Contribution à l'étude des mesures de l'intérêt des règles d'association et à leurs propriétés algorithmiques.

Yannick Le Bras

N° d'ordre : 2011telb0189

Sous le sceau de l'Université européenne de Bretagne

Télécom Bretagne

En habilitation conjointe avec l'Université de Bretagne Sud

École Doctorale – SICMA

Contribution à l'étude des mesures de l'intérêt des règles d'association et à leurs propriétés algorithmiques

---

Thèse de Doctorat

Mention : « Sciences et Technologies de l'Information et de la Communication »

Présentée par **Yannick Le Bras**

Département : LUSSI

Laboratoire : LabSTICC Pôle : CID

Directeur de thèse : Philippe Lenca Co-directeur : Stéphane Lallich

Soutenue le Mardi 5 Juillet 2011

Jury :

Rapporteurs :	José Luis Balcázar, Professeur, Universidad de Cantabria
	Amedeo Napoli, Directeur de Recherche au CNRS, Loria Nancy
Examinateurs :	Marc Boullé, Directeur de Recherche, France Télécom R&D, Lannion
	Bruno Crémilleux, Professeur, Université de Caen
	Thierry Dhorne, Professeur, Université de Bretagne-Sud
Directeurs :	Philippe Lenca, Maître de Conférences, Télécom Bretagne, Brest
	Stéphane Lallich, Professeur, Université Lumière - Lyon 2

Remerciements

Je remercie tout le monde...

Ces trois années ont été riches de rencontres. Je tiens en premier lieu à remercier Philippe Lenca et Stéphane Lallich pour m'avoir proposé ce sujet, et m'avoir encadré avec efficacité et liberté tout au long de ces années. Leurs avis, leur culture et leur recul m'ont permis de m'améliorer sur de nombreux points et notamment sur mon approche pédagogique des problèmes, ce qui est pour moi l'un des points les plus importants. J'ai beaucoup appris aussi au contact de Patrick Meyer et Sorin Moga, ainsi que de tout le département LUSSI qui m'a accueilli à bras ouverts durant ces trois ans, Yvon Kermarrec en tête. Je tiens en particulier à remercier l'ensemble des thésards d'avoir supporté mes élucubrations durant nos nombreuses pauses déjeuner. J'ai une pensée particulière pour ceux qui ont partagé mon bureau, notamment pour Vanea Chiprianov, Éric Le Pors, Clément Pira et Komate Amphawan.

Je voudrais remercier José Luis Balcázar et Amedeo Napoli d'avoir accepté d'être rapporteurs sur ma thèse, ainsi que Marc Boullé, Bruno Crémilleux et Thierry Dhorne d'avoir bien voulu faire partie de mon Jury. La qualité de leur travaux fait que leur point de vue m'est précieux et je suis fier que mon travail soit soumis à leur jugement.

Cette thèse a pu voir le jour grâce à un financement sous la forme d'une Allocation de Recherche comme il était possible d'en obtenir en 2008. Cette Allocation est couplée avec un Monitorat que les investissements de Alain Hillion, Godefroy Dang-Nguyen et particulièrement Luc Bougé, m'ont permis de réaliser à Télécom Bretagne, au sein même de l'École qui m'a accueilli pour ma Thèse. Sans ces personnes, rien n'aurait été possible. Je tiens aussi à remercier l'ensemble des personnes qui agissent dans les différentes administrations pour rendre la vie des doctorants moins monotone. Je pense particulièrement à toute l'équipe du CIES Grand Ouest, et particulièrement à Lydia Coléou pour leur travail riche et efficace.

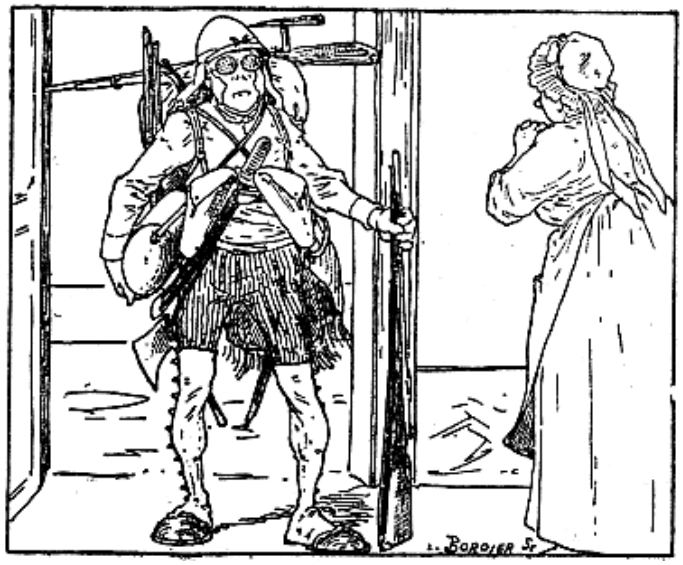
Pour le côté humain, je tiens à remercier certains membres du département, en priorité Ghislaine Le Gall pour son soutien et sa bonne humeur à tout moment, mais aussi Julie Le Diraison, Philippe Tanguy, Gilles Coppin, Sébastien Bigaret et bien d'autres pour leur bonne humeur et les rires partagés ensemble. J'ai une pensée particulière pour ceux qui ont traversé cette aventure avec moi depuis le début : Santiago Ruano Rincón, Vanea Chiprianov et Dusan Iorgovan. Sur un plan plus personnel, je veux aussi adresser un immense Merci plein de reconnaissance et de respect à mes parents, qui m'ont toujours soutenu et accompagné, et grâce à qui j'ai compris très tôt qu'il était important d'avoir des ambitions mais aussi de faire ce que l'on aime. J'aimerai qu'ils sachent que je les remercie chaque jour de tout ce qu'ils m'ont apporté. J'en profite évidemment pour remercier mes frères, Florent et Julien, pour leur soutien permanent. Je ne serais pas le même non plus sans ma seconde famille, tous ces amis que j'ai rencontrés au Lycée Franco-Allemand, et que je remercie d'avoir compris et accepté mon exil en Bretagne. C'est en partie grâce à eux que le retour en Région Parisienne sera moins dur.

Je tiens aussi à remercier d'une manière plus générale l'ensemble des Bretons qui m'ont accueilli à Brest. Je me suis senti chez moi dès les premiers jours. En particulier, je remercie l'ensemble du club de Hand Ball de Locmaria-Plouzané qui fera toujours partie de ma famille. Je ne citerai qu'Olivier Bonnaud, pour qui j'ai un profond respect et qui m'a tant appris, mais à travers lui, c'est l'ensemble des joueurs qui ont été un temps mes coéquipiers, que je tiens à remercier.

Ces trois années ont été parsemées de rencontres et il est compliqué de penser à toutes. Que ceux ou celles dont le nom n'apparait pas ici me pardonnent : je pense à eux.

Enfin, à Sheila qui a illuminé ces trois années et qui illumine toujours mon quotidien, j'adresse des remerciements pleins de tendresse.

À vous tous, merci.



« Scholastique ! — Et ousque va Monsieur avec toute cette quincaillerie ? — Je vais sonder les abîmes de l'Océan, combattre le Huron et soumettre le Pied-Noir ; je vais de ma semelle triomphante fouler les cimes orgueilleuses des monts… Enfin, je vais faire [une thèse], quoi ! »

Figure ⊥ : D'après *L'idée fixe du Savant Cosinus*, 3è Chant, Christophe, 1893

Table des matières

Table des figures

Liste des tableaux

Introduction

La fouille de données est un domaine nouveau qui trouve sa source dans les années 80, avec l'avènement de technologies toujours plus avancées pour le stockage des données. Les entreprises stockent, mais n'exploitent pas et l'on commence alors à prendre conscience de la quantité de connaissances perdues à laisser dormir ces informations. De nombreuses techniques sont nées pour extraire des informations de ces données, et découvrir des connaissances inconnues. Parmi ces techniques, la recherche d'associations connait depuis le début des années 90 un fabuleux essort.

Il existe de nombreuses techniques pour la recherche de connaissances dans une base de données. Lorsque les données ne sont pas trop nombreuses, les outils statistiques permettent de formuler ou de valider des hypothèses. Lorsque la taille des données croît, on peut alors faire appel au domaine de l'analyse de données, dont l'une des principale technique est l'analyse en composantes principales. Seulement cette technique demande l'extraction de valeurs propres, qui devient rapidement impossible. La fouille de données entre alors en jeu, avec des techniques telles que la construction d'arbres de décision, les machines à vecteurs de support ou encore la recherche d'associations.

La recherche d'associations dans une base de données consiste à mettre en évidence des liens entre les colonnes (attributs) de la base. L'exemple historique concerne la mise en évidence par la chaine de produits pharmacologiques Osco Drug aux États-Unis d'un lien entre l'achat de couches pour bébé et l'achat de bière dans ses magasins distributeurs. Ce lien a été découvert après l'analyse de 1.2 millions de tickets de caisse. La vérité sur cette anecdote peut être lue dans un texte de Daniel J. Power [1]. Ce qu'il faut en retenir, c'est que parfois, des informations insoupçonnées se cachent dans les données.

La recherche d'associations, et plus particulièrement de règles d'associations, c'est-à-dire des associations orientées, ou de règles de classe, dont la cible est prédéterminée, est le sujet principal de cette thèse. Les algorithmes permettant cette recherche présentent de grands avantages et ont été rendu toujours plus efficaces grâce à des évolution techniques. Ils permettent d'exploiter de grandes bases de données, mais avec l'accroissement des bases de données, leurs inconvénients prennent chaque jour un peu plus le dessus. Les deux principaux problèmes rencontrés sont tout d'abord un problème de quantité : les règles d'association dans une base de données de n colonnes sont au nombre de 3^n et le problème de décision sous-jacent fait partie de la très réputée classe des problèmes NP-complets ; ensuite un problème de qualité : les règles extraites ne sont pas toujours pertinentes, et il nous faut trouver des moyens de les évaluer.

La plupart des méthodologies pour la fouille de données prévoient deux étapes distinctes : modélisation, et évaluation, qui correspondraient d'une part à la génération des règles d'association, et d'autre part à l'évaluation de ces règles par des mesures. Cette méthodologie a longtemps été respectée dans le cadre des règles d'association. Le point de mire de cette thèse est la réunification de ces deux étapes, c'est-à-dire l'extraction directe de règles pertinentes, soit au sens d'une mesure, soit au sens d'un concept donné. Le chemin vers ce but nous a mené dans différentes directions : du point de vue de la qualité, nous nous sommes intéressés à la robustesse des règles d'association ; concernant la quantité, nous nous sommes focalisés sur la découverte de pépites et de règles optimales qui sont des sous-ensembles pertinents de l'ensemble des règles ; enfin, pour réunir les aspects qualité et quantité, nous avons introduit de nouvelles heuristiques de découverte s'appuyant sur des mesures d'intérêt.

Dans un premier temps, il a fallu se familiariser avec la recherche de règles d'association existante, et par conséquent avec la recherche de motifs fréquents. La partie I décrit ces deux étapes en

1. http://www.dssresources.com/newsletters/66.php

s'appuyant sur un état de l'art du domaine. Nous définissons à la fin de cette partie un cadre formel d'étude des règles d'association et des mesures de leur intérêt s'appuyant sur la table de contingence d'une règle d'association. Les trois degrés de liberté permettent de voir une règle comme un point de $\mathbb{R}^3$ et les mesures d'intérêt comme des fonctions de trois variables. Nous adoptons donc une vision géométrique des règles.

Dans le cadre de l'évaluation des règles d'association, nous introduisons dans la partie II une nouvelle notion de robustesse qui caractérise la résistance de la mesure d'une règle par rapport à des perturbations de la base de données. Cette notion utilise notre cadre formel et notamment la vision projetée des règles qui en découle. Nous verrons qu'évaluer une règle d'association par rapport à une mesure revient à étudier des lignes de niveau dans un espace à trois dimensions, et nous ferons ainsi un parallèle entre la fouille de données et la géométrie euclidienne. Nous définissons la robustesse d'une règle comme la distance de son projeté à la ligne de niveau définie par le seuil de mesure.

Nous nous intéressons ensuite à la phase d'extraction des règles, et notamment aux propriétés algorithmiques lui permettant d'être efficace. La partie III sera consacrée à ces propriétés algorithmiques, parmi lesquelles nous verrons que la plupart se limitent à l'utilisation d'une unique mesure. Notre but est de trouver des propriétés algorithmiques compatibles avec un grand nombre de mesures. Nous proposerons ainsi des généralisations de trois propriétés d'élagage de l'espace de recherche, et nous étudierons le lien entre propriétés algorithmiques et analytiques des mesures afin d'établir des conditions nécessaires ou/et suffisantes d'existence de ces propriétés pour une mesure donnée. Nous montrerons que ces propriétés définies initialement pour un ensemble restreint de mesures ont souvent un grand pouvoir de généralisation.

Enfin, pour clôturer cette thèse, nous présenterons nos travaux sur les propriétés d'anti-monotonie des mesures d'intérêt. Dans la partie IV, nous verrons que l'aspect géométrique de notre cadre formel permet de définir des propriétés d'anti-monotonie pour au moins 10 mesures d'intérêt dans le cadre des règles de classe. Ces propriétés d'anti-monotonie donnent naissance à un algorithme d'élagage efficace, du bas vers le haut, et sans aucune contrainte de fréquence. Nous verrons que son efficacité et sa qualité sont supérieures aux techniques actuelles. Nous offrons ainsi une contribution algorithmique, et non pas purement technique.

PREMIÈRE PARTIE : RECHERCHE ET ÉVALUATION DES RÈGLES - VERS UN CADRE FORMEL D'ÉTUDE.

Rechercher des motifs et des comportements dans une base de données est une tâche coûteuse à laquelle de nombreux travaux se sont intéressés. Nous détaillons dans cette partie les principaux axes de Recherche dans le domaine des motifs d'abord (chapitre 1), puis dans le domaine des règles d'association (chapitre 2). Nous verrons que les nombreuses améliorations techniques ne peuvent plus aujourd'hui suffire, et que la qualité des résultats est une question profonde. Nous introduirons alors les fondements de cette thèse par la définition d'un cadre formel d'étude des règles d'association et des mesures d'intérêt (chapitre 3). Ce chapitre est une compilation de nos publications [Le Bras et al. 09a, Le Bras et al. 10a, Le Bras et al. 09b]

TELECOM Bretagne

1 Sur la recherche de motifs dans une base de données

Dans ce chapitre, nous nous intéressons principalement à la recherche de motifs, sans introduire de notion d'intérêt. Le chapitre sera articulé autour de la notion de bases de données et de leurs différentes représentations (horizontale ou verticale). Il sera l'occasion de présenter les deux principales philosophies de recherche de motifs *via* les algorithmes historiques Apriori et Eclat, ainsi que leurs améliorations. De nombreux états de l'art existent sur la recherche de motifs, notamment fréquents. Nous faisons le choix ici de nous intéresser essentiellement, mais en détail, aux fondements qui nous serviront au cours de cette thèse. On pourra par exemple consulter [Goethals 05] ou [Pasquier 00] pour un état de l'art plus étendu.

1.1 Représentation des bases de données

Une base de données est constituée d'un ensemble fini d'attributs A, d'un ensemble fini de transactions T et d'une relation $\mathcal{R}$ entre ces deux ensembles. Cette relation peut prendre diverses formes suivant le type de base que l'on étudie. Elle pourra être binaire afin de spécifier naturellement la présence ou l'absence d'un attribut dans une transaction. Pour un attribut $\mathtt{A}$ et une transaction $\mathtt{T}$, on aura alors $\mathtt{T}\mathcal{R}\mathtt{A} \in \{0, 1\}$. Classiquement, les transactions d'un magasin pourront être représentées sous cette forme : si $client_i \mathcal{R} pain$ alors le client numéro i a acheté du pain. La base de données ainsi formée pourra être appelée base de données binaire et sera le plus souvent représentée horizontalement, en ne mentionnant pour chaque transaction que les attributs présents. La table 1.1 montre une telle base, où TID signifie *Transaction ID* et permet d'identifier la transaction.

TID	attributs
1	A C
2	B C D
3	A C D
4	C
5	A D

Table 1.1 : Base de données binaire, avec $T = \{1, 2, 3, 4, 5\}$ et $A = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}\}$.

On pourra aussi considérer une relation $\mathcal{R}$ discrète, qui traduira une base de données catégorielle. Dans ce contexte, chaque attribut $\mathtt{A}_i$ prend une valeur parmi un ensemble possible $\mathcal{A}_i$. On a donc $\mathtt{T}\mathcal{R}\mathtt{A}_i \in \mathcal{A}_i$ et pour chaque transaction, une valeur doit être spécifiée pour chaque attribut. Exceptionnellement, on pourra avoir à faire à une valeur manquante dans une transaction, sur un attribut. La gestion de ces valeurs manquantes ne fera pas l'objet de ce texte et nous considérerons une valeur manquante sur l'attribut $\mathtt{A}$ comme étant une valeur particulière de l'ensemble $\mathcal{A}_i$. Une telle base de données pourra par exemple servir à relever des particularités physiques d'un échantillon de personnes : yeux bleus, verts, noirs... ; cheveux blonds, roux, chatains, bruns... ;

peau jaune, blanche, mate, noire... La table 1.2 montre un exemple d'une telle base. Il est possible

TID	yeux	cheveux	sexe	taille
1	bleus	blonds	M	grand
2	vert	bruns	F	moyen
3	bleus	blonds	F	petit
4	marrons	roux	M	moyen
5	noirs	chatains	M	petit

TABLE 1.2 : Base de données catégorielle

d'effectuer une opération de *binarisation* d'une base de données catégorielles en créant la relation binaire $\mathcal{R}'$ sur les ensembles $T' = T$ et $A' = \bigcup_i (\{\mathtt{A}_i\} \times \mathcal{A}_i)$, c'est-à-dire en créant autant de nouveaux attributs que de couples $(attribut, valeur)$ possibles. La table 1.3 montre le résultat de la binarisation de la table 1.2. Le lecteur attentif remarquera que le processus de binarisation produit un ensemble de transactions contenant toutes le même nombre d'attributs, ce qui n'est pas le cas pour toutes les bases de données binaires (voir par exemple la table 1.1).

TID	attributs
1	(yeux,bleus) (cheveux,blonds) (sexe,M) (taille,grand)
2	(yeux,vert) (cheveux,bruns) (sexe,F) (taille,moyen)
3	(yeux,bleus) (cheveux,blonds) (sexe,F) (taille,petit)
4	(yeux,marrons) (cheveux,roux) (sexe,M) (taille,moyen)
5	(yeux,noirs) (cheveux,chatains) (sexe,M) (taille,petit)

TABLE 1.3 : Base de données catégorielles binarisée

Nous détaillons ici un troisième et dernier type de base de données, continues, dont les attributs peuvent prendre une valeur dans un espace continu. Pour chaque attribut $\mathtt{A}_i$, notons $\mathcal{A}_i$ son domaine de variation, que nous considérons pour plus de facilité, comme un sous-ensemble de $\mathbb{R}$ de mesure non nulle. La relation continue entre T et A est alors telle que $\mathtt{TRA}_i \in \mathcal{A}_i$ et représentera par exemple une base de données contenant des relevés de mesures continues. La table 1.4 montre un exemple d'une telle base pour des relevés de type environnementaux. Ce type de base de données peut être utilisé tel quel pour des tâches d'analyse de données comme l'analyse en composante principales ou la construction d'arbres de décision, mais on peut aussi le relier aux deux types précédemment énoncés par le biais d'une opération de discrétisation des attributs [Dougherty et al. 95]. Il existe de nombreuses stratégies de discrétisation, parmi lesquelles la plus simple consiste à dé-

TID	température	ensoleillement	pluviométrie	vent
1	12.4	4.1	115	60.2
2	6.5	5.2	214	23.2
3	15.7	2.5	146	23.8
4	14.3	6.4	175	0.5
5	14.2	6.7	357	67.8

TABLE 1.4 : Base de données continue

couper l'intervalle de variation de l'attribut en intervalles de même taille, dont le nombre doit être pré-déterminé. On obtient alors, après discrétisation, une base catégorielle classique, que l'on peut ensuite binariser. Les techniques de discrétisation ne sont pas anodines et sont indispensables pour de nombreux algorithmes de fouille de données. Le choix de la bonne technique (du bon nombre d'intervalles...) peut faire largement varier les résultats finaux.

Nous verrons par la suite que la plupart des algorithmes concernant la recherche de motifs fréquents et de règles d'association s'appuient sur les bases de données binaires, mais aujourd'hui, la grande majorité des bases de données contiennent des attributs catégoriels ainsi que des attributs continus. Nous remarquerons alors l'avènement de quelques algorithmes intéressants se focalisant sur la recherche de règles dans des bases de données catégorielles (partie III).

Le site de l'UCI Machine Learning Repository [1] est une référence en termes de bases de données de tests. À titre d'illustration, sur les 198 bases référencées au 1er février 2011, 36 sont décrites comme catégorielles et 82 comme contenant des attributs numériques (le reste contenant des attributs variés). Les bases utilisées pour illustrer les résultats de cette thèse figurent parmi les bases de l'UCI.

Après avoir défini ce que nous allons considérer tout au long de cette thèse comme étant une base de données, intéressons-nous maintenant précisément aux domaines que nous allons étudier. Dans un premier temps, nous nous focaliserons sur la recherche de motifs fréquents, pour laquelle nous présenterons les deux principales méthodes d'extraction, puis nous étendrons cette tâche à la recherche de règles d'association et finalement à l'évaluation de ces règles. Pour chacune de ces tâches, nous détaillerons les solutions algorithmiques et leur optimisation. Les principales ont eu lieu dans le domaine de la recherche de motifs fréquents.

1.2 Recherche de motifs fréquents

Nous avons donné en introduction une définition intuitive de la recherche de motifs fréquents. Nous allons ici définir clairement le problème sous-jacent ainsi que les enjeux et les difficultés rencontrées lors de cette recherche.

1.2.1 Définitions

Afin de définir la recherche de motifs fréquents, nous devons passer par trois étapes. La première consiste à donner la définition d'un motif. Nous verrons ensuite ce que l'on entend par motif fréquent, puis nous présenterons le problème informatique de la recherche de motifs.

Nous avons vu précédemment qu'une base de données peut être décrite par des attributs et des transactions.

> **Définition 1 – motif – :** Soit une base de données $\mathcal{DB}$ binaire définie par un ensemble d'attributs A et un ensemble de transactions T, reliés par une relation binaire $\mathcal{R}$. Un motif $\mathtt{I}$ est un sous-ensemble de A : $\mathtt{I} \subset A$. On dira qu'une transaction $\mathtt{T}$ contient un motif $\mathtt{I}$ et on notera $\mathtt{I} \subset \mathtt{T}$ *ssi* pour tout attribut $\mathtt{A}$ dans $\mathtt{I}$, on a $\mathtt{T}\mathcal{R}\mathtt{A}$:
>
> $$\mathtt{I} \subset \mathtt{T} \Leftrightarrow \forall \mathtt{A} \in \mathtt{I},\ \mathtt{T}\mathcal{R}\mathtt{A}.$$

La notation que nous utiliserons est la suivante : si $\mathtt{A}$ et $\mathtt{B}$ sont deux attributs, nous dénoterons par $\mathtt{AB}$ le motif $\{\mathtt{A}, \mathtt{B}\}$. De même si $\mathtt{I}_1$ et $\mathtt{I}_2$ sont deux motifs, nous noterons $\mathtt{I}_1\mathtt{I}_2$ le motif résultant de l'union de ces deux motifs.

Naturellement, nous pourrons voir une transaction $\mathtt{T}$ comme étant le motif constitué de l'ensemble des attributs qui lui sont associés par la relation $\mathcal{R}$. Ainsi, la notion d'inclusion valable pour les transactions est en fait une relation d'inclusion classique sur les ensembles. Dès lors, il est possible d'introduire une vision des motifs qui sera très utile par la suite, sous la forme d'un treillis dont l'ordre est donné par l'inclusion ensembliste. Sur la figure 1.1 est présenté le treillis des motifs

1. http://archive.ics.uci.edu/ml/

de la base de données de la table 1.1, sur lequel sont encadrés en vert les motifs effectivement présents dans au moins une transaction de la base et en rouge ceux qui sont absents. Sur cette figure,

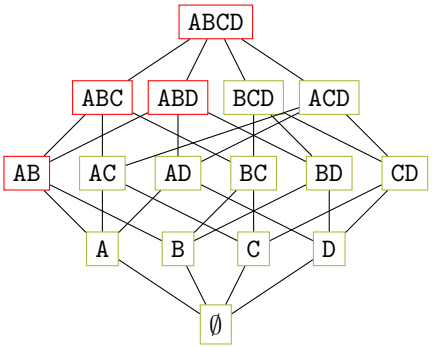


FIGURE 1.1 : Exemple de treillis de motifs

les motifs encadrés en vert sont donc les motifs qui ont une fréquence strictement plus grande que 0 dans la base de données. On pourrait par exemple s'intéresser aux motifs qui apparaissent au moins deux fois dans la base, ce qui nous donnerait la figure 1.2. D'une manière générale, on peut

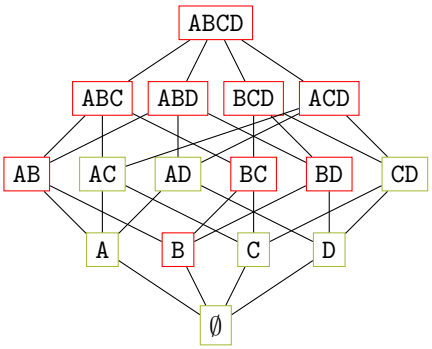


FIGURE 1.2 : Exemple de treillis de motifs apparaissant au moins deux fois

s'intéresser à n'importe quelle valeur minimale d'apparition. Ce choix est traduit par l'utilisation d'une mesure appelée le **support**.

> **Définition 2 – support – :** Soit une base de données $\mathcal{DB}$ binaire définie par un ensemble d'attributs A et un ensemble de transactions T, reliés par une relation binaire $\mathcal{R}$. Soit $\mathtt{I}$ un motif dans cette base. On définit le **support** absolu de $\mathtt{I}$ dans $\mathcal{DB}$ par :
>
> $$supp_{\mathcal{DB}}(\mathtt{I}) = \#\{\mathtt{T} \in T | \mathtt{I} \subset \mathtt{T}\}.$$
>
> Lorsqu'il n'y a pas de risque d'ambiguité, on omettra de préciser la base de données.

Dans la base 1.1 on a donc $supp(\mathtt{C}) = 4$, $supp(\mathtt{CD}) = 2$ et $supp(\mathtt{ABD}) = 0$. Cette définition du **support** est une définition absolue. Il possède une variante relative, obtenue en normalisant le **support** par rapport à la base, c'est-à-dire en divisant sa valeur par le nombre de transactions contenues dans la base de données. On se ramène ainsi à une fréquence d'apparition du motif dans la base. Par la suite, nous ferons preuve d'un abus de langage en parlant de probabilité plutôt que de fréquence. L'analogie dans ce domaine est simple, et permet d'introduire des notions classiques du calcul des probabilités.

Le problème de la recherche de motifs fréquents énoncé par [Agrawal et al. 93] consiste à fixer un seuil de **support** minimal σ et à rechercher, dans la base de données, tous les motifs qui apparaissent plus de σ fois lorsque l'on considère la version absolue du **support**.

> **Définition 3 – recherche de motifs fréquents – :** Soit une base de données $\mathcal{DB}$ binaire définie par un ensemble d'attributs A et un ensemble de transactions T, reliés par une relation binaire $\mathcal{R}$. On se donne un seuil de **support** σ et on définit l'ensemble des motifs fréquents $\mathcal{F}req_{\mathcal{DB}}(\sigma)$par :
>
> $$\mathcal{F}req_{\mathcal{DB}}(\sigma) = \{\mathtt{I} \subset A | supp_{\mathcal{DB}}(\mathtt{I}) \geq \sigma\}.$$
>
> Le problème de la *recherche de motifs fréquents* est de déterminer, étant donnés une base et un seuil, l'ensemble complet $\mathcal{F}req_{\mathcal{DB}}(\sigma)$.

Les problèmes rencontrés lors de la recherche de motifs fréquents sont nombreux et sont liés à la complexité du problème. En effet, si l'on note p le nombre d'attributs binaires dans la base de données étudiée, le nombre de motifs possible est 2^p, c'est-à-dire exponentiel. Or il n'est pas rare d'avoir à faire à des bases contenant plusieurs dizaines d'attributs. Il suffit, pour s'en persuader, de considérer le cas d'une base de données commerciales contenant un attribut par article vendu dans un magasin et une transaction par client. Le nombre d'articles dans un grand magasin dépasse largement la centaine et le nombre de motifs possibles dépasse alors rapidement... le nombre d'atomes dans notre galaxie. Il n'est donc absolument pas raisonnable d'espérer tester tous les motifs pour savoir s'ils sont fréquents ou non. Cette intuition est renforcée par l'existence de résultats théoriques montrant que le problème de l'existence d'un motif fréquent est un problème NP-complet [Zaki 00] et que la recherche des motifs fréquents est par conséquent un problème NP-difficile. La nécessité de mettre en place des heuristiques de découverte des motifs fréquents est dès lors évidente et nous allons étudier dans la partie suivante deux algorithmes particuliers de découverte, s'appuyant chacun sur une représentation informatique différente de la base de données.

1.2.2 Base de données horizontale : l'exemple de Apriori

Apriori est un algorithme introduit par [Agrawal et Srikant 94] qui permet de trouver l'ensemble des motifs fréquents d'une base de données à partir d'un seuil de **support** σ fixé par l'utilisateur. Cet algorithme s'appuie sur une propriété dite d'*anti-monotonie* de la mesure de **support** et a popularisé l'utilisation du **support** en fouille de données. Il utilise pour cela une représentation horizontale de la base de données, c'est-à-dire que les lignes sont représentées par les transactions, chaque ligne contenant l'ensemble des attributs en relation avec la transaction. Cette représentation s'oppose à la représentation verticale où les lignes sont les attributs et que nous verrons dans la section suivante.

1.2.2.1 Heuristique d'élagage

Nous allons tout d'abord expliciter la propriété d'anti-monotonie du **support** [Agrawal et Srikant 94], afin de pouvoir en déduire un algorithme efficace de recherche des motifs fréquents.

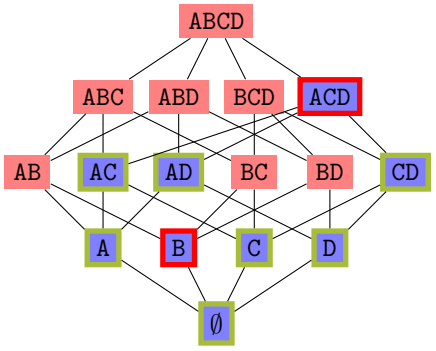


FIGURE 1.3 : Recherche des motifs fréquents niveau par niveau. Un fond bleu représente un candidat (tous ses sous motifs sont fréquents), un fond rouge indique un motif dont au moins un sous-motif est non fréquent. Le pourtour rouge indique un motif non fréquent dont le **support** a dû être calculé, un pourtour vert indique un motif fréquent dont le **support** a dû être calculé.

Propriété 1 : Soit une base de données $\mathcal{DB}$ binaire définie par un ensemble d'attributs A et un ensemble de transactions T, reliés par une relation binaire $\mathcal{R}$. Soit $\mathtt{I}$ et $\mathtt{I}'$ deux motifs dans cette base, d'intersection nulle et σ un seuil de **support** fixé au préalable. Alors :

$$supp(\mathtt{I}) \geq supp(\mathtt{II}')$$

et par conséquent

$$supp(\mathtt{I}) \leq \sigma \Rightarrow supp(\mathtt{II}') \leq \sigma.$$

Démonstration. En effet, si $\mathtt{T}$ est une transaction telle que $\mathtt{II}' \subset \mathtt{T}$, alors évidemment $\mathtt{I} \subset \mathtt{T}$. Ainsi, on a l'inclusion ensembliste suivante : $\{\mathtt{T} \in T | \mathtt{II}' \subset \mathtt{T}\} \subset \{\mathtt{T} \in T | \mathtt{I} \subset \mathtt{T}\}$, ce qui implique l'inégalité entre les cardinaux $\#\{\mathtt{T} \in T | \mathtt{II}' \subset \mathtt{T}\} \leq \#\{\mathtt{T} \in T | \mathtt{I} \subset \mathtt{T}\}$, ce que l'on peut traduire en terme de **support** par $supp(\mathtt{II}') \leq supp(\mathtt{I})$. Ainsi, si $supp(\mathtt{I}) \leq \sigma$, alors nécessairement $supp(\mathtt{II}') \leq \sigma$. □

Cette propriété est d'une importance capitale dans le cadre de la recherche de motifs fréquents, car elle nous dit que si l'on parcourt le treillis des motifs du bas (le motif vide) vers le haut (l'ensemble complet des attributs), alors dès que l'on rencontre un motif non fréquent, il est inutile de calculer le **support** des motifs situés au dessus. Cette heuristique va nous permettre d'orienter et de diriger la recherche des motifs fréquents. Cette propriété se lit bien sur les treillis proposés précédemment. Prenons l'exemple de la figure 1.2, résumé sur la figure 1.3 avec un seuil σ fixé à 2 : dans une stratégie niveau par niveau, pour le niveau 1, les 4 motifs `A`, `B`, `C` et `D` seront testés et l'on découvrira que `B` n'est pas fréquent. Au niveau 2, il est donc inutile de tester les motifs `AB`, `BC` et `BD`, dont on sait, grâce à l'anti-monotonie du **support**, qu'ils n'auront pas un **support** supérieur à 2. Les seuls candidats à ce niveau sont donc `AC`, `AD` et `CD` dont on découvre qu'ils ont un **support** de 2, donc sont fréquents. On passe ainsi au niveau 3 : inutile de tester `ABC`, `ABD` ni `BCD` puisque, entre autres, ni `AB` ni `BD` ne sont fréquents. Il ne reste qu'à tester `ACD`, seul motif dont tous les sous-motifs sont fréquents. Il se trouve que `ACD` a un **support** de 1 et n'est pas fréquent. Ainsi, aucun motif du niveau 3 n'est fréquent et par conséquent, l'unique motif de niveau 4, `ABCD`, non plus.

