit sur les données aléatoires nous a permis de montrer d'une part qu'il est difficile de simplement minimiser l'impact du bruit et d'autre part qu'il est important de prendre en considération cet aspect dans la qualité des connaissances puisque des règles en apparence très fiables s'avèrent aussi très sensibles au bruit.

Lorsque nous travaillons avec des données aléatoires, la modélisation du bruit par un phénomène purement aléatoire venant altérer les données est suffisante. Par contre, dès que nous travaillons avec des données réelles, cette modélisation peut s'avérer insuffisante voire inappropriée.

Dans le cadre de cette thèse, une collaboration avec Sylvie Guillaume et Philippe Castagliola nous a permis d'avoir accès à des données réelles : des données bancaires. Les données ont été analysées par Sylvie Guillaume qui s'est intéressé à l'extraction de règles d'association ordinales [Guillaume 2002a, Guillaume et al. 1998, Guillaume 2002b]. La collaboration mise en place a permis d'une part d'étudier la résistance au bruit des techniques utilisées pour extraire les règles d'association ordinales et d'autre part, d'étudier des formes de bruit plus réalistes et mieux adaptées aux données.

Dans les sections suivantes, nous présenterons les données bancaires, puis les techniques utilisées pour extraire les règles d'association ordinales, enfin nous détaillerons les méthodes proposées pour bruiter les données ainsi que les résultats obtenus.

5.1 Présentation des données bancaires

La base de données se compose de 47112 clients décrits par 44 variables quantitatives. Ces variables peuvent être classées en trois catégories :

1. les informations relatives aux clients (*âge, ancienneté*)

2. les différents produits financiers proposés par la banque (*actions, obligations, etc.*)
3. les statistiques sur les différents comptes ouverts par les clients (*montant des ressources, montant des encours prêt, etc.*)

Les variables relatives aux produits financiers proposés par la banque peuvent également être répertoriées en deux catégories :

(a) les variables mémorisant les encours de chaque produit financier pour tous les clients de la base

(b) les variables comptabilisant le nombre de comptes ouverts par le client pour chaque produit financier

À partir de ces données, Sylvie Guillaume s'est intéressé à l'extraction des meilleures règles d'association ordinales permettant d'aider le banquier à mieux comprendre ses données.

La particularité de ces données est liée au fait que les attributs décrivant les individus ne sont ni booléens, ni discrets mais quantitatifs. Ainsi, aucune des mesures de qualité présentées dans le chapitre II ne peut être utilisée sur ces données. Nous allons donc commencer par présenter les différentes mesures quantitatives utilisées par S. Guillaume pour extraire, à partir de ces données, les associations ou règles d'association ordinales.

5.2 Mesures Quantitatives

Dans cette section, nous présentons deux mesures quantitatives de similarité (le coefficient de corrélation linéaire significatif et la mesure de vraisemblance du lien) permettant de détecter des associations entre deux variables (X, Y) et une mesure implicative (l'intensité d'inclination) permettant de détecter des règles $(X_1...X_p \rightarrow Y_1...Y_q)$, c'est-à-dire des associations orientées entre p et q variables. Dans la suite cette thèse, nous utiliserons le terme *association* pour désigner des associations ou des règles d'association.

5.2.1 Coefficient de Corrélation Linéaire Significatif

Le coefficient de corrélation linéaire r est une mesure de liaison linéaire entre deux variables quantitatives X et Y. Lorsque r est proche de 0, les deux variables sont indépendantes ; lorsque r est proche de 1, les deux variables évoluent dans le même sens selon approximativement une droite de pente positive et pour finir lorsque r est proche de -1, les deux variables évoluent cette fois-ci en sens inverse selon approximativement une droite de pente négative. La caractéristique de l'extraction de connaissances à partir des données étant de travailler avec des données volumineuses, nous retenons l'approximation faite par Saporta dans [Saporta 1990], c'est-à-dire pour une population Ω de taille N supérieure à 100, la variable aléatoire R, dont le coefficient de corrélation r est une valeur observée, suit approximativement la loi normale de moyenne 0 et d'écart-type $\frac{1}{\sqrt{N-1}}$. Comme nous recherchons les liaisons significatives entre variables (*liaisons selon une droite de pente positive et liaisons selon une droite de pente négative*), nous retenons l'indice significatif de dépendance suivant :

$$CCLS = P(|R| \leq |r|)$$

5.2.2 Mesure de Vraisemblance du Lien

La mesure de vraisemblance du lien [Lerman 1981] évalue si le nombre des associations positives (*c'est-à-dire le nombre des individus vérifiant de fortes valeurs pour X et de fortes valeurs*

pour Y) est significativement élevé comparativement à ce que l'on obtiendrait si X et Y étaient indépendantes. L'indice brut de similarité est défini par :

$$v_0 = \sum_{i=1}^{N} x_i y_i$$

où x_i et y_i sont respectivement les valeurs prises par les variables X et Y pour l'individu t_i. I.C. Lerman a démontré que la variable aléatoire V dont cet indice brut v_0 est une réalisation suit asymptotiquement la loi normale $\mathcal{N}(\mu, \sigma)$ de moyenne $\mu = N \times \mu_X \times \mu_Y$ et de variance $\sigma^2 = \frac{N^2}{N-1} \times v_X \times v_Y$ avec N le nombre de transactions[2], μ_X et μ_Y respectivement les moyennes des variables X et Y, et v_X et v_Y respectivement les variances des variables X et Y. De plus, il a démontré que l'indice brut normalisé $v_{on} = \frac{v_0 - \mu}{\sigma}$ est égal à $\sqrt{N-1}r$, où r est le coefficient de corrélation linéaire défini précédemment (voir section 5.2.1). La mesure de similarité locale $S_L(X, Y)$ est donc définie de la façon suivante :

$$S_L(X, Y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\sqrt{N-1} \times r} e^{-\frac{1}{2}t^2} dt$$

Lorsque la valeur de r est négative (*respectivement positive*) et la taille de la population N est importante, la valeur de $z = \sqrt{N-1} \times r$ tends vers $-\infty$ (*respectivement* $+\infty$), et par conséquent la valeur de $S_L(X, Y)$ tends vers 0 (*respectivement 1*). Pour finir, lorsque r vaut 0, la valeur de $S_L(X, Y)$ est égale à $0,5$. Ainsi, dans le cas de données volumineuses, cette mesure locale n'est pas sélective car elle ne peut prendre que trois valeurs : 0, 1 et $0,5$. Afin de remédier à ce problème, I.C. Lerman a proposé une mesure de similarité globale $S_G(X, Y)$ définie de la façon suivante :

$$S_G(X, Y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{z - \mu_z}{\sigma_z}} e^{-\frac{1}{2}t^2} dt$$

où μ_z et σ_z sont respectivement la moyenne arithmétique des valeurs de z extraites sur la population étudiée et la variance de ces mêmes valeurs.

5.2.3 Intensité d'inclination

L'intensité d'inclination [Guillaume 2002b] évalue si le nombre des individus ne vérifiant pas fortement la règle $X \to Y$ (*c'est-à-dire le nombre des individus vérifiant de fortes valeurs pour* X *et de faibles valeurs pour* Y) est significativement faible comparativement à ce que l'on obtiendrait si X et Y étaient indépendantes. Soient X et Y deux variables quantitatives prenant respectivement leurs valeurs x_i et y_i $(i = 1, .., N)$ dans les intervalles $[x_{min}..x_{max}]$ et $[y_{min}..y_{max}]$ et soit q_0 la mesure brute de non-inclination définie de la façon suivante :

$$q_0 = \sum_{i=1}^{N} (x_i - x_{min})(y_{max} - y_i)$$

Soient μ_X, μ_Y les moyennes arithmétiques de respectivement X et Y et v_X, v_Y les variances de X et Y. L'intensité d'inclination est donnée par la formule suivante :

$$\varphi(X \to Y) = \frac{1}{\sigma\sqrt{2\pi}} \int_{q_0}^{+\infty} e^{\frac{(t-\mu)^2}{2\sigma^2}} dt$$

[2] ou observations ou encore individus.

$$avec \begin{cases} \mu = n(\mu_X - x_{min})(y_{max} - \mu_Y) \\ \sigma^2 = n[\sigma_X \sigma_Y + \sigma_Y(\mu_X - x_{min})^2 + \sigma_X(y_{max} - \mu_Y)^2] \end{cases}$$

Cette mesure garantit que les effectifs observés s'écartent significativement des effectifs théoriques et particulièrement en présence de fortes valeurs pour X et de faibles valeurs pour Y.

5.3 Extraction des associations et règles d'association ordinales

Ces mesures de qualité quantitatives ont été utilisées pour extraire des associations ou règles d'association vérifiant les critères suivants :

- pas de contrainte minimale pour le support
- seuil d'élagage $= 0,9$ pour le Coefficient de Corrélation Linéaire Significatif et la Mesure de Vraisemblance du Lien
- seuil d'élagage $= 0,85$ pour l'Intensité d'Implication Ordinale

Ces choix sont liés aux expérimentations précédemment réalisées par Sylvie Guillaume. Nous avons voulu reproduire le même protocole expérimental pour pouvoir évaluer la qualité des connaissances obtenues en présence de données bruitées.

Les seuils d'élagage varient par pas de $0,01$ de leur borne inférieure jusqu'à 1 et pour chaque seuil, nous obtenons trois ensembles de règles (un pour chaque mesure étudiée). Ces règles sont comparées aux règles obtenues lorsque les données sont bruitées et nous mesurons le taux de règles perdues et le taux de règles nouvelles. Cette démarche est similaire à celle utilisée pour évaluer l'impact du bruit sur les données aléatoires. La section suivante présente le protocole retenu pour bruiter les données, puis la section 5.6 présente les résultats obtenus.

5.4 Nature du bruit introduit

Nous avons choisi de répartir le bruit dans les données de manière aléatoire (comme pour les expérimentations précédentes sur les données aléatoires). La différence majeure réside dans la modification apportée à la valeur prise par le couple bruité (individu, attribut).

Nous définissons donc dans cette section, la technique retenue pour modifier la valeur d'un couple (individu, attribut).

Soit une base de données composée de n individus t_i $(i = 1, .., n)$ décrits par p variables quantitatives X_1, ... , $X = X_j$, .. , X_p $(j = 1, .., p)$ et soit x_i la valeur de la variable X prise par l'individu t_i. Nous supposons que la variable X prend ses valeurs dans l'intervalle $[x_{min}..x_{max}]$.

Soient $F_X(x, a, b, c, d)$ la fonction de répartition de la variable X et $F_X^{-1}(y, a, b, c, d)$ la fonction inverse de F_X avec $x \in [x_{min}..x_{max}]$ et $y \in [0..1]$. Les paramètres a et b déterminent la forme de ces deux fonctions, le paramètre c correspond à la valeur minimale x_{min} de la variable X et le paramètre d représente le taux de couverture[3] pour cette valeur minimale $(X = x_{min})$.

Afin d'obtenir la valeur bruitée x'_i pour la valeur observée x_i, nous modifions les paramètres a, b et d en choisissant une nouvelle valeur a' (*respectivement* b' *et* d') dans l'intervalle $[a(1-\gamma), .., a(1+\gamma)]$ (*respectivement dans les intervalles* $[b(1-\gamma), .., b(1+\gamma)]$ *et* $[d(1-\gamma), .., d(1+\gamma)]$). Afin d'effectuer des tests réalistes, nous ne pouvons changer la valeur du paramètre c sinon nous obtiendrions une nouvelle fonction de répartition trop différente de la fonction initiale. N'oublions pas que l'injection de bruit a pour but de simuler des erreurs dans la base et que cette nouvelle base doit être assez proche de la base initiale (non bruitée). Ce choix est lié à la contrainte suivante : nous ne voulons pas modifier le domaine de variation de la variable bruitée mais seulement la répartition

[3] ou support.

de quelques valeurs (choisies aléatoirement dans l'intervalle $[x_{min}..x_{max}]$). En effet, nous pensons qu'il est raisonnable de considérer que le bruit, ayant pu altérer les données, n'a pas modifié le domaine de variation de celles-ci.

La nouvelle valeur x_i' pour l'individu t_i est égale à $F_X^{-1}(y_i, a', b', c, d')$ avec $y_i = F_X(x_i, a, b, c, d)$ comme le montre la Figure IV.9.

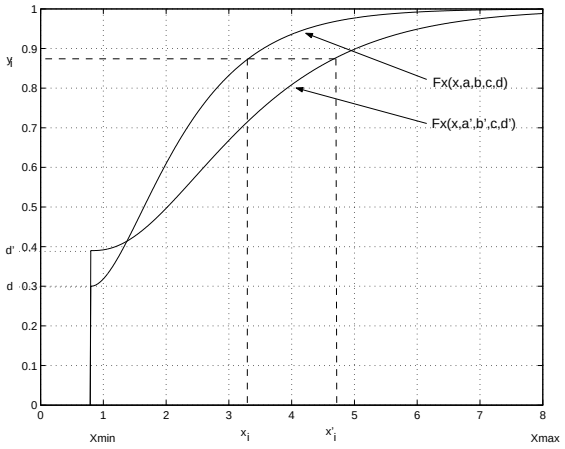


FIG. IV.9 – Injection du bruit dans la base pour une variable X.

Dans la Figure IV.9, $F_X(x, a, b, c, d)$ est la fonction de répartition gamma $F_X(x; 2; 0,8; 0,8; 0,3)$ et la valeur de γ retenue est égale à $0,3$.

Nous souhaitons introduire du bruit pour s $(s < p)$ variables X_1 ,.. , $X = X_k$, .., X_s $(k = 1, .., s)$. Soit α le pourcentage de bruit injecté dans la base de données. Plus généralement, soit $x_i^{(k)}$ la valeur de la variable X_k pour l'individu t_i, nous allons modifier de façon aléatoire $\alpha \times n$ valeurs $x_i^{(k)}$ par le processus précédent.

5.5 Définition et estimation de lois hybrides

Cette section décrit la technique qui a été utilisée pour trouver les fonctions de répartition des variables de la base de données bancaires.

Les échantillons correspondants aux 41 variables retenues pour l'étude peuvent se mettre sous la forme $\{x_{(i)}, f_{(i)}\}$, $i = 1, \ldots, n$, où $x_{(i)}$ sont des occurences ordonnées et $f_{(i)}$ sont des fréquences. La principale caractéristique de ces données est d'avoir une fréquence $f_{(1)}$ très élevée $(f_{(1)} > 0,9)$. Afin d'obtenir une modélisation paramétrique qui prenne en compte cette particularité, nous avons développé un ensemble de lois hybrides à quatre paramètres (a, b, c, d), avec $d \in [0, 1[$, basées sur des lois de probabilité classiques (gamma, lognormale et Weibull) à trois paramètres (a, b, c). Nous allons expliciter dans ce qui suit comment ces lois hybrides sont obtenues, comment générer des nombres aléatoires selon ces lois et comment estimer les paramètres (a, b, c, d). Soit X une variable aléatoire (v.a.) continue définie sur $[c, +\infty[$, dont la fonction de répartition est $F_X(x, a, b, c)$ et vérifie $F_X(c, a, b, c) = 0$. On souhaite définir, à partir de la fonction de répartition $F_X(x, a, b, c)$ de la v.a. X, une nouvelle fonction de répartition $F_Y(y, a, b, c, d)$ définie sur $[c, +\infty[$, ayant pour propriété $F_Y(c, a, b, c, d) = d$. Pour cela, on propose de définir la fonction de répartition $F_Y(y, a, b, c, d)$ de la manière suivante :

$$F_Y(y, a, b, c, d) = \{d + (1 - d)F_X(y, a, b, c)\}1_{y \geq c}$$

On voit clairement que $F_Y(y,a,b,c,d) = 0$ pour $y < c$, $F_Y(c,a,b,c,d) = d$ et $F_Y(y,a,b,c,d) = d + (1-d)F_X(y,a,b,c)$ pour $y > c$. Puisque $F_X(c,a,b,c) = 0$, la distribution de probabilité de la v.a. Y est

$$f_Y(y,a,b,c,d) = (1-d)f_X(y,a,b,c)1_{y \geq c} + d1_{y=c}$$

On montre facilement que le moment non centré d'ordre r de la v.a. Y est égal à $m_r(Y) = (1-d)m_r(Y) + dc^r$ et en « inversant » la fonction de répartition $F_Y(y,a,b,c,d)$, on obtient la fonction de répartition inverse définie pour $d < \alpha < 1$

$$F_Y^{-1}(\alpha,a,b,c,d) = F_X^{-1}\left(\frac{\alpha - d}{1-d}, a, b, c\right)$$

On déduit donc que pour générer aléatoirement une v.a. Y de fonction de répartition $F_Y(y,a,b,c,d)$, il suffit de tirer une v.a. U uniforme sur $(0,1)$ et de calculer

$$Y = \begin{cases} c & \text{si } U \leq d \\ F_X^{-1}\left(\dfrac{U-d}{1-d}, a, b, c\right) & \text{si } U > d \end{cases}$$

À partir d'un échantillon $\{x_{(i)}, f_{(i)}\}$, $i = 1, \ldots, n$, on peut proposer comme estimateurs initiaux pour c et d, $\hat{c} = x_{(1)}$ et $\hat{d} = f_{(1)}$. Pour ce qui est de l'estimation des paramètres a et b nous allons étudier les trois cas suivants, dans lesquels $\hat{m}_1$ et $\hat{\mu}_2$ sont respectivement les moments d'ordre 1 et 2 estimés à partir des données.

– Si X est une v.a. gamma de paramètres $(a > 0, b > 0, c)$ de distribution de probabilité

$$f_X(x,a,b,c) = \frac{1}{b} f_\gamma\left(\frac{x-c}{b}, a\right) = \frac{\exp\{-(x-c)/b\}(x-c)^{a-1}}{b^a \Gamma(a)}$$

alors, pour obtenir $\hat{a}$ et $\hat{b}$ il suffit de calculer

$$\hat{a} = \frac{(\hat{m}_1 + c)^2}{\hat{\mu}_2 - \hat{d}\{\hat{\mu}_2 - (\hat{m}_1 + \hat{c})^2\}} \qquad \text{et} \qquad \hat{b} = \frac{\hat{\mu}_2 - \hat{d}\{\hat{\mu}_2 - (\hat{m}_1 + \hat{c})^2\}}{(\hat{m}_1 + c)(1 - \hat{d})}$$

– Si X est une v.a. lognormale de paramètres $(a, b > 0, c)$ de distribution de probabilité

$$f_X(x,a,b,c) = \left(\frac{b}{x-c}\right) \phi\{a + b\ln(x-c)\}$$

alors pour obtenir $\hat{a}$ et $\hat{b}$ il suffit de calculer

$$\hat{a} = \frac{-\ln(\hat{v})}{\sqrt{2\ln(\hat{u})}} \qquad \text{et} \qquad \hat{b} = \frac{1}{\sqrt{2\ln(\hat{u})}}$$

avec

$$\begin{aligned} \hat{u} &= \frac{\sqrt{(1-\hat{d})(\hat{\mu}_2 - (\hat{m}_1 + \hat{c})^2)}}{\hat{m}_1 - \hat{c}} \\ \hat{v} &= \frac{(\hat{m}_1 - \hat{c})^2}{(1-\hat{d})\sqrt{(1-\hat{d})(\hat{\mu}_2 - (\hat{m}_1 + \hat{c})^2)}} \end{aligned}$$

– Si X est une v.a. de Weibull de paramètres $(a > 0, b > 0, c)$ de distribution de probabilité

$$f_X(x, a, b, c) = \frac{a}{b}\left(\frac{x-c}{b}\right)^{a-1} \exp\left\{-\left(\frac{x-c}{b}\right)^{a}\right\}$$

alors pour obtenir $\hat{a}$ il faut résoudre numériquement l'équation en a ci-dessous

$$\hat{\mu}_2 = (\hat{m}_1 - \hat{c})^2 \left\{\frac{\Gamma(2/a+1)}{(1-\hat{d})\Gamma^2(1/a+1)} - 1\right\}$$

et $\hat{b}$ s'obtient en calculant

$$\hat{b} = \frac{\hat{m}_1 - \hat{c}}{(1-\hat{d})\Gamma(1/\hat{a}+1)}$$

Une fois que les estimateurs initiaux $\hat{a}$, $\hat{b}$, $\hat{c}$ et $\hat{d}$ sont calculés, nous proposons d'utiliser un algorithme d'optimisation pour trouver les estimateurs $\hat{a}^*$, $\hat{b}^*$, $\hat{c}^*$ et $\hat{d}^*$ qui minimisent la distance de Kolmogorov $D = \max|F_Y(y, a, b, c, d) - \hat{F}(y)|$ où $\hat{F}(y)$ est la fonction de répartition empirique. La distance D la plus faible indique quelle distribution hybride doit être choisie pour modéliser les données.

La Figure IV.10 donne la fonction de répartition gamma
$F_X(x; 1,16822; 0,627758; 0; 0,875767)$ de la variable *« nombre de comptes SICAV »*. La valeur minimale de cette variable (*respectivement maximale*) est égale à 0 (*respectivement* 8).

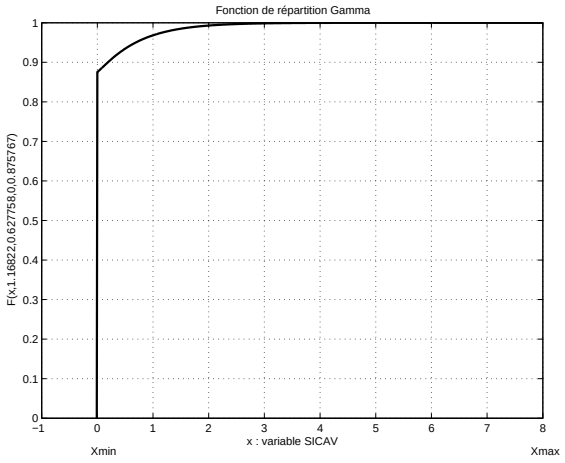


FIG. IV.10 – Fonction de répartition gamma pour la variable « SICAV ».

Le taux de couverture des variables de la base pour la valeur minimale ($X = x_{min}$) est élevé à l'exception de la variable *« âge »*. Trois variables ont une valeur pour le taux de couverture inférieure à 20% (ces variables sont répertoriées dans la troisième catégorie et ont les valeurs suivantes : 12%, 18% et 19%), 10 variables ont un support compris entre 60% et 85% et pour finir, 27 variables ont un support supérieur à 85%. Nous pouvons vérifier l'importance de ne pas changer la valeur du paramètre c (*valeur minimale de* X) pour la fonction de répartition.

5.6 Résultats obtenus

Les expérimentations ont été faites en posant les valeurs suivantes pour les paramètres : $\alpha = 0,10$ (*pourcentage du bruit introduit dans la base*), $\gamma = 0,01$ (*valeur déterminant l'amplitude des intervalles pour les paramètres a, b et d*) et $k = 20$ (*nombre d'itérations*).

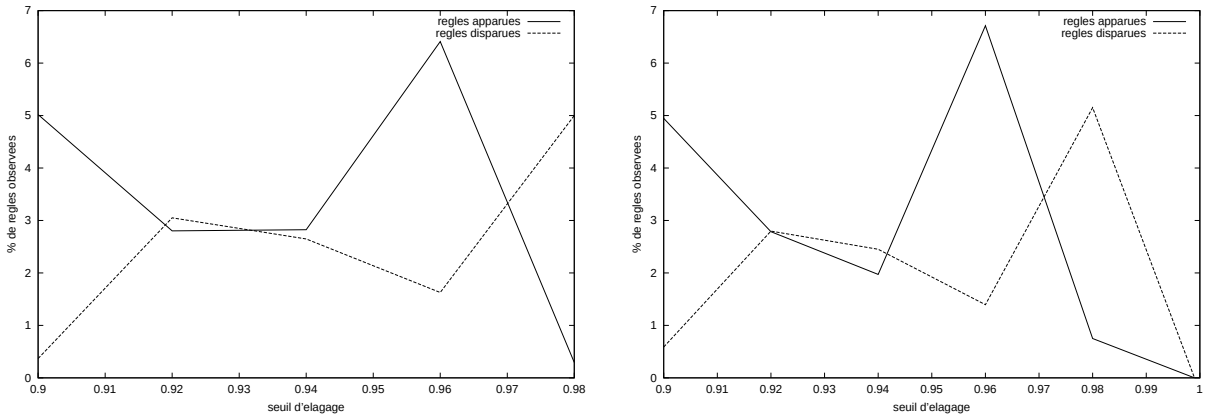


FIG. IV.11 – Effet du bruit sur le coefficient de corrélation linéaire significatif (courbe de droite) et l'indice de vraisemblance du lien (courbe de gauche).