Si l'on considère que l'action d'évaluer le **support** est coûteuse (il faut en effet ici parcourir la base pour tester l'appartenance du motif à chaque transaction), on se rend compte que l'heuristique nous permet de ne calculer que 8 valeurs de **support** sur les 15 possibles (sans compter le **support** du motif vide). Sans pour autant rien changer à la complexité de la tâche, le travail a été divisé par deux sur cet exemple. Cet algorithme, qui s'appuie sur la propriété d'anti-monotonie

du **support** a été nommé APRIORI dans [Agrawal et Srikant 94]. C'est le premier algorithme efficace de recherche de motifs fréquents, s'appuyant sur une heuristique d'élagage et une génération de candidats efficaces. Il marque le véritable commencement de la Recherche dans le domaine des motifs fréquents et des règles d'association. Il est aux fondements d'un grand nombre d'algorithmes appelés *algorithmes de type* APRIORI dont on pourrait résumer la structure générale par l'algorithme 1.

```
   Données : Base de données 𝒟ℬ, seuil σ
   Sorties : L'ensemble des motifs fréquents ℱ
 1 ℱ ← ∅;
 2 Générer tous les motifs de taille 1;
 3 Mettre les motifs fréquents dans C_1;
 4 Garder les motifs fréquents dans ℱ;
 5 tant que C_l ≠ ∅ faire
 6 |   C_{l+1} ← Générer le niveau l + 1 à partir de C_l ;
 7 |   C_{l+1} ← Vérifier l'hérédité entre C_{l+1} et C_l;
 8 |   C_{l+1} ← Vérifier les valeurs de support dans C_{l+1};
 9 |   Garder les motifs fréquents de C_{l+1} dans ℱ;
10 fin
11 retourner ℱ
```

ALGORITHME 1: Algorithme de type APRIORI

La ligne 7 contient la phase d'élagage : c'est là qu'est mise en œuvre la propriété d'anti-monotonie du **support**, puisque l'on vérifie l'hérédité entre les niveaux l et $l+1$, c'est-à-dire que tous les éléments de C_{l+1} ont tous leurs sous-motifs dans C_l. Dans le cas contraire, les motifs concernés seront supprimés car ils ont au moins un sous-motif non fréquent, et ne sont par conséquent pas fréquents. Nous nommons ici *pré-candidats* l'ensemble obtenu après l'exécution de la ligne 6 et *candidats* l'ensemble obtenu après la ligne 7. La ligne 8 renvoie, quant à elle, les motifs fréquents de taille $l + 1$.

Les différences entre les implémentations se jouent donc sur les sous-modules permettant la génération du niveau suivant et la vérification des valeurs de **support**. Nous verrons par la suite que la vérification des valeurs de **support** peut être réalisée d'une manière plus efficace qu'en parcourant à chaque fois la base de données. En effet cette dernière méthode impose un grand nombre de lectures de la base alors que certains algorithmes proposent des solutions réalisant seulement deux passages sur la base de données.

1.2.2.2 Génération des candidats

La génération du niveau $l+1$ (ligne 6 de l'algorithme 1) peut être réalisée naïvement, en prenant chaque paire de motifs fréquents du niveau l $(\mathtt{I}_1, \mathtt{I}_2)$, en réalisant l'union $\mathtt{I}_1\mathtt{I}_2$ et en vérifiant si le résultat obtenu est de taille $l + 1$. Si c'est le cas, on obtient un pré-candidat au niveau $l + 1$, sinon, on passe à la paire suivante. Il faut donc réaliser $\binom{|C_l|}{2}$ tests pour calculer le tout. Il faudra ensuite vérifier que tous les sous-motifs de chaque motif obtenu sont bien dans C_l pour obtenir les candidats. Cette méthode présente le désavantage de générer plusieurs fois les mêmes motifs.

On peut alors faire appel à une méthode définissant des classes d'équivalences sur les motifs. Cette méthode est proposée dans [Agrawal et Srikant 94], sans notion de classe d'équivalence et est développée par exemple dans [Zaki et al. 97]. Considérons un ordre quelconque sur les attributs de la base de données, on impose que les motifs soient triés en fonction de cet ordre (on pourra, par

exemple, prendre naturellement l'ordre lexicographique). Soit I un motif de taille $l-1$, on définit la classe d'équivalence de I sur C_l, notée [I], par :

$$[\mathtt{I}] = \{\mathtt{I}_l \in C_l | \mathtt{I} = \mathtt{I}_l[1 \dots l-1]\}$$

où $\mathtt{I}_l[1 \dots l-1]$ représente le motif constitué des $l-1$ premiers attributs de $\mathtt{I}_l$ au sens de l'ordre défini.La classe d'équivalence [I] contient tous les motifs de taille l dont I est un préfixe. On génère donc les motifs de taille $l+1$ ayant I pour préfixe à partir, uniquement, des motifs de la classe d'équivalence de I. En effet, si un motif IAB de taille $l+1$ est un bon candidat, c'est-à-dire s'il résiste à la vérification de l'hérédité, alors en particulier, $\mathtt{IA} \in C_l$ et $\mathtt{IB} \in C_l$, c'est-à-dire étaient fréquents. On ne manque donc pas d'information et cette méthode évite des tests inutiles. Elle est parfaitement résumée sous la forme d'une requête SQL dans [Agrawal and Srikant 94] :

insert into C_{l+1}
select $p.item_1, p.item_2, \dots, p.item_l, q.item_l$
from C_l p, C_l q
where $p.item_1 = q.item_1, \dots, p.item_{l-1} = q.item_{l-1}, p.item_l < q.item_l$;

Cette version de la génération de pré-candidats est efficace et demande moins de tests que la version naïve. C'est celle qui est implémentée dans l'APRIORI classique, mais nous allons aussi voir que la notion de classe d'équivalence d'un motif peut aussi servir à générer les motifs fréquents d'une manière bien différente.

1.2.3 Base de données verticales : l'exemple de ECLAT

L'algorithme ECLAT est introduit dans [Zaki et al. 97]. A l'inverse de APRIORI, ECLAT n'utilise pas explicitement la propriété d'anti-monotonie du **support**. Il s'appuie essentiellement sur la notion de base conditionnelle que nous allons développer et sur une représentation verticale des données, en ce sens que les lignes représentent les attributs et que l'on trouve dans chaque ligne l'ensemble des TID des transactions contenant l'attribut en question. Pour un attribut donné A, nous appellerons cet ensemble de transactions la *couverture* de A. L'avantage principal d'une telle représentation est de faciliter le calcul du **support** d'un motif I : il suffit en effet de calculer le cardinal de l'intersection des couvertures des attributs contenus dans I.

$$supp(I) = \#\bigcap_{\mathtt{A} \in \mathtt{I}} couv(\mathtt{A})$$

1.2.3.1 Notion de base conditionnelle

Ici encore, nous allons présupposer l'existence d'un ordre $\prec$ sur les attributs (par exemple, l'ordre lexicographique). Donnons-nous une base de données $\mathcal{DB}$ caractérisée par un ensemble d'attributs A, un ensemble de transactions T et une relation binaire $\mathcal{R}$ sur $T \times A$. D'un point de vue horizontal, la base de donnée $\mathcal{DB}$ conditionnée par l'attribut A est formée par les transactions contenant A et dont on retire tous les attributs plus petits que A, ainsi que A lui-même. D'un point de vue vertical, les lignes sont les attributs plus grands que A et chaque ligne représentée par un attribut B contient l'ensemble des transactions couvrant à la fois A et B. La figure 1.4 met en parallèle les deux cas.

Dans le cas vertical, comme nous l'avons déjà dit, la base permet de calculer les valeurs de **support** rapidement. La base conditionnée par A (Table 1.4(d)) permet de retrouver facilement les valeurs de **support** des motifs commençant par A : le **support** de AB est 0, celui de AC et AD est de 2. Les motifs commençant par B seront construits à partir de la base conditionnée par B, qui est indépendante de la base conditionnée par A.

TID	attributs
1	A C
2	B C D
3	A C D
4	C
5	A D

(a) base horizontale

attr	transactions
A	1 3 5
B	2
C	1 2 3 4
D	2 3 5

(b) base verticale

TID	attributs
1	C
3	C D
5	D

(c) base horizontale conditionnée par A

attr	transactions
B	
C	1 3
D	3 5

(d) base verticale conditionnée par A

FIGURE 1.4 : Représentation horizontale et verticale d'une même base et leur conditionnement par rapport à A.

D'une manière générale, on peut définir la base conditionnée par un itemset I. Nous la noterons $\mathcal{DB}^{\mathtt{I}}$, c'est la base formée par les transactions qui contiennent I et dont on a retiré tous les attributs plus petits (au sens de $\prec$) qu'un attribut de I. D'un point de vue vertical, c'est la base formée par les attributs plus grands que tous les attributs de I et contenant pour chaque attribut A les transactions qui contiennent à la fois I et A.

Tentons de formaliser le concept de base conditionnée. Soit donc $\mathcal{DB}$ une base de données définie par le triplet $(A, T, \mathcal{R})$. Soit un motif $\mathtt{I} \subset A$, on définit la base de données $\mathcal{DB}^{\mathtt{I}}$ grâce au triplet $(A^{\mathtt{I}}, T^{\mathtt{I}}, \mathcal{R}^{\mathtt{I}})$ défini lui même de la manière suivante :

Définition 4 – Base conditionnée – :

$$A^{\mathtt{I}} = \{\mathtt{B} \in A | \forall \mathtt{A} \in \mathtt{I},\ \mathtt{A} \prec \mathtt{B}\}$$

$$T^{\mathtt{I}} = \{\mathtt{T} \in T | \forall \mathtt{A} \in \mathtt{I},\ \mathtt{A}\mathcal{R}\mathtt{T}\}$$

$$\mathcal{R}^{\mathtt{I}} = \mathcal{R}_{|T^{\mathtt{I}} \times A^{\mathtt{I}}}$$

Nous l'avons compris, l'avantage d'une base conditionnée par un item I est de pouvoir donner directement tous les motifs fréquents commençant par I, selon un seuil σ, de manière efficace. Le dernier obstacle est donc de construire ces bases conditionnées, mais cet obstacle peut être surmonté par la définition incrémentale des bases conditionnées. On peut aisément vérifier que les composants de la base conditionnée vérifient les relations :

$$A^{\mathtt{IA}} = \{\mathtt{B} \in A^{\mathtt{I}} | \mathtt{A} \prec \mathtt{B}\}$$

$$T^{\mathtt{IA}} = \{\mathtt{T} \in T^{\mathtt{I}} | \mathtt{A}\mathcal{R}^{I}\mathtt{T}\}$$

$$\mathcal{R}^{\mathtt{IA}} = \mathcal{R}^{\mathtt{I}}_{|T^{\mathtt{IA}} \times A^{\mathtt{IA}}}$$

où I est un motif et A un attribut. Ainsi la construction des bases conditionnées est-elle rendue plus simple. Nous allons voir que c'est là la base de l'algorithme ECLAT.

1.2.3.2 Conditionnement par un seuil

Comme la propriété d'anti-monotonie du **support** peut le faire pressentir, plus l'on conditionnera une base, moins il y aura d'attributs possédant une couverture non nulle. Dans l'exemple

précédent (table 1.4(d)), l'attribut B n'est associé à aucune transaction. Il n'est donc pas pertinent de rechercher, dans ce cas, la base conditionnée par AB, dont on sait qu'elle sera vide. Plus généralement, l'algorithme Eclat propose de fixer un seuil de **support** σ, permettant un élagage non pas des attributs de couverture nulle, mais des attributs dont la couverture aurait une cardinalité inférieure au seuil. On introduit donc la base $\mathcal{DB}^{\mathtt{I}}_{\sigma}$ conditionnée par un motif et un seuil de **support** σ, définie incrémentalement de la manière suivante, où $couv_{\mathtt{I},\sigma}(\mathtt{A})$ représente la couverture de l'attribut A dans $\mathcal{DB}^{\mathtt{I}}_{\sigma}$:

$$A^{\mathtt{IA}}_{\sigma} = \{\mathtt{B} \in A^{\mathtt{I}}_{\sigma} | \mathtt{A} \prec \mathtt{B} \ \& \ |couv_{\mathtt{I},\sigma}(\mathtt{A}) \cap couv_{\mathtt{I},\sigma}(\mathtt{B})| \geq \sigma\}$$

$$T^{\mathtt{IA}}_{\sigma} = \{\mathtt{T} \in T^{\mathtt{I}}_{\sigma} | \mathtt{A}\mathcal{R}^{\mathtt{I}}_{\sigma}\mathtt{T}\}$$

$$\mathcal{R}^{\mathtt{IA}}_{\sigma} = \mathcal{R}^{\mathtt{I}}_{\sigma|T^{\mathtt{IA}}_{\sigma} \times A^{\mathtt{IA}}_{\sigma}}$$

L'initialisation est faite avec $\mathcal{DB}^{\emptyset}_{\sigma} = \mathcal{DB}$.

Ces définitions ouvrent la porte à l'écriture d'un algorithme récursif s'appuyant sur la représentation verticale des données. L'arrêt de l'algorithme est naturellement assuré grâce à l'ordre sur les attributs et la complétude est assurée par la propriété d'anti-monotonie du **support**. Nous présentons le résultat sur l'algorithme 2 extrait de [Goethals 05].

Données : Base de données $\mathcal{DB}^{\mathtt{I}}_{\sigma}$, seuil σ, motif I
Sorties : L'ensemble des motifs fréquents $\mathcal{F}^{\mathtt{I}}$

1 $\mathcal{F}^{\mathtt{I}} \leftarrow \emptyset$;
2 **pour chaque** *A dans* A^{I}_{σ} **faire**
3 $\quad \mathcal{F}^{\mathtt{I}} \leftarrow \mathcal{F}^{\mathtt{I}} \cup \{\mathtt{IA}\}$;
4 $\quad A^{\mathtt{IA}}_{\sigma} \leftarrow \emptyset$;
5 $\quad T^{\mathtt{IA}}_{\sigma} \leftarrow \emptyset$;
6 $\quad$ **pour chaque** *B dans* A^{I}_{σ} *tel que* $\mathtt{A} \prec \mathtt{B}$ **faire**
7 $\quad\quad C \leftarrow couv_{\mathtt{I},\sigma}(\mathtt{A}) \cap couv_{\mathtt{I},\sigma}(\mathtt{B})$;
8 $\quad\quad$ **si** $|C| \geq \sigma$ **alors**
9 $\quad\quad\quad A^{\mathtt{IA}}_{\sigma} \leftarrow A^{\mathtt{IA}}_{\sigma} \cup \{\mathtt{B}\}$;
10 $\quad\quad\quad T^{\mathtt{IA}}_{\sigma} \leftarrow T^{\mathtt{IA}}_{\sigma} \cup C$;
11 $\quad\quad$ **fin**
12 $\quad$ **fin**
13 $\quad$ Construire $\mathcal{F}^{\mathtt{IA}}$ récursivement avec $\mathcal{DB}^{\mathtt{IA}}_{\sigma} = (A^{\mathtt{IA}}_{\sigma}, T^{\mathtt{IA}}_{\sigma}, \mathcal{R}^{\mathtt{I}}_{\sigma|T^{\mathtt{IA}}_{\sigma} \times A^{\mathtt{IA}}_{\sigma}})$;
14 $\quad \mathcal{F}^{\mathtt{I}} \leftarrow \mathcal{F}^{\mathtt{I}} \cup \mathcal{F}^{\mathtt{IA}}$;
15 **fin**
16 **retourner** $\mathcal{F}^{I}$

Algorithme 2: Algorithme Eclat

Dans une version évoluée [Zaki 00], Eclat est utilisé en tirant bénéfice de l'indépendance des bases conditionnées pour réaliser les calculs sur des processeurs en parallèle. Par rapport à Apriori, Eclat dispose d'une facilité de calcul des valeurs de **support**. Cependant, l'hérédité n'est jamais vérifiée et le test de la ligne 8 qui correspond au calcul de **support** est réalisé plus souvent que ne l'est le test de fréquence dans l'algorithme Apriori. Pourtant, les résultats expérimentaux montrent un grand avantage en termes de temps de calcul pour l'algorithme Eclat [Bodon 06]. Il est de plus remarquable que, si l'algorithme Apriori parcourt le treillis des motifs en largeur d'abord, l'algorithme Eclat quant à lui offre un parcours en profondeur d'abord. C'est pourquoi l'hérédité ne peut en aucun cas être vérifiée : en effet, au moment de la création d'un motif de taille l, tous les motifs de taille $l-1$ ne sont pas connus.

Une question reste cependant légitime, concernant l'algorithme Eclat. En effet, nous avons jusqu'à présent supposé l'existence d'un ordre sur les attributs, en suggérant l'ordre lexicographique...

TID	attributs
1	f a c d g i m p
2	a b c f l m o
3	b f h j o
4	b c k s p
5	a f c e l p m n

TABLE 1.5 : base de données exemple pour FP-GROWTH

Mais les résultats sont-ils différents en fonction de l'ordre que l'on choisit ? Nous allons voir dans la section suivante le cas d'un algorithme tirant avantageusement parti des deux représentations, verticales et horizontales, ainsi que de la définition d'un bon ordre sur les attributs.

1.2.4 Mixer les deux : l'algorithme FP-GROWTH

L'algorithme FP-GROWTH a été introduit en 2000 par [Han et al. 00] et propose une véritable avancée dans la recherche de motifs fréquents. Il combine vision horizontale et vision verticale et s'appuie sur la structure de données FP-Tree des arbres préfixes pour stocker efficacement la base de données. Son principe algorithmique est le même que celui de ECLAT, mais la structure de donnée choisie, ainsi que l'ordre utilisé pour les attributs permettent une implémentation plus efficace. Nous allons, dans un premier temps, décrire cette structure de données puis nous verrons les détails de l'algorithme. Dans cette section, nous utiliserons l'exemple jouet original [Han et al. 00] et un **support** minimum de 3 pour illustrer la construction d'un FP-Tree et le déroulement de FP-GROWTH. Cet exemple est donné dans le tableau 1.5.

1.2.4.1 Structure de données FP-Tree

Un FP-Tree est une structure de données sous forme d'arbre préfixe codant les transactions (à partir d'une vue horizontale) tout en conservant une information sur les valeurs de **support**. On commence par définir un ordre sur les attributs en parcourant une première fois la base de données pour calculer le **support** de chaque attribut. On classe ensuite les attributs fréquents (les attributs non fréquents sont inutiles en raison de la propriété d'anti-monotonie du **support**) de chaque transaction par ordre croissant de **support** (ici $f \prec c \prec a \prec b \prec m \prec p$, deux motifs ayant le même **support** sont donnés dans un ordre arbitraire) et l'on construit une liste appelée table de liens, qui est la liste triée par ordre décroissant de **support** des attributs, auxquels sont associés leur **support** et un pointeur. Ce pointeur est initialement vide, mais pointera ensuite vers la première occurence de l'attribut dans l'arbre. L'apparition de l'attribut dans la table de liens représente symboliquement l'occurence 0 de l'attribut dans l'arbre. On remplit ensuite la structure d'arbre. Chaque nœud de l'arbre possède trois valeurs :

- un attribut ;
- un entier représentant une valeur de **support** ;
- un lien vers la prochaine occurence de l'attribut dans l'arbre.

On insère récursivement un motif `AI` dans l'arbre selon la procédure décrite dans l'algorithme 3. Cette procédure est suivie pour chaque transaction de la base de données et l'on obtient pour la base exemple 1.1 l'arbre de la figure 1.5. Les liens qui partent de la table de liens permettent une vision *verticale* de la base en autorisant l'accès, pour chaque attribut, à l'ensemble des transactions qui le contiennent. Les auteurs montrent que cette structure est compacte et complète au sens des motifs fréquents en ce sens que l'on a accès à toutes les informations concernant la fréquence des motifs et que la taille de l'arbre est nécessairement bornée par la taille de la base de données initiale. Nous avons donc un gain de place sans perte d'information pour la tâche qui nous intéresse.

Données : F arbre (on débute avec la racine, d'attribut $\emptyset$), `AI` motif commençant par `A`
F.support $\leftarrow$ F.support+1;
si *`AI` n'est pas le motif vide* **alors**
  si *F a un fils f d'attribut `A`* **alors**
    insérer récursivement `I` dans f
  sinon
    ajouter $f = (\mathtt{A}, 0, null)$ aux fils de F;
    relier la dernière occurence de `A` à f;
    insérer récursivement `I` dans f;
  fin
fin

ALGORITHME 3: Construction d'un FP-tree.

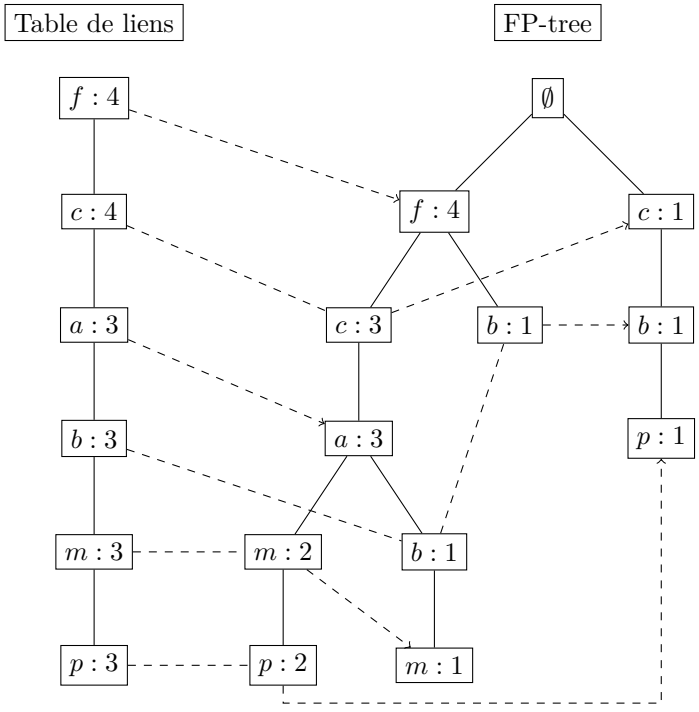


FIGURE 1.5 : FP-tree et table de liens sur un exemple.

(a) base élaguée et triée

TID	attributs
1	p m a c f
2	m b a c f
3	b f
4	p b c
5	p m a c f

(b) base conditionnée par m

TID	attributs
1	a c f
2	b a c f
5	a c f

TABLE 1.6 : Conditionnement d'une base de données cohérent avec le résultat de FP-GROWTH.

La structure de données étant bien définie nous allons maintenant nous intéresser à l'algorithme en lui même.

1.2.4.2 Arbres conditionnés

Comme annoncé précédemment, l'algorithme FP-GROWTH utilise le même principe que l'algorithme ECLAT en s'appuyant sur un ordre bien défini et sur une bonne structure de données. Considérons le FP-tree de la figure 1.5 et intéressons nous plus particulièrement au cas de l'attribut m. L'information fournie par l'arbre sur l'attribut m est multiple : m apparait 3 fois dans la base, mais on sait aussi qu'il apparait deux fois dans une transaction contenant les attributs f, c et a et une fois dans une transaction contenant f, c, a et b. Pour construire la base conditionnée par m, il nous faut l'ensemble des transactions contenant m dont on aurait supprimé les motifs *plus petits* selon l'ordre donné. Considérons cette fois l'ordre croissant des valeurs de **`support`**. Si m apparait dans un nœud, alors, d'après la construction de l'arbre, seuls les attributs ayant une valeur de **`support`** plus faible peuvent se trouver plus profondément dans la même branche. Ainsi, en remontant les branches de l'arbre à partir des nœuds m, on récupère l'ensemble de ces transactions conditionnées. On leur adjoint un ordre de multiplicité égal au **`support`** enregistré dans le nœud m : pour cet attribut, on obtiendra donc la branche $< f : 2, c : 2, a : 2 >$ et la branche $< f : 1, c : 1, a : 1, b : 1 >$ signifiant que, dans la base conditionnée par m, la transaction a, c, f apparait deux fois et la transaction b, a, c, f apparait une fois. En effet, réalisons la tâche de conditionnement sur la base exemple, après avoir élagué et trié les attributs, et observons le résultat sur les tableaux 1.6. On obtient bien deux fois la transaction a, c, f et une fois la transaction b, a, c, f. Ainsi, la structure d'FP-tree permet de calculer rapidement la base conditionnée par m. On applique alors récursivement l'algorithme FP-GROWTH sur la base conditionnée obtenue, de la même manière que pour ECLAT. Le principal avantage par rapport à ECLAT est que l'opération d'intersection est inutile, ce qui enlève une opération ensembliste non négligeable. De plus, les transactions sont dépersonnalisées pour opérer par classes d'équivalences : deux transactions couvrant les mêmes attributs sont identifiées (elles sont sur la même branche), en retenant cependant le cardinal de la classe d'équivalence. Finalement, la figure 1.6 montre l'arbre obtenu par conditionnement suivant m. Par récursions successives, on trouvera finalement que les motifs fréquents débutant par m sont mf, mc, ma, mac, maf, mcf et $macf$, tous avec un **`support`** de 3.

Ainsi, l'algorithme FP-GROWTH procèdera à la recherche des motifs fréquents de la même manière qu'ECLAT, en profondeur d'abord, commençant par les motifs dont le premier attribut est le moins fréquent de la base, puis en remontant vers les attributs les plus fréquents. L'importance de cet ordre se trouve dans le temps de calcul final. En effet, la structure d'arbre privilégiant le positionnement des attributs les plus fréquents au plus près de la racine, on est assuré d'avoir un arbre de taille minimale. Ensuite, commencer par la recherche des attributs les plus rares assure une profondeur de récursion minimale. L'algorithme FP-GROWTH optimise donc l'algorithme ECLAT sur bien des points.

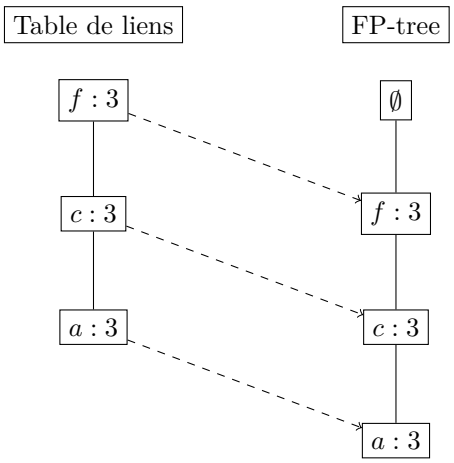


FIGURE 1.6 : FP-tree conditionné par m et table de liens.

Cependant, l'algorithme FP-GROWTH souffre toujours de défauts de complexité et de qualité de résultats. En effet, avec le temps, on s'est aperçu qu'en ne s'intéressant qu'aux motifs fréquents, on pouvait perdre de l'information. De nombreux travaux ont donc été menés, d'un point de vue technique, pour pouvoir repousser le seuil de **support** au plus bas, mais aussi d'un point de vue algorithmique, pour proposer des motifs particuliers permettant de réduire l'espace de recherche. Nous allons détailler dans la section suivante quelques travaux qui ont permis de faire un pas en avant dans la recherche de motifs.

1.3 AMÉLIORATIONS DANS LA RECHERCHE DE MOTIFS

Pour améliorer la recherche de motifs, on peut s'intéresser à plusieurs points bien distincts : proposer des améliorations techniques, comme l'usage de structures de données spécifiques, ou alors la parallélisation de certains processus ; essayer d'améliorer la qualité des résultats, en essayant de trouver des motifs non fréquents (ce qui peut être fait en abaissant les seuils, mais aussi en s'intéressant aux motifs non découverts par APRIORI) ; réduire l'espace de recherche en se focalisant sur une famille de motifs biens précis. Nous allons ici détailler quelques travaux s'appuyant sur ces idées.

1.3.1 Améliorations techniques

L'un des principaux problèmes de la recherche de motifs se situe dans la quantité des motifs générés, mais surtout la quantité de motifs parcourus. Cela est principalement dû à la taille toujours plus importante des bases de données, qui parfois ne tiennent même pas en mémoire et posent donc problèmes aux algorithmes APRIORI ou ECLAT. Pour pallier ce problème, l'algorithme PARTITION est proposé dans [Savasere et al. 95]. Cet algorithme propose de découper la base de donnée, trop grande pour tenir en mémoire, en sous-bases de données d'intersection vide et calibrées pour tenir en mémoire. Si l'on note $\mathcal{DB}$ la base de données originale et $(\mathcal{DB}^1, \ldots, \mathcal{DB}^p)$ la partition effectuée et que l'on considère un seuil de **support** σ, alors on définit pour chaque sous-base $\mathcal{DB}^i$ le seuil $\sigma_i = \frac{\sigma}{|\mathcal{DB}|} * |\mathcal{DB}^i|$. Pour qu'un motif soit fréquent par rapport à σ dans $\mathcal{DB}$, il faut qu'il soit fréquent par rapport à σ_i dans au moins une base $|\mathcal{DB}^i|$. On obtient par cette procédure un ensemble de candidats, fréquents dans au moins une sous-base, dont il faudra vérifier le **support** dans la base complète. L'avantage est donc de pouvoir traiter la base de données même si celle-ci ne tient

pas en mémoire et de pouvoir traiter les sous-bases en parallèle. Le résultat d'un tel algorithme sera évidemment complet. Ce n'est pas le cas d'une autre approche, le sampling [Toivonen 96], qui s'appuie sur les probabilités pour générer les motifs fréquents. Le principe est de tirer aléatoirement un sous-ensemble de l'ensemble des transactions et d'appliquer un algorithme de recherche de motifs fréquents, en adaptant le seuil. Plus le sous ensemble tiré aléatoirement est grand, plus les résultats sont proches de la réalité. Cependant, il n'est pas exclut de découvrir des motifs qui ne sont pas vraiment fréquents, ou bien de ne pas découvrir des motifs qui étaient pourtant fréquents.

Une autre façon de faire des progrès est d'améliorer au mieux les implémentations, comme nous l'avons vu pour FP-Growth, qui améliore Eclat en changeant de structure de données. On peut utiliser cette même structure de données d'arbres préfixes dans une implémentation classique de Apriori. Elle servira au moment du calcul des valeurs de **support** et évitera en particulier un passage sur la base complète à chaque recherche de **support**. C'est l'un des choix qui a été fait dans l'implémentation réalisée dans [Borgelt 03] et que l'on peut retrouver sur le site de l'auteur Christian Borgelt [2]. Cette implémentation est l'une des plus efficaces que nous ayons trouvée, mais n'est malheureusement que difficilement adaptable à cause de la complexité de son code, ce qui nous a encouragé à développer nos propres implémentations. Elle sert cependant de référence grâce à l'accessibilité et à la simplicité de paramétrisation du programme. Elle permet notamment d'utiliser ou non les arbres préfixes dans Apriori et d'utiliser librement Apriori ou FP-Growth.

1.3.2 Réduire l'espace de recherche

Afin d'accélérer la recherche des motifs fréquents, l'une des idées très développées dans la littérature consiste à se concentrer sur la découverte de motifs qui pourraient servir de base pour la génération des motifs fréquents. Il faut évidemment, pour que cela soit intéressant, que cet ensemble de motifs générateurs permette non seulement de générer tous les motifs fréquents, mais aussi leur **support**. C'est dans ce but qu'ont été introduits les motifs fermés [Pasquier et al. 99]. Les motifs fermés s'appuient sur une vision galoisienne du treillis et des opérateurs de fermeture. Au final, la clôture d'un motif est son plus grand surmotif qui possède le même **support**. Un motif est fermé s'il est égal à sa clôture. On peut aussi définir la clôture d'un point de vue ensembliste puisque la clôture d'un motif est en fait l'intersection de toutes les transactions contenant ce motif, c'est-à-dire le plus grand motif commun à ces transactions. La figure 1.7 montre les motifs fermés pour notre exemple jouet. On se rend compte que les motifs peuvent être fermés et fréquents (bord vert, fond vert), fermés et non fréquent (bord vert, fond rouge), non-fermés et fréquents (bord rouge, fond vert) ou bien non-fermés et non-fréquents (bord rouge, fond rouge). La tâche que l'on considère est la recherche de motifs fermés fréquents, dont les auteurs ont montré qu'ils permettent de retrouver tous les motifs fréquents, ainsi que leur **support**. L'algorithme Close s'appuie sur une propriété d'anti-monotonie des motifs fermés fréquents pour permettre de manière efficace leur découverte : soit $h(\mathtt{I})$ la clôture d'un motif $\mathtt{I}$, alors

$$\mathtt{I} \subset \mathtt{I}' \subset h(\mathtt{I}) \Rightarrow h(\mathtt{I}') = h(\mathtt{I}).$$

Lors d'un parcours du bas vers le haut, cette relation permet de ne pas calculer la clotûre de certains motifs. L'algorithme s'appuie cependant sur une phase supplémentaire pour générer effectivement tous les motifs fréquents et leur **support**. On le comprend bien, cette propriété de fermeture s'appuie toujours sur le **support**. Nous allons donc énoncer une autre représentation des motifs s'appuyant cette fois sur un calcul de **support** différent.

Notons qu'une transaction contient le motif $\mathtt{I}\neg\mathtt{A}$ si elle contient effectivement $\mathtt{I}$ mais ne contient pas $\mathtt{A}$. Si δ est un entier, on définit les motifs δ-*libres* de la manière suivante [Boulicaut et al. 03] :

2. http://www.borgelt.net/fpm.html

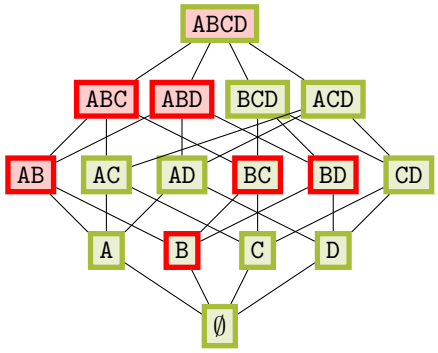


FIGURE 1.7 : Exemple de treillis de motifs indiquant les motifs fermés (bord vert) et les motifs fréquents (fond vert)

Définition 5 – motif δ-libre – : $\mathtt{I}$ est un motif δ-libre s'il n'existe pas d'attribut $\mathtt{A}$ dans $\mathtt{I}$ ni de sous-motif $\mathtt{I}' \subset \mathtt{I} - \{\mathtt{A}\}$ tels que $supp(\mathtt{I}' \neg \mathtt{A}) \leq \delta$. C'est-à-dire si l'on ne peut dériver de $\mathtt{I}$ aucune règle $\mathtt{I}' \rightarrow \mathtt{A}$ qui soit contredite moins de δ fois.

Cette propriété de δ-liberté a le grand avantage, encore une fois, d'être anti-monotone : si $\mathtt{I}$ est un motif δ-libre, alors tout sous-motif de $\mathtt{I}$ est δ-libre. Elle permet donc un élagage efficace du treillis des motifs. De plus, il est possible d'approcher les valeurs de **support** des motifs fréquents non-libres à partir des motifs fréquents δ-libres : cette approximation est exacte seulement si $\delta = 0$. Bien sûr, ce sont les motifs non-libres qui nous intéressent, puisqu'ils permettent de déduire des relations fortes, peu contredites. Cet exemple permet de montrer l'importance que peut avoir l'étude de contre-exemples lors d'une recherche d'information dans les bases de données. Nous reviendrons là dessus dans les parties suivantes.

1.3.3 Changer d'objectif

Une méthode naïve de réduction de l'espace de recherche consiste à augmenter le seuil de **support** pour rendre plus efficace l'élagage. Cette méthode n'a cependant pas que des avantages car elle oblige à se focaliser sur des motifs très fréquents et donc sur des coprésences d'attributs évidentes. On peut par exemple être à la recherche d'associations rares, qui échapperaient à la connaissance humaine. Dans ce but, on pourra se poser la question de l'extraction des motifs rares, le complémentaire des motifs fréquents. Un algorithme de recherche de ces motifs est présenté dans [Szathmary et al. 06, Szathmary et al. 07]. Il faut bien y faire la différence entre motifs rares et motifs de **support** nul, mais les auteurs proposent une représentation de l'ensemble des motifs rares par les motifs rares minimaux (figure 1.8), ce qui permet d'améliorer la qualité selon le critère de la rareté tout en gardant une représentation concise. Il faudra cependant se poser la question du véritable intérêt de tels motifs, dans le sens où une information trop rare peut être due au hasard, ou bien être le complémentaire d'une information fréquente, plus intéressante et correspondre à du bruit.

On peut aussi se demander s'il n'est pas possible d'ajouter de nouvelles contraintes à la recherche, qui viendraient s'ajouter à la contrainte de **support**. On peut évidemment demander à l'utilisateur si certains attributs peuvent être retirés de la recherche grâce aux connaissances du domaine. Mais la notion de contrainte convertible [Pei et al. 01] apporte réellement un plus dans la qualité des résultats, tout en permettant un élagage supplémentaire de l'espace de recherche. Une contrainte est convertible anti-monotone s'il existe un ordre sur les attributs tel que, pour cet ordre, la contrainte soit anti-monotone. Par exemple, la contrainte $moy(I) \geq \sigma$ est convertible

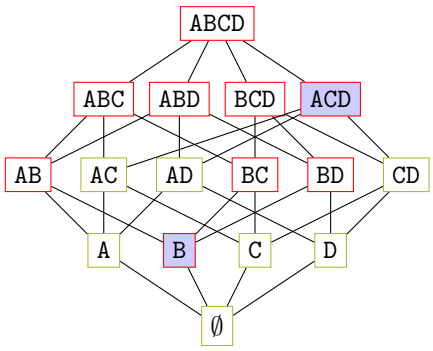


FIGURE 1.8 : Mise en évidence des motifs rares pour un **support** de deux. Les motifs rares apparaissent en rouge, les motifs rares minimaux ont un fond bleu.

anti-monotone, où *moy* correspond à la moyenne de valeurs des attributs (par exemple la moyenne des valeurs de **support** de chaque attribut). L'ordre à choisir est alors l'ordre croissant des valeurs. Les expériences sur l'algorithme de recherche de motifs par contraintes convertibles montrent de bons résultats par rapport même à FP-Growth. Il faut cependant bien se rendre compte que l'ensemble des motifs fréquents n'est pas obtenu et que donc le résultat n'est pas complet en ce sens. L'objectif est différent et ne se borne pas à des motifs fréquents.

Finalement, on peut vouloir aller plus loin que les motifs et vouloir que cette coprésence d'attributs ne soit pas la seule information extraite du motif. Nous allons, dans le prochain chapitre, détailler la notion de règle d'association qui permet de scinder le motif en 2 pour donner un sens, une direction et caractériser ainsi une dépendance entre deux motifs.

CONCLUSION

La découverte de motifs fréquents est un domaine de recherche en pleine évolution, dont les prémisses se retrouvent au début des années 90. Depuis l'algorithme APRIORI, de nombreuses évolutions ont eu lieu, menant vers des implémentations beaucoup plus performantes et des algorithmes toujours plus efficaces. Cependant, on se heurte sans cesse aux mêmes problèmes de quantité et de qualité des résultats. On peut dès lors ajouter des contraintes, se focaliser sur des motifs rares, réduire l'espace de recherche, mais le besoin principal consiste à être capable d'utiliser les connaissances extraites. Ce sera le rôle des règles d'association, qui permettent, comme nous allons le voir, de détecter des liens orientés entre motifs.

2 Règles d'association : extraction et évaluation

Les règles d'association sont un modèle répandu de connaissances au sein d'une base de données. Elles ont été introduites en même temps que l'algorithme Apriori et permettent de caractériser des dépendances présentes dans la base. Comme les motifs fréquents, elles peuvent apparaître en très grand nombre et se pose alors la question de leur évaluation. Nous verrons, après avoir défini formellement la notion de règle, que cette question de l'évaluation, aussi bien pour le classement des règles que pour le filtrage, a fait l'objet de nombreuses recherches et que beaucoup de mesures de l'intérêt d'une règle ont vu le jour, qu'elles fassent appel à l'utilisateur final ou bien qu'elles permettent une évaluation automatique.

2.1 Qu'est ce qu'une règle d'association ?

Une règle d'association s'appuie sur une base de données $\mathcal{DB}$ et un motif $\mathtt{I}$. Dans sa définition originale [Agrawal et al. 93], une règle est définie comme étant un objet $\mathtt{A} \rightarrow_{\mathcal{DB}} \mathtt{C}$ où $\mathtt{C}$ est un attribut, $\mathtt{I} = \mathtt{AC}$ et $\mathtt{C} \notin \mathtt{A}$. D'une manière plus générale, on peut aussi considérer, comme cela a été fait plus tard, qu'une règle d'association peut avoir un ensemble d'attributs dans sa partie droite, en conséquent.

Nous noterons une telle règle $\mathtt{A} \rightarrow_{\mathcal{DB}} \mathtt{B}$ et les conditions deviennent alors : $\mathtt{B}$ est un motif et $\mathtt{A} \cap \mathtt{B} = \emptyset$. Si le nombre de motifs possibles sur une base de données à n attributs est de 2^n, le nombre de règles d'association possibles varie entre $n * 2^{n-1}$ et 3^n suivant que l'on considère la définition originale ou la définition étendue, ce qui rend la recherche exhaustive de règles aussi déraisonnable que celle des motifs. Dans un premier temps on fixe donc un seuil de **support** et l'on extrait l'ensemble des motifs fréquents associés à ce seuil de **support**. Puis l'on recherche, de façon exhaustive cette fois, l'ensemble des règles s'appuyant sur ces motifs : encore une fois, cet ensemble peut-être très important et il faudra filtrer les règles pour ne garder que celles qui peuvent avoir un réel intérêt. Historiquement, c'est la mesure de **confiance** qui est utilisée pour évaluer cet intérêt. La **confiance** d'une règle $\mathtt{A} \rightarrow_{\mathcal{DB}} \mathtt{B}$ est définie comme le pourcentage de transactions qui contiennent $\mathtt{B}$ parmi les transactions qui contiennent $\mathtt{A}$. C'est donc la probabilité conditionnelle de $\mathtt{B}$ sachant $\mathtt{A}$ dans la base $\mathcal{DB}$, c'est-à-dire la probabilité, lorsque $\mathtt{A}$ se trouve dans une transaction, que $\mathtt{B}$ s'y trouve aussi. En fixant un second seuil γ de **confiance**, en plus du seuil σ de **support**, l'algorithme Apriori permet de trouver toutes les règles $\mathtt{A} \rightarrow_{\mathcal{DB}} \mathtt{B}$ fréquentes (i.e. $supp(\mathtt{AB}) > \sigma$) et intéressantes (i.e. $conf(\mathtt{A} \rightarrow_{\mathcal{DB}} \mathtt{B}) > \gamma$) grâce à une phase indépendante de génération des règles. Si $\mathtt{A} \subset \mathtt{A}'$, on dira que la règle $\mathtt{A} \rightarrow \mathtt{B}$ est plus générale que la règle $\mathtt{A}' \rightarrow \mathtt{B}$, qui est, elle, plus spécifique. L'algorithme Apriori part des règles plus générales et procède par niveaux vers les règles plus spécifiques.

L'introduction de la notion de **confiance** permet de mettre en évidence l'un des défauts de la méthode et d'introduire le concept de *pépite de connaissance*. Une pépite de connaissance est une règle d'association dont le **support**, c'est-à-dire le **support** du motif sur lequel elle s'appuie, est

TID	Articles
11	PC, Imprimante, PDA
12	Imprimante, Notebook
13	Imprimante, Scanner
14	PC, Imprimante, Notebook
15	PC, Scanner
16	Imprimante, Scanner
17	PC, Scanner
18	PC, Imprimante, Scanner, PDA
19	PC, Imprimante, Scanner

TABLE 2.1 : Articles achetés par des clients d'un magasin d'électronique

faible, mais dont la **confiance** est élevée. Une telle connaissance est importante puisqu'elle caractérise une règle qui arrive rarement, mais qui, lorsqu'elle arrive, est très fiable : c'est typiquement le genre de règle que l'on peut rechercher en médecine, pour mettre en évidence des symptômes de maladies très rares. Malheureusement, les paramètres de l'algorithme APRIORI ne permettent pas la découverte de ce type de règles, car si la hausse du seuil de **confiance** permet de filtrer d'avantage de règles et donc d'avoir moins de règles en sortie, la baisse du seuil **support** crée une explosion opérationnelle en augmentant considérablement le nombre de résultats et le temps d'éxécution. Les pépites de connaissance sont alors noyées dans l'ensemble des règles, parmi lesquelles une grande partie peuvent être inintéressantes. De plus, la **confiance** traduit bien la coprésence de motifs, mais pas nécessairement la corrélation. Reprenons un exemple issu de [Lin et al. 02]. Dans la base de la table 2.1 on s'intéresse à l'influence de l'achat d'un scanner sur l'achat d'une imprimante, c'est-à-dire à la règle $Scanner \rightarrow Imprimante$. Son **support** est de $44,4\%$ et sa **confiance** de $66,7\%$, qui sont deux valeurs élevées. Pourtant le **support** de l'attribut *Imprimante* seul est de $77,8\%$ et la règle ne fait donc pas augmenter la probabilité d'achats d'imprimantes et ne permet pas d'établir de stratégie de vente.

D'autre part, la **confiance** favorise les règles à fort conséquent, car si le conséquent est très fréquent, la probabilité d'avoir `A` et `B` dans la même transaction est proche de la probabilité d'avoir `A` et la **confiance** est donc proche de 1. Or, comme pourrait le montrer une règle du type *caviar* $\rightarrow$ *pain*, cela n'est pas souhaitable, car ne caractérise pas un comportement spécifique à l'achat de caviar, mais simplement la forte présence de pain.

Ainsi donc, bien qu'historique, l'utilisation du couple **support**/**confiance** ne semble pas entièrement satisfaisante. Nous verrons qu'il est pourtant très utilisé, notamment pour construire des classifieurs. Il existe cependant des alternatives à la **confiance**, sous la forme d'autres mesures, toute la difficulté résidant alors dans la définition de ces mesures et leur choix.

2.2 LES RÈGLES DE CLASSE : UN EXEMPLE D'UTILISATION DES RÈGLES D'ASSOCIATION.

Les règles d'association peuvent par exemple être utilisées dans une tâche de classification. Dans ce but, nous introduisons la notion de règle de classe, une particularisation de la notion de règle d'association. Intéressons nous pour cela à la base de données bien connue *mushroom* issue de l'archive UCI. Cette base contient 22 attributs catégoriels et 8124 transactions. Elle regroupe un ensemble de caractéristiques de champignons, parmi lesquels certains sont vénéneux et d'autres non. Ce caractère vénéneux-comestible fait l'objet d'un vingt-troisième attribut et il est clair que ce que l'on peut attendre d'une tâche de fouille de données sur une telle base est la prédiction du caractère vénéneux d'un champignon en fonction des caractéristiques observées. Dans le cadre

de règles d'association, il s'agirait donc de découvrir des règles au format `I` $\rightarrow$ *comestible* ou `I` $\rightarrow$ *vénéneux*. On sort ainsi du cadre classique des règles d'association puisque l'on s'intéresse exclusivement aux règles ayant cet attribut comestible/vénéneux en conséquent. Cet attribut sera appelé un attribut de classe et les règles obtenues seront nommées, de la même manière, règles de classe. Il existe différentes façons de découvrir les règles de classe : l'algorithme Apriori-C [Jovanoski et Lavrac 01] recherche toutes les règles d'association, puis les filtre pour ne retenir que celles qui ont le bon attribut en sortie ; l'algorithme FCP-Growth [Bahri et Lallich 09] adapte FP-Growth pour ne découvrir que les règles de classe en donnant un poids élevé à l'attribut de classe, ce qui le fait remonter dans l'arbre préfixe et crée autant de sous-arbres disjoints que de classes : cela permet notamment de définir un seuil de mesure particulier pour chaque classe ; enfin, l'algorithme CBA [Liu et al. 98] adapte Apriori en ce concentrant sur les attributs qui ne sont pas des attributs de classe et en associant à chaque motif une liste de paires $(classe, support)$ indiquant la représentation du motif dans chaque classe. Nous allons nous intéresser plus en détail à ce dernier algorithme.

En effet, CBA n'est pas simplement un algorithme de recherche de règles de classe, car il est utilisé ensuite pour construire un classifieur, dans la même veine que C4.5 [Quinlan 93], permettant ainsi de prédire une classe pour de nouvelles transactions. Le principe de CBA est le suivant : on commence par extraire de la base d'apprentissage l'ensemble des règles qui sont à la fois fréquentes et intéressantes au sens de la **`confiance`**. Puis un ordre est défini sur les règles : soient deux règles r_i et r_j, r_i précède r_j si

- r_i a une **`confiance`** plus élevée que r_j ;
- r_i et r_j ont la même valeur de **`confiance`** mais r_i a une plus grande valeur de **`support`** que r_j ;
- r_i et r_j ont même valeur de **`confiance`** et de **`support`**, mais r_i est générée avant r_j.

Cet ordre, total, permet de classer les règles. L'ensemble des règles du classifieur est ensuite construit en prenant le plus petit ensemble de règles, choisies dans l'ordre prédéfini, permettant de couvrir l'ensemble des transactions de la base de données. Finalement, la règle de la classe majoritaire permet de faire de ce classifieur un classifieur capable d'affecter une valeur à toute nouvelle transaction. Encore une fois, dans ce cadre, le **`support`** est omniprésent et les auteurs reconnaissent la grande dépendance des résultats par rapport au seuil choisi. En fixant un seuil inférieur à 2%, les expériences montrent que ce type de classifieur est plus efficace que C4.5 [Quinlan 93] en terme de taux d'erreur. D'autres algorithmes ont été présentés tels que CMAR [Li et al. 01] qui présente trois différences majeures avec CBA. La première se trouve dans la génération des règles puisque CMAR utilise une approche fondée sur FP-Growth, quand CBA se contente d'adapter Apriori. La seconde réside dans la sélection des règles qui formeront le classifieur. Dans CBA, une transaction est retirée de la base d'apprentissage dès qu'elle est couverte par une règle. Dans CMAR, la transaction reste jusqu'à ce qu'elle soit couverte par un nombre paramétrable δ de règles. Finalement, la troisième différence consiste à utiliser le χ^2 pour évaluer la corrélation des règles. On diminue ainsi le biais introduit par l'utilisation de la **`confiance`**.

Il apparait ainsi clairement qu'une grande difficulté de la recherche de règles d'association est d'évaluer les résultats. La **`confiance`**, mesure historique, pose des problèmes de qualité des résultats, mais aussi de quantité. Le **`support`** quant à lui peut aussi empêcher l'apparition de certaines règles pourtant intéressantes comme les pépites de connaissance. Face à ces difficultés, la question des mesures d'intérêt a souvent été soulevée : il s'agit d'être capable d'extraire des règles qui soient intéressantes pour l'utilisateur, sans avoir forcément à monopoliser ce dernier. Dans la section suivante, nous allons revenir sur l'évaluation des règles et notamment d'un point de vue objectif (sans intervention de l'utilisateur).

2.3 Évaluation des règles d'association

Nous nous intéressons ici aux méthodes d'évaluation des règles d'association. Celles-ci peuvent se présenter sous deux aspects : les mesures objectives qui sont de classiques formules mathématiques et ne prennent en argument que les caractéristiques statistiques de la règle ; les méthodes subjectives qui font intervenir les connaissances de l'utilisateur final, notamment des connaissances du domaine. Nous allons dans un premier temps décrire deux exemples d'utilisation de méthodes subjectives, pour nous concentrer par la suite sur les mesures objectives qui sont au cœur de cette thèse.

2.3.1 Mesures subjectives

Il serait bien difficile d'être objectif sur le sujet des méthodes subjectives. On pourra par exemple se référer à [Geng et Hamilton 06] pour obtenir des précisions sur ce domaine. Nous décrivons ici deux travaux proposant l'intégration de connaissances du domaine dans la phase de recherche des règles de manières différentes.

Tout d'abord, dans [Padmanabhan et Tuzhilin 99], on propose une gestion des connaissances sous la forme de croyances, qui sont des règles définies par les connaissances du domaine : pour reprendre l'exemple classique de [Reiter 87], considérons la règle *oiseau*→*vole*. Bien que cette règle puisse être contredite (*oiseau,pingouin*→ ¬*vole*) les auteurs supposent que l'on peut converger vers un système de croyance possédant une propriété de monotonie (ici, il faudrait remplacer ces deux règles par : *oiseau,*¬*pingouin*→*vole* et *oiseau,pingouin*→ ¬*vole*). Le but de ce travail est de découvrir des règles surprenantes dans le sens où elles contredisent des croyances du domaine : si `X` → `Y` est une croyance du domaine, on cherchera des règles de la forme `A` → `B` où `B` et `Y` sont contradictoires alors que `A` et `X` peuvent coexister. L'algorithme proposé par les auteurs consiste en deux phases : une phase de recherche s'appuyant sur les croyances (le treillis est donc largement élagué), puis une phase de généralisation, pendant laquelle toutes les règles surprenantes sont généralisées pour obtenir un ensemble complet. Cette approche s'appuie largement sur le concept de **`support`** et de **`confiance`**. Elle est originale dans le sens où l'on se sert de connaissances du domaine pour rechercher dans la base de données des informations les contredisant.

Inversement, il peut aussi être demandé à l'utilisateur final les connaissances qu'il souhaite mettre en avant. Dans [Klemettinen et al. 94], les auteurs proposent de définir des modèles de règles, inspirés des expressions rationnelles. On définit une hierarchie sur les attributs, par exemple

$$\{pigeon, mouette, pingouin\} \subset oiseaux \subset animaux$$

et l'utilisateur peut être intéressé par les règles de la forme $oiseaux+ \rightarrow vole$ dont on sortira probablement $pigeon \rightarrow vole$ et $mouette \rightarrow vole$. Cette approche s'appuie sur un filtrage *a posteriori* des règles et ne permet pas, contrairement à la méthode précédente, une réduction de l'espace de recherche. Cependant, elle permet la réduction de l'ensemble de règles présenté à l'utilisateur final. En plus de définir des règles à respecter, l'utilisateur peut, de plus, définir des règles restrictives, qu'il ne souhaite pas voir apparaitre. Toutes ces conditions, font que l'ensemble de règles présenté ne contient que des règles que l'utilisateur veut visualiser. Mais il sera difficile d'obtenir des règles surprenantes par exemple, ce qui fait de cette méthode une méthode bien différente de la précédente du point de vue de ses résultats. Cependant, l'on doit toujours s'appuyer sur le couple **`support`**/**`confiance`** pour découvrir initialement les règles.

On pourrait aussi citer [Jaroszewicz et Simovici 04] qui modélise les connaissances du domaine sous la forme d'un réseau bayésien afin de définir une mesure d'intérêt, sur les motifs, comme étant la distance du **`support`** observé au **`support`** attendu par le réseau.

Dans notre travail nous allons essentiellement nous intéresser à des mesures d'intérêt, mais objectives celles-là, c'est-à-dire qu'une fois la mesure choisie, l'utilisateur n'a plus d'influence jusqu'à la présentation des règles obtenues.

2.3.2 Mesures objectives et propriétés

Pour évaluer l'intérêt d'une règle d'association et devant le grand nombre d'informations que l'on peut recueillir, il parait difficile de demander une évaluation règle par règle à un expert. On peut cependant lui demander des tendances sur le comportement de la règle : les règles très fréquentes doivent-elles être favorisées, préfère-t-on avoir peu de faux-positifs, peu de faux-négatifs... En fonction de ces critères, on pourra choisir une mesure objective possédant de bonnes propriétés. Mais voyons dans un premier temps comment une mesure objective est définie.

Une règle d'asssociation est définie par un antécédent et un conséquent. Concernant ses propriétés statistiques au sein de la base, il est possible de les visualiser sous la forme d'une table de contingence. Cette table de contingence peut être considérée relative (point de vue des proportions) ou absolue (point de vue des effectifs), suivant que l'on ramène ses valeurs à l'intervalle $[0, 1]$ ou non. La table 2.2 montre ces deux types de tables de contingence. Dans ces tables, $n_{\mathtt{I}}$ représente

(a) contingence absolue

	$\mathtt{A}$	$\neg\mathtt{A}$	
$\mathtt{B}$	$n_{\mathtt{AB}}$	$n_{\neg\mathtt{AB}}$	$n_{\mathtt{B}}$
$\neg\mathtt{B}$	$n_{\mathtt{A}\neg\mathtt{B}}$	$n_{\neg\mathtt{A}\neg\mathtt{B}}$	$n_{\neg\mathtt{B}}$
	$n_{\mathtt{A}}$	$n_{\neg\mathtt{A}}$	n

(b) contingence relative

	$\mathtt{A}$	$\neg\mathtt{A}$	
$\mathtt{B}$	$p_{\mathtt{AB}}$	$p_{\neg\mathtt{AB}}$	$p_{\mathtt{B}}$
$\neg\mathtt{B}$	$p_{\mathtt{A}\neg\mathtt{B}}$	$p_{\neg\mathtt{A}\neg\mathtt{B}}$	$p_{\neg\mathtt{B}}$
	$p_{\mathtt{A}}$	$p_{\neg\mathtt{A}}$	1

TABLE 2.2 : Tables de contingence

le nombre d'occurences du motif $\mathtt{I}$, c'est ce que nous avons appelé jusqu'à présent **support** et qui est en fait le **support** absolu et $p_{\mathtt{I}}$ représente la probabilité de présence du motif $\mathtt{I}$ dans la base, soit en fait $\frac{n_{\mathtt{I}}}{n}$: c'est le **support** relatif de $\mathtt{I}$.

À titre d'exemple, nous pouvons citer deux mesures d'intérêt déjà connues : le **support** et la **confiance**. Ceci nous permet de préciser un point : si nous avons, pour des raisons de compréhension, considéré jusqu'à présent le **support** absolu, c'est-à-dire $supp(\mathtt{AB}) = n_{\mathtt{AB}}$, par la suite, nous utiliserons indifféremment le **support** absolu ou le **support** relatif, c'est-à-dire $supp(\mathtt{AB}) = p_{\mathtt{AB}}$ et ce sans le préciser lorsqu'il n'y aura pas ambiguité (un **support** entier est un **support** absolu, un **support** rationnel entre 0 et 1 est un **support** relatif) ou indifférence (un **support** à 0 est indifféremment relatif ou absolu). Quant à la **confiance**, elle est indifféremment reliée aux effectifs ou aux proportions : $conf(\mathtt{A} \rightarrow \mathtt{B}) = \frac{n_{\mathtt{AB}}}{n_{\mathtt{A}}} = \frac{p_{\mathtt{AB}}}{p_{\mathtt{A}}}$.

Il existe un grand nombre de mesures et l'on pourrait se dire que n'importe quelle fonction de la table de contingence pourrait convenir. Il est donc nécessaire de définir quelques critères assez naturels de *bonne conduite* pour la création de nouvelles mesures. Nous en avons déjà énoncé un lors de l'introduction de la **confiance**. La règle *caviar* $\rightarrow$ *pain* n'est pas intéressante car elle viserait à prédire un attribut trop courant : on trouve du pain dans une grande part des transactions et l'achat de caviar n'a aucune influence significative sur l'achat de pain. La règle symétrique, *pain* $\rightarrow$ *caviar* n'est pas plus intéressante dans le sens où, de la même manière, l'achat de pain est trop fréquent pour être significatif par rapport à l'achat de caviar. De la même manière, en médecine, on pourrait s'interroger sur l'utilité d'une règle *toux* $\rightarrow$ *maladie rare*. Il apparait donc que si l'antécédent ou le conséquent sont trop fréquents, la règle semble moins pertinente. Une autre considération porte sur la fréquence du motif support de la règle. On aura tendance à préférer des règles qui apparaissent plus souvent. Cela peut sembler en contradiction avec la notion de pépite,

mais en fait il s'agit là de comparaison entre deux règles : pour deux règles extraites, suivant un système particulier, on aura tendance à préférer celle qui apparait plus souvent, sans pour autant établir une notion de seuil.

Ces trois principes ont été agrégés par Piatetsky-Shapiro dans [Piatetsky-Shapiro 91] pour décrire ce que pourrait être une bonne mesure d'intérêt. Une quatrième propriété est ajoutée, qui consiste à être capable de repérer l'indépendance des attributs antécédent et conséquent de la règle. Dans sa définition originale, cette valeur de repère est 0 et les principes prennent la forme suivante [Piatetsky-Shapiro 91] : *soit m une mesure d'intérêt objective des règles d'association, alors*

- $m(\mathtt{A} \rightarrow \mathtt{B}) = 0$ si $p_{\mathtt{A}}p_{\mathtt{B}} = p_{\mathtt{AB}}$, i.e. s'il y a indépendance entre les attributs $\mathtt{A}$ et $\mathtt{B}$;
- m croît avec $p_{\mathtt{AB}}$ lorsque tous les autres paramètres restent les mêmes ;
- m décroît avec $p_{\mathtt{A}}$ (ou $p_{\mathtt{B}}$) lorsque tous les autres paramètres restent les mêmes.

La mesure la plus simple répondant à ces critères et citée par l'auteur à titre d'exemple, est la mesure de **levier** définie par $lev(\mathtt{A} \rightarrow \mathtt{B}) = p_{\mathtt{AB}} - p_{\mathtt{A}}p_{\mathtt{B}}$. Cependant, la valeur de référence à l'indépendance peut être différente de 0, l'important étant de pouvoir la repérer facilement et indépendemment de la règle. Ainsi, la mesure de **lift** [Brin et al. 97a] définie par $lift(\mathtt{A} \rightarrow \mathtt{B}) = \frac{p_{\mathtt{AB}}}{p_{\mathtt{A}}p_{\mathtt{B}}}$ vérifie-t-elle aussi les mêmes critères de variation, mais repère l'indépendance par une valeur de 1.

2

Depuis, beaucoup d'autres critères ont été étudiés et beaucoup de mesures définies. Michael Hahsler propose l'étude de quelques mesures sur son site [1], mais nous reviendrons plus en détail sur l'ensemble des mesures. Concernant les critères, dans [Lenca et al. 08, Vaillant 06], 9 sont par exemple étudiés pour comparer les mesures et les ordonner dans une approche orientée aide à la décision multicritère. Ces critères sont les suivants :

- symétrie de la mesure : $\mathtt{A} \rightarrow \mathtt{B}$ et $\mathtt{B} \rightarrow \mathtt{A}$ sont-elles traitées de la même manière ?
- décroissance avec $n_{\mathtt{B}}$;
- situation à l'indépendance (constante, variable) ;
- situation à l'indétermination,
- situation pour les règles logiques (constante, variable) ;
- linéarité avec $p_{\mathtt{A}\neg\mathtt{B}}$ au voisinage de 0 : comportement par rapport à l'arrivée de contre-exemples ;
- sensibilité par rapport à la taille de la base n ;
- facilité de fixer un seuil ;
- intelligibilité.

Ces critères sont repris dans [Geng et Hamilton 06] parmi d'autres pour décrire un grand nombre de mesures. Ils peuvent évidemment guider l'utilisateur dans le choix de sa mesure d'intérêt.

2.3.3 Quelques exemples de mesures

Le but de cette thèse n'est pas d'être exhaustif sur les mesures d'intérêt étudiées, mais plutôt de fournir des outils génériques d'étude des mesures. Nous verrons que ces outils peuvent être dirigés selon trois grands axes : robustesse des règles, adaptabilité des mesures à des propriétés algorithmiques existantes et anti-monotonicité des mesures dans le cadre des règles de classe. Nous allons cependant donner ici une ensemble de mesures sur lequel nous nous appuierons par la suite. Ces mesures sont détaillées dans [Ohsaki et al. 04, Tan et al. 04, Geng et Hamilton 06, Lenca et al. 08]. Nous donnons dans les tables 2.3 et 2.4 l'expression d'un ensemble de 42 mesures et, lorsqu'il nous a été possible de les retrouver, les références originales des mesures.