Ces paramètres ont été choisis de manière à injecter un bruit de nature réaliste dans les données étudiées. Il est raisonnable de considérer que les données peuvent contenir jusqu'à 10% de bruit ($\alpha = 0,10$), et que l'erreur associée à chaque valeur incorrecte n'excède pas 1% ($\gamma = 0,01$). Enfin, nous avons choisi d'observer l'effet moyen du bruit pour 20 itérations car les résultats ne semblent pas se modifier en augmentant le nombre d'itérations (*très faible valeur de la variance pour* 20 *itérations*).

Ces différents paramètres nous permettent de « contrôler » le bruit introduit dans les données. Nous pensons que le bruit obtenu est plus réaliste qu'un bruit uniforme ou gaussien car les données sont modifiées en fonction de leur distribution initiale.

Sur les Figures IV.11 et IV.12, l'axe des abscisses correspond au seuil d'élagage des associations, c'est-à-dire la valeur minimale au-dessous de laquelle l'association ne peut être jugée significative et par conséquent, retenue. L'axe des ordonnées correspond, pour la courbe en trait plein, au pourcentage de nouvelles règles qui sont apparues dans la base de données bruitée et pour la courbe en pointillé, au pourcentage de règles qui ont disparu des données bruitées.

La Figure IV.11 montre l'effet du bruit sur le coefficient de corrélation linéaire significatif (courbe de droite) et sur l'indice de vraisemblance du lien (courbe de gauche) et la Figure IV.12 montre l'effet du bruit sur l'intensité d'inclination.

Ayant introduit 10% de bruit dans les données, nous pensions observer l'apparition d'environ 10% de nouvelles règles (*incorrectes par définition*), ainsi que la disparition de 10% de règles existantes. Les résultats observés pour ces deux mesures sont plutôt encourageants car, même pour des seuils d'élagage élevés, le bruit observé reste relativement inférieur au bruit introduit. De plus, la variance observée est proche de 0 ce qui renforce la qualité des résultats.

Pour l'intensité d'inclination, les résultats obtenus sont nettement meilleurs que ceux observés pour les mesures précédentes car le pourcentage de règles nouvelles (obtenues par introduction du bruit dans les données) est très faible et proche de 0 (voir Figure IV.12).

Le bruit moyen observé pour la disparition des règles existantes est proche de 5%, à l'exception d'une valeur élevée pour un seuil d'élagage égal à 1. Cette augmentation du nombre de disparitions est liée au fait que peu de règles sont extraites lorsque le seuil d'élagage est fixé à 1. Ainsi, la disparition de peu de règles entraîne l'apparition de cette valeur « extrême » sur la courbe.

Rappelons toutefois que ces résultats ont été obtenus sur une seule base de données et que les

attributs étudiés sont de nature ordinale (c'est-à-dire munis d'une structure d'ordre, contrairement aux attributs booléens par exemple). Nous ne pouvons donc pas généraliser ces résultats à d'autres types de données, notamment des données booléennes étudiées dans la première partie de cette thèse.

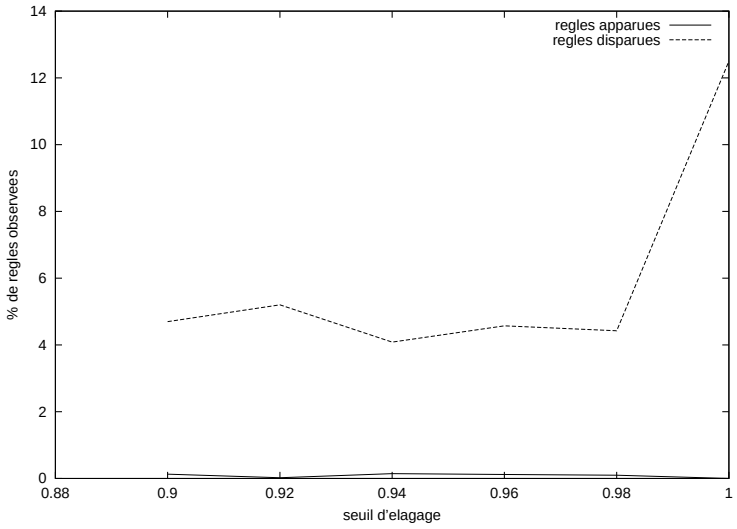


FIG. IV.12 – Effet du bruit sur l'intensité d'inclination.

5.7 Conclusion de l'étude réalisée sur les données bancaires

L'extraction non supervisée de connaissances dans des données volumineuses est particulièrement difficile car l'évaluation de la qualité des résultats obtenus repose essentiellement sur les mesures de qualité utilisées pour obtenir ces résultats. Un des critères majeurs de qualité (selon nous), lorsque les données sont réelles et donc imparfaites, est de minimiser l'impact du bruit sur l'apparition de règles incorrectes. Les mesures ne vérifiant pas ce critère doivent être manipulées avec la plus grande précaution car les résultats fournis à l'expert, seul juge de la qualité des connaissances obtenues, peuvent s'avérer incomplets et surtout incorrects. Nos travaux ont permis de mettre en évidence, pour le corpus considéré et les mesures étudiées, que l'impact du bruit sur les connaissances obtenues n'est pas négligeable (bien que relativement faible) et varie en fonction des mesures de qualité utilisées.

Ces travaux doivent être poursuivis sur d'autres bases de données, de manière à valider les premiers résultats observés. La poursuite de ces travaux nécessite d'une part de pouvoir estimer efficacement les fonctions de distribution des variables des bases de données étudiées et d'autre part, de valider l'approche retenue pour introduire du bruit dans les données.

6 Conclusion sur l'étude du bruit

Les travaux présentés dans ce chapitre montrent l'importance de l'étude de l'impact des données bruitées sur les connaissances que nous pouvons en extraire. Les résultats obtenus diffèrent en fonction de la nature des données étudiées. En effet, lorsque nous avons étudié l'impact du bruit sur des données ordinales, nous avons pu constater que l'extraction de connaissances fiables est possible notamment en utilisant l'intensité d'implication ordinale qui s'avère peu sensible au bruit.

Inversement, l'étude réalisée sur les données booléennes montre qu'aucune mesure n'est totalement insensible au bruit et ce malgré les solutions algorithmiques que nous avons proposées pour diminuer l'impact du bruit.

Nous avons proposé et étudié deux solutions pour réduire l'impact du bruit dans les données sur les connaissances obtenues :

- détecter le support minimal tel qu'aucune connaissance très intéressante ne disparaisse et utiliser ce support pour élaguer les règles d'association recherchées
- calculer et associer à chaque règle d'association un indicateur de fiabilité correspondant à la capacité du système à retrouver cette règle lorsque les données sont bruitées

La solution que nous avons retenue est de proposer à l'expert un indicateur de fiabilité associé à chaque mesure. Les résultats obtenus, lorsque nous avons évalué cet indicateur de fiabilité, ont permis d'isoler deux groupes de mesures :

- la moindre contradiction qui s'avère être la plus fiable des six mesures testées
- l'intensité d'implication classique et normalisée, la confiance, l'indice de Lœvinger et la nouveauté dont les comportements sont comparables.

Les expérimentations réalisées sur les données booléennes utilisent un modèle du bruit peu réaliste car le bruit introduit est purement aléatoire et ne prend donc pas en considération la couverture des différents attributs dans les données initiales. Pour l'étude des données bancaires qui est présentée dans la section 5, nous avons pu déterminer précisément les distributions de la majorité des variables ordinales. Nous avons alors utilisé cette information pour introduire un bruit plus réaliste dans les données et les résultats que nous avons obtenus sont probablement plus fiables, pour ces données, que ceux obtenus sans prendre en considération la nature des attributs.

Cependant, bien que nous soyons convaincus que l'amélioration du modèle utilisé, pour les données booléennes, devrait permettre d'obtenir des résultats plus concluants, l'expérience menée sur les données bancaires a montré que l'estimation des distributions des variables est coûteuse. Il nous semble donc difficile d'obtenir une méthode générale permettant de bruiter de manière réaliste tout type de données.

Notons que pour les expérimentations réalisées sur des données aléatoires, l'utilisation d'un bruit lui aussi aléatoire semble la solution la mieux adaptée.

Dans toutes les expérimentations réalisées, nous avons considéré que les bases de données initiales étaient « saines » et que les connaissances obtenues à partir de ces données pouvaient être considérées comme un oracle lors de l'étude de l'impact du bruit. Cette démarche peut être comparée à un apprentissage supervisé dans lequel nous disposons d'un ensemble de connaissances (dans notre cas des règles d'association) dont l'intérêt est connu. Cet ensemble est ensuite utilisé pour évaluer la capacité de différentes mesures de qualité à extraire, à partir de données bruitées, les « bonnes » connaissances. Dans notre cas, les données initiales nous fournissent cet ensemble de connaissances, mais nous pourrions envisager qu'il soit fourni par une source extérieure.

L'indicateur de fiabilité proposé dans cette thèse ne représente qu'une solution pour évaluer la résistance au bruit d'une règle. Notre tentative pour extraire automatiquement les connaissances fiables à partir de données bruitées n'est pas très concluante et nous pensons que nous ne pouvons que difficilement espérer minimiser l'impact du bruit en apprentissage non supervisé.

Notons que le coût associé à l'utilisation de l'indicateur de fiabilité est relativement élevé puisqu'il est nécessaire, non seulement de bruiter les données plusieurs fois, mais à chaque itération, il faut aussi extraire les règles d'association intéressantes. Or nous avons vu que le coût de notre algorithme peut être élevé ce qui rend l'utilisation de l'indicateur de fiabilité difficilement applicable sur des données fortement corrélées et décrites par de nombreux attributs.

L'étude d'approches issues de l'apprentissage supervisé pour pouvoir estimer la qualité des

connaissances obtenues, au moins en terme de résistance au bruit, pourrait permettre d'obtenir une méthode comparable et utilisable de manière interactive par l'expert.

CHAPITRE V

La fouille de textes

1 Introduction

La recherche de pépites de connaissance représente, comme nous l'avons vu dans les chapitres précédents, un défi en soi pour la communauté de fouille de données. Les outils développés dans ce contexte peuvent être utilisés dans des domaines connexes tels que celui de la fouille de textes.

L'extraction de connaissances à partir d'un corpus de textes peut prendre plusieurs formes : extraction de la terminologie associée au domaine, construction automatique d'une ontologie reliant les concepts découverts dans le corpus, découverte de formes syntaxiques spécifiques au domaine, extraction de règles d'association entre les concepts du domaine, *etc.*

Dans le cadre de cette thèse, nous nous intéresserons principalement à l'aspect « extraction de règles d'association entre les concepts du domaine ».

Les travaux relatifs à la fouille de textes ont été réalisés en étroite collaboration avec Mathieu Roche, qui s'intéresse, dans le cadre de sa thèse, à l'extraction de terminologies et la détection de traces de concepts à partir de textes [Roche 2003, Roche et al. 2004].

L'approche que nous avons retenue se décompose en trois grandes phases :

1. collecter un corpus homogène sur un thème donné
2. réaliser une détection des traces de concepts spécifique au corpus
3. extraire et valider des règles d'association à partir du corpus

La Figure V.1 présente les différentes étapes du processus de fouille de textes.

1.1 Collecter un corpus homogène sur un thème donné

Cette première phase concernant la collecte du corpus ne sera pas abordée en détail dans la suite de cette thèse. Cependant, il est important de noter que cette étape constitue la brique de base de l'ensemble du processus de fouille de textes. Le succès du processus est fonction de la qualité et de l'homogénité du corpus collecté. La constitution du corpus est donc confiée explicitement à un expert.

Nous supposerons dans la suite de nos travaux que les différents corpus étudiés présentent des critères de qualité et d'homogénité nous permettant d'appliquer nos outils de fouille de textes.

En entrée des différents traitements que nous allons décrire dans la suite de ce chapitre et pour vérifier la généricité de notre méthodologie, nous utiliserons cinq corpus de langue et de technicité différentes. Ces corpus sont présentés en section 2.1.

1.2 Réaliser une détection des traces de concepts spécifique au corpus

La deuxième phase du processus de fouille de textes est composée de deux étapes :
- recherche des termes dans les textes (étape A, Figure V.1)
- utilisation des termes pour détecter les traces de concepts (étape B, Figure V.1)

La recherche des termes consiste à isoler des collocations [Halliday 1976] pour le domaine étudié.

Par exemple, les collocations « decision-tree » et « genetic-algorithm » ont été détectées dans un corpus d'introductions d'articles scientifiques du domaine de la fouille de données.

Cette première étape est réalisée de manière automatique par le système présenté dans [Roche 2003]. La liste des termes obtenue permet à l'expert d'associer les termes à des concepts, c'est-à-dire des regroupements de termes ayant la même sémantique (étape B).

Voici deux exemples de concepts de termes utilisés pour illustrer les traces de concepts détectées dans un des corpus étudiés :

$$concept_{NatofOutput} = \{frequent - pattern, interesting - rule, decision - table\}$$
$$concept_{KnownMethods} = \{decision - tree, genetic - algorithm\}.$$

Comme nous le verrons dans la section 2.5, l'expert du domaine n'associe pas seulement les termes à des concepts, il associe également des relations syntaxiques aux concepts. Par exemple, les relations $(find : Verbe, cluster : Objet)$ et $(bi - directional : Adjectif, rule : Nom)$ sont associées au concept « NatofOuput » (ce concept représentant la nature des sorties des systèmes présentés dans les articles). Les relations syntaxiques correspondent aux sorties du Shallow Parser de Xerox. Nous obtenons ainsi une liste de concepts décrits par un ensemble de termes ou de relations syntaxiques.

1.3 Extraire et valider des règles d'association à partir du corpus

La dernière phase est elle aussi divisée en deux étapes :
- extraction de connaissances (étape C, Figure V.1)
- validation (étape D, Figure V.1)

La détection des concepts permet de réécrire le corpus de manière plus compacte. En effet, chaque instance de concept détectée dans les textes est remplacée par le concept lui-même.

Un système d'extraction de connaissances, détaillé dans la section 6, est appliqué au corpus réécrit. Les connaissances extraites sont présentées sous la forme de règles d'association (définies dans le Chapitre II). Ces règles permettent de mieux comprendre les interactions entre certains concepts du domaine. Les connaissances ainsi obtenues sont validées par un expert du domaine (étape D, Figure V.1).

2 Détection des traces de concepts dans les corpus

Les étapes A et B présentées sur la Figure V.1 sont détaillées dans les sections suivantes.

2.1 Description des corpus

Les cinq corpus que nous avons utilisés sont issus de domaines différents et rédigés dans des langues et des styles suffisaments différents pour nous permettre de tester la généricité de la méthodologie retenue.

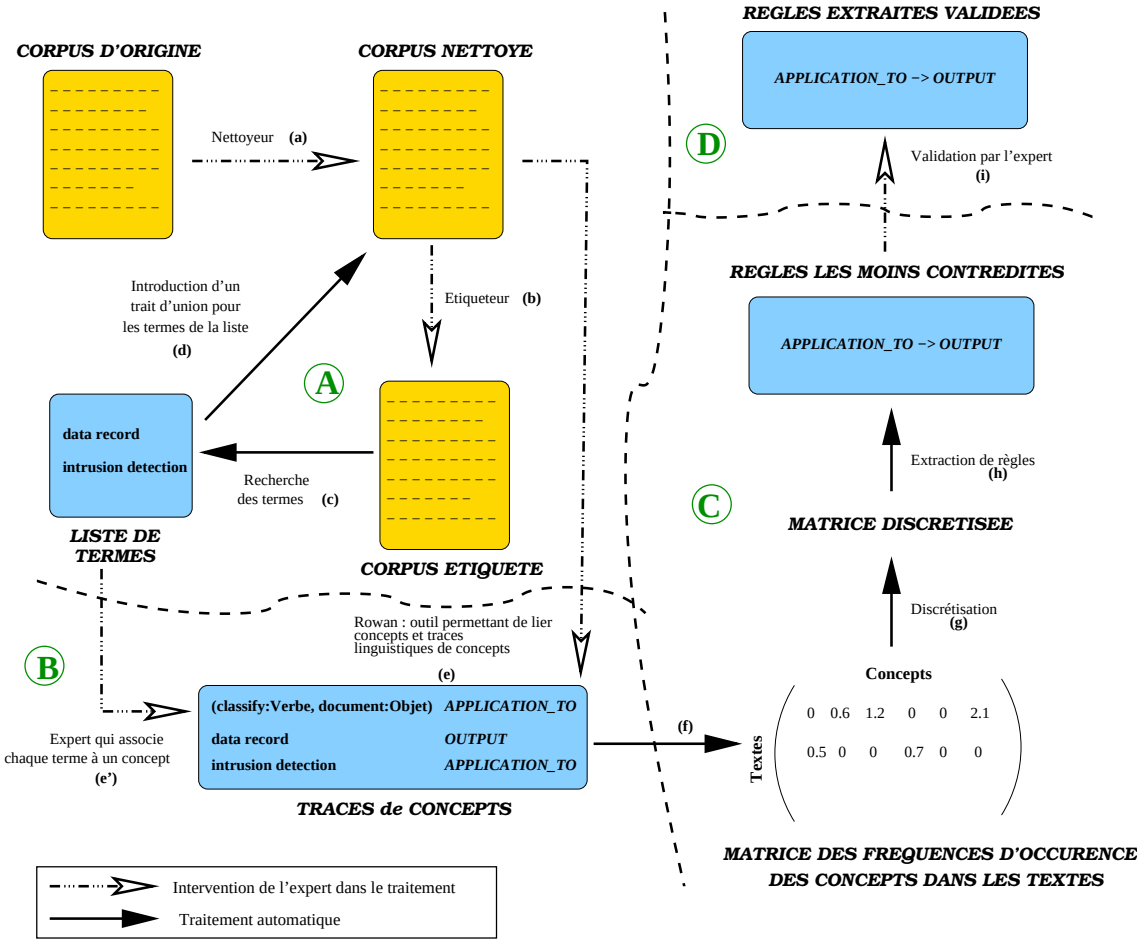


FIG. V.1 – Processus global de la fouille de textes.

- Le corpus de biologie moléculaire (9424 Ko) a été obtenu par une requête au *NIH* sur Medline[1] (PubMed) avec les mots-clés *DNA-binding, proteins, yeast*. Le résultat de cette requête nous a fourni un corpus de 6119 résumés d'articles scientifiques en anglais. Ce corpus illustre le problème du traitement d'un grand nombre de textes écrits dans une langue étrangère fortement technique.
- Le second corpus est constitué de 100 introductions d'articles traitant de la fouille de données. Ce corpus est rédigé en langue anglaise par des anglophones et représente 369 Ko.
- Le troisième corpus est constitué de 31 introductions d'articles traitant de la fouille de données. Ce corpus est rédigé en langue anglaise par des francophones et représente 113 Ko.
- Le corpus en ressources humaines d'une taille de 3784 Ko (fourni par la compagnie PerformanSe[2]), a été rédigé, en français, par un psychologue qui sert d'expert pour ce corpus.
- Le corpus de Curriculum Vitæ, d'une taille de 2470 Ko, contient 1144 CVs (fourni par le groupe VediorBis[3]). Ces textes sont écrits, en langue francaise, dans un mode semi-télégraphique et contiennent beaucoup de fautes d'orthographe.

2.2 Normalisation (ou nettoyage) des corpus

Chaque corpus exige un type particulier de nettoyage. Le nettoyage peut cependant être divisé en deux étapes :

- élimination de portions de textes formatées, voir l'exemple ci-après
- application de règles de nettoyage spécifiques au domaine étudié

La première étape permet d'enlever les noms des auteurs, supprimer le nom des laboratoires, éliminer les codes propres aux bases de données d'où sont extraits certains textes.

La deuxième étape permet d'uniformiser les différents textes du corpus en uniformisant les notations utilisées par les différents auteurs par exemple. Un grand nombre de règles sont appliquées, leur nombre exact et leur variété dependent de chaque corpus. Par exemple, sur le corpus de biologie moléculaire, nous avons appliqué deux grands types de règles. Toutes les occurrences de *amino-terminal*, *amino-termini*, *N-terminal*, *N-termini*, *NH2-terminal* et *NH2-termini* ont été remplacées par « N-term ». Ce type d'opération, consistant à uniformiser le vocabulaire employé, est effectué par environ 100 groupes de règles. Le deuxième type de traitement, représentant 1932 règles, consiste à remplacer les alias de gènes par leur nom générique connu dans le domaine.

L'exemple suivant est issu du corpus de biologie moléculaire. Le nettoyage réalisé correspond aux phrases en gras dans l'exemple qui ont été supprimées à l'aide d'une ou plusieures règles conçues par l'expert.

Exemple :

1 : Biochim Biophys Acta 2001 Dec 30 ;1522(3) :175-86

The modulation of the biological activities of mitochondrial histone Abf2p by yeast PKA and its possible role in the regulation of mitochondrial DNA content during glucose repression.

Cho JH, Lee YK, Chae CB.

Department of Life Science and Division of Molecular and Life Science, Pohang University of Science and Technology, 790-784, Pohang, South Korea

[1] `http ://www.nlm.nih.gov/`
[2] `http ://www.performanse.fr/`
[3] `http ://www.vediorbis.com/`

The mitochondrial histone, Abf2p, of Saccharomyces cerevisiae is essential for the maintenance of mitochondrial DNA (mtDNA) and appears to play an important role in the recombination and copy number determination of mtDNA.

PMID : 11779632 [PubMed - in process]

2.3 Étiquetage

L'étiquetage (*b*, Figure V.1) correspond au fait d'apposer à chaque mot du texte une étiquette grammaticale.

Cette étape permet d'extraire les couples ou triplets de mots, appelés collocations, ayant une étiquette grammaticale spécifique : Nom-Nom, Nom-Préposition-Nom, Adjectif-Nom, Nom-Verbe_gérondif, Formule-Nom[4], Nom-Adjectif[5], *etc.*.

Cependant, une partie non négligeable des mots, en particulier lorsque le domaine est spécialisé, sont inconnus de l'étiqueteur de Brill. L'absence de ces mots dans le lexique de l'étiqueteur provoque inévitablement des erreurs d'étiquetage.

Dans cette étape, nous utilisons l'étiqueteur de Brill [Brill 1994] que nous allons décrire très succintement.

2.3.1 L'étiqueteur de Brill

Cet étiqueteur est gratuit et disponible à l'adresse suivante : `http ://www.cs.jhu.edu/~brill/`. Il est disponible pour deux langues : le français et l'anglais. La version française n'est disponible qu'après avoir établi une convention de travail avec l'auteur.

Cet étiqueteur a été entrainé sur des corpus généralistes : le « wall street journal » pour l'anglais et sur la base textuelle FRANTEXT[6].

Le principe de cet étiqueteur est, comme nous l'avons déjà dit, d'apposer une étiquette grammaticale à chacun des mots d'un texte. L'étiquetage est réalisé par deux modules : le module lexical puis le module contextuel.