1. http://michael.hahsler.net/research/association_rules/measures.html

nom	formule	référence
confiance	$\frac{p_{AB}}{p_A}$	[Cleverdon et al. 66]
confiance centrée	$p_{B\|A} - p_B$	[]
moindre contradiction	$\frac{2p_{AB} - p_A}{p_B}$	[Azé et Kodratoff 02]
conviction	$\frac{p_A p_{\neg B}}{p_{A\neg B}}$	[Brin et al. 97b]
cosine	$\frac{p_{AB}}{\sqrt{p_A p_B}}$	[Salton et McGill 83]
couverture	p_A	[]
Czekanowski	$\frac{2p_{AB}}{p_A + p_B}$	[Czekanowski 13]
facteur bayésien	$\frac{p_{A\|B}}{p_{A\|\neg B}}$	[Jeffreys 35]
force collective	$\frac{p_{AB} + p_{\neg A\neg B}}{p_A \times p_B + p_{\neg A} \times p_{\neg B}} \times \frac{p_{\neg A} \times p_B + p_A \times p_{\neg B}}{p_{\neg AB} + p_{A\neg B}}$	[Aggarwal et Yu 98]
gain	$p_{AB} - \theta p_A$	[Fukuda et al. 96]
gain informationnel	$\log \frac{p_{AB}}{p_A p_B}$	[Church et Hanks 90]
Ganascia	$2\frac{p_{AB}}{p_A} - 1$	[Ganascia 91]
indice de Gini	$\frac{1}{p_A} \times ({p_{AB}}^2 + {p_{A\neg B}}^2) + \frac{1}{p_{\neg A}} \times ({p_{\neg AB}}^2 + {p_{\neg A\neg B}}^2) - {p_B}^2 - {p_{\neg B}}^2$	[Gini 21]
indice d'implication	$\sqrt{n}\frac{p_{AB} - p_A p_B}{\sqrt{p_A p_{\neg B}}}$	[Lerman et al. 81]
intérêt	$\|p_{AB} - p_A p_B\|$	[]
J1-mesure	$p_{AB} \times \log \frac{p_{AB}}{p_A p_B}$	[Wang et al. 98]
Jaccard	$\frac{p_{AB}}{p_A + p_B - p_{AB}}$	[Jaccard 01]
J-mesure	$p_{AB} \times \log \frac{p_{AB}}{p_A p_B} + p_{A\neg B} \times \log \frac{p_{A\neg B}}{p_A p_{\neg B}}$	[Smyth et Goodman 91]
Kappa	$2\frac{p_{AB} - p_A p_B}{p_A p_{\neg B} + p_B p_{\neg A}}$	[Cohen 60]
Klosgen	$\sqrt{p_{AB}} \times (p_{B\|A} - p_B)$	[Klösgen 92]
Kulczynski	$\frac{p_{AB}}{p_{A\neg B} + p_{\neg AB}}$	[Kulczynski 27]

TABLE 2.3 : Ensemble de mesures

nom	formule	référence
Laplace	$\frac{np_{AB}+1}{np_A+2}$	[Good 65]
levier	$p_{AB}-p_Ap_B$	[Piatetsky-Shapiro 91]
lift	$\frac{p_{AB}}{p_Ap_B}$	[Brin et al. 97a]
Loevinger	$\frac{p_{B\|A}-p_B}{1-p_B}$	[Loevinger 47]
odds ratio	$\frac{p_{AB}p_{\neg A\neg B}}{p_{A\neg B}p_{\neg AB}}$	[Yule 00]
one way support	$p_{B\|A}\times\log\frac{p_{AB}}{p_Ap_B}$	[Yao et Liu 97]
coefficient de Pearson	$\frac{p_{AB}-p_Ap_B}{\sqrt{p_Ap_Bp_{\neg A}p_{\neg B}}}$	[Pearson 96]
Piatetsky-Shapiro	$n\times(p_{AB}-p_Ap_B)$	[Piatetsky-Shapiro 91]
précision	$p_{AB}+p_{\neg A\neg B}$	[]
prevalence	p_B	[]
Q de Yule	$\frac{p_{AB}\times p_{\neg A\neg B}-p_{A\neg B}\times p_{\neg AB}}{p_{AB}\times p_{\neg A\neg B}+p_{A\neg B}\times p_{\neg AB}}$	[Yule 00]
rappel	$\frac{p_{AB}}{p_B}$	[Cleverdon et al. 66]
risque relatif	$\frac{p_{B\|A}}{p_{B\|\neg A}}$	[]
Sebag-Shoenauer	$\frac{p_{AB}}{p_{A\neg B}}$	[Sebag et Schoenauer 88]
spécificité	$p_{\neg B\|\neg A}$	[]
spécificité relative	$p_{\neg A\|\neg B}-p_{\neg A}$	[Lavrac et al. 99]
support	p_{AB}	[Agrawal et al. 93]
taux exemples contre-exemples	$1-\frac{p_{A\neg B}}{p_{AB}}$	[]
valeur ajoutée	$\max(p_{B\|A}-p_B, p_{A\|B}-p_A)$	[Tan et al. 04]
Y de Yule	$\frac{\sqrt{p_{AB}\times p_{\neg A\neg B}}-\sqrt{p_{A\neg B}\times p_{\neg AB}}}{\sqrt{p_{AB}\times p_{\neg A\neg B}}+\sqrt{p_{A\neg B}\times p_{\neg AB}}}$	[Yule 00]
Zhang	$\frac{p_{AB}-p_Ap_B}{\max(p_{AB}p_{\neg B}, p_Bp_{A\neg B})}$	[Zhang 00]

TABLE 2.4 : Ensemble de mesures

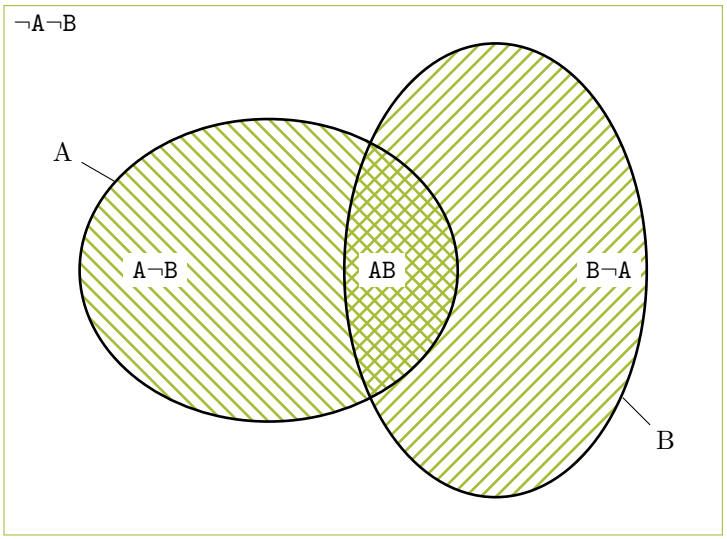


FIGURE 2.1 : Table de contingence ensembliste

Exemple 1 – Mesure de Jaccard – : La mesure de **Jaccard** est définie par $\frac{p_{AB}}{p_A+p_B-p_{AB}}$. Si l'on considère la version ensembliste de la table de contingence (Figure 2.1), on voit que cette mesure représente le rapport entre l'intersection des transactions contenant A et B et l'union de ces deux ensembles : $Jacc(A \rightarrow B) = \frac{A \cap B}{A \cup B}$. Ainsi, la mesure de **Jaccard** va favoriser les motifs fortement corrélés et proposant peu de contre-exemples et/ou faux positifs. À l'indépendance, elle ne présente pas une valeur fixe et sa valeur pour les règles logiques est aussi variable. Elle prend ses valeurs dans l'intervalle $[0, 1]$ et sa valeur maximale 1 n'est atteinte que pour des motifs A et B égaux.

La mesure de **Jaccard** a donc une interprétation ensembliste. D'autres mesures s'apparentent aux fonctions d'entropie utilisées en théorie de l'information, typiquement celles faisant apparaitre un logarithme.

Exemple 2 – One Way Support – : La mesure du **One Way Support** est définie par $p_{B|A} \times \log \frac{p_{AB}}{p_A p_B}$. Elle peut être interprétée comme la quantité d'information apportée par le motif A dans la règle $A \rightarrow B$ par rapport à B. En effet, on remarque que $oneway(A \rightarrow B) = -p_{B|A} \times (\log(p_B) - \log(p_{B|A}))$. C'est le gain obtenu sur l'information B en possédant l'information supplémentaire A. Cette mesure va largement favoriser les règles $A \rightarrow B$ dont les motifs A et B sont fortement positivement corrélés. Elle est nulle à l'indépendance, négative lorsque les variables sont décorrélées et prend ses valeurs dans $[-e^{-1}, +\infty]$.

Parmi le grand nombre de mesures d'intérêt existantes, certaines ont une écriture plus complexe que d'autres. On peut cependant en relever quelques-unes qui ont une définition très intuitive. Le

support et la **confiance** en font bien entendu partie.

> **Exemple 3 – précision – :** La mesure de **précision** est définie par $p_{AB} + p_{\neg A \neg B}$. Elle représente donc la probabilité de *vrais* cas, c'est-à-dire de vrais-positifs et vrais-négatifs. Elle peut être par exemple utilisée en médecine pour s'assurer de la diminution des faux-positifs et faux-négatifs, qui sont les cas problématiques des tests cliniques. Elle est aussi largement utilisée pour la validation des modèles de classification. Elle va favoriser les règles à fort **support**, mais aussi les règles dont l'antécédent et le conséquent sont peu fréquents, puisque l'on a alors beaucoup de vrais négatifs. Elle n'a pas de valeur fixe à l'indépendance et prend ses valeurs dans l'intervalle $[0, 1]$.

Il apparait donc que l'ensemble des mesures est vaste et varié. Les mesures changent suivant les domaines d'applications et les besoins des utilisateurs. Nous n'en définissons ici qu'un petit nombre, par rapport à leur totalité. Notons enfin qu'il est également possible de les agréger (voir par exemple [Barthélemy et al. 06, Le et al. 08]).

2 Conclusion

L'obtention de règles d'association, c'est-à-dire d'objets de la forme $A \rightarrow B$ est le but principal de cette thèse. Nous avons cependant vu que la recherche de règles posait de nombreux problèmes, aussi bien algorithmiques que qualitatifs. Plusieurs solutions ont été proposées, parmi lesquelles nous retiendrons l'utilisation de mesures d'intérêt objectives. L'étude une à une des mesures est une tâche laborieuse et il est impératif de mettre en place un cadre général et rigoureux d'étude des mesures d'intérêt objectives.

3 Un cadre formel d'étude des mesures d'intérêt

L'évaluation des règles d'association peut se faire par le biais de mesures d'intérêt objectives toujours plus nombreuses. Il est intéressant de pouvoir trouver des caractéristiques communes à ces mesures, afin de décrire des méthodes générales pour leur étude. Quelques travaux ont été menés dans ce but, que nous décrivons dans la première partie de ce chapitre. En s'appuyant sur ces travaux, nous définissons un cadre formel d'étude qui consiste à considérer les mesures d'intérêt objectives comme des fonctions de trois variables. Nous verrons que toute la difficulté réside dans la définition correcte de ces trois variables et dans la description de leur domaine de variation conjoint.

3.1 État de l'art

Nous allons ici décrire trois travaux qui nous semblent les plus similaires à notre approche et dont nous nous sommes inspirés. Nous verrons qu'ils ont en commun le fait de s'intéresser au lien entre les propriétés analytiques des mesures et la génération de méthodes liées à ces mesures.

3.1.1 Étude du comportement des mesures par les contre-exemples

La notion de contre-exemple est importante dans l'évaluation des règles d'association. Il est en effet assez naturel de penser qu'une bonne règle est une règle ayant peu de contre-exemples. Il faut cependant aller plus loin, car il n'est pas raisonnable d'espérer que les règles n'auront jamais de contre-exemple. C'est ce problème qui est abordé dans [Vaillant et al. 06] où trois modèles de variation des contre-exemples sont étudiés. Les contre-exemples sont représentés par la quantité $p_{\mathtt{A}\neg\mathtt{B}}$ et l'on peut établir un lien avec les autres quantités de la table de contingence, en considérant les marges comme étant fixées. La table 3.1(a) résume ces relations. Les auteurs étudient la variation de différentes mesures objectives par rapport à la variation des contre-exemples et notamment le comportement de ces mesures face à l'arrivée de contre-exemples : linéaires, convexes, concaves. Cependant, que l'on considère la présence de bruit, ou l'arrivée de nouvelles transactions, il n'est pas nécessairement justifié de penser que les valeurs de marge ($p_{\mathtt{A}}$ et $p_{\mathtt{B}}$) restent fixes. Deux autres modèles sont donc proposés, qui autorisent la variation du motif `A` ou du motif `B`. Ces deux modèles de variation sont exposés tables 3.1(b) et 3.1(c). Le modèle 2 permet d'étudier le cas où des exemples deviendraient des contre-exemples et plus particulièrement la résistance de la mesure à une forme de bruit. Dans le cadre de règles de classe, on pourrait penser par exemple à des erreurs de classement : que se passe-t-il si certaines transactions n'ont en fait pas la bonne classe ? Cette étude a été menée dans [Lenca et al. 06] pour définir une notion de robustesse, notion sur laquelle nous reviendrons plus tard. Les auteurs considèrent qu'un comportement décroissant par rapport à l'apparition de contre-exemples est un critère d'éligibilité et étudient un ensemble de mesures répondant à ce critère. Ils montrent que, suivant les modèles, certaines mesures résistent

TABLE 3.1 : Modèles de variation de la table de contingence par rapport aux contre-exemples

(a) Modèle 1

	A	¬A	
B	$p_{\mathtt{A}} - p_{\mathtt{A}\neg\mathtt{B}}$	$p_{\mathtt{B}} - p_{\mathtt{A}} + p_{\mathtt{A}\neg\mathtt{B}}$	$p_{\mathtt{B}}$
¬B	$p_{\mathtt{A}\neg\mathtt{B}}$	$1 - p_{\mathtt{B}} - p_{\mathtt{A}\neg\mathtt{B}}$	$p_{\neg\mathtt{B}}$
	$p_{\mathtt{A}}$	$p_{\neg\mathtt{A}}$	1

(b) Modèle 2

	A	¬A	
B	$p_{\mathtt{A}} - p_{\mathtt{A}\neg\mathtt{B}}$	$p_{\neg\mathtt{A}\mathtt{B}}$	$p_{\neg\mathtt{A}\mathtt{B}} + p_{\mathtt{A}} - p_{\mathtt{A}\neg\mathtt{B}}$
¬B	$p_{\mathtt{A}\neg\mathtt{B}}$	$p_{\neg\mathtt{A}\neg\mathtt{B}}$	$p_{\neg\mathtt{A}\neg\mathtt{B}} + p_{\mathtt{A}\neg\mathtt{B}}$
	$p_{\mathtt{A}}$	$p_{\neg\mathtt{A}}$	1

(c) Modèle 3

	A	¬A	
B	$p_{\mathtt{A}\mathtt{B}}$	$p_{\neg\mathtt{A}\mathtt{B}}$	$p_{\mathtt{B}}$
¬B	$p_{\mathtt{A}\neg\mathtt{B}}$	$p_{\neg\mathtt{B}} - p_{\mathtt{A}\neg\mathtt{B}}$	$p_{\neg\mathtt{B}}$
	$p_{\mathtt{A}\mathtt{B}} + p_{\mathtt{A}\neg\mathtt{B}}$	$p_{\neg\mathtt{B}} + p_{\neg\mathtt{A}\mathtt{B}} - p_{\mathtt{A}\neg\mathtt{B}}$	1

bien à l'introduction de contre-exemples en présentant un caractère concave, alors que d'autres sont convexes et résistent donc moins bien à cette introduction. Cependant, la grande variabilité des résultats en fonction du modèle choisi montre les insuffisances de cette méthode, mais aussi la complexité du problème.

Les deux travaux étudiés en 3.1.2 et 3.1.3 sont assez similaires dans leur esprit, puisqu'ils visent tous les deux à considérer que, probablement, un grand nombre de mesures se comportent de la même manière, ce que l'on pourrait présupposer en considérant les principes de Piatetsky-Shapiro. Le premier, en considérant le lien entre les mesures et la **`confiance`**, montre qu'un certain nombre de mesures génèrent les mêmes règles intéressantes. Le second montre que certaines mesures peuvent être simultanément minorées et que cette minoration peut être utilisée dans une stratégie d'élagage.

3.1.2 Étude des ressemblances entre les mesures par la `confiance`

Les principes de Piatetsky-Shapiro [Piatetsky-Shapiro 91] encadrent clairement le comportement souhaitable des mesures d'intérêt et certains auteurs se sont donc naturellement demandé s'il n'était pas possible de regrouper un ensemble de mesures présentant un comportement similaire. Dans [Bayardo et Agrawal 99], on s'intéresse ainsi aux *meilleures* règles, c'est-à-dire aux règles extraites d'une base de données et qui maximisent un certain critère, ce critère pouvant être une mesure d'intérêt. L'idée principale de cet article est de convertir le problème de la recherche de règles en un problème d'optimisation. On définit donc un ordre partiel sur les règles : soient deux règles d'association r_1 et r_2, on dit que $r_1 <_{sc} r_2$ si et seulement si

- $supp(r_1) \leq supp(r_2)$ et $conf(r_1) < conf(r2)$;
- ou $supp(r_1) < supp(r_2)$ et $conf(r_1) \leq conf(r2)$,

et l'on considère le problème d'optimisation défini par le quintuplet $\langle U, D, <_{sc}, C, N\rangle$ où U est un ensemble de contraintes sur l'antécédent (par exemple, quels attributs considérer), D est la base de données, C contient l'ensemble des conséquents et N est un ensemble de contraintes sur les règles (par exemple, une contrainte de **`support`** minimum). On voit, dans la définition de ce problème, que les conséquents sont préfixés : c'est le cas des règles de classe. Ainsi, dans la table de contingence, $p_{\mathtt{B}}$ est connu d'avance.

Un autre ordre partiel noté $<_{s\neg c}$ est aussi défini, sur lequel nous ne nous attarderons pas. La démarche qui nous intéresse est la suivante. Les auteurs font le lien entre la *sc*-optimalité, c'est-à-dire le fait pour une règle d'être optimale pour l'ordre partiel $<_{sc}$ dans le problème $\langle U, D, <_{sc}, C, N\rangle$ et l'optimalité dans $\langle U, D, <_t, C, N\rangle$ pour tout ordre total $<_t$ (par exemple, l'ordre induit par une mesure) qui serait induit par $<_{sc}$. Par induit, on entend que si $r_1 <_{sc} r_2$ alors $r_1 <_t r_2$. Le problème

étant de savoir quelles sont les mesures qui définissent un ordre total induit par l'ordre $<_{sc}$. Pour répondre à cette question, un lemme est proposé dans [Bayardo et Agrawal 99] qui prend une forme que l'on retrouvera souvent au cours de ce manuscrit :

> **Lemme 1 :** Soit m une fonction d'évaluation des règles, si m est croissante par rapport au **support** pour des règles de même **confiance** et si m est croissante par rapport à la **confiance** pour des règles de même **support** alors l'ordre total $<_m$ défini par m est induit par $<_{sc}$.

Ici la croissance indique que, par exemple pour deux règles ayant la même mesure de **confiance**, la mesure m devra être plus élevée sur la règle ayant un support plus élevée. Ainsi, si l'on peut exprimer la fonction m comme étant une fonction du **support** de la règle, de sa **confiance** et finalement du **support** du conséquent (qui est fixe puisque prédéterminé), nous avons ici une condition permettant de dire si l'on peut se concentrer sur le problème $\langle U, D, <_{sc}, C, N\rangle$ pour résoudre $\langle U, D, <_m, C, N\rangle$. C'est évidemment le cas pour le **support** et la **confiance**, mais aussi pour d'autres mesures, comme par exemple la mesure de **conviction** qui peut s'écrire sous la forme $\frac{1-p_{\mathtt{B}}}{1-conf(\mathtt{A}\to\mathtt{B})}$. Les auteurs expliquent donc que la résolution du problème $\langle U, D, <_{sc}, C, N\rangle$ renvoie un ensemble de règles parmi lesquelles on trouve les meilleures règles pour toutes ces mesures.

Cette démarche est, à notre avis, très intéressante et porteuse de bien des espoirs. On fournit en effet ici un algorithme qui va retourner un ensemble de règles, mais sans jamais s'appuyer sur d'autres mesures que le **support** et la **confiance**, c'est-à-dire que la mesure de **Laplace**, de **conviction**, ou d'autres n'interviennent pas réellement au cœur de l'algorithme, si ce n'est par la cohérence des résultats. En effet, on sait que parmi les règles retournées, se trouve la meilleure règle pour ces mesures ; en revanche, pour une autre mesure, on ne peut pas l'affirmer (ni l'infirmer). Nous avons donc là une propriété générale, c'est-à-dire qu'elle retourne des résultats utiles pour un ensemble particulier de mesures (celles du lemme 1).

3.1.3 Étude de propriétés partagées par les mesures

Une approche similaire est développée dans [Hébert et Crémilleux 06, Hébert et Crémilleux 07] pour l'identification d'un ensemble de mesures simultanément optimisables. L'idée sur laquelle sont fondés ces deux articles est que l'on peut écrire un grand nombre de mesures objectives en fonction du **support** de l'antécédent, du conséquent et du motif d'une règle d'association (quelques mesures échappent à cette règle, comme par exemple la mesure de **Laplace**). Dès lors, les auteurs s'intéressent à la variation des mesures pour un nombre de contre-exemples fixé, dans le cadre des règles de classe pour [Hébert et Crémilleux 06] et dans le cadre général des règles d'association pour [Hébert et Crémilleux 07]. En effet, pour une règle de classification, le conséquent est prédéterminé et si l'on fixe le nombre de contre-exemples, la mesure n'a plus qu'un degré de liberté, par exemple le **support** de l'antécédent. Dans le cas d'une règle d'association générale, la fixation du nombre de contre-exemples laisse à la mesure deux degrés de liberté. Ces remarques seront primordiales pour la suite de ce document.

Ainsi donc, si l'on considère une mesure m, on peut lui associer la fonction $\Psi_m(x, y, z)$ telle que pour toute règle d'association $r : \mathtt{A} \to \mathtt{B}$, on ait $m(r) = \Psi_m(supp(\mathtt{A}), supp(\mathtt{B}), supp(\mathtt{AB}))$. Et si le nombre de contre-exemples est fixé, par exemple à un nombre δ, on peut écrire le changement de variable $z = x - \delta$ et définir la fonction $\Psi_{m,\delta}(x, y) = \Psi_m(x, y, x - \delta)$. Si une mesure m est telle que la fonction $\Psi_{m,\delta}$ est croissante par rapport à sa première variable et si m vérifie les principes de variations énoncés par Piatetsky-Shapiro, alors elle fait partie de l'ensemble des mesures simultanément optimisables. Cet ensemble est ainsi nommé car si l'on a une règle $r : \mathtt{A} \to \mathtt{B}$ telle que $supp(\mathtt{A}) \geq \gamma$, $supp(\mathtt{B}) \leq \eta$ et qu'elle admet moins de δ contre-exemples, pour toute mesure de cet ensemble, $m(r) \geq \Psi_{m,\delta}(\gamma, \eta)$.

	A	¬A	
B	p_{AB}	$p_{\neg AB}$	p_B
¬B	$p_{A\neg B}$	$p_{\neg A\neg B}$	$p_{\neg B}$
	p_A	$p_{\neg A}$	1

TABLE 3.2 : Table de contingence relative

Cette propriété est intéressante car elle s'applique à un grand nombre de mesures, comme la mesure de **support**, la **confiance**, mais aussi le **lift**, la **conviction** ou **Jaccard**. Les auteurs en recensent 17, mais précisent que toute combinaison linéaire positive de mesures simultanément optimisables est une mesure simultanément optimisable. Cependant, cette démarche ne serait pas intéressante si elle ne permettait pas la recherche efficace de ces règles qui maximisent toutes les mesures. Les conditions formées sur les contre-exemples rappellent la notion de motif δ-libre et l'on possède toujours une contrainte de **support**, qui permet d'utiliser des algorithmes de type APRIORI. Les auteurs préfèrent quant à eux se limiter aux règles s'appuyant sur des motifs fermés et dont l'antécédent est un motif libre, ces règles étant appelées règles informatives. Elle sont informatives car, comme nous l'avons déjà vu, la notion de clôture ainsi que la notion de liberté portent toutes les informations nécessaires, aussi bien pour générer l'ensemble des motifs fréquents que l'ensemble des valeurs de **support** associés. La recherche de règles parmi ces motifs se fait alors de manière exhaustive.

Si le résultat final est intéressant, cette méthode ne permet cependant pas la définition d'un algorithme orginal ou d'une propriété d'anti-monotonie comme le permettaient par exemple les motifs libres. Les règles retournées sont optimales pour un grand nombre de mesures et la recherche est, encore une fois, mesure-indépendante. Elle est en ce sens proche de la recherche des meilleures règles du paragraphe précédent.

Ce qui fait l'intérêt de ces travaux par rapport à cette thèse, c'est la considération des mesures comme de *simples* fonctions continues de trois variables. Cela implique notamment la considération des principes de Piatetsky-Shapiro sous forme analytique, ainsi par exemple, *m croît avec p_{AB} lorsque tous les autres paramètres restent les mêmes* devient *Ψ_m croît par rapport à la troisième variable*. Nous allons dans la suite reprendre cette idée et la formaliser, notamment en étudiant le co-domaine de Ψ_m, c'est-à-dire la façon dont x, y et z évoluent, mais aussi en étudiant d'autres paramétrisations, pas uniquement en fonction des exemples comme ici, mais aussi en fonction des contre-exemples, ou de la **confiance**, comme dans les deux travaux décrits précédemment.

3.2 SYSTÈME DESCRIPTEUR ET DOMAINE ADAPTÉ.

Nous proposons ici de poser les bases formelles du reste de cette thèse et donc de mettre en avant notre vision des règles *via* des projections dans $\mathbb{R}^3$ et notre vision des mesures comme de simples fonctions réelles de trois variables. Pour ce faire, nous reviendrons tout d'abord sur la notion de table de contingence, afin de faire le lien entre les règles et $\mathbb{R}^3$ puis nous consoliderons cette méthode de projection.

3.2.1 Système Descripteur

Ainsi que nous l'avons précisé précédemment, nos travaux s'inspirent des techniques utilisées dans les différents articles détaillés en 3.1. Principalement, nous considérons les mesures pouvant s'exprimer en fonction des valeurs de la table de contingence en fréquences, que nous rappelons sur la table 3.2. Cette table de contingence possède trois degrés de libertés, dans le sens où trois de

	A	¬A	
B	p_{AB}	$p_B - p_{AB}$	p_B
¬B	$p_A - p_{AB}$	$1 - p_B - p_A + p_{AB}$	$1 - p_B$
	p_A	$1 - p_A$	1

TABLE 3.3 : Table de contingence relative en fonction du triplet (p_{AB}, p_A, p_B).

	A	¬A	
B	$conf \times ant$	$cons - conf \times ant$	$cons$
¬B	$ant \times (1 - conf)$	$1 - cons - ant \times (1 - conf)$	$1 - cons$
	ant	$1 - ant$	1

TABLE 3.4 : Table de contingence relative en fonction du triplet $(conf, ant, cons)$.

ses valeurs (bien choisies) suffisent à la décrire, alors que deux valeurs ne suffisent pas. Par décrire, on entend que toutes les valeurs de la table peuvent être retrouvées en combinant trois valeurs particulières. Par exemple, le triplet (p_{AB}, p_A, p_B) suffit à décrire la table de contingence comme le montre la table 3.3. D'un autre côté, il n'est pas possible de choisir trois valeurs quelconques, comme le montre le triplet $(p_A, p_{\neg A}, p_B)$. En effet, il est impossible de calculer une valeur croisée, par exemple p_{AB} et cela est principalement dû au fait que le couple $(p_A, p_{\neg A})$ est lié par la relation $p_A + p_{\neg A} = 1$. Pour décrire la table de contingence en tant qu'objet fonctionnel, c'est-à-dire libérée de l'instanciation à une règle, il suffit donc de choisir un bon triplet de fonctions. Si l'on définit les fonctions suivantes :

$$ex(A \to B) = p_{AB} \; ; \quad ant(A \to B) = p_A \; ; \quad cons(A \to B) = p_B$$

alors le triplet de fonctions $(ex, ant, cons)$ permet de décrire toute table de contingence. Nous décidons d'appeler un tel "bon" triplet de fonctions un *système descripteur* de la table de contingence. Pour plus de liberté, nous étendrons même cette notion à un triplet de fonctions des règles d'association quelconque, comme par exemple le triplet $(conf, ant, cons)$ où $conf$ désigne la mesure de `confiance`.

Définition 6 – système descripteur – : Nous nommons *système descripteur de la table de contingence* tout triplet de fonction (s_1, s_2, s_3) sur les règles d'association qui permet de décrire entièrement la table de contingence.

Ainsi donc, si (s_1, s_2, s_3) est un système descripteur, alors toute cellule de la table de contingence peut être calculée à partir de ce système, par exemple il existe une fonction u telle que $p_{AB} = u(s_1(A \to B), s_2(A \to B), s_3(A \to B))$. Dans le cas du triplet $(conf, ant, cons)$ décrit plus tôt, la table de contingence prend la forme de la table 3.4. Il serait ainsi possible de définir un grand nombre de systèmes descripteurs, mais dans notre travail, nous ne nous servirons que des deux précédents, à savoir $(ex, ant, cons)$ et $(conf, ant, cons)$ et d'un troisième, celui mettant en jeu les contre-exemples $(cex, ant, cons)$, où $cex(A \to B) = p_{A\neg B}$. Dans ce système descripteur, la table de contingence prend la forme de la table 3.5.

	A	¬A	
B	$ant - cex$	$cons - ant + cex$	$cons$
¬B	cex	$1 - cons - cex$	$1 - cons$
	ant	$1 - ant$	1

TABLE 3.5 : Table de contingence relative en fonction du triplet $(conf, ant, cons)$.

La plupart des mesures d'intérêt objectives peuvent s'écrire en fonction des valeurs de la table

de contingence et donc finalement, ces mesures d'intérêt objectives peuvent s'écrire en fonction de n'importe lequel des systèmes descripteurs. Parfois, comme pour la mesure de **Laplace**, la variable n, qui désigne le nombre de transactions de la table, peut intervenir. Lorsque celle-ci est fixe (par exemple, si une unique base est considérée), cela ne pose pas de problème. Par contre, dans le cas d'étude de la robustesse par exemple, où l'on sera amené à se poser la question de la stabilité de la mesure lors de variations de la base, ces mesures ne pourront être considérées comme s'écrivant en fonction d'un système descripteur, car alors il y aura 4 degrés de liberté et non 3.

3.2.2 Domaine adapté

Soit m une mesure d'intérêt objective des règles d'association pouvant s'exprimer en fonction des cellules de la table de contingence. Soit S un système descripteur de la table de contingence composé des trois fonctions s_1, s_2 et s_3. On note Φ_m^S la fonction réelle de trois variables telle que pour toute règle d'association $r : \mathtt{A} \rightarrow \mathtt{B}$ on ait $m(r) = \Phi_m^S(s_1(r), s_2(r), s_3(r))$. La question que nous allons nous poser ici concerne le codomaine de Φ_m^S, ou encore l'espace dans lequel le triplet $(s_1(r), s_2(r), s_3(r))$ évolue lorsque r décrit l'ensemble des règles d'association et plus particulièrement dans les trois systèmes descripteurs que nous avons rencontrés jusque là. Inspirés par les conditions de Piatetsky-Shapiro, nous commençons naturellement par le système que nous noterons $S_{ex} = (ex, ant, cons)$.

3.2.2.1 Domaine des exemples

Les lois du calcul des probabilités imposent les contraintes suivantes concernant les cellules de la table de contingence :

$$p_{\mathtt{AB}} \leq p_{\mathtt{A}} \; ; \qquad p_{\mathtt{AB}} \leq p_{\mathtt{B}} \; ; \qquad p_{\mathtt{AB}} \geq p_{\mathtt{A}} + p_{\mathtt{B}} - 1$$

ainsi que les contraintes

$$0 \leq p_{\mathtt{AB}}, p_{\mathtt{A}}, p_{\mathtt{B}} \leq 1.$$

Ces contraintes peuvent se résumer simplement dans le cadre de notre système descripteur.

L'ensemble des valeurs prises par le triplet $(ex, ant, cons)$ sur l'espace des règles d'association est

$$D_{ex} = \left\{ \begin{pmatrix} ex(r) \\ ant(r) \\ cons(r) \end{pmatrix} \middle| \, r \text{ règle d'association} \right\} = \left\{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{Q}^3 \middle| \, \max(0, y+z-1) \leq x \leq \min(y, z) \right\} \tag{3.1}$$

Il faut remarquer que les inégalités de la partie droite de l'équation 3.1 contiennent les inégalités $0 \leq y, z \leq 1$ que l'on aurait pu s'attendre à voir apparaître. Nous appelons le sous-ensemble de $\mathbb{R}^3$ ainsi défini le *domaine adapté* au système descripteur S_{ex} et nous le notons D_{ex}. Il reste cependant à démontrer l'égalité de l'équation 3.1, car si l'inclusion directe découle trivialement des contraintes de probabilité énoncées précédemment, l'inclusion réciproque demande d'associer à tout triplet (x, y, z) une règle d'association. Il faut pour ce faire définir une bonne base de données.

Démonstration. Considérons donc un point (x, y, z) de l'ensemble

$$\left\{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{Q}^3 \middle| \, \max(0, y+z-1) \leq x \leq \min(y, z) \right\}.$$

Comme x, y et z sont des nombres rationnels, il existe un entier n tel que $x \times n$, $y \times n$ et $z \times n$ soient des nombres entiers. Construisons la base de données de la table 3.6. Cette base de données

	1		$(y-x)\times n$			$y\times n$			$(y-x+z)\times n$			n
A	1	$\cdots$	$\cdots$	$\cdots$	$\cdots$	1	0	$\cdots$	$\cdots$	$\cdots$	$\cdots$	0
B	0	$\cdots$	0	1	$\cdots$	$\cdots$	$\cdots$	$\cdots$	1	0	$\cdots$	0

TABLE 3.6 : Base de données pour les domaines adaptés.

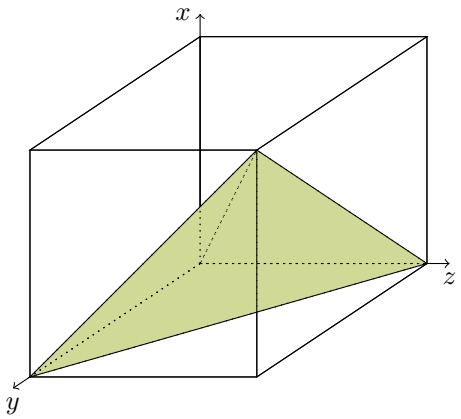


FIGURE 3.1 : Domaine adapté au système descripteur S_{ex}

peut être construite car d'après la définition du domaine auquel appartient le triplet (x, y, z), on a $0 \leq y - x \leq y \leq y - x + z \leq 1$. Elle est donc licite et dans cette base de données, la règle d'association $\mathtt{A} \rightarrow \mathtt{B}$ vérifie

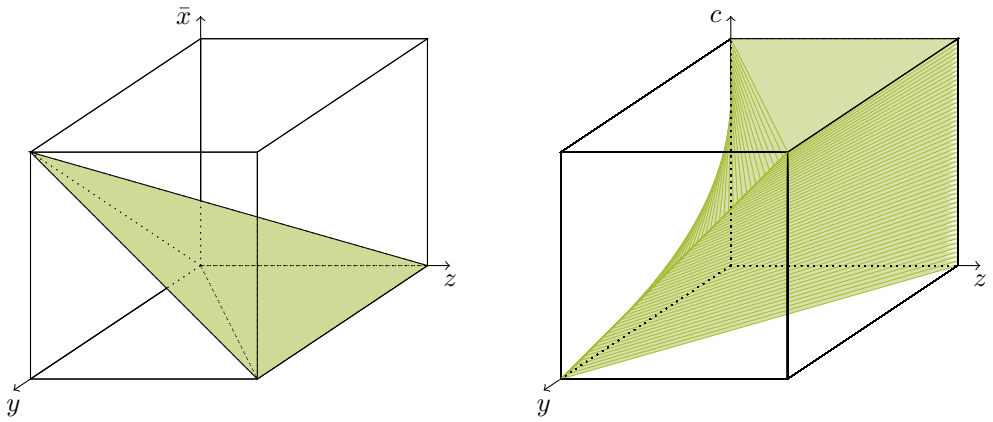
$$\begin{aligned} ex(\mathtt{A} \rightarrow \mathtt{B}) &= x \\ ant(\mathtt{A} \rightarrow \mathtt{B}) &= y \\ cons(\mathtt{A} \rightarrow \mathtt{B}) &= z \end{aligned}$$

Ce qui démontre donc l'inclusion réciproque. □

Nous avons identifié précisément le domaine adapté au système descripteur S_{ex}. Ce domaine est représenté Figure 3.1. Il est très important, car il permet de n'étudier que des cas réels lors de l'étude de la fonction associée à une mesure. En effet, il est inutile, lors d'une paramétrisation par les exemples, d'aller étudier ce qui se passe au point $(0.7, 0.3, 0.2)$ où la mesure pourrait avoir un comportement exotique et perturber les analyses : ce point n'appartient pas à D_{ex} et ne correspond donc à aucune règle.

3.2.2.2 Domaines des contre-exemples et de la `confiance`

Après avoir détaillé le cas de la paramétrisation par les exemples, nous allons voir la définition du domaine adapté au système descripteur des contre-exemples $S_{cex} = (cex, ant, cons)$ et au système descripteur de la **`confiance`** $S_{conf} = (conf, ant, cons)$ sans en donner les preuves, car pour celles-ci, la base de données utilisée est la même que pour les exemples. Il ne s'agit en fait que d'un simple changement de variable. Le domaine associé aux contre-exemples est illustré sur la figure 3.2(a) et



(a) (b)

FIGURE 3.2 : Domaines adaptés aux contre-exemples et à la **confiance**

est défini par

$$D_{cex} = \left\{ \begin{pmatrix} cex(r) \\ ant(r) \\ cons(r) \end{pmatrix} \middle| \, r \text{ règle d'association} \right\} = \left\{ \begin{pmatrix} \bar{x} \\ y \\ z \end{pmatrix} \in \mathbb{Q}^3 \middle| \, \max(0, y-z) \leq \bar{x} \leq \min(y, 1-z) \right\}. \tag{3.2}$$

Le domaine adapté à la paramétrisation par la **confiance** est quant à lui défini par :

$$D_{conf} = \left\{ \begin{pmatrix} conf(r) \\ ant(r) \\ cons(r) \end{pmatrix} \middle| \, r \text{ règle d'association} \right\} = \left\{ \begin{pmatrix} c \\ y \\ z \end{pmatrix} \in \mathbb{Q}^3 \middle| \, \max(0, 1 - \frac{1-z}{y}) \leq c \leq \min(1, \frac{z}{y}) \right\}. \tag{3.3}$$

Il est illustré Figure 3.2(b).

Une remarque est encore nécessaire. Nous allons nous appuyer sur cette notion de domaine adapté, que nous associerons donc à l'expression d'une mesure dans ce domaine, notamment pour des études de variations ou des problèmes d'optimisation. Dans ce cadre, la définition du domaine dans $\mathbb{Q}^3$ au lieu de $\mathbb{R}^3$ pourrait poser problème. Cependant, les mesures objectives que nous avons rencontrées ont toutes des propriétés commodes notamment de continuité et pourront être considérées comme des fonctions de $\mathbb{R}^3$ grâce à un argument de densité. Ainsi, aucun problème ne se posera concernant la continuité ou la dérivabilité, ou bien d'autres propriétés.

3.2.3 Fonction de mesure adaptée

Il nous reste donc à regrouper les notions de domaine adapté et de mesure. Une mesure d'intérêt objective sur les règles d'association peut être vue comme une fonction de trois variables fixées en fonction du système descripteur choisi. Ce système descripteur définit par ailleurs le codomaine de la fonction de trois variables et est indispensable si l'on veut faire une étude précise et juste de la

mesure dans celui-ci. Nous intégrons donc ces deux notions dans une seule et même définition.

> **Définition 7 – fonction de mesure adaptée – :** Soit un système descripteur S de la table de contingence et m une mesure d'intérêt objective des règles d'association. Il existe une fonction Φ_m^S de $\mathbb{Q}^3$ dans $\mathbb{R}$ telle que pour toute règle d'association r, on ait $\Phi_m^S(s_1(r), s_2(r), s_3(r)) = m(r)$. Le codomaine de cette fonction est donné par D_S le domaine adapté à S. Nous appelons *fonction de mesure S-adaptée à m* le couple (Φ_m^S, D_S).

Lorsque cela ne portera pas à ambiguité, nous nous permettrons des facilités de notation, notamment en omettant la précision du système descripteur lorsque celui-ci aura été clairement identifié. La table 3.7 donne, pour chacune des mesures identifiées précédemment, leur écriture dans les trois systèmes descripteurs.

Le domaine adapté est primordial, comme le montre par exemple l'étude des variations de la **spécificité** dans le domaine D_{ex} par rapport à la deuxième variable (celle représentant le **support** de l'antécédent). En effet, on a $\partial_2 \Phi_{spé}^{D_{ex}}(x, y, z) = -\frac{z-x}{(1-y)^2}$ et cette dérivée pourrait être positive si l'on avait $z - x < 0$. Cependant, la définition de D_{ex} (équation 3.1) assure que l'on a $x < z$, donc la mesure de **spécificité** est décroissante dans D_{ex} par rapport à sa deuxième variable et respecte donc l'une au moins des conditions de Piatetsky-Shapiro. C'est bien ce genre de situation que la notion de fonction de mesure adaptée doit résoudre.

Conclusion

Nous avons proposé un cadre d'étude des règles d'association. La base de contingence d'une règle possédant 3 degrés de liberté, nous pouvons projeter les règles dans $\mathbb{R}^3$ dans un domaine défini par les variations conjointes des éléments. Le choix des variables de description de la table de contingence sera primordial par la suite. Dans ce contexte, les mesures d'intérêts sont de simples fonctions de trois variables, définies sur le domaine adapté. La fonction de mesure adaptée à une mesure d'intérêt permet ainsi une étude analytique rigoureuse du comportement des mesures et sera le fondement théorique de la suite de cette thèse.

nom	exemples D_{ex}	contre-exemples D_{cex}	confiance D_{conf}
confiance	$\frac{x}{y}$	$\frac{y-x}{y}$	c
confiance centrée	$\frac{x}{y} - z$	$1 - z - \frac{x}{y}$	$c - z$
moindre contradiction	$\frac{2x-y}{z}$	$\frac{y-2x}{z}$	$y \times \frac{2c-1}{z}$
conviction	$\frac{y\times(1-z)}{y-x}$	$\frac{y\times(1-z)}{x}$	$\frac{1-z}{1-c}$
cosine	$\frac{x}{\sqrt{y\times z}}$	$\frac{y-x}{\sqrt{y\times z}}$	$c \times \sqrt{\frac{y}{z}}$
couverture	y	y	y
Czekanowski	$\frac{2x}{y+z}$	$\frac{2(y-x)}{y+z}$	$\frac{2c}{1+\frac{z}{y}}$
facteur bayésien	$\frac{x\times(1-z)}{z\times(y-x)}$	$\frac{(y-x)\times(1-z)}{z\times x}$	$\frac{c\times(1-z)}{z\times(1-c)}$
force collective	$\frac{1+2x-y-z}{y\times z+(1-y)\times(1-z)} \times \frac{y\times(1-z)+z\times(1-y)}{y+z-2x}$	$\frac{1+y-z-2x}{y\times z+(1-y)\times(1-z)} \times \frac{y\times(1-z)+z\times(1-y)}{2x+z-y}$	$\frac{1+2c\times y-y-z}{y\times z+(1-y)\times(1-z)} \times \frac{y\times(1-z)+z\times(1-y)}{y+z-2c\times y}$
gain	$x - \theta y$	$y \times (1-\theta) - x$	$y \times (c - \theta)$
gain informationnel	$\log \frac{x}{y\times z}$	$\log \frac{y-x}{y\times z}$	$\log c \times z$
Ganascia	$2\frac{x}{y} - 1$	$1 - \frac{x}{y}$	$2c - 1$
indice de Gini	$\frac{1}{y} \times (x^2 + (y-x)^2) + \frac{1}{1-y} \times ((z-x)^2 + (1-y-(z-x))^2) - z^2 - (1-z)^2$	$\frac{1}{y} \times ((y-x)^2 + x^2) + \frac{1}{1-y} \times ((z-y+x)^2 + (1-z-x)^2) - z^2 - (1-z)^2$	$y \times (c^2 + (1-c)^2) + \frac{1}{1-y} \times ((z - c\times y)^2 + (1-z-(1-c)\times y)^2) - z^2 - (1-z)^2$
indice d'implication	$\sqrt{n}\frac{x-y\times z}{\sqrt{y\times(1-z)}}$	$\sqrt{n}\frac{y\times(1-z)-x}{\sqrt{y\times(1-z)}}$	$(c-z) \times \sqrt{\frac{ny}{z}}$
intérêt	$\lvert x - y \times z\rvert$	$\lvert y \times (1-z) - x\rvert$	$\lvert y \times (c-z)\rvert$
J1-mesure	$x \times \log \frac{x}{y\times z}$	$(y-x) \times \log \frac{y-x}{y\times z}$	$c \times y \times \log \frac{c}{z}$
Jaccard	$\frac{x}{y+z-x}$	$\frac{y-x}{z+x}$	$\frac{c}{1-c+\frac{z}{y}}$
J-mesure	$x \times \log \frac{x}{y\times z} + (y-x) \times \log \frac{y-x}{y\times(1-z)}$	$(y-x) \times \log \frac{y-x}{y\times z} + x \times \log \frac{x}{y\times(1-z)}$	$c \times y \times \log \frac{c}{z} + y \times (1-c) \times \log \frac{1-c}{1-z}$
Kappa	$2\frac{x-y\times z}{y+z-2y\times z}$	$2\frac{y\times(1-z)-x}{y+z-2y\times z}$	$2\frac{y\times(c-z)}{y+z-2y\times z}$
Klosgen	$\sqrt{x} \times (\frac{x}{y} - z)$	$\sqrt{y-x} \times (1 - z - \frac{x}{y})$	$\sqrt{c\times y} \times (c-z)$
Kulczynski	$\frac{x}{y+z-2x}$	$\frac{y-x}{z+2x-y}$	$\frac{c}{1+\frac{y}{z}-2c}$
Laplace	$\frac{n\times x+1}{n\times y+2}$	$1 - \frac{n\times x+1}{n\times y+2}$	$\frac{n\times c\times y+1}{n\times y+2}$
levier	$x - y \times z$	$y \times (1-z) - x$	$y \times (c-z)$
lift	$\frac{x}{y\times z}$	$\frac{y-x}{y\times z}$	$\frac{c}{z}$
Loevinger	$1 - \frac{y-x}{y\times(1-z)}$	$1 - \frac{x}{y\times(1-z)}$	$1 - \frac{1-c}{1-z}$
odds ratio	$1 + \frac{x-y\times z}{(y-x)\times(z-x)}$	$1 + \frac{y\times(1-z)-x}{x\times(z-y+x)}$	$1 + \frac{c-z}{(1-c)\times(z-c\times y)}$
one way support	$\frac{x}{y} \times \log \frac{x}{y\times z}$	$\frac{y-x}{y} \times \log \frac{y-x}{y\times z}$	$c \times \log \frac{c}{z}$
coefficient de Pearson	$\frac{x-y\times z}{\sqrt{y\times z\times(1-y)\times(1-z)}}$	$\frac{y\times(1-z)-x}{\sqrt{y\times z\times(1-y)\times(1-z)}}$	$\frac{y\times(c-z)}{\sqrt{y\times z\times(1-y)\times(1-z)}}$
Piatetsky-Shapiro	$n \times (x - y \times z)$	$n \times (y \times (1-z) - x)$	$n \times y \times (c-z)$
précision	$2x + 1 - y - z$	$y - z - 2x + 1$	$y \times (2c-1) + 1 - z$
prevalence	z	z	z
Q de Yule	$\frac{x\times(1+x-y-z)-(y-x)\times(z-x)}{x\times(1+x-y-z)+(y-x)\times(z-x)}$	$\frac{(y-x)\times(1-x-z)-x\times(z-y+x)}{(y-x)\times(1-x-z)+x\times(z-y+x)}$	$\frac{c\times(1-y\times(1-c)-z)-(1-c)\times(z-c\times y)}{c\times(1-y\times(1-c)-z)+(1-c)\times(z-c\times y)}$
rappel	$\frac{x}{z}$	$\frac{y-x}{z}$	$\frac{y}{z} \times (1-c)$
risque relatif	$\frac{x}{y} \times \frac{1-y}{z-x}$	$\frac{y-x}{y} \times \frac{1-y}{z-y+x}$	$c \times \frac{1-y}{z-c\times y}$
Sebag-Shoenauer	$\frac{x}{y-x}$	$\frac{y-x}{x}$	$\frac{1-c}{c}$
spécificité	$1 - \frac{z-x}{1-y}$	$\frac{1-z-x}{1-y}$	$1 - \frac{z-c\times y}{1-y}$
spécificité relative	$y - \frac{y-x}{1-z}$	$y - \frac{x}{1-z}$	$y\frac{c-z}{1-z}$
support	x	$y - x$	$c \times y$
taux exemples contre-exemples	$1 - \frac{y-x}{x}$	$1 - \frac{x}{y-x}$	$1 - \frac{c}{1-c}$
valeur ajoutée	$\max(\frac{x}{y} - z, \frac{x}{z} - y)$	$\max(1 - z - \frac{x}{y}, \frac{y-x}{z} - y)$	$\max(c-z, y \times (\frac{c}{z} - 1))$
Y de Yule	$\frac{\sqrt{x\times(1+x-y-z)}-\sqrt{(y-x)\times(z-x)}}{\sqrt{x\times(1+x-y-z)}+\sqrt{(y-x)\times(z-x)}}$	$\frac{\sqrt{(y-x)\times(1-x-z)}-\sqrt{x\times(z-y+x)}}{\sqrt{(y-x)\times(1-x-z)}+\sqrt{x\times(z-y+x)}}$	$\frac{\sqrt{c\times(1-y\times(1-c)-z)}-\sqrt{(1-c)\times(z-c\times y)}}{\sqrt{c\times(1-y\times(1-c)-z)}+\sqrt{(1-c)\times(z-c\times y)}}$
Zhang	$\frac{x-y\times z}{\max(x\times(1-z), z\times(y-x))}$	$\frac{y\times(1-z)-x}{\max((y-x)\times(1-z), z\times x)}$	$\frac{c-z}{\max(c\times(1-z), z\times(1-c))}$

TABLE 3.7 : Fonctions de mesures adaptées à différents domaines.

Conclusion sur les règles d'association

Dans cette partie, nous avons passé en revue les grands algorithmes de la recherche de motifs et de la recherche de règles. Il apparaît que ces méthodes ne capturent pas toujours une bonne notion d'intérêt des règles et sont souvent peu efficaces d'un point de vue algorithmique. Nous avons notamment mis en évidence les lacunes de ces approches, notamment dans l'utilisation de mesures d'intérêt objectives, qui sont presque exclusivement utilisées pour évaluer les règles a posteriori. Nous avons donc eu l'idée de nous focaliser sur l'étude des règles et des mesures d'un point de vue analytique afin de faire le lien entre analyse mathématique et fouille de données. La projection des règles dans $\mathbb{R}^3$ permet une étude topologique des règles, comme nous le verrons dans la partie suivante, mais aussi une étude fonctionnelle des mesures et notamment de leurs variations. La notion de fonction de mesure adaptée va ainsi nous permettre de faire le lien entre propriétés analytiques des mesures et propriétés algorithmiques pour utiliser ou établir des heuristiques de recherche de règles.

«Je plie et ne romps pas»

Le Chêne et le Roseau
Jean De La Fontaine

DEUXIÈME PARTIE : ÉTUDE DE LA ROBUSTESSE DES RÈGLES D'ASSOCIATION

La robustesse des règles d'association est un sujet primordial, bien que sa définition ne soit pas encore universelle. On peut par exemple vouloir imposer une valeur de mesure (robustesse du chêne) ou bien accorder une souplesse à la règle et demander simplement à ce qu'elle reste au dessus d'un seuil fixé (robustesse du roseau). Dans cette dernière optique nous proposons une nouvelle notion de robustesse s'inspirant de notions déjà existantes et s'appuyant sur notre cadre formel (chapitre 4). Notre définition est générale, cohérente avec les outils statistiques usuels et s'applique à toute mesure d'intérêt objective. Nous nous concentrons ici sur deux grandes famille de mesures : les mesures planes (chapitre 5) et les mesures quadratiques (chapitre 6). Le texte que nous proposons est en grande partie inspiré de nos publications [Le Bras et al. 10b, Le Bras et al. 10c].

TELECOM Bretagne

4 Robustesse des règles d'association

La notion de robustesse d'une règle d'association a déjà été étudiée sous différentes formes, et nous essayons d'en donner ici une version plus générale et plus formalisée, par rapport aux définitions antérieures, souvent expérimentales ou peu générales. La notion de robustesse des règles d'association telle que nous allons la définir est une conséquence logique de la vision des règles comme des éléments de $\mathbb{R}^3$. Elle consiste à observer ce qui se passe autour de la règle pour savoir s'il y a un risque ou non que, soumise à du bruit dans les données, la règle soit en fait non-intéressante malgré une valeur de mesure observée supérieure au seuil fixé. Nous allons, dans ce chapitre, détailler cette notion, les chapitres suivants étant consacrés à l'étude de la robustesse dans le cas particulier de deux familles de mesures que nous définirons. Dans toute cette partie, dans le but d'alléger un peu les formules, nous notons $m_{\downarrow}$ le seuil d'intérêt d'une mesure m, et de manière générale, le symbole $\downarrow$ en indice traduira un élément minimum.

4.1 Différentes visions de la robustesse

Malgré la variété des travaux au sujet de la robustesse, les différents auteurs se rejoignent sur un point : étudier la robustesse des règles d'association, c'est s'intéresser à la validité d'une règle et à la variation de sa mesure lorsque les données changent. On pourra par exemple parler de bruit, ce bruit pouvant avoir bien des origines. Dans ce cas, il s'agit de savoir si une règle extraite des données est très fragile au bruit. Une règle est extraite des données par rapport à une mesure d'intérêt et à partir d'un seuil, mais la base de données observée n'est qu'un échantillon de la réalité et cet échantillon peut évidemment être plus ou moins représentatif. La *véritable* règle, c'est-à-dire la règle issue d'une base de données parfaitement représentative, sans bruit, est-elle intéressante par rapport à cette mesure et à ce seuil ? Nous allons tout d'abord citer un travail qui considère une notion de robustesse assez éloignée de la notre, puis nous reviendrons sur les travaux qui ont inspiré notre notion de robustesse et sur lesquels nous nous appuyons ici.

4.1.1 La robustesse du chêne

Lorsque l'on fait la recherche de règles d'association l'on s'intéresse à l'ensemble des règles dont la valeur de mesure est supérieure à un seuil donné. Cette notion de seuil prévaut et c'est de loin la méthode la plus répandue. Cependant, il est aussi envisageable que la valeur de cette mesure soit importante et dans le cadre de la robustesse, que l'on veuille extraire des règles dont la mesure ne risque pas de bouger malgré les modifications de la base. On est ici dans la fable *Le chêne et le roseau* de La Fontaine : est-il préférable d'avoir une règle dont la mesure ne bougera pas malgré les perturbations, ou bien peut-on se contenter de règles qui, certes, bougeront, mais dont on pourra s'assurer qu'elles restent malgré tout au dessus d'un seuil donné ? Dans [Gay et Boullé 11], c'est la première idée qui est retenue. Une nouvelle mesure est introduite, appelée le `level` qui est comparée avec la `confiance` et le `facteur bayésien` dans des contextes bruités. La détailler ici

entrainerait la définition d'un trop grand nombre de concepts dont le propos n'est pas réellement pertinent dans notre contexte, et nous laissons le soin au lecteur intéressé de consulter l'article en question. Les expériences montrent que la **confiance** et le **facteur bayésien** ne résistent pas à l'introduction de bruit dans les données puisque la valeur calculée sur les données bruitées et très différente de la valeur calculée sur les données originales. À l'opposé, le **level** est quant à lui très stable dans les contextes bruités : ainsi, la valeur de **level** dans la base bruitée est très proche de la valeur de **level** dans la base originale. Pour bien cerner l'intérêt de ces travaux, il faut raisonner dans le sens inverse et considérer que la base de données que l'on peut observer, celle à laquelle on a accès, est une base de données bruitée, à cause d'erreurs de saisie, d'erreurs de mesure, de calcul... et que malgré ce bruit, on veut évaluer l'intérêt d'une règle. La mesure de **level** assure que la valeur que l'on obtient sur les données observées est proche de la valeur de mesure réelle des règles.

Cette vision de la robustesse des règles d'association est assez éloignée des contextes classiques d'extraction de règles, qui sont de manière générale fondés sur la fixation de seuils et donc une vision binaire de l'intérêt. Cependant, certaines applications mettent en œuvre un classement des règles en fonction de leur valeur de mesure d'intérêt, par exemple pour la construction de classifieurs [Liu et al. 98, Li et al. 01, Yin et Han 03], ou la recherche de règles optimales au sens de la mesure [Webb et Zhang 05]. C'est dans ce contexte qu'une mesure robuste aux variations de la base peut être intéressante. Malgré tout, nos travaux s'orientent vers une autre vision de la robustesse déjà abordée par quelques auteurs.

4.1.2 La robustesse du roseau

Comme nous venons de le faire remarquer, les outils classiques d'extraction de règles d'association s'appuient sur l'utilisation de seuils de mesure et une vision binaire de la validation des règles. Dans [Ragel et Crémilleux 98], le constat est fait que la présence de valeurs manquantes dans une base peut modifier largement l'ensemble des règles que l'on peut y trouver pour un seuil de **support** et un seuil de **confiance** donnés. Les auteurs y proposent une version modifiée de ces deux mesures, mettant en jeu la notion de *base désactivée* par rapport à un motif donné. Les comptes sont fait dans l'ensemble des transactions contenant le motif. La validation expérimentale de cette approche est faite en observant la différence des règles extraites dans une première base, puis dans différentes bases dans lesquelles a été introduit du bruit : on évalue les règles disparues et les règles apparues. Cette approche expérimentale sera réutilisée par la suite dans nos travaux.

Comme on peut le constater, cette approche de la robustesse donne une certaine liberté aux règles. Il ne s'agit pas d'espérer des règles dont la mesure (ici de **confiance**) ne changera pas avec l'introduction de bruit, la présence de valeurs manquantes ou toute autre perturbation, mais plutôt de s'assurer que, malgré les perturbations, la règle que l'on a extraite possède bien une valeur de mesure au dessus du seuil préfixé. Cela est évidemment difficile à juger, car la quantité de bruit dans les données n'est pas une chose facile à évaluer bien que l'on puisse en avoir une estimation, mais peut cependant être utile dans d'autres contextes, notamment le cas des valeurs manquantes ou encore l'apparition de nouvelles transactions. On peut alors se poser la question de la quantité de perturbation qu'une règle pourrait accepter avant que la valeur de sa mesure n'atteigne – et ne passe sous – le seuil fixé.

Une fois cette notion de robustesse bien assimilée, on est en droit de se demander s'il est bien raisonnable de s'arrêter à la mesure de **confiance**. Dans [Lenca et al. 06, Vaillant et al. 06] les auteurs s'intéressent à la variation de la table de contingence supportée par une règle avant de passer sous le seuil de mesure fixé, pour un ensemble de 9 mesures étudiées par rapport à leurs dérivées. Ce qui est étudié, c'est la perte d'exemples dans la base, la question étant de savoir ce que deviennent ces exemples perdus. Ils peuvent devenir des contre-exemples, comme étudié dans [Lenca et al. 06], mais la variation peut se faire différemment et c'est ce qu'apporte [Vaillant et al. 06]. Une fois

fixée la variation de la table de contingence, on peut étudier chaque mesure comme une fonction analytique d'une variable et estimer la variation de chacune en fonction de ces exemples perdus. Ces études sont une première avancée importante dans l'étude de la robustesse car elles offrent un cadre général d'étude qui peut s'appliquer à n'importe quelle mesure. Cependant, elles sont limitées par les modèles de variations définis. En effet dans la réalité, on ne contrôle pas le bruit susceptible de figurer dans la base de données et surtout on ne sait pas ce que deviennent les exemples. Pour faire le lien avec la partie précédente, on sent bien que dans le contexte de notre cadre formel, introduire du bruit dans la base de données revient à considérer ce qui se passe autour de la projection d'une règle. On voit déjà apparaitre les premiers outils qui nous permettront de définir formellement une notion cohérente et générale de la robustesse des règles d'association.

En l'absence d'outils formels pour contrer ce problème de modèles de variations, il est possible de faire une étude expérimentale de l'influence du bruit sur les mesures d'intérêt. C'est l'étude qui a été menée dans [Azé et al. 07] pour quatre mesures : `Jaccard`, `Loevinger`, `moindre contradiction`, `TEC`. Du bruit est inséré dans les paramètres des règles intéressantes en fonction de l'un des trois modèles de variations choisis par les auteurs, que nous avons déjà rencontrés dans le chapitre 3 sur la figure 3.1 : marges fixées, antécédents fixés ou conséquents fixés. Le modèle est tiré aléatoirement pour chaque règle et l'on étudie les règles perdues et gagnées lors de l'insertion du bruit. Cette étude montre que le comportement des mesures est varié et difficilement prévisible. D'autres études expérimentales ont été menées [Cadot 05], qui interdisent de la même manière toute utilisation systématique de cette notion de robustesse, car le comportement d'une mesure sur une base de données ne peut pas être généralisé à toutes les situations. Les auteurs précisent d'ailleurs bien que cette étude est limitée par l'utilisation des modèles, mais qu'elle ouvre la voie à une définition générale, unifiée, de la notion de robustesse. C'est dans cette direction que s'inscrivent nos travaux.

4.2 UNE DÉFINITION DE LA ROBUSTESSE

Supposons que l'on cherche à évaluer les règles d'association extraites d'une base $\mathcal{DB}$ à l'aide d'une mesure d'intérêt objective m. L'utilisateur aura fixé un seuil, $m_{\downarrow}$, au dessus duquel les règles sont jugées intéressantes. Les règles ainsi sélectionnées sont cependant dépendantes de plusieurs paramètres parmi lesquels :

- le seuil $m_{\downarrow}$: l'utilisateur peut à tout moment modifier le seuil et faire ainsi apparaître, ou disparaître, un grand nombre de règles ;
- le bruit : la règle ne survivra peut-être pas à une variation des données, comme l'introduction de nouvelles transactions (taille de l'échantillon), ou bien la présence d'erreurs.

Nous proposons ici d'apporter une contribution à l'étude du second point, la fragilité de la règle par rapport aux variations des données, dans la lignée du travail de [Vaillant et al. 06]. Nous apportons une solution au problème des multiples modèles en nous concentrant sur une étude topologique des règles. Principalement, nous nous appuyons sur l'étude des contre-exemples en ne considérant que le domaine D_{cex}, que nous noterons simplement $\mathcal{D}$ dans cette partie.

Notre vision de la robustesse [Le Bras et al. 10c] est différente des définitions existantes et s'appuie sur la notion de règle limite. Ces règles limites peuvent être abstraites, au sens où elles ne sont pas nécessairement réalisées dans la base $\mathcal{DB}$. Nous définissons une distance sur les règles, $d_2(r, r')$, qui est la distance euclidienne entre les deux projections de r et r' dans $\mathcal{D}$.

Définition 8 – Règle limite – : Une règle limite est une règle d'association $r_{\downarrow}$, éventuellement abstraite, telle que $m(r_{\downarrow}) = m_{\downarrow}$. Soit r une règle d'association, on note r^* une règle limite qui minimise $d_2(r, r_{\downarrow})$ dans $\mathbb{R}^3$. Formellement,

$$r^* \in argmin\{d_2(r, r_{\downarrow}) | r_{\downarrow} \text{ règle limite}\}$$

Ce sont des règles qui, si elles étaient réalisées, seraient sélectionnées de justesse (par rapport au seuil $m_\downarrow$). Pour une règle r donnée, r^* n'est pas unique, mais son choix n'est pas déterminant pour la notion de robustesse que nous allons définir par la suite.

Puisqu'une règle limite est une règle d'association, $r_\downarrow$, projetée sur un triplet (x, y, z), celui-ci est nécessairement un élément de $\mathcal{D}$. Ainsi, $d_2(r, r^*)$ n'est pas simplement la distance de r à la surface $\mathcal{S}$ d'équation $m = m_\downarrow$, mais la distance à $\mathcal{S} \cap \mathcal{D}$. La figure 4.1 montre les deux cas à prendre en compte dans le calcul de la robustesse.

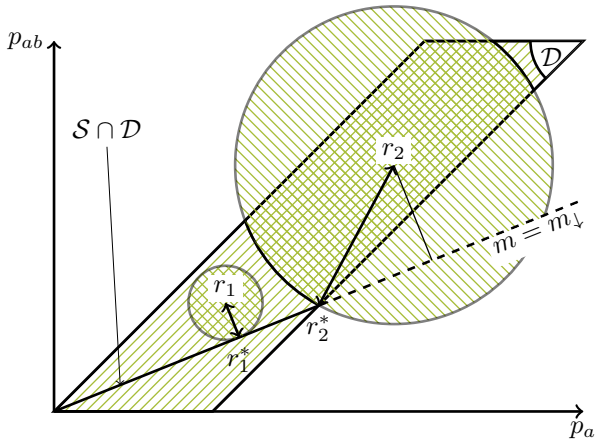


FIGURE 4.1 : Visualisation de la robustesse pour deux règles r_1 et r_2 à p_b fixé pour le cas particulier du plan $\mathcal{S}$ défini par la **confiance**.

Définition 9 – Robustesse d'une règle – : Soit m une mesure d'intérêt des règles d'association et $m_\downarrow$ un seuil prédéfini. Soit une base de données $\mathcal{DB}$ et une règle d'association r sur cette base telle que $m(r) > m_\downarrow$. On définit la robustesse de r par rapport à m et $m_\downarrow$ par :

$$\text{rob}_m(r, m_\downarrow) = \frac{d_2(r, r^*)}{\sqrt{3}}$$

Le facteur important est le numérateur $d_2(r, r^*)$, la division par $\sqrt{3}$ est une normalisation de cette quantité pour la ramener à l'intervalle $[0, 1]$. D'autres normalisations sont évidemment envisageables. S'il n'y a pas d'ambiguïté, nous noterons cette robustesse $\text{rob}(r)$. Le choix des contre-exemples comme domaine d'étude est uniquement guidé par l'historique de la robustesse des règles d'association. On pourrait tout aussi bien étudier la robustesse dans le cadre des exemples, mais les résultats seraient similaires : le passage d'un domaine à un autre se fait par un changement de repère, non orthonormé, ce qui induirait une constante multiplicative supplémentaire dans les calculs de distance. Notons cependant que dans le cas d'une étude dans le domaine lié à la **confiance**, il s'agit d'un changement de variable plus complexe, pour lequel il serait difficile d'établir un lien avec notre étude. Nous allons montrer dans le paragraphe suivant en quoi cette définition est une notion de robustesse et nous donnerons quelques propriétés de la robustesse ainsi définie.

4.3 PROPRIÉTÉS DE LA ROBUSTESSE

Commençons par justifier l'appellation de robustesse. Considérons une base $\mathcal{DB}$ et une règle d'association $r : \mathtt{A} \rightarrow \mathtt{B}$ dans $\mathcal{DB}$ telle que $m(r) > m_\downarrow$. On note $(p_{\mathtt{A}\neg\mathtt{B}}, p_{\mathtt{A}}, p_{\mathtt{B}})$ ses valeurs de **support**

associées. Introduisons du bruit dans la base $\mathcal{DB}$ afin d'obtenir une base $\mathcal{DB}'$ dans laquelle la règle r' : $\mathtt{A} \rightarrow \mathtt{B}$ est caractérisée par $(p'_{\mathtt{A}\neg\mathtt{B}}, p'_{\mathtt{A}}, p'_{\mathtt{B}})$: les motifs restent identiques, mais leur **support** change. On suppose que l'on a des connaissances sur le bruit qui nous permettent d'assurer :

$$\begin{aligned} |p'_{\mathtt{A}\neg\mathtt{B}} - p_{\mathtt{A}\neg\mathtt{B}}| &\leq \frac{d_2(r, r^*)}{\sqrt{3}} \\ |p'_{\mathtt{A}} - p_{\mathtt{A}}| &\leq \frac{d_2(r, r^*)}{\sqrt{3}} \\ |p'_{\mathtt{B}} - p_{\mathtt{B}}| &\leq \frac{d_2(r, r^*)}{\sqrt{3}} \end{aligned}$$

Ainsi, $d_2(r, r') = \sqrt{|p'_{\mathtt{A}\neg\mathtt{B}} - p_{\mathtt{A}\neg\mathtt{B}}|^2 + |p'_{\mathtt{A}} - p_{\mathtt{A}}|^2 + |p'_{\mathtt{B}} - p_{\mathtt{B}}|^2} \leq d_2(r, r^*)$ et donc, par définition de r^*, $m(r') > m_{\downarrow}$.

rob(r) traduit donc la quantité de bruit acceptée par la règle tout en restant de qualité. C'est une notion de sécurité, qui permet d'affirmer que si le bruit est suffisamment contrôlé, la règle restera intéressante. L'inverse n'est cependant pas vrai car une règle peu robuste pourra évoluer de manière à devenir plus intéressante.