Le module lexical utilise un lexique de mots appris à partir des corpus d'entraînement. Les différents mots du lexique peuvent prendre plusieurs étiquettes grammaticales différentes en fonction du contexte dans lesquels ils sont employés. Par exemple, en anglais, le mot « run » peut être étiqueté *Verbe* (verbe courir) ou *Nom Commun* (une course). Lorsqu'un mot peut prendre plusieurs étiquettes grammaticales, l'étiqueteur de Brill commence par choisir l'étiquette la plus fréquente et l'associe au mot. Puis, dans une deuxième phase, un module contextuel est utilisé pour vérifier les étiquettes retenues par le module lexical et corriger les éventuelles erreurs en utilisant le contexte. Le principe de fonctionnement de ce module est fondé sur l'application successive de règles contextuelles prenant en considération un mot du texte et son voisinage immédiat (une fenêtre de un à trois mots avant et après le mot étudié).

Enfin, certains mots spécifiques au domaine étudié sont absents du lexique. Ainsi, sur le corpus de fouille de données, plus de 28% des mots sont inconnus du lexique standard de l'étiqueteur de Brill. Et sur le corpus de biologie moléculaire, plus de 70% des mots sont inconnus par le lexique de l'étiqueteur de Brill.

[4] uniquement pour le corpus de Biologie Moléculaire

[5] uniquement pour les corpus francophones

[6] `http ://zeus.inalf.fr/frantext.htm`

Dans ce cas, le comportement du module lexical de l'étiqueteur est divisé en deux étapes. Dans un premier temps, le module lexical associe au mot inconnu l'étiquette *Nom Propre* si le mot commence par une majuscule, et *Nom Commun* sinon. Puis, afin d'améliorer la précision de cet étiquetage arbitraire, le module lexical applique un ensemble de règles lexicales (apprises de manière supervisée). Les règles suivantes illustrent le type de transformations lexicales permises par l'étiqueteur.

- étant donné un mot étiqueté *Nom Commun*, changer son étiquette par *Verbe au Participe Passé* si le mot possède le suffixe « -ed »
- étant donné un mot étiqueté *Nom Commun*, changer son étiquette par *Nombre* si le symbole « $ » apparaît immédiatement à gauche du mot

L'intervention d'un expert du domaine est nécessaire pour écrire ces règles lexicales. Leur utilisation permet de réduire significativement le taux d'erreur en étiquetage. Par exemple, dans le domaine de la fouille de données, le fait d'ajouter quatorze règles lexicales aux règles existantes permet de réduire le pourcentage de mots inconnus et de le ramener à moins de 8%. De même, en biologie moléculaire, l'ajout d'une trentaine de règles lexicales a permis de ramener le pourcentage de mots inconnus à 15%.

Après l'application de ces règles, le module contextuel décrit précédemment est appliqué sur le texte.

La phrase suivante issue du corpus des ressources humaines illustre le travail réalisé par l'étiqueteur de Brill.

Exemple :

Phrase à étiqueter : En effet, vous tirez l'essentiel de votre fierté de la fiabilité de vos réalisations.

Phrase étiquetée : En/**PREP** effet/**SBC :sg** ,/**,** vous/**PRV :pl** tirez/**VCJ :pl** l'/**DTN :sg** essentiel/**SBC :sg** de/**PREP** votre/**DTN :sg** fierté/**SBC :sg** de/**PREP** la/**DTN :sg** fiabilité/**SBC :sg** de/**PREP** vos/**DTN :pl** réalisations/**SBC :pl** ./**.**

Le Tableau V.1 précise le sens associé aux étiquettes utilisées dans l'exemple.

étiquette	**sens**
PREP	préposition
SBC	nom commun
PRV	pronom verbal
VCJ	verbe conjugé
DTN	déterminant
sg	singulier
pl	pluriel

TAB. V.1 – Signification de quelques unes des étiquettes de Brill.

2.4 Acquisition des termes

Cette étape permet d'extraire les termes les plus pertinents pour le domaine. La mesure de pertinence se fonde sur une mesure d'association favorisant l'association des mots qui sont aussi rarement associés que possible à tous les autres mots.

Par exemple, dans le corpus de biologie, examinons le terme *single-strand-DNA*, et dans le corpus d'introductions d'articles, le terme *data-mining*. Dans leurs corpus respectifs, les mots *DNA* et *data* sont liées à beaucoup d'autres mots, et ceci diminue la pertinence des termes. En revanche, le mot *single-strand* est pratiquement toujours associe à *DNA*, et le mot *mining* toujours précédé de *data*.

M. Roche a donc proposé une mesure permettant de prendre cette particularité en considération afin d'augmenter la pertinence des termes.

Nous détaillerons le choix de la mesure dans la section 2.4.2.

2.4.1 Mesure d'évaluation pour l'extraction

La qualité des termes extraits automatiquement peut être évaluée en utilisant deux mesures classiques : la Précision et le Rappel.

Les résultats obtenus avec ces deux mesures peuvent être visualisés à l'aide de deux types de courbes : les courbes d'élévation pour la Précicion et les courbes ROC (Receiver Operating Characteristic) pour le Rappel.

Avant de présenter la méthode retenue pour évaluer la qualité de nos termes, rappelons la définition de ces deux mesures de qualité.

Notons $\mathcal{E}$ la liste des termes extraits et $\mathcal{L}$ la liste des termes appartenant à la classification conceptuelle. La liste$\mathcal{L}$ est constituée par l'expert du domaine.

$$Précision = \frac{|\mathcal{E} \cap \mathcal{L}|}{|\mathcal{E}|}$$

$$Rappel = \frac{|\mathcal{E} \cap \mathcal{L}|}{|\mathcal{L}|}$$

Les courbes d'élévation consistent à donner la variation de la précision en fonction du nombre de termes extraits par le système.

Les courbes ROC, introduites dans le domaine du traitement du signal, permettent de visualiser le rappel des termes corrects par rapport au rappel des termes incorrects. En effet, sur ces courbes, le pourcentage de termes négatifs (ou incorrects) identifiés par le système est indiqué sur l'axe des abscisses et l'axe des ordonnées correspond au pourcentage de termes corrects (le rappel).

Pour évaluer la qualité de nos termes, nous n'avons utilisé que la Précision et les courbes d'élévation. Ce choix est contraint par le fait que notre approche est non supervisée et il nous est donc impossible de connaître de manière exhaustive la liste $\mathcal{L}$. Par conséquent, le calcul exact du rappel et l'utilisation des courbes ROC s'avèrent impossibles.

2.4.2 Mesure utilisée pour l'extraction

Il existe un certain nombre de mesures dans la littérature et nous en avons examiné plusieurs [Church et Hanks 1990, Dunning 1993, Daille et al. 1998]. Notre observation principale, voir [Roche et al. 2003] pour une comparaison détaillée de plusieurs mesures, est que la courbe d'élévation obtenue en employant le rapport de vraisemblance (« Loglike ») [Dunning 1993] donne les meilleurs résultats sur l'ensemble de nos corpus, c'est pourquoi nous l'avons retenue.

La Figure V.2 présente les courbes d'élévation obtenues avec les deux mesures traditionnelles pour l'extraction de la terminologie du domaine : l'information mutuelle [Church et Hanks 1990] et le rapport de vraisemblance [Dunning 1993]. Les expérimentations présentées ici ont été effectuées sur le corpus des Ressources Humaines avec la relation nom-adjectif ayant un nombre d'occurences supérieur à trois.

Au rapport de vraisemblance, nous pouvons ajouter différents paramètres, présentés dans [Roche 2003], afin d'améliorer la précision obtenue. Notre algorithme d'extraction de la terminologie s'effectue en plusieurs itérations. À chaque itération, les termes binaires extraits (nom-nom, adjectif-nom, *etc.*) sont introduits dans le corpus avec un trait d'union. Lors des itérations suivantes, ces termes sont

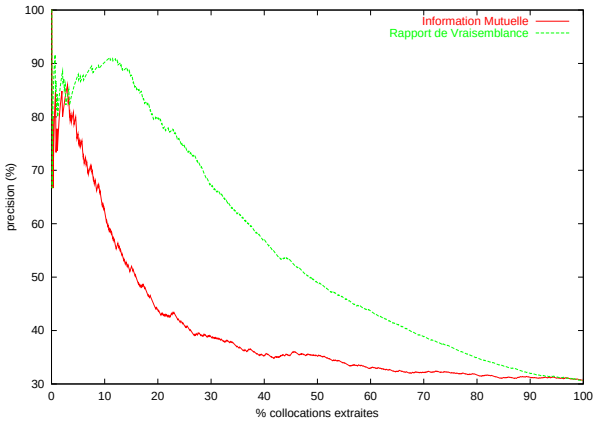


FIG. V.2 – Courbe d'élévation avec la relation nom-adjectif pour le corpus des Ressources Humaines. Nous n'avons sélectionné ici que les termes qui apparaissent plus de 3 fois.

donc reconnus comme des mots à part entière et permettent alors la formation d'autres termes. Un des paramètres essentiels que nous avons ajouté dans notre approche consiste à privilégier les termes formés avec les mots inclus dans les termes des itérations précédentes. Un tel paramètre permet de privilégier plus spécifiquement le vocabulaire du domaine.

2.5 Association des termes et des relations syntaxiques aux concepts

Cette étape de validation de la liste de termes trouvée automatiquement est essentielle pour notre travail (Figure V.1). Cette étape qui est entièrement exécutée de manière « manuelle » par un expert du domaine consiste à estimer si chacun des termes extraits appartient ou non à un concept. Dans le cas positif, l'expert précise le nom du concept auquel les termes appartiennent. Ainsi, l'ensemble des descriptions de concepts construit à partir d'un corpus spécialisé est composé de termes mais également de relations syntaxiques. Cette connaissance nous permet de reformuler le corpus de manière plus compacte. À partir de cette reformulation nous recherchons des connaissances spécifiques au corpus étudié.

3 Réécriture des textes sous une forme plus compacte

Les concepts associés par l'expert aux ensembles de termes et de relations syntaxiques sont utilisés pour reformuler les textes. L'objectif étant d'extraire des règles d'association à partir du corpus, chaque texte est représenté par un vecteur de n valeurs (n étant le nombre de concepts créés par l'expert). Une valeur représentant la fréquence d'occurence du concept concerné dans le texte. Par exemple, si pour un corpus de 10 textes, l'expert a regroupé les termes et relations syntaxiques

en 4 concepts, le corpus peut être représenté par la matrice suivante :

$$\begin{pmatrix} 0 & 0,1 & 0,02 & 0,132 \\ 0,1 & 0,07 & 0,06 & 0 \\ 0,01 & 0,05 & 0 & 0,4 \\ 0,181 & 0,172 & 0,043 & 0 \\ 0,24 & 0 & 0 & 0,214 \\ 0 & 0,01 & 0,302 & 0,184 \\ 0,114 & 0,06 & 0,12 & 0 \\ 0,08 & 0,078 & 0,013 & 0,34 \\ 0 & 0,14 & 0,1 & 0,41 \\ 0,071 & 0,098 & 0,06 & 0 \end{pmatrix}$$

La fréquence de chaque concept dans un texte donné est calculée en divisant le nombre d'occurence des instances du concept dans le texte par le nombre total de mots trouvés dans ce même texte. Ainsi, la fréquence du concept c_j dans le texte t_i est égale à :

$$frequence(c_j, t_i) = \frac{|instances(c_j, t_i)|}{|mots(t_i)|}$$

Ces données doivent être transformées avant de pouvoir leur appliquer notre méthode de détection de pépites de connaissance. Cette transformation consiste à réaliser une discrétisation de chaque concept puis une transformation « triviale » de chaque concept en attributs booléens.

4 Discrétisation de la matrice d'occurence des concepts

Avant de présenter la méthode que nous avons retenue pour effectuer la discrétisation des données manipulées, nous allons présenter un état de l'art des techniques de discrétisation.

4.1 État de l'art en discrétisation

De nombreuses méthodes de discrétisation ont été proposées dans la littérature. Les travaux de [Dougherty et al. 1995, Liu et al. 2002] présentent un état de l'art relativement complet des diverses techniques de discrétisation. Nous présenterons donc ici un tour d'horizon similaire complété par quelques techniques plus récentes.

Les diverses techniques de discrétisation présentées dans [Dougherty et al. 1995, Liu et al. 2002] se répartissent dans les catégories suivantes :

- Discrétisation supervisée vs non-supervisée
- Approche statique vs dynamique
- Approche locale vs globale
- Approche fusionner vs diviser
- Approche descendante vs ascendante
- Approche monothétique vs polythétique

Nous présentons en détail dans les sections suivantes ces diverses catégories, puis nous présenterons plusieurs méthodes de discrétisation appartenant à ces catégories.

4.2 Discrétisation supervisée vs non-supervisée

Lorsque les données étudiées sont telles que chaque individu est associé à une classe, il est alors possible d'utiliser des techniques de discrétisation dites supervisées. Dans ce cas, la classe associée

à chaque individu est utilisée pour réaliser au mieux la discrétisation. En effet, si deux individus o_1 et o_2 ont été associés à la même classe c alors il est probable qu'ils partagent des propriétés. La discrétisation supervisée utilise ce principe pour discrétiser les différents attributs.

Lorsqu'aucune information de type classe n'est indiquée dans les données, seules les techniques de discrétisation non supervisées sont utilisables. Ces techniques sont souvent fondées sur l'utilisation de distances entre les valeurs des attributs ou sur un découpage arbitraire des valeurs en k intervalles.

4.3 Approche statique vs dynamique

La plupart des méthodes de discrétisation sont de type statique, c'est-à-dire que pour chaque attribut à discrétiser, le nombre d'intervalle k à créer est déterminé en une passe sur les données. Cette valeur k est spécifique à chaque attribut et est souvent déterminée indépendamment des valeurs précédentes (associées aux autres attributs).

Les méthodes de discrétisation dynamiques parcourent l'espace de toutes les valeurs possibles de k pour tous les attributs en même temps et ce faisant elles peuvent « capturer » des dépendances entre les différents attributs.

4.4 Approche locale vs globale

Les techniques de discrétisation locales (par exemple C4.5) n'utilisent qu'un sous-ensemble des individus pour effectuer la discrétisation. Les choix effectués sur ce sous-ensemble ne sont donc pas toujours valides pour toute la population.

Par opposition, les techniques dites globales utilisent tous les individus pour réaliser la discrétisation d'un ou plusieurs attributs. Ces méthodes sont donc *a priori* moins sensibles aux phénomènes « locaux » pouvant apparaître dans les données.

4.5 Approche descendante vs ascendante

Les méthodes de discrétisation ascendantes associent à chaque individu un intervalle centré sur l'individu et ne contenant initialement que cet individu. Puis de proche en proche les intervalles sont fusionnés selon des critères de distances, de taille et de nombre d'intervalles (voir figure V.3).

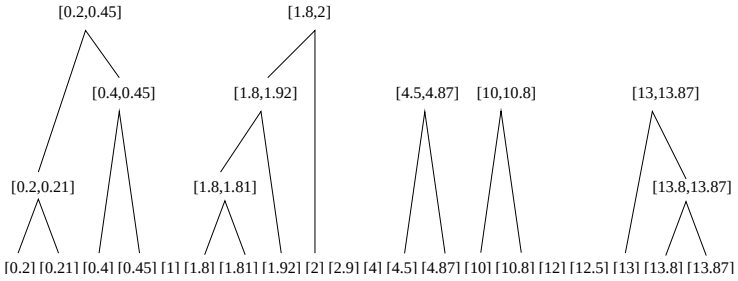


FIG. V.3 – Approche ascendante.

Les méthodes de discrétisation descendantes considèrent que les individus appartiennent initialement tous au même intervalle, puis cet intervalle est divisé en sous-intervalles selon des critères semblables à ceux utilisés par les approches ascendantes (voir figure V.4).

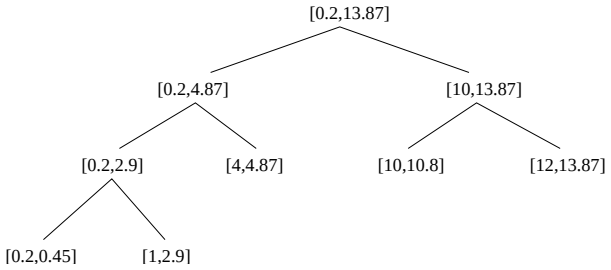


FIG. V.4 – Approche descendante.

$0,1$	$\oplus$	$1,81$	$\ominus$	6	$\oplus$	$13,8$	$\ominus$
$0,2$	$\oplus$	$1,92$	$\oplus$	8	$\ominus$	$13,87$	$\ominus$
$0,21$	$\oplus$	2	$\oplus$	10	$\ominus$	15	$\oplus$
$0,4$	$\ominus$	$2,9$	$\ominus$	$10,8$	$\ominus$	$15,7$	$\oplus$
$0,45$	$\ominus$	4	$\ominus$	12	$\oplus$	$16,1$	$\oplus$
1	$\oplus$	$4,5$	$\ominus$	$12,5$	$\oplus$	$16,9$	$\oplus$
$1,8$	$\oplus$	$4,87$	$\oplus$	13	$\ominus$	18	$\ominus$

TAB. V.2 – Valeurs d'un attribut pour un jeu de données contenant deux classes :$\ominus$ et $\oplus$.

4.6 Approche monothétique vs polythétique

Les méthodes de discrétisation monothétiques discrétisent chaque attribut individuellement. Les éventuelles interactions entre les différents attributs n'influent donc pas sur la discrétisation.

Par opposition, les méthodes de discrétisation polythétiques effectuent la discrétisation des attributs en tenant compte des interactions entre les différents attributs.

Dans la suite de cet état de l'art, nous présentons en détail plusieurs méthodes de discrétisation. Ces méthodes sont regroupées en deux grandes familles : supervisée et non supervisée.

4.7 Approches supervisées

Nous utiliserons l'exemple présenté dans le Tableau V.2 pour illustrer les différentes méthodes de discrétisation supervisées.

4.7.1 *ChiMerge*

[Kerber 1992] propose une méthode de discrétisation supervisée, ascendante, et monothétique.

Initialement, pour chaque attribut, toutes les valeurs sont associées à un intervalle. Le principe de la méthode est de fusionner les intervalles adjacents minimisant le test du χ^2. L'algorithme 10 implante la méthode du *ChiMerge*.

Il est recommandé de fixer le niveau de test entre 0.9 et 0.99, et de fixer $max_{interval}$ entre 10 et 15 pour éviter que l'algorithme ne crée trop d'intervalles.

Algorithme 10 *ChiMerge*$(\mathcal{B}_n^p, level, max_{interval})$

Entrée :

$\mathcal{B}_n^p$: base de données dont les p attributs sont continus

$level$: niveau de validité retenu pour le test du χ^2

$max_{interval}$: nombre maximal d'intervalles créés par la méthode

Sortie :

$\{I_j, 1 \leq j \leq p\}$: liste des intervalles trouvés

Début

Pour $(j \leftarrow 1 ; j \leq p ; j++)$ **Faire**

$D \leftarrow \{(\alpha_i^j, etiq(i)), 1 \leq i \leq n\}$ *-- sélection des valeurs de l'attribut j pour tous les individus*

trier les valeurs de D selon α_i^j

-- chaque valeur est considérée comme un intervalle

$I_j \leftarrow \emptyset$

Pour tout $(c \in D)$ **Faire**

$I_j \leftarrow I_j \cap \{([c.\alpha, c.\alpha], c.etiq)\}$

Fin Pour

Répéter

$E \leftarrow \emptyset$

Pour $(k \leftarrow 1 ; k < |I_j| ; k++)$ **Faire**

$i_1 \leftarrow I_j[k]$; $i_2 \leftarrow I_j[k+1]$; *-- i_1 et i_2 sont deux intervalles adjacents*

calculer $\chi^2(i_1, i_2)$

Si $(\chi^2(i_1, i_2) \leq level)$ **Alors**

$E \leftarrow E \cap (i_1, i_2)$

Fin Si

Fin Pour

$E' \leftarrow \{(a, b) \in E, \chi^2(a, b) = min(\chi^2(E))\}$

Pour tout $((a, b) \in E')$ **Faire**

$nb_{interval} \leftarrow nb_{interval} - 1$

$I_j \leftarrow I_j - \{(a, b)\}$; $I_j \leftarrow I_j \cap \{(a.inf, b.sup)\}$ *-- fusion des intervalles a et b*

Fin Pour

Jusqu'à ce que $((E = \emptyset)$ ou $(nb_{interval} > max_{interval}))$

Fin Pour

Fin

$(2\ \oplus)$	$(6\ \ominus)$	$(12,5\ \ominus)$	$(16,9\ \oplus)$	$(18\ \ominus)$
$k_{min}=4$				

$(1,8\ \oplus)$	$(4,5\ \ominus)$	$(12,5\ \oplus)$	$(16,9\ \oplus)$	$(18\ \ominus)$
$k_{min}=6$				

TAB. V.3 – Discrétisation obtenue avec *1R* pour $k_{min}=4$ et $k_{min}=6$.

4.7.2 *1R*

[Holte 1993] présente une méthode de discrétisation supervisée, globale et monothétique. Cette méthode consiste à découper le domaine de variation de chaque attribut en intervalles les plus « purs » possibles, c'est-à-dire ne contenant idéalement qu'une seule classe.

La seule contrainte imposée à l'utilisateur par cette méthode réside dans le choix du nombre minimal d'individus (k_{min}) dans chaque intervalle créé (à l'exception du dernier intervalle pouvant contenir moins d'individus). Holte préconise de choisir $k_{min}=6$, cette valeur ayant été déterminée empiriquement.

Cette méthode de discrétisation est très sensible au choix de k_{min} comme le montre les deux discrétisations présentées dans le tableau V.3 réalisées sur les mêmes données mais avec des valeurs différentes pour k_{min}. De plus, les résultats obtenus pour la discrétisation diffèrent souvent selon que celle-ci est réalisée en fonction des valeurs croissantes ou décroissantes de l'attribut.

Définissons les fonctions suivantes utilisées dans l'algorithme 11.

$$val(\mathcal{D},i)=\begin{cases}\mathcal{D}(i).\alpha_i^j & si\ i\leq n\\ code\ d'erreur & sinon\end{cases}$$

$$etiq(\mathcal{D},i)=\begin{cases}\mathcal{D}(i).etiq & si\ i\leq n\\ code\ d'erreur & sinon\end{cases}$$

$$etiquette\ majoritaire=\begin{cases}\oplus & si\ n_\ominus<n_\oplus\\ \ominus & sinon\end{cases}$$

Lorsqu'à l'issue de la discrétisation, deux intervalles contigus, ou plus, ont la même étiquette, ils sont fusionnés. L'extension de l'algorithme 11 aux données multiclasses est relativement simple.

4.7.3 *ID3*

[Quinlan 1990] propose une méthode de discrétisation supervisée et locale. L'objectif d'*ID3* est de créer un arbre de classification (ou d'induction). Lors de la création de l'arbre, si besoin (condition d'arrêt locale non atteinte), un attribut est discrétisé. L'attribut et le point de coupure sont déterminés en utilisant les propriétés de l'entropie.

Commencons donc par rappeler brièvement quelques notations liées à l'entropie.

Étant donné un attribut A, l'ensemble des individus S et un point de coupure possible t, l'entropie des partitions engendrées par t est égale à :

$$E(A,t;S)=\frac{|S_1|}{|S|}Ent(S_1)+\frac{|S_2|}{|S|}Ent(S_2)$$

où $|S_i|$ = nombre d'individus dans la partition S_i.

$Ent(S_i)=\sum_{c\in C}-P(c|S_i)log_2(P(c|S_i))$ avec C : l'ensemble des classes possibles et $P(c|S_i)$: la probabilité d'observer la classe c dans la population S_i.