Cette notion de robustesse est particulièrement facile à comprendre dans le cadre de bruit inséré par transactions. En effet, si l'on insère le bruit dans moins de rob(r)% des transactions, la règle r restera intéressante par rapport à $m_{\downarrow}$. Cela peut correspondre à une base de données qui évolue, avec de nouvelles transactions, ou bien à des données insérées avec une possibilité d'erreur sur chaque transaction. Lorsque le bruit est inséré par attributs [Azé et Kodratoff 02, Azé et al. 03], le contrôle est plus difficile à assurer.

Inversement, si l'on sait que la base de données contient un certain pourcentage de bruit et que l'on extrait de cette base bruitée des règles dont la robustesse assure l'intérêt pour ce pourcentage de bruit, alors l'utilisateur est assuré que ces règles sont effectivement intéressantes dans la base *idéale* non bruitée. On peut par exemple penser à un système de capteurs physiques qui possèderaient des marges d'erreurs garanties par le constructeur, ou bien au cas de bases dont la qualité a été évaluée [Berti-Equille 07].

Propriété 2 : La robustesse $rob(r)$ présente des caractéristiques analytiques intéressantes :
- la robustesse d'une règle est un réel de $[0, 1]$ [1] ;
- $\text{rob}_m(r, m_{\downarrow}) = 0$ si r est une règle limite, c'est-à-dire si $m(r) = m_{\downarrow}$;
- si la mesure m, vue comme fonction de 3 variables, est continue de $\mathcal{D} \subset \mathbb{R}^3$ dans $\mathbb{R}$, alors la robustesse est décroissante par rapport à $m_{\downarrow}$;
- la robustesse est continue par rapport à r.

Ces propriétés permettent de déduire des comportements attendus de la notion de robustesse. Ainsi, plus le seuil est fixé haut, moins les règles sont robustes et plus il est important d'avoir des données fiables. D'autre part, deux règles dont les projetés sont proches ont des robustesses équivalentes.

4.4 Applications pratiques de la robustesse

La robustesse définie précédemment peut avoir trois applications immédiates. La première concerne la classification de règles : cette mesure permet de comparer deux règles entre elles et

1. Il faut noter que la valeur $\text{rob}_m(r, m_{\downarrow}) = 1$ est une valeur théorique qui correspond à une configuration très particulière de r, $m_{\downarrow}$ et de m. En pratique dans nos expériences nous n'avons pas rencontré cette valeur.

donc d'établir un pré-ordre sur l'ensemble de règles concerné. La seconde concerne le filtrage des règles situées au-dessus d'un certain seuil fixé par l'utilisateur. Cependant, au même titre que le seuil de mesure, le seuil de robustesse peut s'avérer délicat à fixer. La figure 4.2 montre que l'on

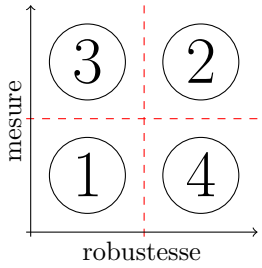


FIGURE 4.2 : Zones remarquables entre robustesse et mesure

peut distinguer quatre comportements types. S'il est facile d'arbitrer entre deux règles dont l'une est robuste et intéressante (2) et l'autre fragile et peu intéressante (1), la tâche est moins évidente entre deux règles (robuste/peu intéressante) (4) et (fragile/intéressante) (3). Vaut-il mieux avoir une règle très intéressante, mais très dépendante du bruit dans les données, ou bien est-il préférable d'avoir une règle très robuste, qui supportera des changements dans les données, mais dont la mesure est proche du seuil fixé ? Les réponses à cette question dépendent évidement de la situation pratique et de la confiance que l'utilisateur a dans la qualité des données.

Enfin, dans le cas de mesures non statistiques, c'est-à-dire des mesures qui ne dépendent que des fréquences relatives observées de `A`, `B` et `AB`, notre mesure de robustesse, qui indique une sécurité, permet aussi d'évaluer le nombre de nouvelles transactions que la base peut accepter de telle sorte que la règle reste intéressante, comme le montre le raisonnement suivant.

Soit une règle $r : \texttt{A} \rightarrow \texttt{B}$, étudions la variation de ses paramètres lorsque la taille des données varie d'une quantité δ_n. On note par exemple n'_a la nombre de lignes contenant le motif `A` dans la nouvelle base, et n_a la quantité originale.

$$\begin{array}{ccccc} n_a & \leq & n'_a & \leq & n_a + \delta_n \\ p_a & \leq & \frac{n+\delta_n}{n} p'_a & \leq & p_a + \frac{\delta_n}{n} \\ -\frac{\delta_n}{n} p'_a & \leq & p'_a - p_a & \leq & \frac{\delta_n}{n}(1 - p'_a) \end{array}$$

On en déduit que $|p_a - p'_a| \leq \frac{\delta_n}{n} \max\{p'_a, 1 - p'_a\}$ et donc que $|p_a - p'_a| \leq \frac{\delta_n}{n}$. Cette inégalité vaut aussi pour les autres quantités ce qui nous permet d'écrire que

$$d(r, r') = \sqrt{|p_{ab} - p'_{ab}|^2 + |p_a - p'_a|^2 + |p_b - p'_b|^2} \leq \sqrt{3}\frac{\delta_n}{n}$$

Il suffit donc que l'on ait l'inégalité $\frac{\delta_n}{n} \leq rob(r)$ pour assurer que r', c'est-à-dire la règle $\texttt{A} \rightarrow \texttt{B}$ dans la base augmentée, soit intéressante.

CONCLUSION

Nous avons défini une nouvelle notion de robustesse s'appuyant sur notre cadre formel d'étude. Cette notion de robustesse est très visuelle et intuitive, et possède de bonnes propriétés que l'on est en droit d'attendre de la robustesse des règles d'association. Cependant, son calcul n'est pas simple dans le cas général, car le problème sous-jacent est un problème d'optimisation quadratique sous des contraintes quelconques (les mesures d'intérêt peuvent prendre bien des formes !). Nous allons cependant voir dans les chapitres suivants que certaines familles de mesures se laissent mieux aborder par le calcul de la robustesse.

5 Cas des mesures planes

Parmi les mesures qui se prêtent bien au calcul se trouvent les mesures que nous appellerons planes. Ces mesures associées à un seuil divisent l'espace des règles en deux par un plan et nous autorisent donc une étude formelle de la robustesse des règles. Nous proposons ici une étude en profondeur de ces mesures planes.

5.1 Évaluer la robustesse

La méthode que nous proposons fait naturellement appel à un calcul de distance à une surface sous certaines contraintes. Il existe un certain nombre de mesures pour lesquelles le calcul de la distance se ramène à un calcul de distance à un plan. Nous nous intéressons ici uniquement à ces mesures, que nous appelons mesures planes. Les mesures plus complexes (e.g. **`Klosgen`**, **`force collective`**, **`spécificité relative`**) demandent de recourir à des techniques d'analyse numérique plus poussées et seront étudiées, pour certaines, dans le chapitre 6.

> **Définition 10 – Mesure plane – :** Une mesure d'intérêt m est dite plane si la surface définie par $m(r) = m_{\downarrow}$ est un plan.

C'est en particulier le cas de mesures telles que **`Sebag-Shoenauer`**, **`TEC`**, **`Jaccard`**, **`moindre contradiction`**, **`précision`**, **`rappel`**, **`spécificité`**. Ces mesures possèdent donc des lignes de niveau qui forment des plans. La géométrie euclidienne permet alors de calculer la distance à un plan $\mathcal{P} : ax + by + cz + d = 0$ d'une règle r de coordonnées (x_1, y_1, z_1) :

$$d(r_1, \mathcal{P}) = \frac{|ax_1 + by_1 + cz_1 + d|}{\sqrt{a^2 + b^2 + c^2}}$$

Il reste cependant à prendre en compte que r^* doit appartenir au domaine $\mathcal{D}$. C'est donc en fait la distance au polygone intersection $\mathcal{P} \cap \mathcal{D}$ qui nous intéresse réellement. Encore une fois, ce calcul est aisément réalisable : il suffit pour cela de déterminer les points formant les sommets de ce polygone (convexe), puis de calculer la distance à chaque côté (en tant que segment). La distance au périmètre du polygone sera la plus petite de ces distances. On obtient donc l'algorithme suivant de calcul de la robustesse dans le cas d'une mesure plane :

- Trouver $r^{\perp}$, projection orthogonale de r sur $\mathcal{P}$;
- Si $r^{\perp} \in \mathcal{D}$, $r^* = r^{\perp}$ et renvoyer $d_2(r, r^*)$;
- Sinon, renvoyer la distance au périmètre du polygone intersection.

Nous avons décidé lors de nos expérimentations de nous concentrer sur ce type de mesures, car elles permettent d'obtenir des résultats précis ne faisant pas appel à des algorithmes d'approximation. Il existe un grand nombre de mesures planes. Dans [Le Bras et al. 10c] nous en étudions 5. Nous étendons ce nombre aux 15 mesures planes que se trouvent parmi les 42 mesures de cette thèse. Le tableau 5.1 détaille l'ensemble de ces mesures, ainsi que le plan qu'elles forment.

mesure	plan
confiance	$x - (1 - m_0) \cdot y = 0$
moindre contradiction	$2x - y + m_0 z = 0$
couverture	$y - m_0 = 0$
Czekanowski	$2x - (2 - m_0) \cdot y + m_0 z = 0$
gain	$x - (1 - m_0) \cdot y + m_0 = 0$
Ganascia	$2x - (1 + m_0) \cdot y = 0$
Jaccard	$(1 + m_0) \cdot x - y + m_0 z = 0$
Kulczynski	$(2 + m_0) \cdot x - (1 + m_0) \cdot y + m_0 z = 0$
précision	$2x - y + z + m_0 - 1 = 0$
prevalence	$z - m_0 = 0$
rappel	$x - y + m_0 z = 0$
Sebag-Shoenauer	$(1 + m_0) \cdot x - y = 0$
spécificité	$x - m_0 y + z - (1 - m_0) = 0$
support	$x - y + m_0 = 0$
TEC	$(2 - m_0) \cdot x - (1 - m_0) \cdot y = 0$

TABLE 5.1 : Ensemble des mesures planes de notre échantillon.

Approfondissons le cas de la **confiance**. Dans une paramétrisation par les contre-exemples, le plan défini par le seuil de **confiance** $conf_\downarrow$ est $\mathcal{P} : x - (1 - conf_\downarrow)y = 0$. La distance à ce plan d'une règle r_1 de projection (x_1, y_1, z_1) et de **confiance** $conf(r_1) > conf_\downarrow$ sera alors donnée par

$$d(r_1, \mathcal{P} = y_1 \frac{conf(r_1) - conf_\downarrow}{\sqrt{1 + (1 - conf_\downarrow)^2}} \tag{5.1}$$

La robustesse dépend donc, à $conf_\downarrow$ fixé, de deux paramètres : y_1, le **support** de l'antécédent et $conf(r_1)$, la mesure de la règle. Ainsi, deux règles ayant la même **confiance** peuvent avoir des robustesses très différentes. De même, deux règles ayant la même robustesse peuvent avoir des valeurs de **confiance** différentes. Il ne sera donc pas étonnant d'observer des règles de mesure faible et de robustesse élevée, tout comme des règles de mesure élevée, mais de robustesse très faible. En effet, il est possible de découvrir une règle qui soit très intéressante, mais très fragile.

Exemple 4 – Robustesse – : Considérons une base fictive de 100000 transactions. On note n_x le nombre d'occurrences du motif `X`. Dans cette base, on trouve une première règle r_1 : `A` $\rightarrow$ `B` telle que $n_a = 100$ et $n_{a\bar{b}} = 1$. Sa **confiance** est de 99%. Mais sa robustesse comme définie précédemment au seuil de **confiance** 0.8 est $\text{rob}(r_1) = 0.0002$. Une seconde règle r_2 : `C` $\rightarrow$ `D` présente les caractéristiques suivantes : $n_c = 50000$ et $n_{c\bar{d}} = 5000$. Sa **confiance** n'est que de 90% mais sa robustesse de 0.05. Elle présente pourtant plus de contre-exemples proportionnellement à son antécédent que r_1 et pourrait être jugée, à tort selon la robustesse, moins fiable.
Dans le premier cas, la règle limite la plus proche a comme caractéristiques $n^*_a = 96$ et $n^*_{a\bar{b}} = 19$. La règle originale ne supporte donc que de très petites variations sur les lignes de la base de données. La seconde règle a quant à elle une plus proche règle limite de paramètres $n^*_c = 49020$ et $n^*_{c\bar{d}} = 9902$ et la règle originale accepte donc de l'ordre du millier de changements. La règle r_2 est donc beaucoup moins sensible au bruit que la règle r_1. Pourtant, c'est la règle r_1 qui présentait le plus fort intérêt selon la mesure de **confiance**.

Il convient donc de se poser la question du réel intérêt d'une règle : comment doit-on arbitrer entre une règle d'association très bien évaluée par les mesures, mais de robustesse très faible et une

règle moins bien évaluée, mais dont la robustesse nous assure une plus grande fiabilité vis-à-vis du bruit ?

5.2 Mise en oeuvre de la robustesse

Nous présentons ici les résultats obtenus sur 4 bases et pour 5 mesures planes parmi les 15. Dans un premier temps, nous présentons le protocole expérimental choisi, puis nous étudions les graphiques obtenus afin de mettre en évidence les liens entre mesure et robustesse. Enfin, nous analysons l'effet du bruit sur les règles d'association.

5.2.1 Protocole expérimental

5.2.1.1 L'extraction des règles

Les 5 mesures planes sélectionnées dans le cadre de nos expériences sont : **`confiance`**, **`Jaccard`**, **`Sebag-Shoenauer`**, **`TEC`** et **`spécificité`**. Le tableau 5.2 rappelle leur écriture en fonction des contre-exemples, ainsi que le plan qu'elles définissent.

nom	formule	plan
`confiance`	$\frac{p_{A}-p_{A\neg B}}{p_{A}}$	$x-(1-m_0)y=0$
`Jaccard`	$\frac{p_{A}-p_{A\neg B}}{p_{B}+p_{A\neg B}}$	$(1+m_0)x-y+m_0z=0$
`Sebag-Shoenauer`	$\frac{p_{A}-p_{A\neg B}}{p_{A\neg B}}$	$(1+m_0)x-y=0$
`spécificité`	$\frac{1-p_{B}-p_{A\neg B}}{1-p_{A}}$	$x-m_0y+z=1-m_0$
`TEC`	$1-\frac{p_{A\neg B}}{p_{A}-p_{A\neg B}}$	$(2-m_0)x-(1-m_0)y=0$

TABLE 5.2 : Les mesures planes retenues avec leur écriture par rapport aux contre-exemples, le plan défini par une valeur m_0

Pour effectuer nos expériences, nous nous sommes appuyés sur 4 bases de données usuelles [Asuncion et Newman 07]. La base *census* a été discrétisée et nous en avons extrait, ainsi que de *mushroom*, des règles de classe, c'est-à-dire où le conséquent est contraint. Les bases *chess* et *connect* ont été binarisées afin d'en extraire des règles d'association sans contrainte. Les règles ont ensuite été extraites, grâce à l'implémentation d'APRIORI de [Borgelt et Kruse 02], de manière à obtenir des règles de **`support`** non nul, de **`confiance`** supérieure à 0.8 et de taille variable en fonction de la base. L'ensemble de ces informations est synthétisé dans la table 5.3. Nous avons ainsi obtenu des règles intéressantes, sans exclure les pépites de connaissance, mais tout en gardant un nombre de règles raisonnable. Notons que [Borgelt 03] présente de manière détaillée les caractéristiques de ces différentes bases.

base	attributs	transactions	type	taille	# règles
census	137	48842	classe	5	244487
chess	75	3196	sans contrainte	3	56636
connect	129	67557	sans contrainte	3	207703
mushroom	119	8124	classe	4	42057

TABLE 5.3 : Bases de données utilisées dans nos expériences. L'avant-dernière colonne fixe la taille maximum des règles extraites.

nom	seuil $m_\downarrow$
`confiance`	0.984
`Jaccard`	0.05
`Sebag-Shoenauer`	10
`spécificité`	0.5
`TEC`	0.95

TABLE 5.4 : Les mesures planes retenues et le seuil choisi.

5.2.1.2 Calcul de la robustesse

Pour chaque ensemble de règles et chaque mesure, nous avons appliqué la même méthode de calcul de la robustesse des règles d'association extraites des bases. Dans un premier temps, nous avons sélectionné uniquement les règles dont la mesure était supérieure à un seuil prédéfini. Nous avons choisi de fixer ce seuil définitivement pour toutes les bases aux valeurs indiquées table 5.4. Ces seuils ont été fixés après observation du comportement des mesures sur les règles extraites de la base *mushroom*, afin d'obtenir des règles intéressantes et des règles inintéressantes dans des proportions équilibrées.

Nous avons ensuite implémenté un algorithme s'appuyant sur la description faite dans la section 5.1 pour le cas spécifique des mesures planes, et calculant ainsi la robustesse d'une règle par rapport à une mesure et son seuil. Nous obtenons en sortie une liste de règles avec leur `support`, leur robustesse et leur mesure. La complexité de cet algorithme dépend essentiellement du nombre de règles à analyser mais nous ne nous attarderons pas ici dessus, car il s'agit simplement d'analyser un ensemble de règles. Ces résultats nous permettent d'obtenir des graphiques mesure/robustesse que nous analyserons dans la partie 5.2.2.

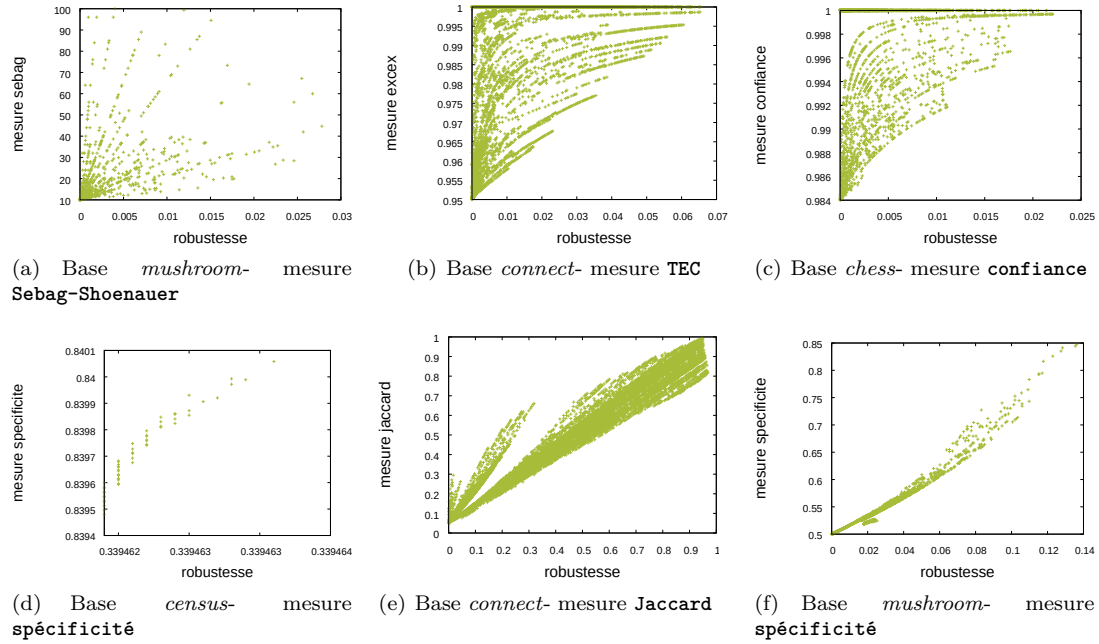
5.2.1.3 Insertion du bruit

Comme indiqué précédemment, nous analysons l'influence du bruit sur les règles en fonction de leur robustesse. Nous avons donc mis en place une procédure d'insertion de bruit dans une base de données. Notre choix s'est porté sur un bruit introduit par ligne. Nous avons décidé d'introduire du bruit dans 5% des lignes de chaque base en sélectionnant les lignes bruitées de manière aléatoire et en modifiant de manière aléatoire les valeurs des attributs de ces lignes (tirage équiprobable parmi les valeurs apparues pour chaque attribut). Une fois le bruit inséré nous calculons les nouvelles valeurs de `support` des règles de l'ensemble initial. Nous extrayons les règles intéressantes au sens des mesures données et évaluons leur robustesse. L'étude du bruit est discutée dans la partie 5.2.3

5.2.2 Analyse de la robustesse

Nous avons obtenu, pour chaque base et chaque mesure, des données nous permettant de visualiser la mesure d'une règle en fonction de sa robustesse. La figure 5.1 propose un échantillon représentatif des résultats, dans le sens ou l'allure des graphiques est sensiblement la même pour toutes les bases, pour une mesure donnée. Plusieurs points peuvent être relevés.

On observe tout d'abord que la mesure possède un caractère globalement croissant avec la robustesse. Cependant si l'on observe précisément, il est bien visible qu'un très grand nombre de règles sont dominées au sens de la mesure par des règles pourtant moins robustes. Cela est particulièrement marqué dans le cas de la mesure de `Sebag-Shoenauer`, puisqu'une règle de mesure de `Sebag-Shoenauer` de valeur 100 peut être beaucoup moins robuste (10^{-4}) qu'une règle de mesure 20 ($2 \cdot 10^{-3}$). La seconde supportera vingt fois plus de changements que la première.



(a) Base *mushroom*- mesure **Sebag-Shoenauer**

(b) Base *connect*- mesure **TEC**

(c) Base *chess*- mesure **confiance**

(d) Base *census*- mesure **spécificité**

(e) Base *connect*- mesure **Jaccard**

(f) Base *mushroom*- mesure **spécificité**

FIGURE 5.1 : Valeur de la mesure en fonction de la robustesse pour différents couples base/mesure.

5

Ensuite, nous observons des lignes de niveau dans la plupart des cas. **Sebag-Shoenauer** et **Jaccard** présentent des droites de niveau, la **confiance** et **TEC** présentent des courbes concaves et la **spécificité** semble présenter des courbes convexes.

Traitons le cas particulier des courbes relatives à la **confiance**. Une telle démonstration peut se faire pour les autres mesures. L'équation (5.1) montre l'écriture de la robustesse en fonction de la mesure, où y représente $p_{\mathtt{A}}$. Exprimons $m(r)$ en fonction de d et de x en utilisant le fait que $p_{\mathtt{A}} = \frac{p_{\mathtt{A}\to\mathtt{B}}}{1-conf(r)}$:

$$m(r) = \frac{m_{\downarrow} + \sqrt{1 + (1 - m_{\downarrow})^2} * \frac{d}{x}}{1 + \sqrt{1 + (1 - m_{\downarrow})^2} * \frac{d}{x}} \tag{5.2}$$

Ainsi, à x constant, c'est-à-dire à nombre de contre-exemples constant, les règles se trouvent sur une courbe bien définie, concave et croissante. Les lignes de niveau observées dans le cas de la **confiance** sont donc formées par des règles ayant le même nombre de contre-exemples.

Un comportement paraît récurrent quel que soit la mesure : il ne semble pas exister de règle qui soit à la fois très proche du seuil de mesure et très robuste. Seule **Sebag-Shoenauer** se distingue un peu de ce comportement. Nous pensons que cela est fortement lié au fait que les mesures étudiées ici sont des mesures planes. En effet, dans ce cas, la variation des lignes de niveau est constante et ne présente pas de fortes pentes.

5.2.3 Étude de l'influence du bruit

Nous allons ici étudier les liens entre l'introduction de bruit et l'évolution des ensembles de règles par rapport à la robustesse. Nous avons décidé de créer 5 bases bruitées à partir de chaque base initiale, puis pour chaque base bruitée (méthode présentée en 5.2.1.3), d'étudier la robustesse des règles qui sont conservées et des règles qui ont disparu. Pour valider notre notion de robustesse,

nous attendons de ces expériences d'observer une robustesse plus faible dans l'ensemble des règles disparues que dans l'ensemble des règles conservées. La table 5.5 montre les résultats obtenus en

(a) mesure **TEC**

base	disparues	conservées
census	0.83e-6	0.79e-6
chess	1.16e-3	0.96e-2
connect	5.26e-4	7.72e-3
mushroom	9.4e-5	6.6e-4

(b) mesure de **Sebag-Shoenauer**

base	disparues	conservées
census	1.53e-6	1.53e-6
chess	1.63e-3	1.72e-2
connect	8.38e-4	1.42e-2
mushroom	1.28e-4	1.22e-3

(c) mesure de **spécificité**

base	disparues	conservées
census	0	0.19
chess	7.23e-5	8.76e-2
connect	0	1.2e-1
mushroom	2.85e-4	1.37e-2

(d) mesure de **confiance**

base	disparues	conservées
census	2.61e-7	2.61e-7
chess	5.59e-4	3.77e-3
connect	2.16e-4	2.73e-3
mushroom	5.51e-5	2.34e-4

(e) mesure de **Jaccard**

base	disparues	conservées
census	0	0
chess	3.2e-4	1.69e-1
connect	1.94e-3	1.43e-1
mushroom	3.20e-4	1.90e-2

TABLE 5.5 : Comparaison entre les robustesses moyennes des règles disparues et conservées pour les différentes mesures

faisant la moyenne des robustesses des règles au sein des différents ensembles de règles obtenus, sur les 5 bruitages.

Dans la plupart des cas apparaît un facteur 10 entre la robustesse des règles conservées et des règles disparues. Seul le cas de la base de données *census* pour les mesures **TEC**, **Sebag-Shoenauer** et **confiance** ne confirme pas ce résultat, mais le comportement de *census* ne contredit pas pour autant notre théorie. En effet, les robustesses initiales issues de la base de données *census* sont de l'ordre de 10^{-6} et sont donc vulnérables à 5% de bruit (de l'ordre de 10^{-2}). Il est donc normal que toutes les règles soient susceptibles de devenir inintéressantes.

A l'opposé, la mesure de **spécificité** fait apparaître un comportement commun à la base *census* et à la base *connect*. Pour ces deux bases, aucune règle ne disparaît lorsque l'on introduit 5% de bruit. Si l'on regarde la moyenne de la robustesse des règles conservées, on s'aperçoit qu'elle est bien supérieure aux 5%, ce qui signifie que toutes les règles sont protégées. Dans le cas de la base *census*, la plus petite mesure de **spécificité** relevée est de 0.839, donc bien au dessus du seuil fixé. Il n'est donc pas étonnant que les règles de la base *census* soient protégées du bruit. Dans le cas de la base *connect*, la moyenne des mesures observées est de 0.73 avec un écart type de 0.02. la plus petite mesure de **spécificité** est de 0.50013 et correspond à une robustesse de $2.31e-5$. Pourtant elle a bien été sauvée dans les 5 tirages de bruit effectués. Cela permet de souligner le fait que notre définition de la robustesse correspond à la définition d'un périmètre de sécurité autour de la règle. Si la règle change et sort de ce périmètre, son évolution peut se faire librement dans l'espace sans atteindre la surface seuil. Cependant, le risque persiste.

5.3 ROBUSTESSE ET STATISTIQUES

Dans les sections précédentes, nous avons défini la robustesse d'une règle par sa capacité à résister aux variations des données, comme par exemple à une apparition/disparition d'exemples, de manière à ce que son évaluation $m(r)$ à l'aide d'une mesure d'intérêt reste au-dessus d'un seuil fixé $m_{\downarrow}$. Cette définition est très similaire à la notion de significativité en statistiques. Dans cette section, nous tentons d'étudier la relation entre ces deux notions.

5.3.1 Règle significative

L'utilisation des statistiques nous oblige à distinguer différentes notions : $m(r)$ est la valeur empirique de la mesure de la règle, calculée sur un échantillon de données précis, c'est donc la valeur observée d'une variable aléatoire $M(r)$, dont $\mu(r)$ est la valeur théorique. Une règle statistiquement significative r est une règle pour laquelle on peut considérer que $\mu(r) > m_{\downarrow}$.

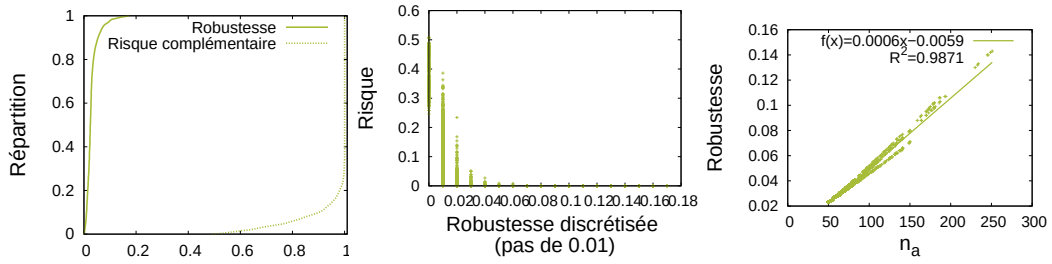
Généralement, pour chaque règle, l'hypothèse nulle $H_0 : \mu(r) = m_{\downarrow}$ est testée contre l'hypothèse alternative $H_1 : \mu(r) > m_{\downarrow}$. Une règle est considérée comme significative au niveau α_0 (erreur de première espèce, caractérisation des faux positifs) si sa p-valeur est au plus α_0. Nous rappelons que la p-valeur d'une règle dont la mesure empirique est $m(r)$ est définie par $\mathbb{P}(M(r) \geq m(r)|H_0)$.

Cependant et comme l'ont montré [Dudoit et van der Laan 07] dans le cas général et [Lallich et al. 07a] dans le cas particulier des règles d'association, étant donné le grand nombre de tests qui doivent être menés et le grand nombre de fausses découvertes qui en découlent, les p-valeurs doivent être adaptées. L'écriture algébrique de la p-valeur ne peut être déterminée que si la loi de $M(r)$ sous H_0 est, au moins approximativement, connue. C'est le cas de la mesure de `confiance`, pour laquelle [Lerman et al. 81, Lallich et al. 07b] ont établi un modèle de la distribution de $M(r)$ sous H_0. Cependant ce modèle suppose que les marges soient fixées, ce qui peut paraître simpliste et pour une grande majorité de mesures, il est impossible d'établir cette loi. La stratégie que nous avons retenue afin d'évaluer la p-valeur dans ces cas précis est d'avoir recours au bootstrap [Efron 79]. Cette technique permet, à partir d'un échantillon restreint, d'estimer la distribution réelle par tirage successifs et indépendants dans la base observée et de même taille que celle-ci. Pour nos expériences, nous tirons avec remise 400 échantillons de taille n dans la population observée, de taille n également. Le risque est estimé par la proportion d'échantillons dans lesquels la règle tombe sous le seuil d'intérêt fixé. Dans notre cas, nous avons de plus décidé de lisser la distribution grâce à la loi normale. Seules les règles dont le risque est évalué à moins de α_0 sont considérées comme significatives.

Le seuil de risque indique que, dans une base possédant n transactions, $n\alpha_0$ règles peuvent être sélectionnées sans être réellement significatives. Dans le cas d'une base de données contenant 10000 transactions et un seuil de risque à 5%, on obtiendrait donc 500 fausses découvertes. Afin de résoudre le problème des fausses découvertes, la méthode de Benjamini et Liu [Benjamini et Liu 99] mérite notre attention. Les valeurs de risque sont triées par ordre croissant et nommées $p_{(i)}$ où i est le rang. Une règle est retenue si son risque est tel que $p_{(i)} \leq i\frac{\alpha_0}{n}$. Cette procédure permet de contrôler la proportion de règles sélectionnées par erreur sous l'hypothèse de l'indépendance des données et est compatible avec des données positivement corrélées.

5.3.2 Comparaison des deux approches sur un exemple

Afin de mieux cerner les différences entre ces deux approches de la robustesse des règles, nous comparons la robustesse des règles avec le risque complémentaire issu du bootstrap. Nos expériences s'appuient sur la base de données *solar flare* issue du site de l'UCI. Nous détaillons ici le cas de deux mesures particulières : `confiance` et `Jaccard`, pour lesquelles une approche algébrique est soit problématique (fixation des marges), soit impossible. Dans un premier temps, nous extrayons les règles grâce à une implémentation de l'algorithme classique Apriori dont nous avons fixé les seuils arbitrairement à 0.13 pour le `support` et 0.85 pour la `confiance`. Cela produit 9890 règles que nous obtenons avec leur `confiance` et leur robustesse. La technique du bootstrap sur 400 itérations permet de calculer le risque, pour chaque règle, de passer sous le seuil d'intérêt. Il est important de remarquer que si, parmi les 9890 règles, 8481 ont un risque bootstrap de moins de 5%, seules 8373 d'entre elles sont conservées en appliquant la stratégie de Benjamini et Liu.