Algorithme 11 $1R(\mathcal{B}_n^p, k_{min})$

Entrée :

$\mathcal{B}_n^p$: base de données dont les p attributs sont continus

k_{min} : nombre minimal d'individus par intervalle

Sortie :

$decoup(j)$: ensemble des intervalles discrèts pour chaque attribut

Début

Pour $(j = 1 ; j \leq p ; j++)$ **Faire**

$\mathcal{D} \leftarrow \{(\alpha_i^j, etiq(i)), 1 \leq i \leq n\}$ *-- sélection des valeurs de l'attribut j pour tous les individus*

trier les valeurs de $\mathcal{D}$ selon α_i^j

-- initialisation

$decoup(j) \leftarrow \emptyset$; $n_\oplus \leftarrow 0$; $n_\ominus \leftarrow 0$; $k \leftarrow 0$

Pour $(i \leftarrow 1 ; i \leq n ; i++)$ **Faire**

Si $((k < k_{min})$ ou $(etiq(\mathcal{D}, i) = etiq(\mathcal{D}, i+1))$ ou $(n_\oplus = n_\ominus))$ **Alors**

$n_{etiq(\mathcal{D},i)} \leftarrow n_{etiq(\mathcal{D},i)} + 1$

$k \leftarrow k + 1$

Sinon

$decoup(j) \leftarrow decoup(j) \cup \{(val(\mathcal{D}, i), etiquette\ majoritaire)\}$

$n_\ominus \leftarrow 0$; $n_\oplus \leftarrow 0$; $k \leftarrow 0$

Fin Si

Fin Pour

$decoup(j) \leftarrow decoup(j) \cup \{(val(\mathcal{D}, n), etiquette\ majoritaire)\}$

Fin Pour

Fin

Étant donnée la population S, le gain réalisé lors de la discrétisation de l'attribut A par le point de coupure t est égal à :

$$gain(A, t; S) = Ent(S) - E(A, t; S)$$

Le critère d'arrêt d'*ID3* est : obtenir des feuilles « pures », c'est-à-dire ne contenant que des individus de la même classe.

Ayant défini ceci, nous pouvons donner les grandes lignes de l'algorithme *ID3*. Pour chaque nœud de l'arbre en cours de construction, trouver le couple (A, t) maximisant le gain $gain(A, t; S)$. La discrétisation des attributs est donc binaire et réalisée au fur et à mesure de la construction de l'arbre d'induction.

L'algorithme *ID3* n'est donc pas vraiment une méthode de discrétisation puisque seuls les attributs liés aux classes seront discrétisés. Ainsi, si des attributs sont indépendants des classes, ils ne seront pas discrétisés.

4.7.4 *D2*

[Catlett 1991] propose une méthode de discrétisation supervisée et globale. La méthode proposée est monothétique et fondée sur la mesure d'entropie.

Le principe de l'algorithme *D2* est le suivant : pour chaque attribut, réaliser récursivement la discrétisation en choisissant le point de coupure t_{min} retenu qui maximise le gain $gain(A, t; S)$.

Ainsi, à chaque pas de la méthode, un découpage binaire de l'attribut A est réalisé.

Les critères d'arrêt de l'algorithme *D2* sont les suivants :

- arrêt si le nombre d'individus dans un intervalle est inférieur à 14.
- arrêt si le nombre d'intervalles est supérieur à 8.
- arrêt si quelque soit le point de coupure choisi, les gains informationnels sont tous identiques.
- arrêt si tous les individus de l'intervalle à découper appartiennent à la même classe.

L'inconvénient de cette approche est liée aux critères d'arrêt qui sont définis de manière *ad hoc* et donc indépendante des données.

4.7.5 Le principe du MDL

[Fayyad et Irani 1993] proposent une méthode similaire à celle proposée par *D2*.

Le principe du MDL (Minimum Description Length) est utilisé pour arrêter la discrétisation.

Le découpage se termine lorsque la condition d'arrêt suivante est atteinte :

$$gain(A, t; S) < \frac{log_2(|S| - 1)}{|S|} + \frac{\Delta(A, t; S)}{|S|}$$

où $\Delta(A, t; S) = log_2(3^k - 2) - [kEnt(S) - k_1 Ent(S_1) - k_2 Ent(S_2)]$ avec k_i : nombre de classes différentes représentées dans la population S_i.

4.7.6 *StatDisc*

[Richeldi et Rossotto 1995] proposent une méthode de discrétisation supervisée et globale semblable à la méthode *ChiMerge*. Cette méthode est elle aussi ascendante et utilise la mesure Φ (au lieu du χ^2) pour fusionner les intervalles. Lors de la fusion, plusieurs intervalles peuvent être fusionnés (contrairement au *ChiMerge* ou seuls deux intervalles sont fusionnés).

La seule contrainte imposée est (comme pour *ChiMerge*) de fixer un seuil pour la mesure Φ permettant de contrôler la fusion des intervalles.

4.7.7 *HyperCluster Finder*

[Muhlenbach et Rakotomalala 2002] proposent une nouvelle méthode de discrétisation supervisée et polythétique.

Cette nouvelle méthode est fondée sur l'utilisation de graphes de voisinage [Toussaint 1980] et d'amas. Un graphe de voisinage permet de regrouper des objets proches étant donné un attribut.

Le graphe des voisins relatifs est défini de la manière suivante [Muhlenbach et Rakotomalala 2002].

Définition 9 *Le Graphe des Voisins Relatifs (GVR) est un graphe connexe dans lequel, si deux points a et b sont reliés par une arête, alors ils vérifient la propriété ci-dessous où $d(a,b)$ est la distance euclidienne entre deux points a et b dans $\mathbb{R}^p$*

$$d(a,b) \leq Max(d(a,c), d(b,c)) \forall c, c \neq a, b$$

À partir du graphe des voisins relatifs, [Muhlenbach et Rakotomalala 2002] proposent d'utiliser la notion d'amas pour discrétiser les p attributs continus.

Définition 10 *Un amas est un sous-graphe connexe du graphe de voisinage dont tous les sommets appartiennent à la même classe.*

La figure V.5 présente un graphe des voisins relatifs (V.5.a) et les amas associés (V.5.b). Les données sont visualisées selon deux attributs X_1 et X_2.

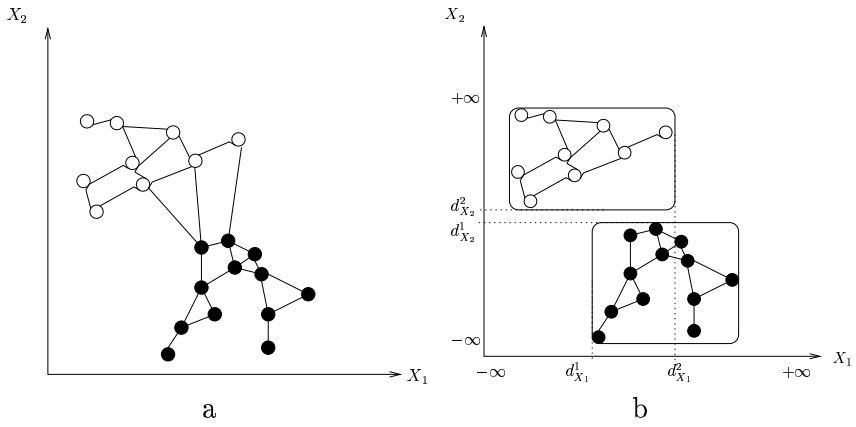


FIG. V.5 – Graphe des voisins relatifs (a) et amas associés (b).

L'algorithme de discrétisation 12 implante la méthode proposée par [Muhlenbach et Rakotomalala 2002].

L'un des avantages de cette méthode est lié à la discrétisation polythétique. En effet, une relation entre deux attributs a et b telle que $XOR(a,b)$ peut être détectée avec cette nouvelle méthode alors que les méthodes classiques échouent sur ce type de problème.

4.8 Approches mixtes : supervisées et non supervisées

Quelques approches combinent des méthodes supervisées et non supervisées pour réaliser la discrétisation.

Algorithme 12 *HyperCluster Finder*($\mathcal{B}_n^p$)

Entrée :

$\mathcal{B}_n^p$: base de données dont les p attributs sont continus

Sortie :

$\mathcal{B}^d$: base de données dans laquelle les attributs ont été discrétisés

Début

génération d'un graphe de voisinage : GVR

coupure des arêtes reliant des points de classes différentes $\Rightarrow \mathcal{A} = \{$ensemble des amas$\}$

sélection des amas les plus pertinents : suppression des amas contenant peu d'individus ($taille(amas) < seuil$ par exemple)

Pour tout $(X_j, 1 \leq j \leq p)$ **Faire**

$val(X_j) \leftarrow \emptyset$

Fin Pour

−− Recherche, pour chaque amas, des valeurs extrêmes pour chaque attribut

Pour tout $(A \in \mathcal{A})$ **Faire**

Pour tout $(X_j, 1 \leq j \leq p)$ **Faire**

déterminer les valeurs minimales et maximales de $X_j \Rightarrow \{min(X_j), max(X_j)\}$

$val(X_j) \leftarrow val(X_j) \cap \{min(X_j), max(X_j)\}$

Fin Pour

Fin Pour

Pour $(j \leftarrow 1\,;\, j \leq p\,;\, j++)$ **Faire**

remplacer la valeur minimale (resp. maximale) de X_j par $-\infty$ (resp. $+\infty$).

trier les valeurs contenues dans $val(X_j)$.

les valeurs triées de $val(X_j)$ définissent les intervalles recherchés pour X_j

Fin Pour

recoder les données en utilisant les intervalles définis par les valeurs $val(X_j)$

Fin

4.8.1 Apprentissage d'un ensemble de règles de classification

[Chan et al. 1991] proposent une méthode de discrétisation décomposée en deux étapes :
- une première étape non supervisée dans laquelle les individus sont partitionnés en deux groupes de tailles égales.
- une seconde étape supervisée consise, à partir des données discrétisées grâce à l'étape précédente, à apprendre un ensemble de règles de classification (en utilisant un algorithme tel qu'*ID3*). La précision des règles obtenues est ensuite évaluée et la partition ayant la précision minimale est alors repartitionnée.

Tant que les conditions d'arrêt liées à la qualité des partitions obtenues ne sont pas atteintes, le procédé est répété.

Les inconvénients de cette approche sont multiples, d'une part le temps de calcul nécessaire pour réaliser la discrétisation est relativement élevé car il faut apprendre un classifieur pour chaque nouvelle partition créée. D'autre part, la méthode présentée suppose qu'une précision minimale est accessible sur les données étudiées, or si les données sont issues d'un processus aléatoire, de nombreuses partitions seront inutilement créées.

4.8.2 Critère de Contraste Monothétique

[Merckt 1993] propose deux méthodes fondées sur le principe général de « Critère de Contraste Monothétique ». La première méhode, considérée par l'auteur comme non supervisée, utilise un algorithme non supervisé de classification qui détecte les partitions créant « le plus grand contraste » étant donné une fonction de contraste. Cette méthode est objective, c'est-à-dire qu'elle n'utilise pas la classe des individus pour réaliser la discrétisation. La fonction de contraste proposée est définie de la manière suivante, où C_1 et C_2 sont les partitions obtenues après découpage et μ_i^j est la moyenne de l'attribut j sur la partition C_i.

$$Contrast(C_1, C_2, A_j) = \frac{|C_1||C_2|}{|C_1 \cup C_2|}(\mu_1^j - \mu_2^j)^2$$

Exemple : Considérons que les valeurs $(2, 2, 5, 5, 5, 8, 10, 15, 15, 20)$ représentent les valeurs prises par un attribut à discrétiser. Les valeurs $2, 5, 8, 10, 15$ et 20 représentent toutes des points de coupure possibles.

Pour chacun de ces points de coupure, nous obtenons deux partitions (voir Figure V.6). Les partitions sont telles que pour un point de coupure p_c, les valeurs appartenant aux partitions C_1 et C_2 sont telles que :

$$\forall x \in C_1,\ x \leq p_c$$

$$\forall x \in C_2,\ x > p_c$$

Le Tableau V.4 présente les valeurs de contraste associées aux différentes partitions obtenues. La meilleure partition (maximisant le contraste) est celle obtenue pour le point de coupure de valeur 10 (voir le Tableau V.4).

Cette méthode favorise l'apparition de partitions contrastées, même lorsqu'il existe des partitions pures. Lorsque deux points de coupure candidats fournissent la même valeur de contraste, il est impossible de les distinguer.

La seconde méthode, qui est fondée sur la précédente, redéfinit la fonction de contraste utilisée en divisant la valeur $Contrast(C_1, C_2, A_j)$ par l'entropie de la partition étudiée. Puisque le calcul de l'entropie impose d'utiliser la classe des individus, la méthode devient alors supervisée.

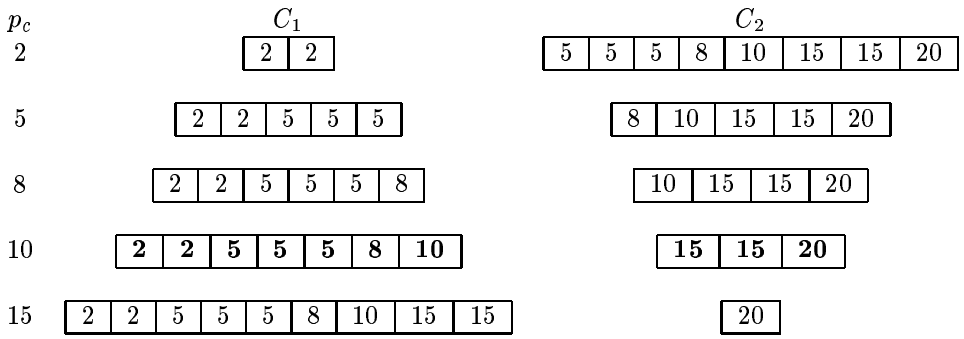


FIG. V.6 – Illustration du principe du Critère de Contraste Monothétique sur les données (2, 2, 5, 5, 5, 8, 10, 15, 15, 20).

p_c	μ_1	μ_2	$Constraste$
2	2	10,375	112,225
5	3,8	13,6	240,1
8	4,5	15	264,6
10	**5,28**	**16,67**	**272,44**
15	7,44	20	141,98

TAB. V.4 – Contraste obtenu pour chaque partition.

La nouvelle fonction de contraste est définie par :

$$CE(C_1, C_2, A_j) = \frac{Contrast(C_1, C_2, A_j)}{Ent(C_1, C_2)}$$

Lorsque les données sont bruitées, la méthode utilisant le critère de contraste CE est moins sensible que les méthodes utilisant uniquement l'entropie.

4.9 Approches non supervisées

Les méthodes de discrétisation non supervisées utilisent seulement, pour un attribut donné, les valeurs des individus.

4.9.1 Intervalles de largeur égale

Le principe de cette méthode est le suivant :
- trier les valeurs à discrétiser
- déterminer le minimum et le maximum : min et max
- découper les valeurs en k intervalles de taille $\left(\frac{max-min}{k}\right)$

La figure V.7.a montre les discrétisations obtenues avec différentes valeurs de k.

Le principal inconvénient de cette méthode est lié aux valeurs extrèmes et isolées prises par certains individus. La figure V.7.b permet de visualiser l'impact des valeurs isolées sur cette méthode.

4.9.2 Intervalles de fréquence égale

Cette méthode permet d'éviter le problème posé par les valeurs isolées.

Le principe de cette méthode est le suivant :

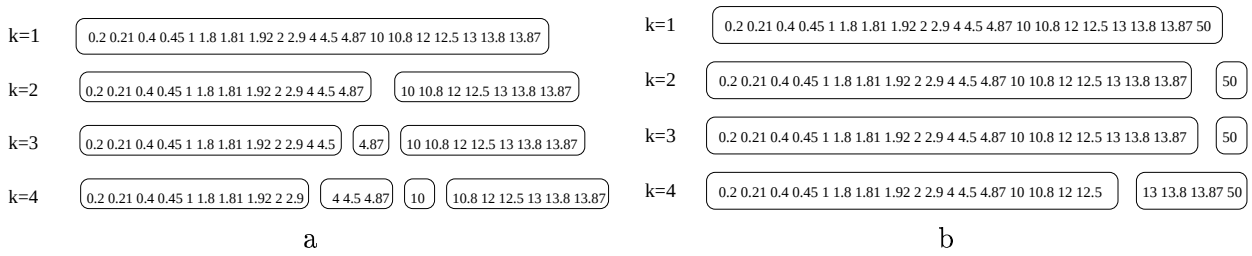


FIG. V.7 – intervalles de largeur égale : influence de k et impact des valeurs isolées.

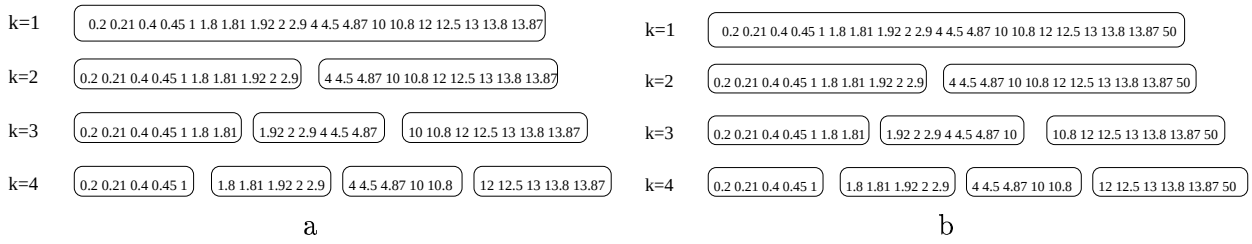


FIG. V.8 – intervalles de fréquence égale : influence de k et impact des valeurs isolées.

- trier les valeurs à discrétiser
- découper les valeurs en k intervalles contenant tous $\frac{n}{k}$ individus

La figure V.8 présente l'impact de k (V.8.a) et celui des valeurs isolées (V.8.b).
Nous pouvons voir qu'en utilisant cette méthode, les intervalles obtenus sont plus « cohérents ».

4.9.3 K-tile, distributions uniforme et gaussienne

[Chickering et al. 2001] proposent de discrétiser les données en supposant qu'elles sont issues d'une distribution soit gaussienne, soit uniforme. Ils proposent aussi une méthode nommées *K-tile* qui est très proche de la technique de discrétisation en intervalles de fréquence égale.

K-tile
Si les données doivent être discrétisées en k intervalles, il faut trouver $k-1$ points de coupure $c_i, 1 \leq i \leq k-1$ tel que $\forall i$, $n \times \frac{i}{k}$ individus aient une valeur inférieure ou égale à c_i et $n \times \frac{k-i}{k}$ individus aient une valeur supérieure à c_i.

La différence notable entre cette méthode et celle des intervalles de fréquence égale est que les c_i sont choisis parmi les valeurs des individus. Les conditions strictes peuvent ne pas être toujours respectées, par exemple s'il n'existe pas d'individu ayant la valeur pivot attendue.

Distribution uniforme
Cette méthode est identique à celle des intervalles de largeur égale. En effet, dans ce cas, les données sont supposées issues d'une distribution uniforme. Ainsi, étant donné un attribut X_j, les points de coupure c_i obtenus sont tels que

$$c_i = min(X_j) + i \times \left(\frac{max(X_j) - min(X_j)}{k} \right)$$

Distribution gaussienne

Dans cette méthode, les valeurs de chaque attribut X_j sont considérées comme issues d'une distribution gaussienne de moyenne $\mu(X_j)$ et d'écart type $\sigma(X_j)$. Les points de coupure c_i sont donc tels que

$$c_i = \mu(X_j) + \sigma(X_j)\Phi^{-1}\left(\frac{i}{k}\right)$$

où Φ^{-1} est la fonction réciproque de la loi normale centrée et réduite.

4.9.4 Algorithme des $K-moyennes$

L'algorithme des $K-moyennes$ [MacQueen 1967] est un algorithme non supervisé permettant de créer, à partir de valeurs numériques, K groupes de valeurs les plus homogènes possibles. Cet algorithme est très répandu dans le domaine de l'apprentissage et son succès est essentiellement dû à sa simplicité et à son efficacité. Les utilisations de cet algorithme sont très variées, par exemple en segmentation d'images [Ray et Turi 1999] ou en catégorisation de documents [Steinbach et al. 2000, Sinka et Corne 2002].

Le principe des $K-moyennes$ est, étant donné un ensemble de valeurs numériques, de créer K groupes de valeurs les plus compacts possibles. Un groupe, aussi appelé *cluster*, est décrit par son centre (noté noyau ou centroïde). Le nombre de groupes que doit créer l'algorithme est un paramètre extérieur au système et qui est contrôlé par l'utilisateur.

Notons $\mathcal{V}$ l'ensemble des valeurs numériques que nous étudions. Étant donnés K et $\mathcal{V}$, l'algorithme peut se résumer de la manière suivante :

1. la première étape de l'algorithme consiste à créer K clusters en choisissant une valeur aléatoire pour le noyau de chaque cluster.
2. Ensuite, toutes les valeurs de $\mathcal{V}$ sont associées au cluster le plus proche. Soit $v \in \mathcal{V}$ une des valeurs étudiées, la distance euclidienne entre cette valeur et chacun des noyaux :

   $$d(v, noyau(i)) = ||v - noyau(i)||^2$$

   est calculée et v est associée au cluster minimisant cette distance.
3. La dernière phase de l'algorithme consiste à recalculer le centre de chacun des clusters. Le noyau de chaque cluster est égal à la moyenne des valeurs contenues dans le cluster :

   $$noyau(i) = \frac{1}{card(cluster(i))} \times \sum_{v \in cluster(i)} v$$

Les étapes 2 et 3 sont répétées tant que les clusters ne sont pas stables.

La première étape est une des étapes principale. En effet, les groupes obtenus à la fin de l'algorithme dépendent beaucoup des centres initiaux des clusters.

De nombreuses méthodes ont été proposées pour initialiser les centres des clusters. Nous ne citerons que trois méthodes étudiées dans la thèse de [Courtine 2002].

- Les centres sont initialisés avec une valeur choisie aléatoirement (et potentiellement absente de $\mathcal{V}$).
- Les centres sont initialisés avec une valeur choisie aléatoirement parmi les valeurs de $\mathcal{V}$.
- chaque valeur de $\mathcal{V}$ est assignée aléatoirement à un cluster.

Ces trois méthodes d'initialisation ont été testées par M. Courtine et la troisième méthode est celle qui lui a permis d'obtenir les meilleurs résultats en termes de temps de convergence vers une solution stable.

L'algorithme des $K-moyennes$ que nous présentons page 129 implante la deuxième méthode d'initialisation. Il est relativement aisé de modifier cet algorithme pour qu'il utilise une autre méthode d'initialisation des noyaux.

Lorsque les K clusters ont été trouvés par l'algorithme, il est nécessaire d'évaluer la validité du découpage des valeurs numériques induit par ces clusters. De nombreuses mesures ont été définies pour déterminer la validité de la partition obtenue [Dunn 1973, Davies et Bouldin 1979, Merckt 1993, Milligan 1996, Bezdek et Pal 1998]. L'objectif commun de toutes ces mesures est de trouver, étant donné un ensemble de partitions, les partitions qui contiennent les clusters les plus compacts et les mieux séparés.

Nous avons vu dans la section 4.8.2 que la mesure de Contraste proposée par [Merckt 1993] favorise les partitions les plus contrastées. De manière générale, une bonne discrétisation est caractérisée par une partition dans laquelle le rapport entre la distance *intra-cluster* et la distance *inter-cluster* est minimale [Ray et Turi 1999].