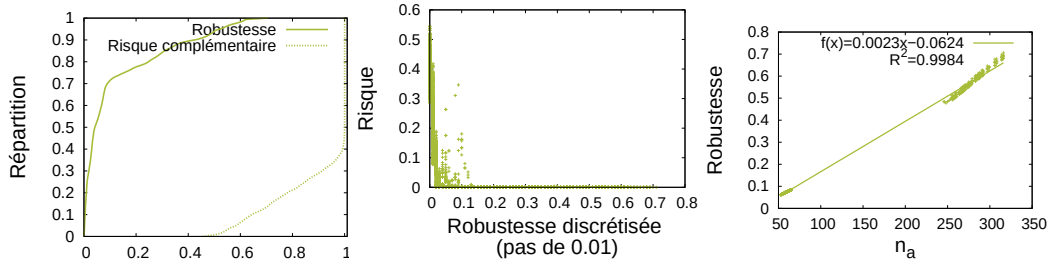


(a) Fonctions de répartition empiriques (b) Risque et robustesse (c) Robustesse et n_a

FIGURE 5.2 : Cas de la **confiance**

La figure 5.2(a) montre la fonction de répartition empirique de la robustesse et le risque complémentaire issu du bootstrap. Elle met en évidence le fait que la robustesse est clairement plus discriminante que le risque complémentaire, spécialement pour les règles ayant une valeur de mesure élevée. La figure 5.2(b) représente le risque par rapport à la robustesse (discrétisée par pas de 0.01). Elle montre que le risque est globalement corrélé à la robustesse.

Cependant, il faut insister sur le fait que les résultats des deux approches sont bien différents. D'un côté, le processus de Benjamini et Liu retourne 1573 règles non significatives dont la robustesse est évaluée à moins de 0.025. D'un autre côté, 3616 règles parmi les règles significatives ont une robustesse inférieure à 0.05. En complément, notons que les 2773 règles logiques prennent des valeurs de robustesse très variées entre 0.023 et 0.143. Finalement, comme le montre la figure 5.2(c), la robustesse d'une règle semble linéairement corrélée avec sa couverture.



(a) Fonctions de répartition empiriques (b) Risque et robustesse (c) Robustesse et n_a

FIGURE 5.3 : Cas de la mesure de **Jaccard**

Les résultats obtenus avec la mesure de **Jaccard** sont similaires. Le seuil de **support** est fixé à 0.0835 et le seuil de mesure est fixé à 0.3. Nous obtenons 6066 règles dont 4059 sont déclarées significatives à 5% par la technique de bootstrap (400 itérations) et 3933 par le processus de Benjamini et Liu (il y a donc 2133 règles non significatives). Encore une fois, l'étude des fonctions de répartition de la figure 5.3(a) montre que la robustesse est plus discriminante que le risque complémentaire du bootstrap pour les règles les plus intéressantes. De la même manière, la figure 5.3(b) met en évidence que le risque pour la mesure de **Jaccard** est globalement corrélé à la robustesse, tout en montrant, encore une fois, des différences significatives entre les deux approches. Les règles significatives pour le processus de Benjamini et Liu ont une robustesse inférieure à 0.118 tandis que les règles significatives au seuil de 5% ont une robustesse répartie entre 0.018 et 0.705, ce qui représente un intervalle considérable.

Il y a 533 règles ayant une valeur de mesure de **Jaccard** supérieure à 0.8. Toutes ont un risque

complémentaire égal à 0 et leur robustesse varie entre 0.062 et 0.705. La figure 5.3(c) montre que la robustesse de la mesure de `Jaccard` est linéairement corrélée à la couverture de la règle pour ses grandes valeurs de mesure.

En conclusion, l'étude par la technique du bootstrap de l'erreur de première espèce a l'inconvénient de ne pas être aussi discriminante que la robustesse, en particulier pour de grandes valeurs de n, ce qui est le cas en fouille de données. De plus, une étude statistique demande d'accepter l'hypothèse que les données soient un échantillon tiré aléatoirement à partir de la population entière, ce qui n'est pas souvent le cas en fouille de données. Finalement, la robustesse que nous proposons pour une mesure donnée représente de manière plus précise la stabilité d'une règle intéressante.

Conclusion

Les mesures planes permettent une étude formelle de la robustesse des règles d'association et nous ont permis de valider expérimentalement nos outils théoriques en attestant de leur utilité dans la prévision de la stabilité d'une règle d'association. Cependant, l'ensemble des mesures planes reste assez restreint, et nous allons donc proposer une méthode de calcul de la robustesse pour un autre ensemble de mesures, contenant en particulier les mesures planes. Cet ensemble est détaillé dans le chapitre suivant et sera appelé l'ensemble des mesures quadratiques.

6 Cas des mesures quadratiques

Dans le cas général, il est difficile de résoudre le problème du calcul de la robustesse qui se résume à un problème d'optimisation sous des contraintes d'égalité quelconques. Comme nous l'avons vu, le cas des mesures planes permet de fournir une solution algébrique exacte. Nous allons ici détailler le cas d'une autre famille de mesures que nous appelons mesures quadratiques. Ces mesures demandent de faire appel à un certain nombre d'outils mathématiques.

6.1 Introduction

Les calculs que nous présentons ici ont été développés avant d'en trouver une confirmation d'une partie sur le site de Dave Eberly [1] et dans son livre [Schneider et Eberly 02] [2]. Notre contribution par rapport à ces travaux consiste à prendre en compte la contrainte supplémentaire du domaine. Il s'agit de proposer une méthode de résolution de la robustesse dans le cas des mesures quadratiques, c'est-à-dire des mesures pour lesquelles la surface d'équation $m = m_{\downarrow}$ définit une quadrique d'équation

$$ {}^t\vec{x}.A.\vec{x} + {}^t\vec{b}.\vec{x} + c = 0 \tag{6.1} $$

où x et b sont des vecteurs d'un espace réel à trois dimensions, A est une matrice réelle $(3,3)$ et c est un nombre réel. A, b et c sont des éléments caractéristiques d'une mesure donnée associée au seuil $m_{\downarrow}$. Par exemple, le cas du **`facteur bayésien`** peut être détaillé sous la forme suivante :

$$ \begin{aligned} BF(r) = m_{\downarrow} \quad &\Rightarrow \quad \frac{y-x}{x} \times \frac{1-z}{z} = m_{\downarrow} \\ &\Rightarrow \quad (y-x)\cdot(1-z) = x\cdot z\cdot m_{\downarrow} \\ &\Rightarrow \quad y - y\cdot z - x + (1-m_{\downarrow})\cdot x\cdot z = 0 \\ &\Rightarrow \quad {}^t\vec{x}.\begin{pmatrix} 0 & 0 & \frac{1-m_{\downarrow}}{2} \\ 0 & 0 & -1/2 \\ \frac{1-m_{\downarrow}}{2} & -1/2 & 0 \end{pmatrix}.\vec{x} + \begin{pmatrix} -1 & 1 & 0 \end{pmatrix}.\vec{x} = 0 \end{aligned} $$

Pour une règle donnée, le problème que nous cherchons à résoudre peut s'écrire :

$$ \min_{\vec{x}} d(\vec{x}, r) \ s.c. \begin{cases} m(\vec{x}) & = & m_{\downarrow} \\ \vec{x} & \in & \mathcal{D} \end{cases} \tag{6.2} $$

Puisque nous avons décidé de nous concentrer sur la métrique euclidienne, la minimisation de $d(\vec{x}, r)$ se ramène à la minimisation de la quantité ${}^t\vec{x}.\vec{x} - 2{}^tr.\vec{x}$. Le domaine des contre-exemples

1. `http://www.geometrictools.com/`
2. Je remercie l'auteur pour les échanges constructifs que nous avons pu avoir à ce sujet.

$\mathcal{D}$ étant délimité par des plans, nous cherchons donc à résoudre le problème quadratique sous contraintes d'inégalités quadratiques, nommé $\mathcal{P}$, suivant :

$$\min_{\vec{x}}\{{}^t\vec{x}.\vec{x} - 2{}^tr.\vec{x}\} \ s.c. \begin{cases} {}^t\vec{x}.A_m(m_\downarrow).\vec{x} + {}^t\vec{b}_m(m_\downarrow).\vec{x} + c_m(m_\downarrow) & = & 0 \\ D.\vec{x} & \leq & d \end{cases} \tag{6.3}$$

Dans l'équation 6.3, D et d sont caractéristiques du domaine des contre-exemples et sont définis par :

$$D = \begin{pmatrix} -1 & 1 & -1 \\ -1 & 0 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \ d = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \tag{6.4}$$

Dans la suite, nous noterons D_i la $i$ème ligne de D et d_i la $i$ème ligne de d. Étant donné qu'aucune hypothèse n'est faite sur la nature de $A_m(m_\downarrow)$ (elle n'est a priori pas définie, pas positive...), il n'existe pas de méthode générale de résolution. Nous allons donc procéder par relaxation en réduisant le problème original $\mathcal{P}$ sous contraintes d'inégalités en sous-problèmes $\mathcal{P}_i$ sous contraintes d'égalités.

Soit ainsi r une règle limite minimisant la distance de la règle r à la surface $\mathcal{S}$ dans le domaine $\mathcal{D}$. r^* peut se trouver soit sur l'intérieur $\overset{\circ}{\mathcal{D}}$ de $\mathcal{D}$, soit sur sa frontière $\bar{\mathcal{D}}$. L'idée est de résoudre ces problèmes indépendamment et de ne retenir ensuite que la solution donnant le minimum.

Pour réaliser cela nous avons encore besoin de réduire le second problème concernant la frontière. Si la solution se trouve sur cette frontière, alors elle peut être soit strictement sur un plan, soit sur une droite déterminée par l'intersection de deux plans. Les 4 plans délimitant le domaine des contre-exemples ont pour équations $p_1 : x = 0$, $p_2 : x - y + z = 0$, $p_3 : x - y = 0$ et $p_4 : x - 1 + z = 0$ et forment donc potentiellement 6 intersections. Le problème de la frontière peut ainsi être réduit à 10 sous-problèmes (4 plans et 6 côtés).

Le problème original peut donc se réduire à 11 sous-problèmes, dont nous donnons les définitions ici :

1 Problème Original (voir section 6.2) :

$$(\mathcal{P}_0)\ \min_{\vec{x}}\{{}^t\vec{x}.\vec{x} - 2{}^tr.\vec{x}\} \ s.c. \ \ {}^t\vec{x}.A_m(m_\downarrow).\vec{x} + {}^t\vec{b}_m(m_\downarrow).\vec{x} + c_m(m_\downarrow) = 0 \tag{6.5}$$

4 Problèmes Plans (voir section 6.3) :

$$(\mathcal{P}_i)\ \min_{\vec{x}}\{{}^t\vec{x}.\vec{x} - 2{}^tr.\vec{x}\} \ s.c. \begin{cases} {}^t\vec{x}.A_m(m_\downarrow).\vec{x} + {}^t\vec{b}_m(m_\downarrow).\vec{x} + c_m(m_\downarrow) & = & 0 \\ D_i.\vec{x} & = & d_i \end{cases} \tag{6.6}$$

6 Problèmes Linéaires (voir section 6.4) :

$$(\mathcal{P}_{ij})\ \min_{\vec{x}}\{{}^t\vec{x}.\vec{x} - 2{}^tr.\vec{x}\} \ s.c. \begin{cases} {}^t\vec{x}.A_m(m_\downarrow).\vec{x} + {}^t\vec{b}_m(m_\downarrow).\vec{x} + c_m(m_\downarrow) & = & 0 \\ D_i.\vec{x} & = & d_i \\ D_j.\vec{x} & = & d_j \end{cases} \tag{6.7}$$

Avant de développer les solutions de ces problèmes, nous montrons l'importance de s'y intéresser en exhibant les mesures quadratiques que nous avons relevées. Pour chaque mesure, nous donnons ses éléments caractéristiques A, $\vec{b}$ et c.

confiance centrée :
$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1/2 \\ 0 & -1/2 & 0 \end{pmatrix} \quad \vec{b} = \begin{pmatrix} -1 \\ 1 - m_{\downarrow} \\ 0 \end{pmatrix} \quad c = 0$$

facteur bayésien :
$$A = \begin{pmatrix} 0 & 0 & \frac{1-m_{\downarrow}}{2} \\ 0 & 0 & -1/2 \\ \frac{1-m_{\downarrow}}{2} & -1/2 & 0 \end{pmatrix} \quad \vec{b} = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \quad c = 0$$

conviction :
$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1/2 \\ 0 & -1/2 & 0 \end{pmatrix} \quad \vec{b} = \begin{pmatrix} -m_{\downarrow} \\ 1 \\ 0 \end{pmatrix} \quad c = 0$$

cosine :
$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & m_{\downarrow}^2/2 \\ 0 & m_{\downarrow}^2/2 & 0 \end{pmatrix} \quad \vec{b} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad c = 0$$

kappa :
$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 - m_{\downarrow} \\ 0 & 1 - m_{\downarrow} & 0 \end{pmatrix} \quad \vec{b} = \begin{pmatrix} 2 \\ 2 - m_{\downarrow} \\ -m_{\downarrow} \end{pmatrix} \quad c = 0$$

levier :
$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix} \quad \vec{b} = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \quad c = m_{\downarrow}$$

lift :
$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & m_{\downarrow}/2 \\ 0 & m_{\downarrow}/2 & 0 \end{pmatrix} \quad \vec{b} = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \quad c = 0$$

Loevinger :
$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \frac{1-m_{\downarrow}}{2} \\ 0 & \frac{1-m_{\downarrow}}{2} & 0 \end{pmatrix} \quad \vec{b} = \begin{pmatrix} 1 \\ m_{\downarrow} - 1 \\ 0 \end{pmatrix} \quad c = 0$$

odds ratio :
$$A = \begin{pmatrix} 1 - m_{\downarrow} & -\frac{1-m_{\downarrow}}{2} & \frac{1-m_{\downarrow}}{2} \\ -\frac{1-m_{\downarrow}}{2} & 0 & -1/2 \\ \frac{1-m_{\downarrow}}{2} & -1/2 & 0 \end{pmatrix} \quad \vec{b} = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \quad c = 0$$

risque relatif :
$$A = \begin{pmatrix} 0 & \frac{1-m_{\downarrow}}{2} & 0 \\ \frac{1-m_{\downarrow}}{2} & -(1 - m_{\downarrow}) & -\frac{m_{\downarrow}}{2} \\ 0 & -\frac{m_{\downarrow}}{2} & 0 \end{pmatrix} \quad \vec{b} = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \quad c = 0$$

spécificité relative :
$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \vec{b} = \begin{pmatrix} 1 \\ -(2 + m_{\downarrow}) \\ 1 \end{pmatrix} \quad c = 0$$

Q de Yule :
$$A = \begin{pmatrix} 2m_{\downarrow} & -m_{\downarrow} & m_{\downarrow} \\ -m_{\downarrow} & 0 & \frac{1-m_{\downarrow}}{2} \\ m_{\downarrow} & \frac{1-m_{\downarrow}}{2} & 0 \end{pmatrix} \quad \vec{b} = \begin{pmatrix} 1 - m_{\downarrow} \\ m_{\downarrow} - 1 \\ 0 \end{pmatrix} \quad c = 0$$

Nous avons donc mis en avant 12 mesures quadratiques : la surface définie par la fixation d'un seuil est une quadrique et calculer la robustesse revient donc à résoudre le programme quadratique

que nous avons détaillé précédemment. Pour résoudre ce problème comprenant des contraintes d'inégalités quadratiques, nous le réduisons à 11 sous-problèmes sous contraintes d'égalités en utilisant la méthode des lagrangiens. Cette méthode indique que pour minimiser une fonction f sous les contraintes $g_i(x) = c_i$, nous devons résoudre l'équation

$$\nabla f = \sum_i \lambda_i \nabla g_i. \tag{6.8}$$

6.2 Résoudre le problème original

Dans cette section et les suivantes, nous fixons une mesure m et un seuil d'intérêt $m_\downarrow$. Nous prenons alors la liberté d'écrire A, $\vec{b}$ et c plutôt que $A_m(m_\downarrow)$, $\vec{b}_m(m_\downarrow)$ et $c_m(m_\downarrow)$ dès lors qu'il n'y a plus aucune ambiguïté.

Concentrons nous alors sur le premier problème $(\mathcal{P}_0)$ (équation 6.5). La technique des multiplicateurs de Lagrange nous pousse à résoudre l'équation d'inconnue $\vec{x}$ définie par :

$$\vec{x} - r = 2\lambda A.\vec{x} + \lambda\vec{b} \Leftrightarrow \vec{x} = (I - 2\lambda A)^{-1}(\lambda\vec{b} + r). \tag{6.9}$$

Nous supposons ici que $\frac{1}{2\lambda}$ n'est pas une valeur propre de la matrice A et que l'inversion se passe donc sans encombre. Nous verrons plus tard comment traiter en théorie les cas où le multiplicateur de Lagrange pose problème. Il faut cependant savoir que lors de nos expériences, le cas ne s'est jamais présenté. Cela peut avoir deux causes : la première est que si le multiplicateur de Lagrange induit une valeur propre de la matrice A, cela rajoute des contraintes supplémentaires au problème et sur quelques exemples, nous avons observé qu'il était impossible de trouver une solution ; la seconde est due à la précision de calcul, qui ne pose cependant pas de problème car la notion de robustesse que nous avons définie est continue et si l'on tombe un peu à côté, le résultat est toujours cohérent.

On définit donc $K_\lambda \triangleq Com(I - 2\lambda A)$ la matrice des cofacteurs de $I - 2\lambda A$ qui intervient dans la formule $\zeta_A(2\lambda)(I - 2\lambda A)^{-1} = K_\lambda$ où $det(I - 2\lambda A) = \zeta_A(2\lambda)$ est un polynôme de degré au plus 3. Nous combinons alors cette écriture avec l'équation 6.9 en introduisant $\vec{x}$ dans l'équation de la contrainte et en multipliant tout par $(det(I - 2\lambda A))^2$:

$${}^t(\lambda\vec{b} + r).{}^tK_\lambda.A.K_\lambda.(\lambda\vec{b} + r) + \zeta_A(2\lambda){}^t\vec{b}.K_\lambda.(\lambda\vec{b} + r) + (\zeta_A(2\lambda))^2 c = 0 \tag{6.10}$$

L'équation ci-dessus (6.10) est un polynôme de degré au plus 6 en λ et est à ce titre résoluble par les méthodes classiques de Newton que l'on trouve dans tout programme de calcul numérique. Les racines de ce polynôme définissent 6 vecteurs de dimension 3 $(\vec{x_i})_{i=1\ldots6}$ parmi lesquels nous choisissons le vecteur réel qui minimise la distance à r. Si ce vecteur se trouve dans le domaine des contre-exemples $\mathcal{D}$ alors il n'est pas nécessaire d'aller plus loin dans les calculs : nous avons trouvé la robustesse de la règle r. Dans le cas contraire, nous devons nous intéresser aux problèmes plans.

6.3 Résoudre les problèmes plans

Nous nous concentrons maintenant sur le problème plan de l'équation (6.6) pour un i donné. Comme précédemment nous utilisons la méthode des multiplicateurs de Lagrange pour le résoudre. L'équation (6.8) peut donc s'écrire :

$$\vec{x} - r = 2\lambda_1 A.\vec{x} + \lambda_1\vec{b} + \lambda_2{}^tD_i. \tag{6.11}$$

Nous prenons la liberté d'écrire D_i à la place de ${}^tD_i$ (le vecteur ligne devient un vecteur colonne) pour faciliter l'écriture. Comme lors de l'écriture de (6.9), nous pouvons transformer l'équation (6.11) en :

$$\vec{x} = (I - 2\lambda_1 A)^{-1}.(\lambda_1 \vec{b} + \lambda_2 D_i + r) \tag{6.12}$$

Nous allons maintenant procéder en deux étapes en utilisant les mêmes notations que lors de la résolution du problème original. Tout d'abord, nous remplaçons $\vec{x}$ dans la seconde contrainte ${}^tD_i.\vec{x} = d_i$ pour obtenir une expression de λ_2 en fonction de λ_1 :

$$\lambda_2 = -\frac{{}^tD_i.K_{\lambda_1}.(\lambda_1\vec{b} + r) - \zeta_A(2\lambda_1) * d_i}{{}^tD_i.K_{\lambda_1}.D_i} \tag{6.13}$$

Si l'on remplace maintenant λ_2 dans l'écriture de $\vec{x}$ obtenue en (6.12), nous avons :

$$\begin{aligned}\zeta_A(2\lambda_1)({}^tD_i.K_{\lambda_1}.D_i)\vec{x} = K_{\lambda_1}(&\lambda_1({}^tD_i.K_{\lambda_1}.D_i\vec{b} - {}^tD_i.K_{\lambda_1}.\vec{b}D_i)\\ &- {}^tD_i.K_{\lambda_1}.rD_i\\ &+ {}^tD_i.K_{\lambda_1}.D_i r + \zeta_A(2\lambda_1)d_iD_i)\end{aligned} \tag{6.14}$$

Définissons par

$$\vec{v_i}(\lambda_1) = \lambda_1({}^tD_i.K_{\lambda_1}.D_i\vec{b} - {}^tD_i.K_{\lambda_1}.\vec{b}D_i) - {}^tD_i.K_{\lambda_1}.rD_i + {}^tD_i.K_{\lambda_1}.D_i r - \zeta_A(2\lambda_1)d_iD_i$$

la parenthèse du membre droit de cette équation. D'après la définition de K_{λ_1}, $\vec{v_i}$ est un polynôme de degré au plus 3 en λ_1. En injectant $\vec{x}$ dans la contrainte quadratique et en multipliant par $(\zeta_A(2\lambda_1)({}^tD_i.K_{\lambda_1}.D_i))^2$, nous obtenons le polynôme en λ_1 de degré au plus 10 suivant :

$${}^t\vec{v_i}.{}^tK_{\lambda_1}.A.K_{\lambda_1}.\vec{v_i} + \zeta_A(2\lambda_1)({}^tD_i.K_{\lambda_1}.D_i){}^t\vec{b}.K_{\lambda_1}.\vec{v_i} + \left(\zeta_A(2\lambda_1)({}^tD_i.K_{\lambda_1}.D_i)\right)^2 c = 0 \tag{6.15}$$

Encore une fois, les racines de ce polynôme peuvent être approchées grâce aux méthodes de Newton classiques et nous nous intéressons à ses solutions réelles, si elles existent, se situant dans le domaine des contre-exemples $\mathcal{D}$. Si la réponse n'est pas satisfaisante, on se ramène alors aux problèmes linéaires.

6.4 RÉSOUDRE LES PROBLÈMES LINÉAIRES

Nous nous intéressons ici à l'évaluation de la distance entre la projection d'une règle et l'intersection entre une droite et la quadrique. Puisque la droite est de dimension 1, l'intersection avec la quadrique peut être vide, de dimension 0 (un point ou deux) ou de dimension 1 (la droite est incluse dans la quadrique). Si l'on considère le problème $(\mathcal{P}_{ij})$ décrit dans l'équation (6.7), on observe que l'on est en fait intéressé par un point situé sur l'intersection de deux plans $D_i.\vec{x} = d_i$ et $D_j.\vec{x} = d_j$, c'est-à-dire sur une droite que nous nommerons $\mathcal{D}_{ij}$. Son équation paramétrique est définie par $\vec{x} \in \mathcal{D}_{ij}$ ssi $\exists \lambda t.q.\vec{x} = \vec{O} + \lambda\vec{u}$ et si l'on l'injecte dans la première contrainte, quadratique, il vient :

$${}^t\vec{u}.A.\vec{u}\lambda^2 + {}^tu.(2A.\vec{O} + \vec{b})\lambda + c + {}^t\vec{O}.A.\vec{O} = 0. \tag{6.16}$$

Les éléments O et $\vec{u}$ sont facilement déterminés à partir des équations des plans. L'équation obtenue est une équation polynomiale de degré au plus 2 et admet donc 0, 1, 2 ou une infinité de solutions réelles. Les trois premiers cas sont faciles à résoudre et le dernier n'est pas intéressant car il aura été traité lors de la résolution des problèmes précédents (la droite est incluse dans la surface).

6.5 Recombinaison

En suivant ces étapes, on obtient un ensemble de points potentiellement solutions, parmi lesquels il faut choisir celui qui se trouve dans $\mathcal{D}$ et qui minimise la distance à la règle. Ce point représentera une instance possible de r^*. Pour implémenter cette résolution, nous utilisons une combinaison entre le logiciel de calcul formel Yacas [3] et l'utilisation de la librairie Pylab de Python.

Il reste cependant à voir comment, en théorie, gérer les cas pathologiques d'apparition de valeurs propres pour certaines valeurs de Lagrangiens. Nous allons ici étudier un exemple de manière détaillée.

Donnons-nous la mesure de **`lift`** avec un seuil d'intérêt fixé à 2 et une règle r dont la projection dans le domaine des contre-exemples est donnée par $\begin{pmatrix} \frac{3}{20} \\ \frac{3}{10} \\ \frac{1}{5} \end{pmatrix}$. La surface de seuil déterminée par le **`lift`** possède les éléments caractéristiques suivants :

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & m_\downarrow/2 \\ 0 & m_\downarrow/2 & 0 \end{pmatrix} \quad \vec{b} = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \quad c = 0$$

Et dans le cas particulier d'un seuil fixé à 2, cela se simplifie en :

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \quad \vec{b} = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \quad c = 0$$

Puisque A est une matrice symétrique, elle est diagonalisable dans une base orthonormée et il existe une matrice orthogonale P telle que ${}^tPP = I_3$ et ${}^tPAP = D_A$ où D_A est diagonale. Ici, on peut trouver $P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$ et l'on voit que ${}^tP = P$ est symétrique. On en déduit d'ailleurs que $D_A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. Les valeurs propres de A sont donc $\{0, -1, 1\}$. La matrice P étant inversible, l'équation ${}^tx.A.x + {}^tb.x + c = 0$ devient ${}^t\bar{x}.D_A.\bar{x} + {}^t\bar{b}.\bar{x} + c = 0$, où $\bar{x}$ (resp. $\bar{b}$) vaut $P.x$ (resp. $P.b$). De plus, puisque le changement de repère opéré par P est orthonormé, la distance entre x et r est la même que la distance entre $\bar{x}$ et $\bar{r}$. On peut d'ailleurs calculer $\bar{b} = P.b = \begin{pmatrix} 1 \\ \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$ et $\bar{r} = P.r = \begin{pmatrix} \frac{3}{20} \\ -\frac{1}{10\sqrt{2}} \\ \frac{1}{2\sqrt{2}} \end{pmatrix}$. Dans la suite, cette transformation est faite par défaut et nous nous contenterons d'écrire v au lieu de $\bar{v}$ pour tout vecteur v.

A partir de là, nous suivons la méthodologie détaillée précédemment pour le problème général. Nous obtenons donc l'équation :

$$(I - 2\lambda D_A)x = \lambda b + r.$$

Supposons à ce stade que $(I - 2\lambda D_A)$ n'est pas inversible, cela signifie que $\frac{1}{2\lambda}$ est une valeur propre et donc que $\lambda = -1/2$ ou $\lambda = 1/2$ (comme λ est un réel, la valeur propre 0 est exclue). Voyons si cela est possible.

3. `http://yacas.sourceforge.net`

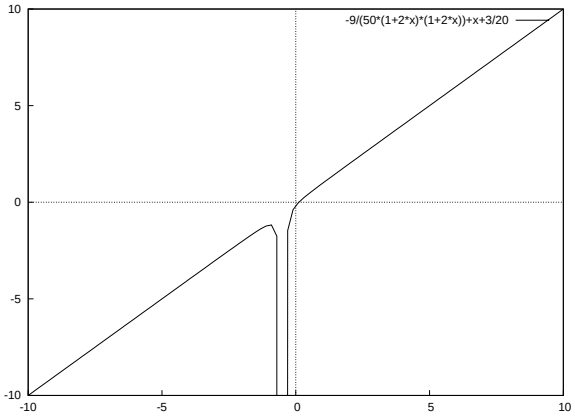


FIGURE 6.1 : fonction rationnelle

Supposons que $\lambda = -1/2$, alors l'équation devient :

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -\frac{7}{10} \\ -\frac{1}{2\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

La seconde composante montre que cela est impossible car $-\frac{1}{2\sqrt{2}} \neq 0$.

Supposons que $\lambda = 1/2$, alors l'équation devient :

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} \frac{13}{20} \\ \frac{2}{5\sqrt{2}} \\ 0 \end{pmatrix}$$

Dans cette équation, il n'y a pas de contradiction et l'on commence par la résoudre suivant les deux premières composantes :

$$\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{13}{20} \\ \frac{2}{5\sqrt{2}} \end{pmatrix} \Longrightarrow \begin{array}{rcl} x_1 & = & \frac{13}{20} \\ x_2 & = & \frac{1}{5\sqrt{2}} \end{array}$$

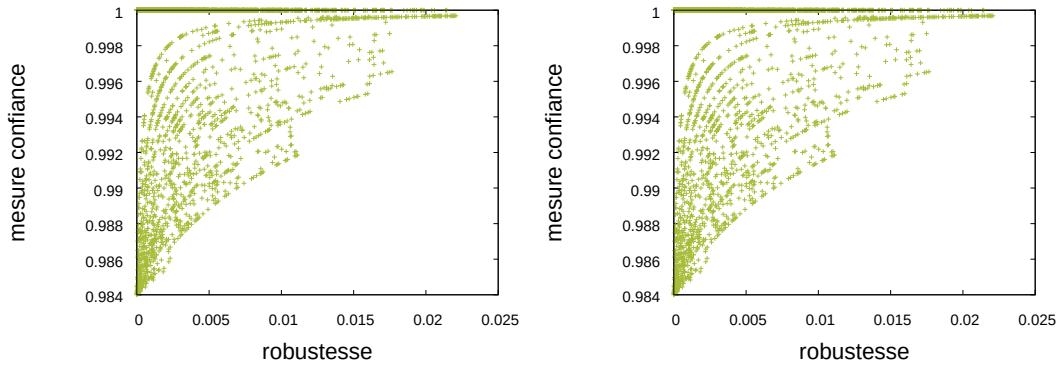
Pour obtenir la troisième composante, nous introduisons ce résultat dans la contrainte quadratique afin d'obtenir un polynôme de degré deux en x_3 : $x_3^2 - \frac{1}{\sqrt{2}}x_3 + \frac{53}{100} = 0$. Ce polynôme n'a aucune racine réelle et donc $\lambda = 1/2$ n'est pas envisageable. Finalement, $(I - 2\lambda D_A)$ est inversible et nous avons :

$$x = (I - 2\lambda D_A)^{-1}(\lambda b + r).$$

En introduisant cela dans l'équation de la quadrique, nous obtenons une fonction rationnelle d'équation :

$$\frac{-\frac{9}{50}}{(1+2\lambda)^2} + \lambda + \frac{3}{20} = 0$$

que nous représentons sur la figure 6.1. Cette fonction admet une racine pour $\lambda = 0.01776$ ce qui permet de calculer une solution $x = \begin{pmatrix} 0.16776 \\ -0.05616 \\ 0.35355 \end{pmatrix}$ correspondant à une distance de 0.02296. Pour savoir si cette solution se trouve dans le domaine des contre-exemples, il suffit de calculer $Px = \begin{pmatrix} 0.16776 \\ 0.28971 \\ 0.2103 \end{pmatrix}$ (changement de repère) et de constater que c'est bien un point du domaine. La robustesse de la règle r est donc bien de 0.02296.



(a) Approche plane

(b) Approche quadratique

FIGURE 6.2 : Résultats de robustesse suivant les deux approches pour la mesure de **confiance** (base *chess*).

6.6 RÉSULTATS D'EXPÉRIENCES

nom	seuil $m_{\downarrow}$
confiance centrée	0.05
facteur bayésien	10
conviction	1.1
cosine	0.8
kappa	0
levier	-0.1

nom	seuil $m_{\downarrow}$
lift	1.1
Loevinger	-0.1
odds ratio	0.9
risque relatif	1.1
spécificité relative	-0.1
Q de Yule	0.2

TABLE 6.1 : Les mesures quadratiques retenues et le seuil choisi.

Nous avons donc implémenté le calcul de la robustesse pour les mesures quadratiques de la table 6.1. Nous avons utilisé le même protocole expérimental que pour les mesures planes pour l'étude de l'influence du bruit. Nous présentons les résultats dans la table 6.2. L'étude des règles conservées et disparues montre que la notion de robustesse est encore une fois validée empiriquement : les règles qui ont disparues sont des règles de faible robustesse, celles qui ont été conservées présentent une robustesse moyenne plus élevée. Encore une fois, une exception est faite pour la base *census* pour laquelle la robustesse est très faible et n'assure donc pas la stabilité des règles. Dans le cas de la mesure de **levier**, nous voyons que le seuil fixé donne des valeurs de robustesse élevé, ce qui explique le fait qu'aucune règle ne disparaisse.