La distance intra-cluster est définie comme la moyenne des distances entre les valeurs et le centre du cluster auquels elles appartiennent.

$$d_{intra} = \frac{1}{card(\mathcal{V})} \sum_{i=1}^{K} \left(\sum_{v \in \mathcal{V}} d(v, noyau(i)) \right)$$

La distance inter-cluster est définie comme la distance minimale entre les centres des clusters.

$$d_{inter} = \min_{1 \leq i < j \leq K} d(noyau(i), noyau(j))$$

La mesure de validité définie par [Ray et Turi 1999] est définie comme le rapport de ces deux valeurs :

$$Validite = \frac{d_{intra}}{d_{inter}}$$

[Courtine 2002] utilise cette mesure pour déterminer la meilleure des solutions trouvées par l'algorithme $K-moyennes$ lors de plusieurs applications de celui-ci avec des initialisations différentes. Elle présente aussi une méthode, fondée sur cette mesure de validité, et permettant de déterminer de manière automatique la valeur optimale de K.

Cette méthode consiste à rechercher la première valeur de K telle que les deux inégalités suivantes soient vérifiées :

$$Validite(K-1) < Validite(K)$$
$$Validite(K+1) < Validite(K)$$

Cette valeur de K correspond au premier minimum local de la mesure de validité. L'utilisation de cette valeur permet d'assurer que la discrétisation proposée par l'algorithme $K-moyennes$ sera optimale pour la mesure de validité sous la contrainte de rechercher le moins de partitions possibles.

Bien que cette méthode globale de discrétisation incluant le calcul automatique de K semble très prometteuse, nous ne l'avons pas retenue car nous n'avons malheureusement pas eu le temps de tester cette méthode de discrétisation pour la fouille de textes. Cependant, nous envisageons de la tester et de la confronter à la méthode retenue pour l'instant et présentée dans la section suivante.

5 Application à la fouille de textes

Parmi les différentes méthodes de discrétisation présentées dans la section précédente, seules les approches non supervisées sont directement exploitables pour notre problématique. Les travaux réalisés par [Chickering et al. 2001] ont montré que, parmi les techniques de discrétisation non

Algorithme 13 *K-moyennes*$(K, \mathcal{V})$

Entrée :

K : nombre de clusters à créer

$\mathcal{V}$: valeurs numériques à découper en clusters

Sortie :

$cluster(1..K)$: les K clusters trouvés par l'algorithme

Début

-- Initialisation des clusters

Pour $(i \leftarrow 1\,;\, i \leq K\,;\, i++)$ **Faire**

$noyau(i) \leftarrow v \in \mathcal{V}$ *-- où v est choisie aléatoirement parmi les valeurs de $\mathcal{V}$*

$cluster(i) \leftarrow \emptyset$

Fin Pour

Répéter

-- Mémoriser les différents clusters pour pouvoir vérifier s'ils sont stables ou non

Pour $(i \leftarrow 1\,;\, i \leq K\,;\, i++)$ **Faire**

$cluster^{-1}(i) \leftarrow cluster(i)$

$cluster(i) \leftarrow \emptyset$

Fin Pour

-- affecter les valeurs de $\mathcal{V}$ aux différents clusters

Pour tout $(v \in \mathcal{V})$ **Faire**

Pour $(i \leftarrow 1\,;\, i \leq K\,;\, i++)$ **Faire**

calculer $d(v, noyau(i))$

Fin Pour

trouver $i \in [1..K]$ tel que $d(v, noyau(i))$ soit minimal

$cluster(i) cluster(i) \cup \leftarrow \{v\}$

Fin Pour

-- calculer les nouveaux centres des clusters

Pour $(i \leftarrow 1\,;\, i \leq K\,;\, i++)$ **Faire**

$noyau(i) \leftarrow \frac{1}{card(cluster(i))} \times \sum_{v \in cluster(i)} v$

Fin Pour

-- vérifier si les clusters sont stabilisés ou non

$stable \leftarrow$Vrai , $i \leftarrow 1$

Tant que $((stable =$Vrai $)$ et $(i \leq K))$ **Faire**

Si $(cluster^{-1}(i) \neq cluster(i))$ **Alors**

$stable \leftarrow$Faux

Sinon

$i \leftarrow i + 1$

Fin Si

Fin Tant que

Jusqu'à ce que $(stable \neq$Faux $)$

retourner $cluster(1..K)$

Fin

supervisées, la méthode des *k-tile* et celle de l'approximation gaussienne fournissent les meilleurs résultats.

La méthode que nous avons retenue pour nos travaux est proche de celle des $k - tile$. Cette méthode est détaillée dans la section suivante.

5.1 Méthode de discrétisation retenue pour la fouille de textes

Reprenons l'exemple précédent lié au corpus « fouille de données ».

Pour un concept donné, la discrétisation la plus simple consiste à diviser les textes en deux sous-ensembles : les textes ne contenant aucune occurence du concept et ceux en contenant au moins une. Le concept étudié est alors divisé en deux sous-concepts : *concept*_0 et *concept*_1 représentant respectivement l'absence du concept et la présence du concept. Ces informations, de nature booléenne, nous permettent obtenir des règles d'association.

La discrétisation obtenue fournit une représentation simple des concepts. Il est possible d'obtenir une représentation plus riche en augmentant le nombre de sous-concepts recherchés. Ainsi, une discrétisation en trois sous-concepts permet d'introduire, pour un concept donné, les notions suivantes : absence du concept, faible présence du concept et forte présence du concept dans les textes concernés.

Pour réaliser cette discrétisation, nous utilisons, après avoir isolé les textes ne contenant pas le concept à discrétiser, la méthode des $k - tile$ sur les autres textes (ceux contenant au moins une occurence du concept).

La nouvelle représentation des concepts obtenue à partir de cette discrétisation est plus précise que la précédente et reste compréhensible. L'augmentation du nombre de sous-concepts recherchés permet encore de raffiner cette représentation, cependant les résultats obtenus sont plus difficiles à interpréter pour le néophyte.

Pour la majorité des corpus étudiés, une discrétisation en au plus trois sous-concepts est suffisante.

L'algorithme 14 représente la méthode retenue pour discrétiser les concepts.

Algorithme 14 *DiscretiseConcepts*(C, T, k)

Entrée :

C : concepts à discrétiser

T : textes du corpus

k : nombre de valeurs discrètes recherchées

Début

Pour tout $(c \in C)$ **Faire**

$\mathcal{T}_c \leftarrow \{t \in T | c \in t\}$ *-- ensemble des textes contenant le concept c*

$\mathcal{T}_{\overline{c}} \leftarrow \{t \in T | \overline{c} \in t\}$ *-- ensemble des textes ne contenant pas le concept c*

découper $\mathcal{T}_c$ en $k - 1$ intervalles

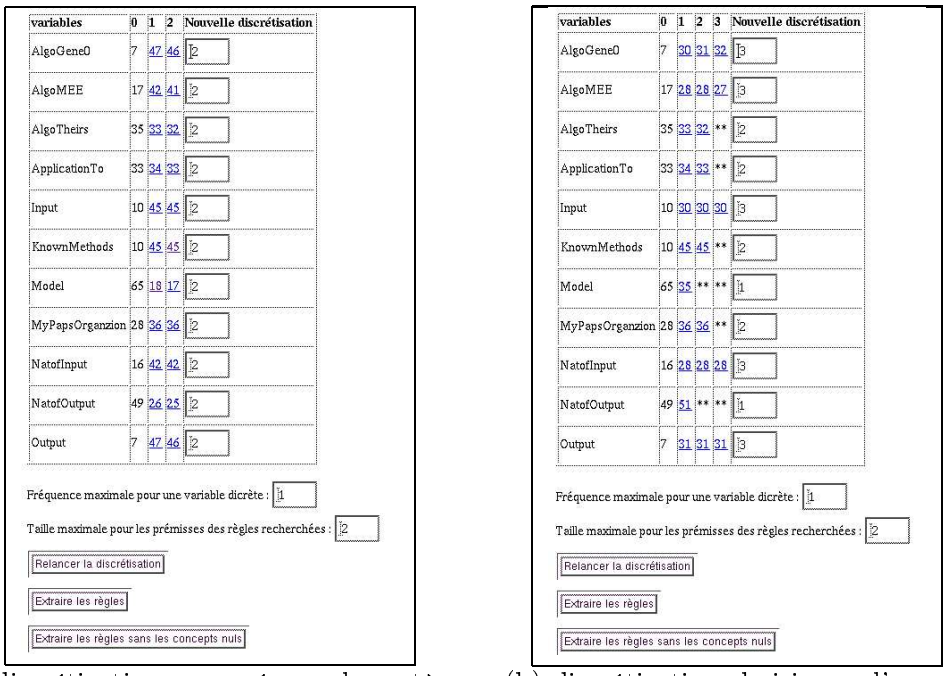
Fin Pour

Fin

5.2 Interface de discrétisation

Nous avons conçu une interface de discrétisation permettant à l'expert de contrôler simplement la discrétisation de chaque attribut indépendamment des autres. Cette interface a été développée et intégrée dans un outil d'extraction de connaissances à partir de textes. Cet outil est écrit en PHP, Perl, C et HTML et est donc utilisable depuis différentes plateformes (Linux, Windows, Sun, *etc.*) car l'outil fonctionne sur un serveur web.

La Figure V.9 représente une capture d'écran de l'interface de discrétisation en cours de fonctionnement sur le corpus des introductions d'articles. Nous pouvons voir sur la Figure V.9.a la discrétisation naïve proposée par le système et sur la Figure V.9.b la discrétisation réalisée par l'expert. Pour modifier le nombre de valeurs discrètes associées à un concept, il suffit à l'expert d'indiquer dans la zone de texte prévue à cet effet le nombre de valeurs discrètes qu'il souhaite obtenir.



variables	0	1	2	Nouvelle discrétisation
AlgoGene0	7	47	46	2
AlgoMEE	17	42	41	2
AlgoTheirs	35	33	32	2
ApplicationTo	33	34	33	2
Input	10	45	45	2
KnownMethods	10	45	45	2
Model	65	18	17	2
MyPapsOrganzion	28	36	36	2
NatofInput	16	42	42	2
NatofOutput	49	26	25	2
Output	7	47	46	2

Fréquence maximale pour une variable dicrète : 1

Taille maximale pour les prémisses des règles recherchées : 2

Relancer la discrétisation

Extraire les règles

Extraire les règles sans les concepts nuls

(a) discrétisation proposée par le système

variables	0	1	2	3	Nouvelle discrétisation
AlgoGene0	7	30	31	32	3
AlgoMEE	17	28	28	27	3
AlgoTheirs	35	33	32	**	2
ApplicationTo	33	34	33	**	2
Input	10	30	30	30	3
KnownMethods	10	45	45	**	2
Model	65	35	**	**	1
MyPapsOrganzion	28	36	36	**	2
NatofInput	16	28	28	28	3
NatofOutput	49	51	**	**	1
Output	7	31	31	31	3

Fréquence maximale pour une variable dicrète : 1

Taille maximale pour les prémisses des règles recherchées : 2

Relancer la discrétisation

Extraire les règles

Extraire les règles sans les concepts nuls

(b) discrétisation choisie par l'expert

FIG. V.9 – Interface de discrétisation : choix du nombre de valeurs discrètes par attribut.

La discrétisation permet, comme nous l'avons vu, d'associer à un intervalle de valeurs (ici les fréquences d'occurence d'un concept dans les textes) une valeur discrète.

L'expert peut choisir de visualiser la répartition des textes dans l'intervalle associé à l'une des valeurs discrètes d'un concept. Il lui suffit de cliquer sur le lien hypertexte associé à la valeur discrète qu'il souhaite visualiser.

L'intérêt de cette option est de permettre à l'expert de s'assurer que l'intervalle créé automatiquement par notre outil de discrétisation est relativement compact (c'est-à-dire qu'il n'existe pas de grande plage de valeurs « vides »). L'existence de plages de valeurs « vides » dans un intervalle associé à une valeur discrète indiquerait que la discrétisation automatique a entrainé la fusion d'intervalles qui pourraient être séparés (sur demande de l'expert).

Par exemple, considérons le concept *AlgoGene0* qui a été initialement discrétisé en trois valeurs discrètes (voir Figure V.9.a) et que l'expert a décider de rediscrétiser en quatre valeurs discrètes

(voir Figure V.9.b).

La valeur discrète 1 de ce concept dans la discrétisation automatique correspond à un intervalle de valeurs telles que pour 47 des 100 textes du corpus, la fréquence d'occurence du concept *AlgoGene0* soit dans cet intervalle. La figure V.10 présente la répartition des fréquences d'occurence du concept *AlgoGene0* pour cette valeur discrète.

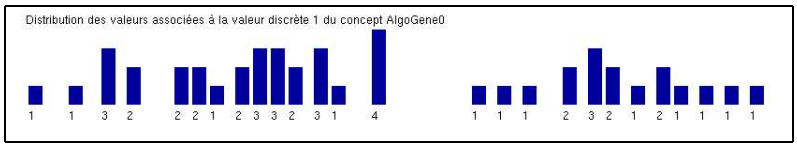


FIG. V.10 – Visualisation de la distribution des textes pour la valeur discrète 1 du concept *AlgoGene0* (discrétisation initiale).

Nous pouvons voir qu'il semble y avoir deux groupes de fréquences relativement homogènes qui pourraient être dissociés et associés à des valeurs discrètes différentes.

Pour cet exemple, l'expert a simplement augmenté le nombre de valeurs discrètes associées au concept posant problème. Ce choix était aussi motivé par le faible nombre de textes ne contenant aucune occurence du concept et l'expert a choisi d'obtenir des discrétisations relativement équilibrées (dans la mesure du possible).

Après rediscrétisation du concept *AlgoGene0*, l'intervalle de valeurs anciennement associé à la valeur discrète 1 a été découpé en deux intervalles qui sont présentés sur les figures V.11 et V.12. Nous voyons alors que le problème a été résolu et que les intervalles obtenus sont beaucoup plus compacts et donc probablement plus significatifs.

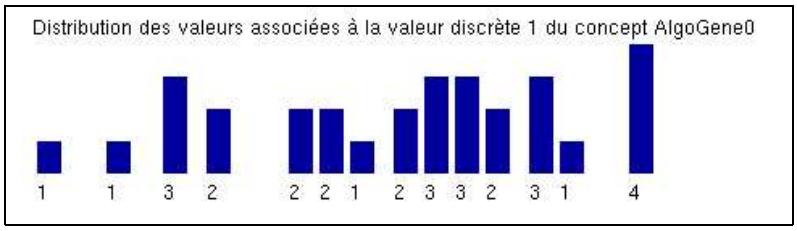


FIG. V.11 – Visualisation de la distribution des textes pour la valeur discrète 1 du concept *AlgoGene0* (nouvelle discrétisation).

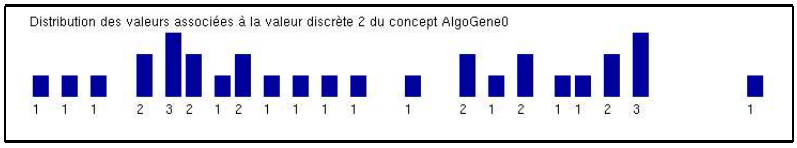


FIG. V.12 – Visualisation de la distribution des textes pour la valeur discrète 2 du concept *AlgoGene0* (nouvelle discrétisation).

5.3 Analyse et reproduction du comportement de l'expert

L'interface de discrétisation (voir Figure V.9) a été utilisée par l'expert pour discrétiser différents corpus. Lors de cette utilisation, le comportement de l'expert a été étudié et nous avons adapté

notre algorithme pour essayer de proposer directement un découpage des données satisfaisant pour l'expert.

Ainsi, lorsque l'expert choisit de découper les concepts en $k > 2$ valeurs discrètes, la discrétisation obtenue pour certains concepts est déséquilibrée car le nombre de textes ne contenant pas le concept est nettement supérieur au nombre de textes contenant le concept. Face à ce type de situation, le comportement de l'expert est de regrouper les textes contenant au moins une occurrence du concept jusqu'à atteindre, si besoin, une discrétisation binaire du concept.

Le Tableau V.5 présente un tel cas où la distribution des fréquences d'occurences du concept C_1 est très différente de celles des concepts C_2, C_3 et C_4. L'expert corrige donc manuellement cette incohérence conduisant à interpréter les textes contenant la valeur discrète 2 du concept C_1 comme des textes ayant une forte présence du concept C_1 au même titre que ceux contenant la valeur 2 du concept C_2. Le Tableau V.6 présente la nouvelle discrétisation obtenue où les tailles des intervalles des valeurs discrètes sont relativement comparables.

Ce comportement est lié au fait que si la discrétisation d'un concept est trop déséquilibrée par rapport aux autres concepts, c'est-à-dire une majorité de valeurs nulles pour un concept et une répartition similaire des valeurs non nulles et nulles pour les autres, alors nous risquons d'une part d'introduire artificiellement des pépites de connaissance dans les données. Notre algorithme a la capacité d'extraire des règles d'association ayant un faible support et une moindre contradiction élevée. Donc, si la discrétisation de certains concepts est telle qu'un intervalle de valeurs est artificiellement associé à deux valeurs discrètes, le pourcentage de contre exemples associé aux règles contenant ces concepts discrets en prémisse va diminuer et donc la moindre contradiction va augmenter. Nous prenons alors le risque de détecter des règles qui refléteraient plus une erreur de discrétisation (donc du bruit) que de vraies pépites de connaissance.

D'autre part il n'est pas très pertinent de discrétiser un concept en plus de deux valeurs si plus de 50% de ses fréquences d'occurences sont nulles.

concept	**discrétisation**		
	0 (absence)	1 (faible présence)	2 (forte présence)
C_1	70%	15%	15%
C_2	20%	40%	40%
C_3	36%	32%	32%
C_4	28%	36%	36%

TAB. V.5 – Incohérence dans la discrétisation automatique.

concept	**discrétisation**	
	0 (absence)	1 (forte présence)
C_1	70%	30%

concept	**discrétisation**		
	0 (absence)	1 (faible présence)	2 (forte présence)
C_2	20%	40%	40%
C_3	36%	32%	32%
C_4	28%	36%	36%

TAB. V.6 – Discrétisation corrigée par l'expert.

Nous avons tenté de reproduire ce comportement en modifiant notre algorithme de discrétisation.

L'algorithme 15 est une version légèrement modifiée de l'algorithme 14 et prenant en considération les observations effectuées lors de l'utilisation de l'interface de discrétisation.

Algorithme 15 *DiscretiseConcepts2*(C, T, k)

Entrée :

C : concepts à discrétiser

T : textes du corpus

k : nombre de valeurs discrètes recherchées

Début

Pour tout $(c \in C)$ **Faire**

$\mathcal{T}_c \leftarrow \{t \in T | c \in t\}$ -- *ensemble des textes contenant le concept c*

$\mathcal{T}_{\overline{c}} \leftarrow \{t \in T | \overline{c} \in t\}$ -- *ensemble des textes ne contenant pas le concept c*

Si $(card(\mathcal{T}_c) > card(\mathcal{T}_{\overline{c}}))$ **Alors**

découper $\mathcal{T}_c$ en $\min\left(k-1, \frac{card(\mathcal{T}_c)}{card(\mathcal{T}_{\overline{c}})}\right)$ intervalles -- *modification permettant d'imiter l'expert*

Fin Si

Fin Pour

Fin

Ce nouvel algorithme permet d'obtenir une première discrétisation des données plus proche de celle réalisée par l'expert et nous pouvons donc lui proposer directement l'écran V.9.b au lieu de V.9.a.

6 Extraction des connaissances et validation par l'expert

Les connaissances que nous recherchons dans les textes sont des règles d'association entre concepts de la forme $concept_1 \rightarrow concept_2$. Lorsque nous présentons les règles à l'expert, nous associons à chaque règle son support et de sa confiance.

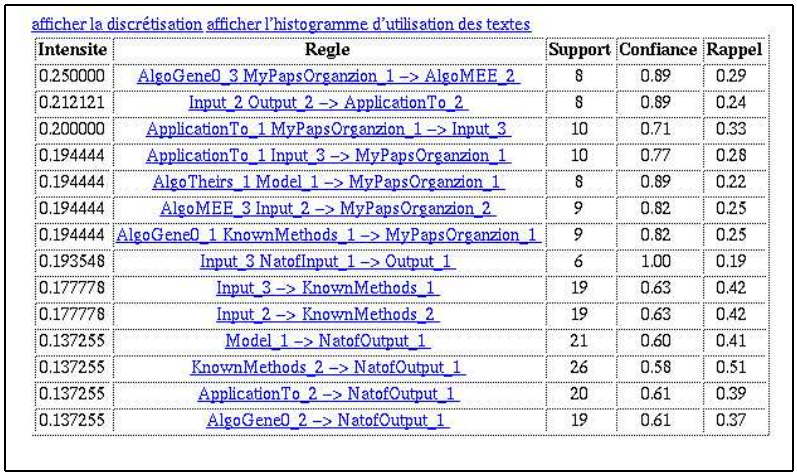
Considérons la règle $concept_{NatofOutput} \rightarrow concept_{KnownMethods}$ (avec $support = 0,16$ et $confiance = 0,64$) obtenue à partir du corpus de fouille de données. Son interprétation est la suivante : lorsque la nature des sorties est évoquée alors, dans 64% des cas, les méthodes connues sont évoquées dans le corpus. Cette règle est vérifiée par 16% du corpus.

L'algorithme d'extraction des pépites de connaissance présenté dans le chapitre III est utilisé pour obtenir, à partir du corpus discrétisé, les règles d'association entre les concepts que nous recherchons.

Parmi les cinq corpus présentés dans la section 2.1, nous n'avons pu appliquer l'extraction des pépites de connaissance que sur le corpus de fouille de données et celui des ressources humaines. Pour les deux autres corpus, la détection des traces de concepts et l'association de celles-ci aux concepts de l'expert n'est pas encore complètement réalisée. Lorsque nous aurons ces informations pour chacun des corpus, nous serons alors en mesure de poursuivre le processus de fouille de textes.

Dans la suite de cette section, nous présenterons les différentes étapes permettant à l'expert d'extraire les règles et de les valider. Nous illustrerons ces différentes étapes sur le corpus d'introductions d'articles.

Une interface semblable à celle présentée pour la discrétisation permet de visualiser les résultats obtenus. La Figure V.13 présente les règles d'association obtenues à partir du corpus d'introductions d'articles. Ces règles ont été obtenues en utilisant la moindre contradiction et en éliminant les concepts nuls (représentant l'absence du concept dans les textes) des données. Comme nous l'avons déjà dit, la présence de ces concepts dans les règles d'association est difficile à interpréter et l'expert peut choisir via l'interface (voir Figure V.9) de les prendre en considération ou non.



afficher la discrétisation afficher l'histogramme d'utilisation des textes

Intensite	Regle	Support	Confiance	Rappel
0.250000	AlgoGene0_3 MyPapsOrganzion_1 –> AlgoMEE_2	8	0.89	0.29
0.212121	Input_2 Output_2 –> ApplicationTo_2	8	0.89	0.24
0.200000	ApplicationTo_1 MyPapsOrganzion_1 –> Input_3	10	0.71	0.33
0.194444	ApplicationTo_1 Input_3 –> MyPapsOrganzion_1	10	0.77	0.28
0.194444	AlgoTheirs_1 Model_1 –> MyPapsOrganzion_1	8	0.89	0.22
0.194444	AlgoMEE_3 Input_2 –> MyPapsOrganzion_2	9	0.82	0.25
0.194444	AlgoGene0_1 KnownMethods_1 –> MyPapsOrganzion_1	9	0.82	0.25
0.193548	Input_3 NatofInput_1 –> Output_1	6	1.00	0.19
0.177778	Input_3 –> KnownMethods_1	19	0.63	0.42
0.177778	Input_2 –> KnownMethods_2	19	0.63	0.42
0.137255	Model_1 –> NatofOutput_1	21	0.60	0.41
0.137255	KnownMethods_2 –> NatofOutput_1	26	0.58	0.51
0.137255	ApplicationTo_2 –> NatofOutput_1	20	0.61	0.39
0.137255	AlgoGene0_2 –> NatofOutput_1	19	0.61	0.37

FIG. V.13 – Visualisation des règles extraites à partir du corpus de la fouille de données.

À partir de l'interface présentée sur la Figure V.13 l'expert peut visualiser l'histogramme présentant le nombre de concepts présents dans chaque texte du corpus. Il lui suffit de cliquer sur le lien *afficher l'histogramme d'utilisation des textes* pour obtenir un nouvel écran dont deux extraits sont présentés sur la Figure V.14.

Cette option permet à l'expert de visualiser pour chaque texte le nombre de concepts qu'il contient (si un concept apparait plusieurs fois dans le même texte, il n'est comptabilisé qu'une seule fois). L'expert peut ainsi vérifier que les textes qui doivent contenir des concepts contiennent bien ces concepts et si ce n'est pas le cas, l'expert peut arrêter le processus d'extraction de connaissances et analyser les textes concernés. Dans une interface plus évoluée, nous pourrons proposer à l'expert de cliquer sur l'histogramme associé à un texte pour afficher les concepts instanciés dans ce texte.

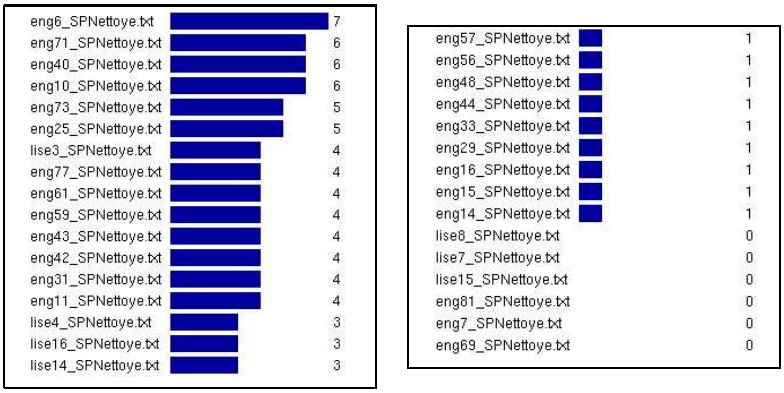


FIG. V.14 – Extraits des histogrammes présentant le nombre de concepts différents qui sont présents dans chaque texte du corpus.

Pour chaque règle $A \rightarrow B$ obtenue, les informations suivantes sont indiquées :
- intensité : la valeur de la mesure
- support : le nombre de texte vérifiant la règle (c'est-à-dire n_{AB} et non pas $P(AB)$)
- confiance et rappel : les deux mesures de qualité présentées dans le chapitre II

Ces informations sont utiles à l'expert pour interpréter les règles obtenues et estimer leur qualité sous différents angles.

Un lien hypertexte est associé à chaque règle. Ce lien permet à l'expert de visualiser les textes vérifiant la règle. Par exemple, pour la règle « ApplicationTo_1, MyPapsOrganzion_1 → Input_3 », la Figure V.15 présente la liste des dix textes vérifiant la règle.

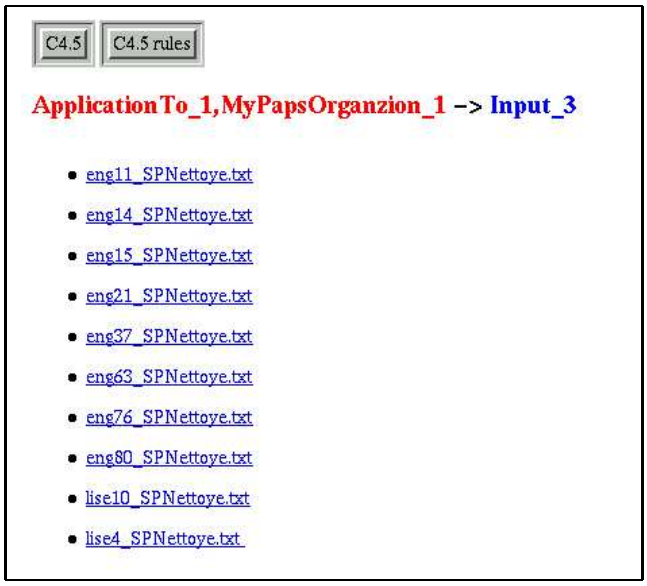


FIG. V.15 – Liste des textes du corpus vérifiant la règle (associés à la classe 1 par C4.5).

Cette interface propose à l'expert d'indexer les textes par rapport à cette règle, c'est-à-dire de créer deux groupes de textes : ceux vérifiant la règle et ceux ne la vérifiant pas. Puis, étant donné ces deux classes, d'effectuer un apprentissage supervisé de règles pour essayer de différencier ces deux classes. Cet apprentissage est réalisé avec C4.5 [Quinlan 1993] et l'expert peut visualiser soit l'arbre de décision appris (bouton *C4.5*), soit les règles obtenues (bouton *C4.5 rules*).

Pour permettre au classifieur d'apprendre les règles, nous utilisons la représentation matricielle des textes, avant discrétisation, et chaque texte est étiqueté 0 ou 1. Étant donné une règle $A \rightarrow B$, la classe 1 regroupe l'ensemble des textes vérifiant la règle, c'est-à-dire ceux pour lesquels A et B sont à $Vrai$ ensembles. Les autres textes sont étiquetés 0 et représentent ceux qui ne vérifient pas la règle $A \rightarrow B$. Les règles obtenues ont souvent un faible support, la classe 0 est donc souvent majoritaire.

La Figure V.16 présente les règles de décision obtenues par C4.5 à partir de la règle « ApplicationTo_1, MyPapsOrganzion_1 → Input_3 ».

Cette règle illustre un ensemble de textes relativement équilibrés. En effet, compte tenu de la discrétisation choisie par l'expert, cette règle indique que les articles qui parlent un peu de leur application et un peu de l'organisation de leur article parlent aussi beaucoup des entrées de leur système. Notre expert considère cette règle intéressante et les articles qui la vérifient méritent une lecture attentive car leur introduction semble bien équilibrée.

Cet outil fournit à l'expert un éclairage différent de celui proposé par les règles d'association.

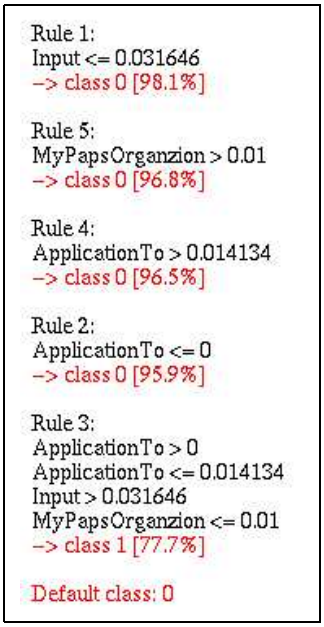


FIG. V.16 – Règles de décision apprises avec C4.5 à partir de la règle d'association « ApplicationTo_1, MyPapsOrganzion_1 → Input_3 ». La classe 0 contient les textes qui ne vérifient pas cette règle, et la classe 1 contient les textes qui vérifient la règle.

De plus, la discrétisation réalisée par C4.5 peut s'avérer différente de celle retenue par l'expert et qui a été utilisée pour obtenir l'ensemble des règles d'association. L'expert peut ainsi comparer la discrétisation qu'il a effectué avec celle proposée par C4.5. Une utilisation plus avancée de cette approche est en cours d'étude et pourrait permettre, par exemple, d'optimiser la discrétisation réalisée de manière non-supervisée.

Enfin, l'expert peut, à partir de l'écran lui proposant la liste des textes vérifiant la règle sélectionnée (voir Figure V.15), visualiser chacun des textes. La Figure V.17 présente un des textes vérifiant la règle sélectionnée. Nous pouvons voir que les concepts détectés dans le texte figurent en capitales dans le texte.

L'utilisateur peut cliquer sur un des concepts de la règle et les différentes instances du concept sont alors mises en relief dans le texte (couleur et indication du concept entre parenthèses) comme le montre la Figure V.18.

7 Validation des règles d'association obtenues à partir de textes

Comme nous l'avons indiqué précédemment, la totalité du processus de fouille de textes n'a pu être appliquée que sur deux des cinq corpus présentés dans ce chapitre.

Notons cependant que les résultats obtenus, concernant la partie extraction des connaissances, sont encourageants. En effet, les règles d'association que nous avons obtenues à partir du corpus d'introductions d'articles et du corpus de la société PerformanSe ont été validées par l'expert qui les a jugées intéressantes et, pour certaines d'entre elles, non triviales.

De manière a évaluer la capacité de notre approche à extraire des pépites de connaissance dans les textes, nous avons reproduit la validation comparative présentée dans le chapitre III, section 5.3. Rappelons que l'objectif de cette validation est de comparer les règles d'association obtenues par

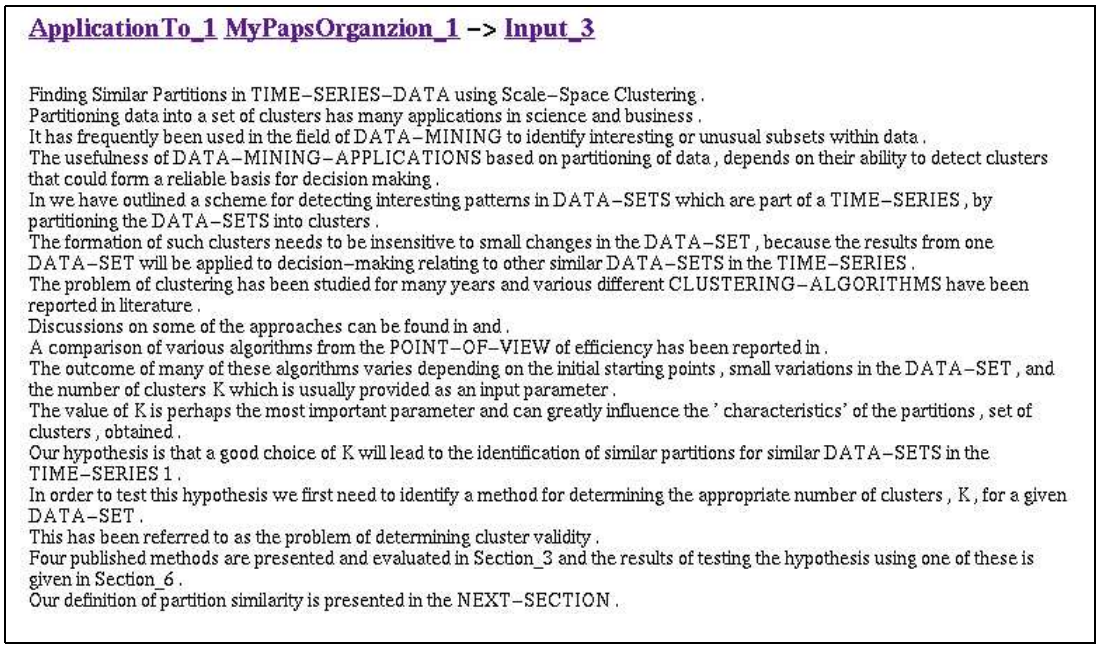


FIG. V.17 – Visualisation d'un texte vérifiant la règle.

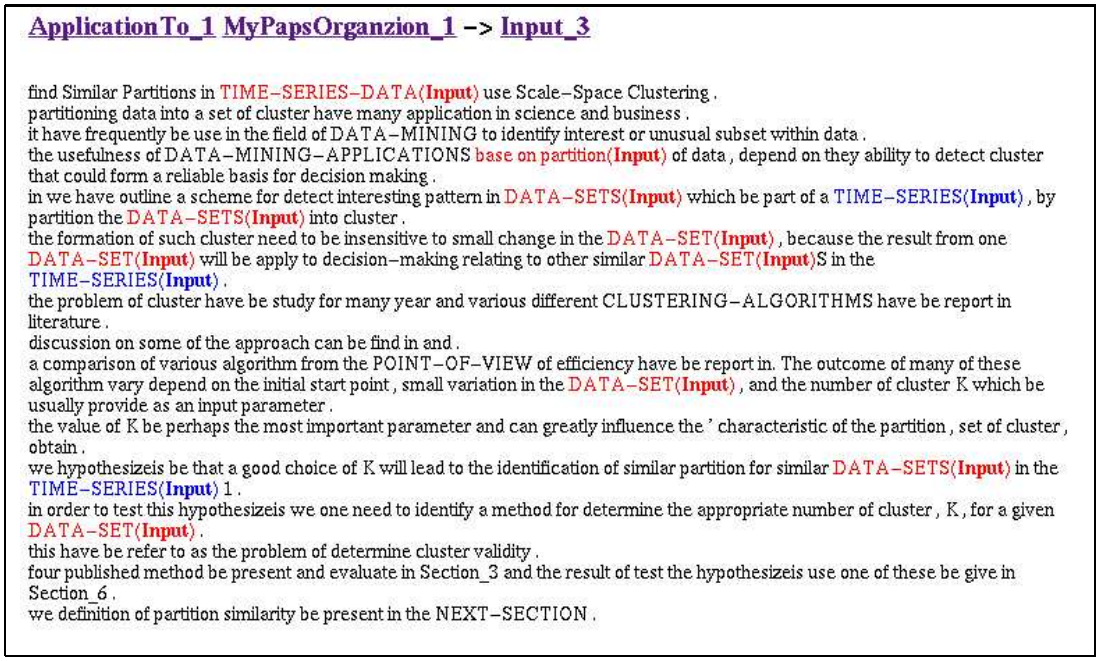


FIG. V.18 – Mise en relief des occurences d'un des concepts présent dans la règle.

deux mesures différentes : la moindre contradiction et l'Intensité d'Implication Normalisée.

Seules les règles de la forme $A \rightarrow B$, ayant une confiance strictement supérieure à 0.5, sont extraites et les mesures de qualité étudiées sont utilisées pour filtrer l'ensemble de ces règles. Ainsi, seules les règles les plus significatives étant donnés un ensemble de règles et une mesures de qualité sont étudiées.

7.1 Résultats relatifs au corpus de PerformanSe.

Sur ces données, la moindre contradiction trouve 25 règles et l'Intensité d'Implication Normalisée 38. Parmi celles-ci, 22 règles sont communes aux deux mesures. Il se trouve qu'aucune de ces règles n'a un support très petit, et donc elles représentent bien les *a priori* utilisés dans les tests du psychologue.

Par exemple, quand le concept de *stress* est fortement évoqué, alors le concept d'*environnement* l'est aussi, ce qui est normal dans la mesure où le stress s'exerce par l'intermédiaire de l'environnement.

Les trois règles trouvées par la moindre contradiction seule sont en fait à la limite des valeurs détectées par l'Intensité d'Implication Normalisée et donc ne sont pas essentiellement uniques à la moindre contradiction. Elles expriment, d'une part, une forme d'équivalence entre le concept *relationnel* et le concept *environnement*, ce qui n'est guère étonnant, et d'autre part le fait que le concept d'*implication dans l'entreprise* implique celui de *relationnel*. Ceci est au contraire une surprise et exprime peut être un *a priori* discutable des tests.

Les règles détectées par l'Intensité d'Implication Normalisée seule présentent toutes la propriété d'être presque autant confirmées qu'infirmées, ce qui les rend peu fiables.

7.2 Résultats relatifs au corpus d'introductions en anglais.

La moindre contradiction trouve une seule règle, en commun avec l'Intensité d'Implication Normalisée, qui en trouve six. La règle en commun affirme que lorsque l'auteur décrit des méthodes connues, alors il parle aussi de la nature des sorties de son système, et ce pour environ 25% des articles concernés. Cette règle est assez surprenante et mérite un examen plus approfondi.

Les cinq règles trouvées par l'Intensité d'Implication Normalisée seule sont, encore, presque autant confirmées qu'infirmées. On ne peut pas en tirer la conclusion que ceci est une propriété de l'Intensité d'Implication Normalisée, mais plutôt un caractère propre à nos corpus. En particulier, nos résultats sur la base « mushrooms » (voir chapitre III, section 5.3) ne présentent pas du tout cette propriété.

7.3 Conclusion sur les connaissances obtenues à partir de textes

Les différents résultats obtenus montrent deux choses, la première est qu'il est possible d'extraire des connaissances intéressantes et utiles pour l'expert à partir de textes, bien que ce type de données soit fortement bruité (ne serait ce que par toute la suite de traitements nécessaires pour aboutir à une représentation booléenne des textes).

L'utilité des règles d'association obtenues a pu être vérifiée, au moins pour le corpus de la société PerformanSe. En effet, lors de précédentes analyses réalisées sur un premier corpus fourni par cette société, nous avons mis en évidence des relations entre concepts que le psychologue pensait avoir clairement séparés. Ces règles lui ont donc permis de reformuler les textes de son corpus pour arriver à bien séparer ces concepts.

Le second point intéressant est que l'utilisation de la moindre contradiction pour extraire les règles d'association les plus pertinentes semble cohérente compte tenu de l'étude comparative réalisée avec l'Intensité d'Implication Normalisée.

Ainsi, la moindre contradiction, qui est une mesure simple et permettant d'extraire des règles dites de bon sens (c'est-à-dire plus souvent confirmées par les données qu'infirmées) s'avère au moins aussi efficace (voire plus) qu'une mesure statistiquement mieux fondée, pour extraire des connaissances dans les textes.

Une des explications de ce comportement est peut être liée aux capacités de la moindre contradiction à extraire des connaissances dans des données qui sont bruitées (voir chapitre IV).

8 Conclusion générale sur la fouille de textes

Nous avons présenté dans ce chapitre un processus complet de fouille de textes. Les travaux réalisés dans le cadre de cette thèse ne concernent qu'un sous-ensemble de ce processus, à savoir l'extraction de connaissances à partir d'une représentation condensée des textes. Les différentes étapes permettant d'obtenir la représentation condensée des textes font l'objet d'autres thèses (M. Roche et A-C. Amrani). La représentation condensée des textes que nous utilisons est obtenue à partir d'une chaîne d'outils dont les résultats sont plus ou moins fiables.

L'introduction de l'expert dans le processus permet de diminuer de manière significative les diverses erreurs pouvant apparaître, notamment au niveau de la phase d'étiquetage des textes où la présence de l'expert permet de réduire le pourcentage de mots inconnus et mal étiquetés.

Dans la première phase du processus de fouille de textes, le système propose à l'expert un ensemble de collocations et de relations syntaxiques. Ensuite l'expert les associe à un ensemble de concepts qui sont pertinents pour le domaine étudié. Ces concepts et leurs descriptions sont utilisés pour reformuler de manière condensée chaque texte du corpus. La forme condensée obtenue correspond à la matrice des fréquences d'occurence de chaque concept dans les textes.

Sachant que nous avons choisi d'extraire des règles d'association entre concepts à partir des textes et que les techniques d'extraction de règles que nous utilisons travaillent avec des données booléeenes ou discrètes, la matrice des fréquences d'occurences doit être discrétisée pour que nous puissions l'exploiter directement pour en extraire des règles d'association.

Plusieurs méthodes de discrétisation ont été présentées dans ce chapitre et nous avons retenu une méthode relativement naïve mais qui permet d'obtenir des résultats qui intéressent l'expert. Un outil permettant d'assister l'expert lors de la phase d'extraction des connaissances a été conçu et testé. Les diverses fonctionnalités intégrées dans cet outil offrent une aide précieuse à l'expert et lui permettent d'interpréter les résultats plus facilement. Dans une version améliorée, nous pensons intégrer un système permettant à l'expert de modifier manuellement la discrétisation associée à un concept en choisissant le ou les points de coupure qui posent problème.

La qualité des connaissances obtenues est très liée à la « qualité » de la discrétisation. Dans notre cas, un des critères de qualité de l'expert concernant la discrétisation est d'obtenir des intervalles de valeurs dont les tailles soient relativement comparables. Cette contrainte permet *a priori* de ne pas introduire artificiellement des relations entre concepts dans les données (voir section 5.3).

L'étude de l'impact de la discrétisation sur les connaissances obtenues devrait nous permettre de définir une technique de discrétisation dédiée à la fouille de textes et probablement moins entâchée des biais de l'expert.

Comme nous l'avons observé dans les expérimentations réalisées, nous obtenons relativement peu de règles et ces règles reflètent, soit les *a priori* du domaine étudié, soit sont suffisament surprenantes pour que l'expert ne soit pas en mesure de les interpréter immédiatement. Bien que le système utilisé

pour extraire les règles n'intègre pas de contrainte sur le support minimal des règles recherchées, nous n'avons pas observé de règles ayant un très faible support et représentant donc potentiellement des véritables pépites de connaissance. L'absence de pépites de connaissance est peut être due aux données qui ne contiennent pas ces pépites ou alors les données analysées sont peut être bruitées. En effet, la représentation condensée des textes est issue de plusieurs processus séquentiels qui sont imparfaits et il y a alors une propagation des erreurs et donc du bruit.

Il serait donc intéressant d'intégrer les travaux concernant l'étude de l'impact des données bruitées sur les connaissances pour évaluer la qualité des règles avec l'indicateur de fiabilité proposé dans le chapitre IV.

Notons malgré tout que nous avons choisi la mesure de qualité qui présente le meilleur comportement en présence de données bruitées.

De plus, les comparaisons expérimentales entre les résultats obtenus par la moindre contradiction et par l'Intensité d'Implication Normalisée montrent que ces deux mesures permettent d'obtenir des règles intéressantes pour l'expert, les « meilleurs » résultats étant obtenus par la moindre contradiction.

Lorsque nous avons obtenu l'ensemble des règles d'association les plus pertinentes entre les concepts du domaine définis par l'expert, une question se pose alors assez naturellement : *que faire de ces règles ?*

De nombreuses utilisations sont envisageables. Ces règles peuvent par exemple être utilisées pour indexer de manière automatique les textes du corpus en fonction des règles qu'ils vérifient et donc des propriétés associées aux règles. Les groupes obtenus peuvent alors être plus significatifs pour l'expert que l'intégralité du corpus.

Nous pouvons aussi envisager d'utiliser les règles pour comparer des corpus issus d'un même domaine mais écrits dans des langues différentes. La présence de relations entre certains concepts d'un corpus écrits en anglais et l'absence de ces mêmes relations dans un corpus écrit en français pourrait être utilisée pour faciliter la rédaction ou la traduction de documents dans une langue étrangère. Nous avons abordé dans l'introduction de cette thèse, la comparaison des deux corpus d'introductions d'articles scientifiques écrits en anglais, l'un par des anglophones et l'autre par des francophones. L'un des objectifs de l'expert était de comparer ces deux corpus pour essayer de mettre en évidence des différences entre les rédacteurs natifs et les rédacteurs non natifs. Nous avons effectivement constaté que certaines relations entre concepts observées chez les anglophones et reflétant une certaine cohérence dans l'écriture des documents, étaient absentes chez les francophones. La connaissance de ces relations pourrait donc être utilisée pour assister les auteurs francophones lors de la rédaction d'articles scientifiques en anglais.

Une autre application possible pour ces règles d'association obtenues à partir de textes qui est relativement proche de l'indexation des textes pourrait être, étant donné un ensemble de relations refletant les *a priori* d'un domaine, d'assister un utilisateur lors de la rédaction d'un document en lui indiquant les relations qui doivent être présentes dans son document.

Un dernier axe de recherche que nous avons relativement peu étudié est lié à la visualisation des règles obtenues. Généralement, les systèmes d'extraction de règles d'association fournissent un ensemble de règles relativement volumineux et il est alors difficile pour l'expert de toutes les visualiser et donc de les analyser. L'outil que nous avons proposé permet de visualiser, pour une règle, les textes qui la vérifient et pour chaque texte, l'expert peut visualiser les diverses occurences des concepts dans les textes. Ce système permet donc de vérifier si les règles sont cohérentes au niveau de chaque texte, mais il n'offre pas une vue globale des règles par rapport au corpus. L'étude d'un système permettant de mieux visualiser les règles et la manière dont elles sont instanciées dans les textes devraient faciliter le travail de l'expert lors de la phase de validation.

CHAPITRE VI

Conclusion

1 Bilan

Dans le cadre de cette thèse, nous nous sommes intéressés à l'extraction de connaissances de manière non supervisée à partir de données, soit booléennes, soit ordinales.

Les données que nous avons étudiées sont des données transactionnelles, c'est-à-dire composées d'un ensemble d'enregistrements décrits par des attributs, soit booléens, soit discrets. Les exemples de données transactionnelles sont multiples (voir l'introduction de cette thèse) et pour mémoire, nous citerons l'exemple fondateur des tickets de caisse d'un magasin où les attributs sont les produits mis en rayon et où chaque enregistrement correspond à la liste des produits achetés par un client.

Nous avons aussi étudié des données transactionnelles pour lesquelles les attributs sont à valeurs ordinales, c'est-à-dire qu'il existe une relation d'ordre entre les différentes valeurs possibles de chaque attribut.

Pour ces deux types de données (booléennes/discrètes et ordinales), les connaissances qui ont retenues notre attention sont, soit des associations entre les attributs décrivant les données, soit des règles d'association entre les attributs. Une règle d'association de la forme $A \rightarrow B$, où A et B sont deux attributs décrivant les données, caractérise une association entre les attributs A et B et indique que la relation liant les deux attributs est orientée de A vers B. L'existence d'une association $A - B$ entre deux attributs A et B indique que lorsque les individus de la base de données possèdent l'attribut A (resp. B) à $Vrai$ (dans le cas de données booléeennes), alors ces individus possèdent aussi fréquemment l'attribut B (resp. A) à $Vrai$.

Les règles d'association sont souvent plus intéressantes car elles indiquent la nature de la relation liant les attributs. En effet, si la règle $A \rightarrow B$ est détectée dans les données, alors nous pouvons en déduire que la présence de A permet de conclure sur la présence de B pour les individus concernés, d'où l'intérêt de ce type de connaissances.

L'extraction des règles d'association à partir d'une base de données est réalisée en utilisant des algorithmes fondés pour la plupart sur l'algorithme APRIORI. Cet algorithme utilise une contrainte imposée par l'utilisateur sur le support (ou couverture) minimal des règles recherchées. Le principe de cet algorithme est indiqué dans le chapitre I.

L'inconvénient majeur du support est qu'il est souvent difficile de demander à l'utilisateur de déterminer un support tel qu'aucune connaissance nouvelle et intéressante ne possède un support inférieur au seuil fixé. Inversement, si cette contrainte n'est pas utilisée alors la majorité des algo-

rithmes sont incapables d'extraire des connaissances utilisables car ils sont « noyés » par le volume de règles trouvé.

Nous avons donc proposé, lors de cette thèse, une solution permettant d'extraire des connaissances à partir de données transactionnelles dont les attributs sont booléens sans utiliser la contrainte du support. Notre approche est fondée sur l'utilisation d'une mesure de qualité, la moindre contradiction, permettant de mesurer le nombre de contradictions associées à une règle dans les données. L'algorithme permettant d'extraire les connaissances dans les données ne propose à l'utilisateur que les règles d'association les moins contredites par les données et ayant la propriété d'être les plus « significatives » parmi l'ensemble des règles les moins contredites (voir chapitre III).

Nous avons testé notre approche sur plusieurs bases de données différentes et les résultats obtenus sont relativement intéressants. Un des avantages de notre approche est lié au volume de règles proposées à l'expert qui est nettement inférieur à celui obtenu par d'autres approches telles qu'Apriori. En contre partie, nous ne sommes pas en mesure d'extraire de manière exhaustive toutes les règles les moins contredites contenues dans les données.

L'algorithme que nous avons proposé permet donc d'extraire des pépites de connaissance, c'est-à-dire des règles d'association intéressantes ayant potentiellement de très faibles supports. D'autres mesures de qualité peuvent être utilisées pour extraire les pépites de connaissance, la seule contrainte que doivent respecter ces mesures est de prendre leurs valeurs dans un intervalle borné et indépendant des données.

En éliminant la contrainte du support pour élaguer les règles les moins intéressantes, nous prenons le risque d'extraire des connaissances ayant de très faibles supports et pouvant donc, soit être dues au bruit inhérent aux données, soit être simplement du bruit.

Pour éviter ce problème et essayer d'évaluer la « pertinence » des connaissances extraites et proposées à l'expert, nous avons étudié l'impact des données bruitées sur les connaissances extraites. L'étude que nous avons réalisée concerne le comportement de notre algorithme lorsque six mesures de qualité différentes sont utilisées pour extraire les règles d'association à partir de données aléatoires dans lesquelles nous avons introduit artificiellement du bruit. Les résultats obtenus montrent que l'impact des données bruitées sur les connaissances extraites est très significatif et ce quelque soit le niveau de bruit introduit dans les données (entre 1 et 10%). Les conclusions obtenues à partir des résultats de cette étude, réalisée sur des données booléennes, ne sont pas généralisables à des données quelconques. Nous avons aussi réalisé une étude de l'impact de données ordinales bruitées sur les connaissances extraites et les résultats obtenus montrent qu'une des mesures étudiées (l'intensité d'implication ordinale) est très peu sensible au bruit. Cette étude a été réalisée sur des données réelles et, avant de conclure que l'intensité d'implication ordinale est quasiment insensible au bruit, nous devons réaliser une étude plus approfondie sur des données ordinales.

Pour essayer de minimiser l'impact du bruit (dans le cas de données booléennes), nous avons proposé et testé deux approches différentes. La première approche consiste à déterminer le support minimal tel qu'aucune règle d'association trouvée par notre algorithme ne disparaisse lorsque les données sont bruitées. Le support minimal ainsi déterminé est alors utilisé pour filtrer l'ensemble des règles d'association que nous proposons à l'expert. L'étude et l'analyse de cette solution s'est révélée quasiment pire que le problème initial. En effet, ce support minimal (lorsqu'il existe) permet bien de réduire le pourcentage de règles perdues lorsque les données sont bruitées. Par contre, le revers innatendu de cette solution est l'augmentation du pourcentage de règles erronées qui apparaissent lorsque les données sont bruitées. Or il semble évident que lorsqu'un ensemble de règles est soumis à l'expert, il est préférable que cet ensemble ne contienne pas de règles erronnées (c'est-à-dire purement dues au bruit).

Dans la deuxième solution que nous avons proposée, nous avons évalué la capacité de chacune

des mesures de qualité testée à retrouver un ensemble de règles obtenu à partir de données « saines » dans ces mêmes données que nous avons artificiellement bruitées. Nous avons ainsi pu proposer un estimateur de fiabilité associé à chaque mesure et reflètant la capacité de la mesure à retrouver les « bonnes » règles dans des données bruitées. Cette évaluation a permis de montrer que la moindre contradiction présente le meilleur comportement (estimation de fiabilité toujours supérieure à celle prise par les autres mesures).

Fort de ce constat, nous avons utilisé cette mesure de qualité pour extraire des connaissances à partir de données réelles. Nous avons étudié deux familles de données réelles : des données académiques présentant l'avantage d'avoir déjà été très étudiées et des données réelles présentant l'avantage de susciter l'intérêt des experts.

Les résultats obtenus sur les données académiques ont montré que notre mesure fournit des ensembles de règles nettement moins volumineux que ceux obtenus par Apriori sans post-élagage. Sachant que notre approche n'utilise pas de support minimal pour extraire les règles, il existe un gain réel par rapport à Apriori en ce sens qu'une partie des règles non proposées à l'expert ne sont pas engendrées, contrairement à Apriori qui engendre tout pour ensuite élaguer les règles.

L'évaluation de l'intérêt des règles obtenues à partir des données académiques n'est pas aisée car peu d'experts s'intéressent vraiment à ces données.

Nous avons donc utilisé des données réelles issues d'un processus de fouille de textes pour lesquels un expert a pu analyser les règles obtenues. Le processus complet de fouille de textes décrit dans cette thèse permet, à partir d'un corpus de textes d'obtenir une représentation condensée de celui-ci à partir de laquelle nous pouvons extraire des connaissances. Une interface permettant d'assister l'expert lors de la phase d'extraction des règles d'association a été conçue. Cette interface permet à l'expert de contrôler l'étape de discrétisation et d'extraction des connaissances. Les règles d'association que nous obtenons sont proposées à l'expert et celui-ci peut utiliser deux outils différents pour les valider. Nous lui proposons de valider les règles en comparant les règles de décision apprises grâce à C4.5 sur la classification induite par chaque règle obtenue (voir chapitre V, section 6, page 134). De plus, pour chaque règle, l'expert peut visualiser les textes qui vérifient la règle ainsi que les différentes occurences des concepts, présents dans la règle, dans le texte. Cet outil lui permet de vérifier si la règle est cohérente pour chaque texte et de voir quelles sont les portions du texte qui vérifient la règle.

Nous avons utilisé deux mesures de qualité différentes pour extraire les règles d'association à partir de deux corpus différents (langues et domaines différents). Les deux mesures utilisées sont la moindre contradiction et l'Intensité d'Implication Normalisée (version améliorée de l'Intensité d'Implication Classique, voir chapitre III, section 5, page 69). La comparaison des règles obtenues par chaque mesure a permis de montrer que la moindre contradiction tend à trouver des règles plus intéressantes que l'Intensité d'Implication Normalisée. Les règles trouvées par les deux mesures sont souvent les règles les plus pertinentes pour l'expert.

2 Perspectives

Nous allons présenter dans cette section plusieurs perspectives de recherche proches des thématiques développées dans cette thèse : la discrétisation, le bruit, les règles d'association, *etc.*.

Avant de présenter les différentes perspectives, nous présentons une extension possible de notre travail concernant l'Intensité d'Implication Normalisée.

2.1 Extension de l'Intensité d'Implication Normalisée

L'Intensité d'Implication Normalisée représente une amélioration de l'Intensité d'Implication Classique que nous avons présenté dans cette thèse. Cette nouvelle mesure conserve les bonnes propriétés de l'Intensité d'Implication Classique et reste discriminante en présence de données volumineuses.

Les expérimentations comparatives que nous avons réalisées avec la moindre contradiction ont montré que l'Intensité d'Implication Normalisée est capable d'extraire des connaissances intéressantes à partir de données réelles.

Sachant que la mesure reste discriminante pour de « gros » volumes de données et que l'exploitation et l'extraction de connaissances dans des données volumineuses est un vrai problème de recherche tant pour le monde académique que pour le monde industriel, nous pensons qu'il serait intéressant de continuer les recherches sur cette mesure.

Dans un premier temps, nous allons étudier le comportement de cette mesure en présence de données bruitées, puis nous chercherons une extension de cette mesure permettant d'extraire des règles d'association plus complexes que les règles de la forme $A \rightarrow B$ où A et B sont de simples attributs décrivant les données.

2.2 Comment évaluer l'influence de la discrétisation sur les règles obtenues ?

Nous avons montré dans cette thèse que l'influence de la discrétisation sur les règles d'association obtenues est très significative et qu'il est difficile de trouver *a priori* la « bonne » discrétisation des attributs continus. Le développement de méthodes de discrétisation dédiées à l'extraction de règles d'association, c'est-à-dire prenant en considération l'utilisation ultérieure des valeurs discrètes trouvées, devrait permettre de réduire l'impact d'une « mauvaise » discrétisation sur les règles obtenues. Dans un premier temps, nous pensons que l'étude de la corrélation entre la discrétisation et les connaissances que nous obtenons à partir des données discrétisées permettrait d'assister l'expert lors de la phase de validation. En effet, si le système peut évaluer, pour chaque règle trouvée ou pour un ensemble de règles, la nature de la relation liant les règles et les choix effectués lors de la discrétisation, alors l'expert aurait une information très utile pour interpréter les règles. De plus, la connaissance des relations entre les règles et les paramètres de la discrétisation pourrait être utilisée pour déterminer un ensemble de règles « stables » étant donné un ensemble de paramètres valides pour la discrétisation.

L'interprétation d'une règle « stable » dépendant beaucoup de l'intérêt de l'expert, il nous semble important d'arriver à intégrer l'expert dans la boucle *discrétisation - extraction des règles*.

2.3 Comment évaluer la fonction d'intérêt de l'expert ?

La conception de méthodes permettant de déterminer (ou d'approximer) la fonction d'intérêt de l'expert en lui demandant de valider régulièrement des règles d'association serait un premier pas vers une extraction automatique et fiable de connaissances.

Pour arriver à « capturer » tout ou partie de la fonction d'intérêt de l'expert, nous pensons qu'il faut commencer par concevoir un système lui permettant d'inter-agir de manière très intuitive avec les différents éléments intervenant dans l'extraction des connaissances. Les quelques propositions que nous avons faites dans cette thèse reflètent le fruit d'une collaboration avec un utilisateur qui, au fur et à mesure de la conception des outils, émettait certains souhaits auxquels nous avons tenté de répondre et inversement, certains résultats suggéraient à l'utilisateur des modes d'interaction différents (notamment pour la validation des règles).

L'intégration de l'utilisateur dans la phase de conception et d'évaluation des outils est non seulement nécessaire mais elle permet aussi de faire émerger des modes d'interactions non triviaux lors de la création des outils.

2.4 Que faire des règles d'association ?

Nous avons déjà évoqué le problème posé par le volume des règles d'association trouvées automatiquement dans la conclusion du chapitre V. En effet, ayant obtenu un ensemble de règles d'association, l'expert peut, dans le meilleur des cas les valider, mais il lui est souvent difficile d'exploiter directement ces règles.

La plupart des méthodes d'extraction de règles d'association engendrent un volume de règles tel qu'aucun expert ne peut raisonnablement l'analyser.

Face à ce problème nous proposons deux solutions :

- concevoir et proposer des systèmes permettant de visualiser les règles pour assister l'expert dans la phase de validation
- définir, en collaboration avec l'expert, une ou plusieurs tâches dont les résultats peuvent être validés automatiquement et pour lesquelles l'utilisation de règles d'association permet d'améliorer les résultats.

De nombreux travaux ont été réalisés dans le cadre de la visualisation des règles. Nous pensons que l'un des objectifs principaux est de définir avec l'expert l'utilisation qu'il compte faire des règles trouvées pour adapter la présentation des résultats en fonction de l'utilisation des règles.

La deuxième solution, très liée à la proposition précédente, permettrait de valider automatiquement les règles trouvées. Toute la difficulté réside alors dans la définition d'une ou plusieurs tâches pour lesquelles l'utilisation de règles d'association pertinentes permet d'optimiser les résultats. Bien évidemment, si nous pouvions déterminer la fonction d'intérêt de l'expert, il serait probablement plus aisé de concevoir de telles tâches.

Nous avons proposé en conclusion du chapitre V plusieurs utilisations potentielles des règles d'association obtenues à partir de textes. Certaines de ces utilisations, telle l'aide à l'indexation automatique de textes, peuvent être validée de manière supervisée. Il serait donc intéressant d'explorer cette voie de recherche pour arriver à définir précisément la manière dont les règles d'association peuvent être utilisées pour pouvoir indexer des textes.

2.5 Réduire l'impact des données bruitées sur les connaissances obtenues

Nous avons malheureusement constaté que l'impact des données bruitées sur les connaissances est très souvent impossible à contrôler. L'estimateur de fiabilité que nous avons proposé dans le chapitre IV permet d'informer l'expert sur la résistance au bruit des règles mais, pour l'instant, nous ne sommes pas en mesure de lui proposer directement les règles les plus stables.

Nous n'avons pas pu tester notre estimateur de fiabilité sur des données réelles et nous ne sommes donc pas en mesure d'indiquer si cet indicateur est suffisant pour réduire l'impact du bruit.

Cependant nous pensons qu'il est difficile (voire impossible) d'obtenir des connaissances fiables en présence de données bruitées dans un cadre non supervisé, même si les résultats obtenus dans le cadre de l'étude des données bancaires tendent à montrer le contraire.

Pour confirmer cet *a priori*, nous proposons d'améliorer la modélisation du bruit pour pouvoir étudier de manière plus réaliste des données non artificielles. La conception d'un générateur de données aléatoires entièrement paramétrable représente aussi une perspective intéressante pour pouvoir étudier de manière « fiable » l'impact du bruit sur les connaissances.

La résistance de l'intensité d'implication ordinale au bruit dans les données permet d'extraire des connaissances fiables à partir de données ordinales. L'analyse de ce comportement et de la stabilité observée de la mesure pourrait permettre de mettre en évidence certaines propriétés utiles pour résister au bruit.

2.6 Étude de la nature des règles recherchées

Dans le cadre de cette thèse, nous nous sommes focalisés sur l'extraction de pépites de connaissance ayant la propriété d'avoir exactement un attribut en conclusion et au plus K_{max} attributs en prémisse. Cette contrainte permet de borner la complexité de notre algorithme d'extraction des pépites de connaissances. De plus, les règles obtenues sont plus facilement interprétables par l'expert.

Cependant, il existe des domaines où de telles règles ne sont pas utiles et où l'expert est plus intéressé par des règles ayant un nombre quelconque d'attributs en prémisse et au plus K_{max} en conclusion. Si l'expert est intéressé par ce type de règles, alors notre approche n'est pas en mesure de lui les proposer.

Il serait donc intéressant d'étudier les différentes familles de règles pouvant intéresser les experts et, à partir de ces familles, définir des approches adaptées à chaque type de règles recherchées.

Les propriétés de certaines familles de règles doivent pouvoir être utilisées pour optimiser les algorithmes.

Table des figures

Liste des tableaux

Liste des Algorithmes

Bibliographie

[Agrawal et al. 1993] R. Agrawal and T. Imielinski and A. N. Swami (1993). « Mining Association Rules between Sets of Items in Large Databases ». Dans *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216.
cité pages 11,16,19,27,28,39,69

[Alphonse et Matwin 2002] E. Alphonse and S. Matwin (2002). « Feature Subset Selection and Inductive Logic Programming ». Dans C. Sammut et A. G. Hoffmann, editors, *Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002)*, pages 11–18, University of New South Wales, Sydney, Australia. Morgan Kaufmann.
cité page 4

[Azé et Kodratoff 2002] J. Azé and Y. Kodratoff (2002). « Évaluation de la résistance au bruit de quelques mesures d'extraction de règles d'association ». *Extraction des connaissances et apprentissage*, 1(4) :143–154.
cité pages 35,39,56

[Azé 2003] J. Azé (2003). « Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances ». *RSTI série RIA-ECA*, 17(1-2-3) :171–182.
cité pages 35,57

[Bastide et al. 2002] Y. Bastide and R. Taouil and N. Pasquier and G. Stumme and L. Lakhal (2002). « Pascal : un algorithme d'extraction des motifs fréquents ». *Techniques et Science Informatiques*, 21(1) :65–95.
cité pages 16,69,72

[Bernard et Charron 1996] J-M. Bernard and C. Charron (1996). « L'analyse implicative bayésienne, une méthode pour l'étude des dépendances orientées : Données binaires ». *Revue Mathématique Informatique et Sciences Humaines*, 134 :5–38.
cité page 31

[Bezdek et Pal 1998] J. C. Bezdek and N. R. Pal (1998). « Some new indexes of cluster validity ». *IEEE Transactions on Systems, Man, and Cybernetics/Part B*, 28(3) :301–315.
cité page 128

[Blake et Merz 1998] C.L. Blake and C.J. Merz (1998). « UCI Repository of machine learning databases ».
cité pages 53,54,66

[Brill 1994] E. Brill (1994). « Some Advances in Transformation-Based Part of Speech Tagging ». Dans *AAAI, Vol. 1*, pages 722–727.
cité page 111

[Brin et al. 1997a] S. Brin and R. Motwani and C. Silverstein (1997a). « Beyond market baskets : generalizing association rules to correlations ». Dans *Proceedings of ACM SIGMOD'97*, pages 265–276.
cité pages 37,39

[Brin et al. 1997b] S. Brin and R. Motwani and J. D. Ullman and S. Tsur (1997b). « Dynamic itemset counting and implication rules for market basket data ». Dans J. Peckham, editor, *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*, pages 255–264. ACM Press.
cité page 19

[Catlett 1991] J. Catlett (1991). « On changing continuous attributes into ordered discrete attributes ». Dans Y. Kodratoff, editor, *Proceedings of the European Working Session on Learning*, pages 164–178. Springer Verlag.
cité page 121

[Chan et al. 1991] C. C. Chan and C. Barmr and A. Srinivasasn (1991). « Determination of Quantization Intervals in Rule Based Model for Dynamic Systems ». Dans *the IEEE Conference on System, Man, and Cybernetics*, pages 1719–1723.
cité page 124

[Chickering et al. 2001] D. M. Chickering and C. Meek and R. Rounthwaite (2001). « Efficient Determination of Dynamic Split Points in a Decision Tree ». Dans N. Cercone, T. Y. Lin, et X. Wu, editors, *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 91–98. IEEE Computer Society.
cité pages 126,128

[Church et Hanks 1990] K. W. Church and P. Hanks (1990). « Word Association Norms, Mutual Information, and Lexicography ». *Computational Linguistics*, 16 :22–29.
cité page 113

[Courtine 2002] M. Courtine (2002). « *Changements de représentation pour la classification conceptuelle non supervisée de données complexes* ». PhD thesis, Université Pierre et Marie Curie - Paris VI.
cité pages 127,128

[Daille et al. 1998] B. Daille and E. Gaussier and J.M. Langé (1998). « An Evaluation of Statistical Scores for Word Association ». Dans *J.Ginzburg, Z. Khasidashvili, C. Vogel, J.-J. Levy, and E. Vallduvi (eds) The Tbilisi Symposium on Logic, Language and Computation : Selected Papers, CSLI Publications*, pages 177–188.
cité page 113

[Daudé 1992] F. Daudé (1992). « *Analyse et justification de la notion de ressemblance entre variables qualitatives dans l'optique de la classification hiérarchique par AVL* ». PhD thesis, - Université de Rennes 1.
cité pages 25,57

[Davies et Bouldin 1979] D. L. Davies and D. W. Bouldin (1979). « A cluster separation measure ». *IEEE Transactions Pattern Anal. MAchine INtell.*, 1 :224–227.
cité page 128

[Dong et Li 1998] G. Dong and J. Li (1998). « Interestingness of Discovered Association Rules in terms of Neighborhood-Based Unexpectedness ». Dans X. Wu, Kotagiri Ramamohanarao, et

K. B. Korb, editors, *Research and Development in Knowledge Discovery and Data Mining, Proc. 2nd Pacific-Asia Conf. Knowledge Discovery and Data Mining, PAKDD*, volume 1394, pages 72–86. Springer.
cité pages 45,47

[Dougherty et al. 1995] J. Dougherty and R. Kohavi and M. Sahami (1995). « Supervised and Unsupervised Discretization of Continuous Features ». Dans *International Conference on Machine Learning*, pages 194–202.
cité page 115

[Dunn 1973] J. C. Dunn (1973). « A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters ». *Journal of Cybernetics*, 3(3) :32–57.
cité page 128

[Dunning 1993] T. E. Dunning (1993). « Accurate Methods for the Statistics of Surprise and Coincidence ». *Computational Linguistics*, 19(1) :61–74.
cité page 113

[Fayyad et Irani 1993] U. M. Fayyad and K. B. Irani (1993). « Multi-interval discretization of continuous-valued attributes for classification learning ». Dans *The 13th International Joint Conference on Artificial Intelligence*, pages 1022–1027.
cité page 121

[Fisher 1987] D. H. Fisher (1987). « Knowledge Acquisition via Incremental Conceptual Clustering ». *Machine Learning*, 2 :139–172.
cité page 4

[Flach et Lavrac 2003] P.A. Flach and N. Lavrac (2003). « *Rule Induction* », pages 229–267. Springer-Verlag, 2nd revised and extended edition edition.
cité page 4

[Frawley et Piatetsky-Shapiro 1991] W. Frawley et G. Piatetsky-Shapiro, editors (1991). *Knowledge Discovery in Databases*. MIT Press.
cité page 5

[Freitas 1998] A. A. Freitas (1998). « On Objective Measures of Rule Surprisingness ». Dans *Principles of Data Mining and Knowledge Discovery*, pages 1–9.
cité page 45

[Goodman et Smyth 1988] R. M. Goodman and P. Smyth (1988). « Information-Theoretic Rule Induction ». Dans *ECAI 1988*, pages 357–362.
cité pages 38,39

[Gras 1979] R. Gras (1979). « Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques ». Master's thesis, Université de Rennes 1.
cité pages 33,39

[Gras et al. 2001] R. Gras and P. Kuntz and H. Briand (2001). « Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille des données ». *Revue Mathématique et Sciences Humaines*, 154-155 :9–29.
cité pages 33,34,39

[Gray et Orlowska 1998] B. Gray and M. E. Orlowska (1998). « CCAIIA : Clustering Categorial Attributed Into Interesting Association Rules ». Dans *Proceedings of PAKDD'98*, pages 132–143.
cité pages 45,49

[Guillaume 2000] S. Guillaume (2000). « *Traitement des données volumineuses. Mesures et algorithmes d'extraction de règles d'association et règles ordinales* ». PhD thesis, Université de Nantes.
cité pages 16,20,22,33,69

[Guillaume 2002a] S. Guillaume (2002a). « Découverte de règles d'association ordinales ». *Extraction et Gestion des Connaissance et apprentissage*, 1(4) :29–40.
cité page 96

[Guillaume 2002b] S. Guillaume (2002b). « Discovery of Ordinal Association Rules ». Dans *PAKDD 2002*, pages 322–327.
cité pages 96,98

[Guillaume et al. 1998] S. Guillaume and F. Guillet and J. Philippe (1998). « Improving the Discovery of Association Rules with Intensity of Implication ». Dans *PKDD 1998*, pages 318–327.
cité page 96

[Hájek 2001] P. Hájek (2001). « The GUHA method and mining association rules ». *Computational Intelligence : Method and Applications*, pages 533–539.
cité page 19

[Hájek et al. 1966] P. Hájek and I. Havel and M. Chytil (1966). « The GUHA method of automatic hypotheses determination ». *Computing*, 1 :293–308.
cité page 19

[Halliday 1976] M. A. K. Halliday (1976). « *System and Function in Language* ». Oxford University Press, London.
cité page 108

[Haykin 1998] S. Haykin (1998). « *Neural Networks - A Comprehensive Foundation* ». Prentice Hall, 2nd edition.
cité page 90

[Heckerman et al. 1995] D. Heckerman and D. Geiger and D. M. Chickering (1995). « Learning Bayesian Networks : The Combination of Knowledge and Statistical Data ». *Machine Learning*, 20(3) :197–243.
cité page 5

[Hilderman et Hamilton 1999] R. Hilderman and H. Hamilton (1999). « Knowledge discovery and interestingness measures : A survey ».
cité page 45

[Holte 1993] R. C. Holte (1993). « Very Simple Classification Rules Perform Well on Most Commonly Used Datasets ». *Machine Learning*, 11 :69–91.
cité page 119

[IBM 1996] IBM (1996). « *IBM Intelligent Miner User's Guide* ». Version 1 Release 1, SH12-6213-00 edition.
cité pages 29,39

[Kerber 1992] R. Kerber (1992). « Chimerge : Discretization for numeric attributes ». *National Conference on Artificial Intelligence*, pages 123–128.
cité page 117

[Klemettinen et al. 1994] M. Klemettinen and H. Mannila and Pi. Ronkainen and H. Toivonen and A. I. Verkamo (1994). « Finding interesting rules from large sets of discovered association rules ». Dans N. R. Adam, B. K. Bhargava, et Y. Yesha, editors, *Third International Conference on Information and Knowledge Management (CIKM'94)*, pages 401–407. ACM Press.
cité pages 45,49,50

[Kodratoff 2000] Y. Kodratoff (2000). « Comparing Machine Learning and Knowledge Discovery in DataBases : An Application to Knowledge Discovery in Texts ».
cité pages 19,30

[Lallich et Teytaud 2004] S. Lallich and O. Teytaud (2004). « Évaluation et validation de l'intérêt des règles d'association ».
cité pages 19,20,29,39,43

[Lavrac et al. 1999] N. Lavrac and P. Flach and B. Zupan (1999). « Rule Evaluation Measures : A Unifying View ». Dans S. Džeroski et P. Flach, editors, *Ninth International Workshop on Inductive Logic Programming (ILP'99)*, volume 1634 of *Lecture Notes in Artificial Intelligence*, pages 174–185. Springer-Verlag.
cité pages 29,38,39,43

[Lehn 2000] R. Lehn (2000). « *Un système interactif de visualisation et de fouille de règles pour l'extraction de connaissances dans les bases de données* ». PhD thesis, Institut de Recherche en Informatique de Nantes.
cité pages 16,72

[Lenca et al. 2003a] P. Lenca and P. Meyer and P. Picouet and B. Vaillant (2003a). « Aide multicritère à la décision pour évaluer les indices de qualité des connaissances ». *RSTI série RIA-ECA*, 17(1-2-3) :271–282.
cité page 50

[Lenca et al. 2003b] P. Lenca and P. Meyer and P. Picouet and B. Vaillant and S. Lallich (2003b). « Critères d'évaluation des mesures de qualité des règles d'association, nř spécial "Entreposage et Fouille des Données" ». *RNTI Revue des Nouvelles Technologies de l'Information, CEPADUES*, pages 123–134.
cité page 50

[Lerman 1981] I. C. Lerman (1981). « *Classification et analyse ordinale des données* ». Dunod.
cité pages 30,39,97

[Lerman 1984] I. C. Lerman (1984). « Justification et validité statistique d'une échelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées ». *Publications de l'Institut de Statistique des Universités de Paris*, 29 :27–57.
cité pages 25,57

[Lerman et Azé 2003] I. C. Lerman and J. Azé (2003). « Une mesure probabiliste contextuelle discriminante de qualité des règles d'association ». *RSTI série RIA-ECA*, 17(1-2-3) :247–262.
cité pages 33,53,69

[Lerman et al. 1981] I. C. Lerman and R. Gras and H. Rostam (1981). « Élaboration et évaluation d'un indice d'implication pour des données binaires I ». *Revue Mathématique et Sciences Humaines*, 75 :5–35.
cité pages 30,31,32,39,69

[Liu et al. 2002] H. Liu and F. Hussain and C. L. Tan and M. Dash (2002). « Discretization : An Enabling Technique ». Dans *Data Mining and Knowledge Discovery*, volume 6, pages 393–423. Kluwer Academic.
cité page 115

[Liu et Motoda 1998] H. Liu and H. Motoda (1998). « *Feature Extraction, Construction, and Selection : A Data Mining Perspective* ». Kluwer Academic.
cité page 5

[Loevinger 1947] J. Loevinger (1947). « A systematic approach to the construction and evaluation of tests of ability ». *Psychological Monographs*, 61 :1–49.
cité pages 35,39

[MacQueen 1967] J. B. MacQueen (1967). « Some methods for classification and analysis of multivariate observations ». Dans *the Fiftf Symposium on Math, Statistics, and Probability*, pages 281–297.
cité page 127

[Merckt 1993] T. V. Merckt (1993). « Decision trees in numerical attribute spaces ». Dans *The 13th International Joint Conference on Artificial Intelligence*, pages 1016–1021.
cité pages 124,128

[Michalski et Chilausky 1980] R. S. Michalski and R. L. Chilausky (1980). « Learning by Being Told and From Examples ». *International Journal of Policy Analysis and Information System*, 4(125–161).
cité page 3

[Milligan 1996] G. W. Milligan (1996). « Clustering validation : Results and implications for applied analyses ». *Clustering and classification*, pages 341–375.
cité page 128

[Mitchell 1977] T. M. Mitchell (1977). « Version Spaces : A Candidate Elimination Approach to Rule Learning ». Dans *the 5th International Joint Conference on Artificial Intelligence*, pages 305–310.
cité page 3

[Mitchell 1997] T. M. Mitchell (1997). « *Machine Learning* ». McGraw Hill.
cité page 3

[Muhlenbach et Rakotomalala 2002] F. Muhlenbach and R. Rakotomalala (2002). « Utilisation d'amas pour la discrétisation de variables ». *IXème Congrès de la Société Francophone de Classification (SFC'02)*, pages 283–286.
cité page 122

[Pearl 1988] J. Pearl (1988). « *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference* ». Morgan Kaufmann.
cité pages 30,39

[Pearl 1999] J. Pearl (1999). « Simpson's Paradox : An Anatomy ».
cité page 46

[Pearl 2000] J. Pearl (2000). « *CAUSALITY. Models, Reasoning and Inference* ». Cambridge University Press.
cité page 46

[Piatetsky-Shapiro 1991] G. Piatetsky-Shapiro (1991). « Discovery, Analysis, and Presentation of Strong Rules ». Dans *Knowledge Discovery in Databases*, pages 229 – 248. AAAI Press / The MIT Press.
cité pages 22,35,39,45

[Provost et Aronis 1996] F. J. Provost and J. M. Aronis (1996). « Scaling Up Inductive Learning with Massive Parallelism ». *Machine Learning*, 23(1) :33–46.
cité page 45

[Quinlan 1990] J. R. Quinlan (1990). « Induction of Decision Trees ». Dans J. W. Shavlik et T. G. Dietterich, editors, *Readings in Machine Learning*. Morgan Kaufmann. Originally published in *Machine Learning* 1 :81–106, 1986.
cité page 119

[Quinlan 1993] J. R. Quinlan (1993). « *C4.5 : Programs for machine learning* ». Morgan Kaufmann.
cité pages 3,136

[Ray et Turi 1999] S. Ray and R. H. Turi (1999). « Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation ». Dans A. K. D. N. R. P. et J. Das, editor, *The 4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'99)*, pages 137–143.
cité pages 127,128

[Richeldi et Rossotto 1995] M. Richeldi and M. Rossotto (1995). « Class-driven statistical discretization of continuous attributes (extended abstract) ». Dans N. Lavrac et S. Wrobel, editors, *European Conference on Machine Learning (ECML'95)*, pages 335–338. Springer Verlag.
cité page 121

[Roche 2003] M. Roche (2003). « Extraction paramétrée de la terminologie du domaine ». *RSTI série RIA-ECA*, 17(1-2-3) :295–306.
cité pages 107,108,113

[Roche et al. 2004] M. Roche and J. Azé and O. Matte-Tailliez and Y. Kodratoff (2004). « Mining texts by association rules discovery in a technical corpus ». Dans *Proceedings of IIPWM'04 (Intelligent Information Processing and Web Mining), Springer Verlag series "Advances in Soft Computing"*, pages 89–98.
cité page 107

[Roche et al. 2003] M. Roche and O. Matte-Tailliez and J. Azé and Y. Kodratoff (2003). « Extraction de la Terminologie du Domaine : Etude de Mesures sur un Corpus Spécialisé Issu du Web ». Dans *Actes des Journées Francophones de la Toile 2003*, pages 279–288.
cité page 113

[Rosenblatt 1958] F. Rosenblatt (1958). « The perceptron : a probabilistic model for information storage and organization in the brain ». *Psychological Review*, 65 :386–408.
cité page 4

[Sahar 1999] S. Sahar (1999). « Interestingness via What is Not Interesting ». Dans *Knowledge Discovery and Data Mining*, pages 332–336.
cité pages 53,55,59

[Saporta 1990] G. Saporta (1990). « *Probabilités, Analyse des Données et Statistique* ». Edition Technip.
cité page 97

[Sebag et Schoenauer 1988] M. Sebag and M. Schoenauer (1988). « Generation of Rules with Certainty and Confidence Factors from Incomplete and Incoherent Learning Bases ». Dans J. Boose, B. Gaines, et M. Linster, editors, *Proc. of the European Knowledge Acquisition Workshop (EKAW'88)*, pages 28–1 – 28–20. Gesellschaft für Mathematik und Datenverarbeitung mbH.
cité pages 37,39

[Sinka et Corne 2002] M. Sinka and D. Corne (2002). « A Large Benchmark Dataset for Web Document Clustering ». Dans *Soft Computing Systems : Design, Management and Applications, Volume 87 of Frontiers in Artificial Intelligence and Applications*, pages 881–890.
cité page 127

[Steinbach et al. 2000] M. Steinbach and G. Karypis and V. Kumar (2000). « A comparison of document clustering techniques ». Technical Report 00-034, Department of Computer Science and Engineering, University of Minesota.
cité page 127

[Suzuki 1997] E. Suzuki (1997). « Autonomous discovery of reliable exception rules ». Dans *the Third International Conference on Knowledge Discovery and Data Mining.*
cité page 27

[Suzuki et Kodratoff 1998] E. Suzuki and Y. Kodratoff (1998). « Discovery of surprising exception rules based on intensity of implication ». Dans Springer-Verlag, editor, *Principles of Data Mining and Knowledge Discovery (PKDD).*
cité page 27

[Toussaint 1980] G. T. Toussaint (1980). « The relative neighborhood graph of a finite planar set ». *Pattern Recognition*, 12(4).
cité page 122

[Weiss et Hirsh 1998] G. M. Weiss and H. Hirsh (1998). « The problem with noise and small disjuncts ». Dans *Proc. 15th International Conf. on Machine Learning*, pages 574–578. Morgan Kaufmann, San Francisco, CA.
cité page 89

Résumé

Cadre général

Cette thèse s'inscrit dans le cadre général de la fouille de données non supervisée et concerne plus précisément l'extraction de connaissances dans des données transactionnelles. La fouille de données non supervisée, par opposition aux approches supervisées, a pour objectif d'extraire automatiquement des informations à partir de données sans l'aide d'un expert.

Des données transactionnelles correspondent à un ensemble de transactions décrites par différents attributs booléens ou discrets. L'exemple classique et fondateur du domaine est celui des tickets de caisse d'un supermarché où un ticket de caisse représente une transaction et les attributs booléens sont les produits mis en rayon. Chaque ticket de caisse contient donc la liste des produits achetés par un client.

Dans le cadre de cette thèse, nous nous sommes focalisés sur l'extraction de connaissances de la forme « règles d'association » à partir de telles données. Ces règles sont de la forme $A \rightarrow B$ où A est la prémisse de la règle et B la conclusion de celle-ci. Par exemple la règle $R_1 =$« pain,beurre $\rightarrow$ confiture » qui s'interprète comme « les clients qui achètent du pain et du beurre achètent aussi de la confiture ». Ces règles sont accompagnées de deux mesures : le **support** et la **confiance**. Le support représente le pourcentage de transactions vérifiant la règle et la confiance représente le pourcentage de transactions vérifiant la prémisse et la conclusion parmi celles qui vérifient la prémisse.

Ainsi, si le support de la règle R_1 est égal à 70% et sa confiance à 95%, alors cette règle est vérifiée par 70% des consommateurs et elle s'interprète comme « 95% des individus achetant du pain et du beurre achètent aussi de la confiture ».

Extraction des connaissances

Notre travail est motivé par l'extraction de « pépites de connaissance ». Ces pépites sont des règles d'association qui sont potentiellement utiles et nouvelles pour un expert du domaine. Leur nouveauté implique qu'elles puissent avoir un très faible support et elles doivent être au moins plus souvent confirmées qu'infirmées par les données pour être intéressantes. La deuxième contrainte se traduit par la sélection des règles d'association ayant une confiance strictement supérieure à 50%. La première contrainte est plus difficile à mettre en place si l'on considère que les approches classiques d'extraction de règles d'association utilisent toutes la notion de support minimal pour élaguer l'espace des règles d'association. Ce support minimal est utilisé pour filtrer les connaissances considérées comme *a priori* non intéressantes par l'expert. L'algorithme Apriori est l'algorithme de référence dans le domaine et utilise les propriétés mathématiques du support pour optimiser l'extraction des règles.

Malheureusement, les experts des données ont souvent des difficultés pour déterminer ce support minimal tel qu'aucune connaissance nouvelle ne possède un support inférieur au seuil retenu. Nous avons donc proposé une méthode permettant de ne pas fixer un support minimal et fondée sur l'utilisation de mesures de qualité.

Ayant choisi de pas utiliser le support pour élaguer l'espace des règles d'association, nous sommes confrontés à un problème majeur qui est le volume des règles à examiner. En effet, l'utilisation de la contrainte du support minimal permet de réduire le nombre de règles mais lorsque cette contrainte est abandonnée, le nombre de règles engendré augmente de manière exponentielle en fonction du nombre d'attributs décrivants les données.

Nous avons donc choisi d'utiliser d'autres mesures de qualité pour filtrer les règles d'association et nous avons étudié différents critères de qualité permettant de mieux cerner les mesures de qualité

proposées dans la littérature. Ces critères permettent aussi à l'utilisateur de mieux comprendre la nature des mesures de qualité utilisées pour extraire les règles d'association.

Les règles d'association obtenues doivent vérifier un ou plusieurs critères de qualité pour être considérées comme intéressantes et donc proposées à l'expert. Nous avons recensé de nombreuses mesures de qualité dans la littérature et la majorité d'entre elles sont souvent difficiles à appréhender pour l'utilisateur néophyte.

Ainsi, nous avons proposé une nouvelle mesure de qualité qui nous semble relativement aisée à expliquer à un utilisateur. Cette mesure, appelée la **moindre contradiction**, permet de ne retenir que les règles qui sont plus souvent vérifiées par les données plutôt qu'infirmées. Elle permet aussi de sélectionner les règles ayant un caractère relativement inattendu.

Puis, nous avons proposé une méthode pour corriger une mesure de qualité statistiquement bien fondée (l'**intensité d'implication**) mais présentant le défaut majeur de ne plus pouvoir distinguer les règles intéressantes des autres règles dès que les données deviennent trop volumineuses. La nouvelle mesure obtenue a été appelée l'**intensité d'implication normalisée**. De manière synthétique, cette mesure permet de sélectionner les règles d'association les plus surprenantes étant donné un ensemble de règles.

Nous avons ainsi pu proposer un algorithme, ExtrairePépites, utilisant la moindre contradiction et permettant d'extraire les pépites de connaissance tant recherchées sans utiliser la contrainte du support minimal. Cette nouvelle approche permet de minimiser l'intervention de l'expert du domaine lors de la phase d'extraction des règles d'association.

Nous avons inclu dans notre algorithme un procédé permettant d'une part de ne pas spécialiser les règles considérées comme intéressantes par la mesure de qualité utilisée. Par exemple, si la règle $R = A \rightarrow B$ est considérée comme intéressante par notre système alors aucune règle de la forme $A \wedge X \rightarrow B$ ne sera étudiée car ces règles répresentent les spécialisations de la règle R.

D'autre part, nous avons aussi inclut un élagage lié aux données et permettant de ne conserver que les règles les plus intéressantes relativement à l'ensemble des règles obtenu. Ainsi, lorsque nous obtenons un ensemble E^m de règles considérées comme intéressantes par la mesure de qualité utilisée, m, les valeurs prises par les règles pour la mesure m sont centrées et réduites. Après centrage-réduction, seules les règles R ayant une valeur $m(R)$ strictement supérieure à 1 sont proposées à l'expert pour validation. La technique du centrage-réduction par rapport aux valeurs prises sur les données est une caractéristique de l'approche AVL (Analyse de la Vraisemblance du Lien).

Nous avons comparé notre approche avec l'algorithme Apriori sur différentes bases de données classiques dans notre domaine et les résultats obtenus sur ces données ont permis de valider l'aspect fonctionnel de notre approche. La validation de la qualité des connaissances obtenues par notre algorithme ne peut être réalisée que par un expert et sur des données qui l'intéresse.

Notre approche a été validée sur des données issues d'un processus de fouille de textes présenté ci-dessous.

Étude de données bruitées

Dans le cadre de cette thèse, nous nous sommes aussi intéressé au comportement de notre algorithme en présence de données bruitées. En effet, lorsque nous travaillons avec des données réelles, il est important de ne pas négliger cet aspect, à savoir le bruit inhérent aux données.

Les données réelles sont souvent imparfaites et peuvent contenir entre autres des données manquantes, des données erronées, *etc.*Nous avons donc étudié le comportement de notre algorithme en présence de données artificielles dans lesquelles nous avons introduit un bruit aléatoire. Les diverses expériences réalisées ont permis de mettre en évidence la difficulté d'extraire automatiquement

des connaissances fiables à partir de données bruitées. Une des solutions que nous avons proposée consiste à évaluer la résistance au bruit de chaque règle et d'en informer l'expert lors de l'analyse et de la validation des connaissances obtenues.

Dans le cadre de notre étude de l'impact des données bruitées sur les connaissances obtenues, nous avons étudié des données réelles fournie par une banque et représentant des informations relatives aux comptes bancaires de divers clients. Les divers attributs décrivant ces données ne sont pas toujours de nature booléenne mais plus souvent de nature ordinale. De tels attributs sont donc muni d'une relation d'ordre. Nous avons étudié l'extraction de règles d'association ordinales et plus précisément l'impact du bruit dans les données sur ces règles. Ces travaux ont été réalisés en collaboration avec S. Guillaume et P. Castagliola.

Fouille de textes

La dernière partie de cette thèse concerne la fouille de textes. Comme nous l'avons précisé précédemment, cette étude nous a permi de valider la qualité des connaissances obtenues avec l'algorithme ExtrairePépites.

Dans le cadre d'un processus de fouille de textes, les connaissances recherchées dans ces textes sont des règles d'association entre des concepts définis par l'expert et propres au domaine étudié.

Le processus de fouille de textes que nous avons présenté est divisé en trois grandes étapes : collecte d'un corpus homogène, détection des traces de concepts et extraction des règles d'association entre concepts. La première étape n'est pas traitée en détail dans cette thèse mais elle doit être réalisée en collaboration étroite avec un expert du domaine car la suite du processus dépend de l'homogénité du corpus collecté.

La phase de détection des concepts est réalisée en collaboration avec l'expert. Le système présenté dans cette thèse et développé dans le cadre de la thèse de M. Roche permet d'extraire automatiquement les collocations les plus pertinentes du corpus pour les proposer à l'expert. Le rôle de l'expert consiste alors à définir un ensemble de concepts dont les traces dans le corpus sont caractérisées par la présence d'une partie de ces collocations et/ou d'un ensemble de relations syntaxiques.

La définition de ces concepts nous permet de réécrire le corpus sous une forme matricielle sur laquelle nous pouvons appliquer notre algorithme d'extraction de règles d'association.

Nous avons proposé un outil permettant d'extraire ces règles et d'assister l'expert lors de la validation de celles-ci. Les différents résultats obtenus montrent qu'il est possible d'obtenir des connaissances intéressantes à partir de données textuelles en minimisant la sollicitation de l'expert dans la phase d'extraction des règles d'association.

Conclusion

Les divers résultats obtenus avec notre algorithme et les deux mesures de qualité présentées dans cette thèse montrent que l'approche est intéressante et que les connaissances engendrées peuvent intéresser les experts. Nous avons aussi montré que l'étude de données réelles et par nature même bruitées est difficile. L'impact du bruit sur les connaissances obtenues n'est pas négligeable. Les diverses solutions algorithmiques présentées dans cette thèse et permettant de prendre en considération l'aspect bruité des données ont montré des limitations évidentes. L'étude amorcée dans cette thèse doit donc être approfondie.

Malgré la difficulté de travailler avec des données bruitées, nous avons intégré notre système d'extraction de connaissances dans un processus de fouille de textes (les textes sont par nature même bruités). La capacité de notre système à extraire peu de règle d'association ayant la particularité

d'être potentiellement des pépites de connaissances nous a permi d'obtenir, sur divers corpus de textes, des résultats intéressants et validés par des experts.

Mots Clés

Règles d'association, Mesures de qualité, Impact du bruit, Fouille de textes