Concernant les graphiques mesure/robustesse, nous avons comparé les résultats de l'approche *quadratique* sur des mesure planes (qui sont un cas particulier de mesures quadratiques) avec les résultats obtenus précédemment en approche *plane* et avons pu vérifier que les résultats sont semblables, par exemple avec la mesure de **confiance** sur la figure 6.2, ce qui nous conforte dans la validité de notre approche. Nous avons par ailleurs observé les résultats sur les mesures strictement quadratiques : si certaines mesures se comportent de manière similaire aux mesures linéaires, certains motifs étonnants sont apparus. Nous en donnons un exemple sur la figure 6.3 avec la mesure de **kappa**. La base *mushroom* 6.3(b) est utilisée comme base de classification, alors que la base *chess* 6.3(a) est utilisée comme base binaire pure (sans attribut de classe). On se rend compte que dans le cas de la base *chess*, une forte concentration se trouve autour de l'axe des abscisses, alors que pour la base *mushroom*, la robustesse semble corrélée à la mesure. Contrairement au cas de la mesure de **confiance** pour laquelle nous étions parvenus à établir un lien entre les lignes apparentes et l'expression de la robustesse (équation 5.2), nous n'avons pas réussi à établir là de

(a) `facteur bayésien`

base	disparues	conservées
census	0	0
chess	0.00039	0.00365
connect	0.00006	0.00120
mushroom	0.00022	0.00214

(b) `confiance`

base	disparues	conservées
census	0	0
chess	0.00096	0.00667
connect	0.00038	0.00471
mushroom	0.00010	0.00040

(c) `conviction`

base	disparues	conservées
census	0.00001	0.00002
chess	0.00125	0.01176
connect	0.00063	0.00442
mushroom	0.00028	0.01046

(d) `cosine`

base	disparues	conservées
census	0	0
chess	0.00553	0.07495
connect	0.00764	0.10738
mushroom	0.00293	0.01807

(e) `kappa`

base	disparues	conservées
census	0.00003	0.00005
chess	0.04040	0.30684
connect	0.02017	0.14940
mushroom	0.00064	0.02788

(f) `levier`

base	disparues	conservées
census	0	0.25908
chess	0	0.18198
connect	0	0.17983
mushroom	0	0.24020

(g) `lift`

base	disparues	conservées
census	0.00001	0.00002
chess	0.00148	0.01603
connect	0.00055	0.00385
mushroom	0.00029	0.01033

(h) `Loevinger`

base	disparues	conservées
census	0.00001	0.00002
chess	0.00134	0.01440
connect	0.00077	0.00753
mushroom	0.00026	0.01231

(i) `odds ratio`

base	disparues	conservées
census	0.00001	0.00002
chess	0.00134	0.01263
connect	0.00077	0.00547
mushroom	0.00029	0.01183

(j) `risque relatif`

base	disparues	conservées
census	0.00001	0.00002
chess	0.00148	0.01665
connect	0.00056	0.00549
mushroom	0.00029	0.01044

(k) `spécificité relative`

base	disparues	conservées
census	0	0
chess	0.37405	0.23677
connect	0.39788	0.28466
mushroom	0.06628	0.06942

(l) `Q de Yule`

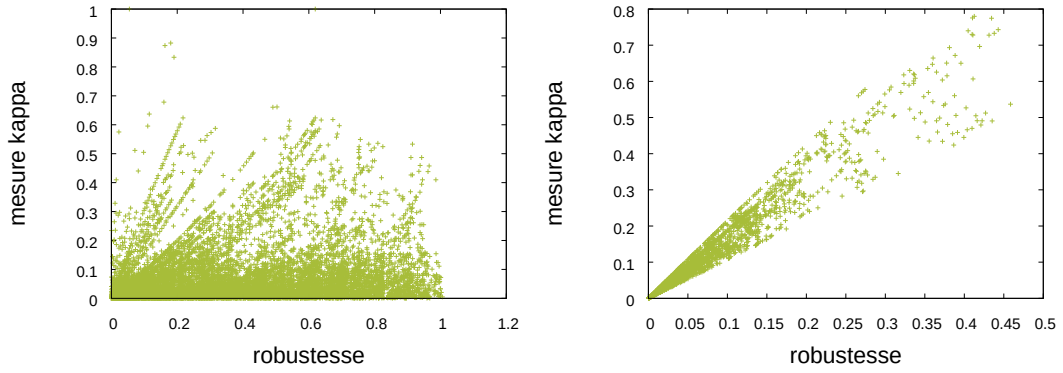
base	disparues	conservées
census	0.00001	0.00002
chess	0.00111	0.01106
connect	0.00045	0.00376
mushroom	0.00031	0.00948

TABLE 6.2 : Comparaison entre les robustesses moyennes des règles disparues et conservées pour les différentes mesures quadratiques

manière formelle l'existence de cette corrélation. Cependant, dans le cas de la base *mushroom* et de la mesure de **`levier`**, nous avons observé deux groupes de règles se superposant : ces deux groupes sont composés des deux classes de la base de données (empoisonné ou comestible). La figure 6.4 met en évidence une dépendance de la robustesse en fonction de la fréquence du conséquent de la règle. De tels comportements apparaissent pour d'autres mesures, mais l'aspect quadratique de ces mesures, associé au fait que la nature de la quadrique définie varie en fonction du seuil de mesure fixé, rendent l'étude formelle délicate. Quitte donc à s'intéresser aux méthodes numériques, nous aurions pu effectuer nos expériences à l'aide d'un solveur général. La nature des contraintes nous a poussé à nous tourner vers des mesures particulières car le solveur que nous avions utilisé (matlab/octave) menait à des erreurs de calculs trop fréquentes.

CONCLUSION

L'étude des mesures quadratiques nous a permis d'étendre un peu plus l'ensemble des mesures pour lesquelles la robustesse peut être évaluée grâce à des outils de calcul numérique classiques. Étendre encore cet ensemble imposerait d'utiliser des outils de calcul plus poussés : les tests que nous avons menés avec les outils que nous avions à notre portée (matlab/pylab) ont menés à de trop nombreuses erreurs de calcul, sans apporter réellement de plus-value à l'étude de la robustesse que nous avons présentée jusqu'à présent. Cela dit, il est tout de même important de préciser que le calcul de la robustesse n'est pas limité aux mesures planes et quadratiques, mais peut s'étendre (à condition d'avoir accès aux bons outils) à toute mesure de l'intérêt des règles d'association.



(a) Base *connect* (b) Base *mushroom*

FIGURE 6.3 : Résultats de robustesse pour la mesure quadratique de **`kappa`**

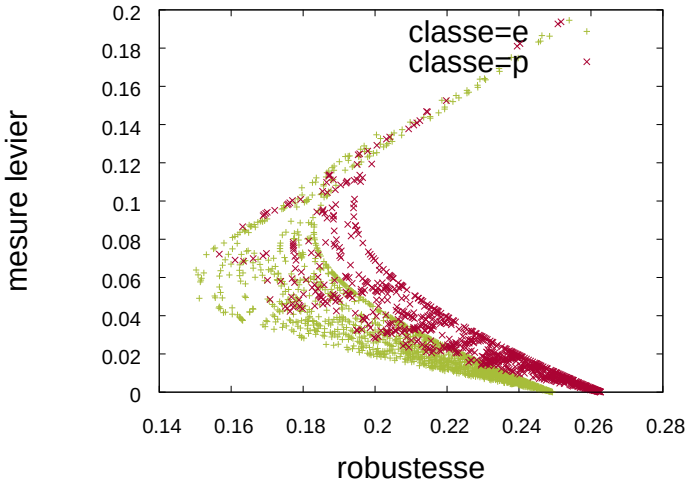


FIGURE 6.4 : Robustesse des règles issues de la base *mushroom* ave la mesure de **`levier`**. La classe e représente les champignons comestibles (edible) et la classe p les champignons vénéneux (poisonous).

Conclusion sur la robustesse

La notion de robustesse que nous avons définie est cohérente au vu des expérimentations menées. Elle permet de bien qualifier une résistance au bruit dans le sens de l'arrivée de nouvelles transactions dans la base de données, ou de la connaissance du bruit présent dans la base. Il est en effet possible, à partir du calcul de la robustesse, de fournir une sécurité sur les règles extraites, assurant du maintien de leur intérêt en cas de légère modification des données, ou de légère modification des paramètres. Son calcul n'est cependant pas aisé et si les mesures planes permettent une définition formelle de la robustesse, les mesures quadratiques quant à elles imposent l'utilisation de méthodes numériques. Nous avons ainsi fourni une méthode de résolution validée par la comparaison au cas des mesures planes. Il reste cependant un grand nombre de mesures à devoir traiter, les mesures *quelconques* pour lesquelles le recours à des solveurs génériques sera indispensable. C'est un travail que nous avons entamé, mais qui a été mis de côté après la découverte de résultats formels sur les mesures planes et les mesures quadratiques, qui couvrent déjà un ensemble non négligeable de mesures.

Une étude de l'influence de la robustesse dans un algorithme de classification du type CBA [Liu et al. 98] a aussi été envisagée : ces travaux ne sont pas suffisamment aboutis pour être présentés ici. Il faut en effet prendre en compte le fait que la validation de classifieurs se fait généralement en utilisant une base d'apprentissage réduite et une base de test plus grande. La robustesse des règles n'assure alors pas la stabilité de l'intérêt. De plus les mesures planes ou quadratiques ne sont pas toujours de bonnes mesures de prédiction et si la robustesse permet d'assurer l'intérêt pour de petites variations, elle ne permet cependant pas d'assurer la prédiction. Un couplage avec les travaux de [Jalali-Heravi et Zaïane 10] devrait être envisagé afin de se pencher sur les couples de mesures élagage/classement permettant une meilleure prédiction. Cependant, d'après cette étude, les couples varient de base en base et il n'est pas possible de mettre en évidence des mesures efficaces en général. Une étude complémentaire pourrait être menée afin d'inclure dans ce processus la robustesse de mesures et d'étudier les couples optimaux pour un ensemble de bases donné.

Dans cette même étude, l'utilisation de mesures d'intérêt prend place dans la construction du classifieur, mais aussi et indépendemment, dans la phase de recherche des règles de manière classique. Les auteurs appellent cela une phase d'élagage, mais il s'agit en réalité d'un filtrage classique. Utiliser les mesures dans les phases d'élagage pures n'est pas aujourd'hui une technique très répandue, car peu de mesures ont de bonnes propriétés de monotonie/anti-monotonie. Le reste de cette thèse sera consacré à l'étude des propriétés d'anti-monotonie des mesures d'intérêt. Nous nous intéresserons tout d'abord à des propriétés existantes, mais qui ne s'appliquent qu'à peu de mesures : nous nous appuierons sur notre cadre formel d'étude des règles pour généraliser ces propriétés à de plus grands ensembles de mesures. Puis, nous définirons une nouvelle propriété d'anti-monotonie sur un grand nombre de mesures d'intérêt dans le cadre de la recherche de règles de classe.

TROISIÈME PARTIE : GÉNÉRALISATION DE PROPRIÉTÉS D'ANTI-MONOTONIE

Nous proposons d'étudier ici le point de vue algorithmique de la recherche de règles d'association. Nous nous appuyons sur des propriétés existantes de quelques mesures afin de les généraliser (chapitre 7). L'apport de notre cadre formel sera de nous permettre de faire un lien entre ces propriétés algorithmiques et les propriétés analytiques des mesures, notamment grâce à l'étude des variations des fonctions de mesure adaptées. Ces travaux s'appuient sur trois propriétés algorithmiques dont nous donnons des conditions nécessaires et/ou suffisantes d'existence (chapitres 8, 9 et 10). Ces propriétés ont l'avantage de se détacher de la notion de `support` et de permettre la recherche de pépites, moyennant quelques concessions. Les travaux que nous présentons dans cette partie sont issus de nos publications [Le Bras et al. 10a, Le Bras et al. 09a, Le Bras et al. 09c].

TELECOM Bretagne

7 Élagage de l'espace de recherche par les mesures

La mesure de **support** possède une propriété d'anti-monotonie qui permet un élagage de l'espace de recherche des motifs très efficace. Nous avons vu dans les parties précédentes qu'au cours de la recherche de règles, il pouvait être intéressant de se concentrer sur les motifs, mais que dans beaucoup d'applications, ce sont les règles d'association qui seront évaluées et utilisées grâce à une mesure d'intérêt. Et le seul moyen d'extraire ces règles d'association est, jusqu'à présent, de tester tous les motifs et pour chacun, toutes les règles qu'il est capable de former. Cette tâche laborieuse (rappelons que le nombre de règles est exponentiel par rapport au nombre d'attributs) devient rapidement inenvisageable, sauf à relever le seuil de **support**. Ce qui encore une fois pose problème dans la recherche de pépites de connaissance. À notre sens, l'unique moyen de s'en sortir serait de découvrir des propriétés d'élagage qui s'appuieraient directement sur les mesures d'intérêt, et qui permettraient d'accéder directement aux règles intéressantes sans passer par les motifs fréquents. Nous proposons dans ce chapitre un état de l'art des propriétés existantes, parmi lesquelles trois nous ont particulièrement intéressé.

7.1 Propriétés d'élagage

Les améliorations techniques et théoriques subies par Apriori sont nombreuses mais elles se concentrent le plus souvent sur la recherche des motifs, sans s'intéresser vraiment à la question de la qualité des règles. Il fallait donc, afin de faire face à l'accroissement des données et aux problèmes de qualité liés au **support** et à la **confiance**, réussir à découvrir des propriétés algorithmiques directement liées aux mesures d'intérêt. C'est l'un des grands challenges proposés par [Yang et Wu 06]. Quelques travaux, très récents, existent à ce sujet et donnent, la plupart du temps, pour une mesure, un algorithme permettant de rechercher des règles d'association en ne fixant que le seuil de cette mesure.

Par exemple, dans [Cohen et al. 01] les auteurs s'intéressent à la mesure de **Jaccard** définie pour une règle $\mathtt{A} \rightarrow \mathtt{B}$ par $\frac{p_{\mathtt{AB}}}{p_{\mathtt{A}}+p_{\mathtt{B}}-p_{\mathtt{AB}}}$. Cette mesure possède une bonne interprétation algébrique, car elle représente le rapport entre l'intersection de deux ensembles et l'union de ces deux ensembles. Le principe de leur approche est de s'appuyer sur des tables de hachage pour évaluer la mesure de **Jaccard**. Ils remarquent en effet qu'étant donnée une fonction de hachage h sur les colonnes de la base de données et deux colonnes c_i et c_j, la probabilité que les deux colonnes aient la même valeur de hachage est donnée par :

$$\mathbb{P}\left(h(c_i) = h(c_j)\right) = Jacc(c_i \rightarrow c_j)$$

Ils établissent alors plusieurs méthodologies s'appuyant sur ce principe pour extraire les règles de taille 2 ayant une forte valeur de mesure de **Jaccard**. Cependant, les algorithmes qui en découlent sont probabilistes, et bien que les expériences montrent une certaine cohérence avec les résultats d'Apriori, ils sont malgré tout incomplets et limités aux règles de taille 2. C'est pourtant une

approche originale des règles d'association et une solution efficace dans le cadre de la mesure de **Jaccard**. Malheureusement, ce principe ne peut s'appliquer à aucune autre mesure d'intérêt.

Une propriété plus récente est la propriété de *loose-anti-monotony* décrite dans [Bonchi et Lucchese 05] et qui s'applique essentiellement à des mesures statistiques telles que la moyenne ou la déviation standard. On rappelle qu'une contrainte est anti-monotone si, lorsqu'un motif X la vérifie, alors tous ses sous-motifs la vérifient aussi. Une contrainte sur les motifs est *loose-anti-monotone* si, lorsqu'un motif X la vérifie, alors *l'un au moins* de ses sous-motifs la vérifie aussi. Cette distinction minime permet d'inclure un grand nombre de mesures dans l'ensemble des mesures statistiques *loose-anti-monotones*. Mais elle n'a pas particulièrement attiré notre attention, étant donné la particularité de ce type de mesure.

Une autre approche originale est celle que l'on peut retrouver dans [Zimmermann et De Raedt 04] adaptant notamment [Morishita et Sese 00] et s'appuyant sur une propriété analytique de la mesure de χ^2 pour les règles d'association. Pour ces travaux, les auteurs se restreignent aux règles de classe (et nous verrons que cela est souvent le cas) ce qui a pour conséquence de fixer l'une des marges de la table de contingence et donc de ne laisser que deux degrés de liberté à la règle. Une fois la classe choisie, la projection des règles se fait dans un plan du domaine adapté. Les auteurs font donc remarquer que la mesure de χ^2, comme quelques autres, est convexe, ce qui permet notamment de majorer la valeur de mesure des règles plus spécifiques que $A \rightarrow c$ par des quantités ne dépendant que de $A \rightarrow c$, comme on le fait pour le **support** dans APRIORI. On trouve ainsi rapidement les règles de χ^2 élevé sans se donner de seuil de **support**, et l'on a là une véritable propriété d'élagage originale. Ce qui nous intéresse ici est le lien entre une propriété algorithmique et des propriétés analytiques des mesures, qui se traduit ici par le lien entre la convexité du χ^2, mais aussi du **gain informationnel** par exemple, et la propriété algorithmique d'élagage, applicable à toute mesure convexe. Ce type de relation correspond exactement à ce que nous cherchons à mettre en place par la suite.

Nous allons proposer notre contribution à l'étude des propriétés d'élagage en démontrant l'avantage que représente notre cadre d'étude dans ce domaine. Nous nous focalisons sur trois propriétés d'élagage que nous avons relevées dans la littérature. Dans un premier temps, nous nous intéresserons au cas de la **all-confidence** (section 7.2), une transcription de la **confiance** dans le monde des motifs qui possède une propriété d'anti-monotonie. Puis nous étudierons la propriété *UEUC* (section 7.3), une propriété de monotonie de la **confiance** encore une fois, qui permet de retrouver toutes les règles situées au-dessus d'un seuil de **confiance** donné sans contrainte de **support**. Cette propriété se focalise sur les règles de classe, tout comme la dernière propriété à laquelle nous nous intéresserons. Il s'agira de la propriété d'anti-monotonie induite par la recherche d'ensembles de règles optimales (section 7.4).

7.2 LA MESURE DE all-confidence

Dans l'article [Xiong et al. 03], une mesure de **h-confidence** est introduite, semblable à la mesure de **all-confidence** décrite plus tôt par [Omiecinski 03]. Cette mesure est une mesure alternative au **support** pour les motifs. Elle est définie pour un motif I à partir de la **confiance** ($conf$) de la manière suivante :

$$all\text{-}conf(\mathtt{I}) = \min_{\mathtt{I}' \subset \mathtt{I}}(conf(\mathtt{I}' \rightarrow \mathtt{I} - \{\mathtt{I}'\}))$$

Ainsi, si l'on se donne un seuil de **confiance** γ, un motif sera intéressant si et seulement si il engendre uniquement des règles intéressantes. Dans [Omiecinski 03], l'auteur montre que cette mesure est anti-monotone, au même titre que le **support**, et peut donc être utilisée dans un algorithme de type APRIORI en ne modifiant que la mesure. L'article [Xiong et al. 03] ajoute une deuxième propriété à cette mesure. En effet, l'élagage par la **all-confidence** permet de plus

d'éliminer tous les motifs dans lesquels les attributs ont des valeurs de **support** trop variées, c'est-à-dire que l'on évitera le phénomène que nous avons déjà mentionné : la règle *caviar* $\rightarrow$ *pain* dans laquelle on retrouve deux motifs très éloignés, et qui forment ainsi une règle qui n'apprend rien (ici, cela est dû à la trop grande fréquence de l'attribut *pain*). Remarquons d'ailleurs que la règle symétrique *pain* $\rightarrow$ *caviar* n'aura certainement pas une confiance très élevée, et le motif $\{pain, caviar\}$ ne passera donc pas le filtre de la **all-confidence**. L'apport du second article réside aussi dans la mise en œuvre de cette mesure, notamment sur des aspects techniques, car un parallèle est fait entre la recherche de motifs intéressants au sens de la **all-confidence** et la recherche de cliques maximales dans un graphe.

Cette nouvelle mesure est intéressante pour une raison bien simple : il s'agit là d'une mesure sur les motifs, qui s'intègre dans des algorithmes déjà existants, et qui ont bénéficié de nombreuses améliorations techniques depuis longtemps. Ainsi, elle pourra être utilisée dans toute implémentation d'Apriori, à condition qu'elle soit simple à calculer. En effet, si le calcul de la mesure d'un motif demande trop de temps, il n'est pas raisonnable de l'implémenter dans un tel algorithme. Heureusement, la mesure de **all-confidence** possède une écriture équivalente plus simple sous la forme suivante [Xiong et al. 03] :

$$allconf(\mathtt{I}) = \min_{i \in \mathtt{I}}(conf(i \rightarrow \mathtt{I} - \{i\}))$$

Sous cette écriture, la mesure peut être rapidement calculée car il n'est plus nécessaire de parcourir tous les sous-motifs de $\mathtt{I}$ pour la calculer : seuls les attributs on besoin d'être parcourus. Apriori renverra donc tous les motifs qui n'engendrent que des règles de valeur de **confiance** au-dessus d'un certain seuil. Bien que très intéressante, cette approche possède quelques défauts. Si l'on recherche les règles de valeur de **confiance** élevée, et plus précisément toutes ces règles, cette technique ne nous le permettra pas car il pourrait bien exister des règles de **confiance** élevée telles que leurs attributs, organisés différemment, forment une règle de **confiance** trop basse. Ces règles ne seront pas présentes en sortie. Le résultat n'est donc pas complet en ce sens.

Nous verrons dans la partie 8 comment généraliser ce mécanisme à d'autres mesures, puis nous étudierons cette technique de généralisation afin de découvrir si d'autres mesures que la **confiance** peuvent présenter un tel caractère, c'est-à-dire si d'autres mesures pourraient être *anti-monotonisées* selon un processus similaire.

7.3 Une propriété de monotonie descendante de la confiance

La mesure de **confiance** focalise la plupart des travaux en recherche de règles et de pépites de connaissance [Omiecinski 03, Balcázar 09, Balcázar et al. 10]. C'est le cas pour la **all-confidence** que nous venons de présenter, mais aussi dans [Wang et al. 01] où une propriété de *monotonie* de la **confiance**, dans le contexte des bases de données catégorielles pour la recherche de règles de classes, est proposée. La différence avec une propriété d'anti-monotonie est que l'on est encore dans le cadre d'une propriété d'élagage, mais que celui-ci se fera cette fois du haut vers le bas. Rappelons que dans le contexte des données catégorielles, les attributs $\mathtt{A}_i$ de la base de données peuvent prendre un ensemble de valeurs $\mathcal{A}_i$, et qu'une règle d'association sera de la forme $\mathtt{A}_{i_1} = a_{i_1}, \ldots, \mathtt{A}_{i_k} = a_{i_k} \rightarrow \mathtt{C} = c$, ce que nous écrirons plutôt $\mathtt{P} = p \rightarrow \mathtt{c}$ pour plus de facilités. Dans ce cas, $\mathtt{P}$ est le k-uplet d'attributs $(\mathtt{A}_{i_1}, \ldots, \mathtt{A}_{i_k})$ et p est le k-uplet de valeurs $(a_{i_1}, \ldots, a_{i_k})$. Si $r : \mathtt{P} = p \rightarrow \mathtt{c}$ est une règle d'association, alors une $\mathtt{A}$-spécialisation de r est une règle de la forme $\mathtt{P} = p, \mathtt{A} = a \rightarrow \mathtt{c}$ où $a \in \mathcal{A}$ est une valeur possible de l'attribut $\mathtt{A}$. Nous allons ici introduire la propriété que nous noterons UEUC, intitulée *Universal Existential Upward Closure* par ses auteurs [Wang et al. 01].

7.3.1 La propriété UEUC

Nous commençons par introduire cette propriété sur l'exemple original de l'article [Wang et al. 01]. Supposons que nous ayons une base de données associant à chaque client d'un magasin un certain nombre de caractéristiques, ainsi que l'ensemble des articles qu'il a acheté. On peut par exemple obtenir ce type de base grâce aux cartes de fidélité des grandes surfaces. Nous aimerions savoir qui sont les clients qui achètent des salsifis. Disons que l'on en vienne à observer les trois règles suivantes :

$$\begin{array}{lrcl} \text{r}: & \hat{\text{A}}ge = jeune & \rightarrow & salsifis = oui \\ \text{r1}: & \hat{\text{A}}ge = jeune,\ Sexe = M & \rightarrow & salsifis = oui \\ \text{r2}: & \hat{\text{A}}ge = jeune,\ Sexe = F & \rightarrow & salsifis = oui \end{array}$$

Nous nous intéressons donc au comportement des jeunes clients (r), en particulier en fonction de leur sexe : masculin (r1) ou féminin (r2). Dans notre formalisme, nous aurions donc les attributs suivants, associés à leur espace de variation :

- $\mathtt{A}_1 = \hat{\text{A}}ge$, $\mathtt{A}_2 = Sexe$, $\mathtt{C} = salsifis$;
- $\mathcal{A}_1 = \{jeune, vieux\}$ $\mathcal{A}_2 = \{M, F\}$ et $\mathcal{C} = \{oui, non\}$.

Dans ces attributs, on remarque par exemple que les valeurs M et F de l'attribut $\mathtt{A}_2$ forment deux spécialisations disjointes. On retrouve donc les égalités suivantes :

$$\begin{aligned} conf(\text{r}) &= \frac{supp(\hat{\text{A}}ge{=}jeune, salsifis{=}oui)}{supp(\hat{\text{A}}ge{=}jeune)} \\ &= \frac{supp(\hat{\text{A}}ge{=}jeune, Sexe{=}M, salsifis{=}oui)}{supp(\hat{\text{A}}ge{=}jeune, Sexe{=}M) + supp(\hat{\text{A}}ge{=}jeune, Sexe{=}F)} \\ &+ \frac{supp(\hat{\text{A}}ge{=}jeune, Sexe{=}F, salsifis{=}oui)}{supp(\hat{\text{A}}ge{=}jeune, Sexe{=}M) + supp(\hat{\text{A}}ge{=}jeune, Sexe{=}F)} \\ &= \frac{supp(\hat{\text{A}}ge{=}jeune, Sexe{=}M)}{supp(\hat{\text{A}}ge{=}jeune, Sexe{=}M) + supp(\hat{\text{A}}ge{=}jeune, Sexe{=}F)} \times conf(\text{r1}) \\ &+ \frac{supp(\hat{\text{A}}ge{=}jeune, Sexe{=}F)}{supp(\hat{\text{A}}ge{=}jeune, Sexe{=}M) + supp(\hat{\text{A}}ge{=}jeune, Sexe{=}F)} \times conf(\text{r2}) \\ &= \alpha_1 \times conf(\text{r1}) + \alpha_2 \times conf(\text{r2}) \end{aligned}$$

où α_1 et α_2 sont des nombres positifs tels que $\alpha_1 + \alpha_2 = 1$. Ainsi donc, nous découvrons que la mesure de **confiance** de la règle r est en fait un barycentre de la mesure de **confiance** des règles r1 et r2. Ce dont nous déduisons qu'au moins l'une des deux règles r1 ou r2 a une mesure de **confiance** au moins aussi élevée que celle de r. Nous voyons donc poindre le nez de l'élagage, puisque si ni r1 ni r2 n'ont une mesure de **confiance** supérieure au seuil γ pré-fixé, alors r ne peut avoir une valeur de **confiance** au dessus de γ : r peut alors être retirée sans crainte de l'espace de recherche.

Nous proposons une généralisation de ce petit exemple [Le Bras et al. 10a]. Notons $\mathtt{P}$ un ensemble d'attributs et $\mathtt{A}$ un attribut ne se trouvant pas dans $\mathtt{P}$, dont les valeurs possibles sont regroupées dans l'ensemble $\mathcal{A}$:

$$\begin{array}{rcl} \exists(\alpha_1, \ldots, \alpha_{|\mathcal{A}|}) & \in & \mathbb{R}^+, \alpha_1 + \cdots + \alpha_{|\mathcal{A}|} = 1, \textit{ tel que} \\ & & conf(\mathtt{X} = x \rightarrow \mathtt{c}) = \sum_{a_i \in \mathcal{A}} \alpha_i \times conf(\mathtt{X} = x, \mathtt{A} = a_i \rightarrow \mathtt{c}). \end{array} \quad (7.1)$$

On connait d'ailleurs la valeur exacte des coefficients barycentriques, puisqu'elle est donnée par : $\alpha_i = \frac{supp(\mathtt{X}=x, \mathtt{A}=a_i)}{supp(\mathtt{X}=x)}$. Cette généralisation nous mène naturellement à la propriété de monotonie de

la confiance énoncée dans [Wang et al. 01].

> **Définition 11 – UEUC – :** Pour tout attribut $\mathtt{A}_i$ n'apparaissant pas dans une règle $\mathtt{P} = p \rightarrow \mathtt{c}$, (i) au moins une $\mathtt{A}_i$-spécialisation de $\mathtt{P} = p \rightarrow \mathtt{c}$ a une valeur de **confiance** au moins égale à celle de $\mathtt{P} = p \rightarrow \mathtt{c}$, (ii) si $\mathtt{P} = p \rightarrow \mathtt{c}$ a une valeur de **confiance** supérieure à un seuil γ, alors il en est de même pour au moins une $\mathtt{A}_i$-spécialisation de $\mathtt{P} = p \rightarrow \mathtt{c}$. Cette propriété est appelée *Universal Existential Upward Closure* (UEUC).

Cette propriété permet l'élaboration d'un algorithme de recherche des règles fondé sur un seuil de **confiance**, et sans contrainte de **support**.

7.3.2 Un algorithme d'élagage efficace

Dans l'article original, la propriété UEUC et l'algorithme qui en découle sont comparés à l'algorithme DenseMiner [Bayardo et al. 99], et l'efficacité de cette propriété semble évidente. Nous allons décrire ici l'algorithme s'appuyant sur cette propriété.

Comme nous l'avons dit précédemment, UEUC suggère un algorithme d'élagage du haut vers le bas. Nous allons donc débuter la recherche en considérant comme règles de départ l'ensemble des transactions, qui forment ainsi des règles de taille m. Nous allons plus particulièrement étudier le passage du niveau $k+1$ au niveau k. Supposons donc que nous ayons déjà trouvé toutes les règles de taille $k+1$ ayant une valeur de **confiance** supérieure à γ.

- dans un premier temps, on projette les règles du niveau $k+1$ sur le niveau k en supprimant un attribut : chaque règle de taille $k+1$ possède $k+1$ projections. Il s'agit là d'une surjection, car par exemple les règles r1 et r2 ont une projection en commun, la règle r. Cette projection permet de prendre en compte la partie *"au moins une"* de la propriété UEUC ;
- dans un second temps, on vérifie que chaque projection obtenue dans l'étape précédente admet bien une $\mathtt{A}_i$-spécialisation de **confiance** supérieure à γ pour chaque attribut $\mathtt{A}_i$ possible. Sur le niveau $k+1$, chaque projection doit admettre ainsi $m-k$ spécialisations. Cette étape permet de prendre en compte la partie *"Pour tout"* de la propriété ;
- finalement, on peut passer à la phase de validation des candidats : toutes les règles de taille k conservées à l'étape suivante et possédant une valeur de **confiance** supérieure au seuil γ pré-fixé sont conservées et seront utilisées pour la génération du niveau $k-1$.

C'est donc un algorithme en trois étapes, tout comme Apriori : il présente une phase de génération des candidats, une phase d'hérédité ainsi qu'une phase de validation. Il peut cependant être très coûteux car pour des bases de grandes dimensions, un niveau de règle peut très bien ne pas tenir en mémoire. L'avantage d'une telle stratégie est que l'on se détache de la notion de **support** ainsi que de ses contraintes. Cependant l'on peut toujours introduire un filtre de **support** ou de toute autre mesure puisque les valeurs de la table de contingence sont évaluées pour le calcul de la **confiance** (le **support** du conséquent est connu puisque les conséquents sont fixés). Il n'y aura donc pas de perte de temps, mais il faudra stocker ces règles filtrées séparemment. Cette stratégie permet de rechercher les pépites de connaissances qui sont en général très présentes : dans une base de données comme *mushroom*, les règles de **support** inférieur à 1% représentent plus de 80% des règles ayant une valeur de **confiance** supérieure à 0.8.

On remarque cependant qu'encore une fois, cette propriété se focalise uniquement sur la mesure de **confiance**, alors qu'un grand nombre de mesures d'intérêt existe [Guillaume et al. 10], parfois plus pertinentes. La propriété que nous allons présenter maintenant est l'une des premières permettant d'impliquer plusieurs mesures d'intérêt objectives au cœur de la phase d'élagage.

TABLE 7.1 : Base exemple pour les règles optimales

P	1	1	1	0	1	0	0	1	1	0	0
X	1	1	1	1	0	0	1	1	0	1	0
Y	1	0	0	1	1	1	0	1	0	0	1
c	0	0	0	1	1	1	1	1	1	1	0

TABLE 7.2 : Évaluation sur les règles extraites de la table 7.1

	force collective	confiance centrée	Q de Yule	Y de Yule	lift	indice de Gini	intérêt	J-mesure
$P \to c : (3/11, 6/11, 7/11)$	0.54	-0.14	-0.6	-0.3	0.78	0.04	0.074	0.021
$PX \to c : (3/11, 4/11, 7/11)$	0.26	-0.39	-0.9	-0.62	0.39	0.17	0.140	0.112
$PY \to c : (1/11, 3/11, 7/11)$	1.07	0.03	0.1	0.04	1.04	0.0006	0.008	0.0005
$PXY \to c : (1/11, 2/11, 7/11)$	0.81	-0.14	-0.3	-0.17	0.78	0.008	0.024	0.007

7.4 RECHERCHE D'ENSEMBLES DE RÈGLES OPTIMALES

Dans [Li 06] est introduite la notion d'ensemble de règles optimales. Une règle n'est pas optimale si elle a un intérêt plus faible qu'une de ses généralisations. En éliminant ces règles de l'ensemble complet des règles, on obtient un ensemble de règles optimales. Cette définition peut être utilisée pour toute mesure d'intérêt objective. En revanche, l'auteur propose une propriété d'anti-monotonie valable, elle, pour un ensemble restreint de mesures. Elle concerne, d'après l'auteur, 12 mesures, et pour chacune d'entre elles, cette propriété est prouvée. À cet ensemble de mesures s'ajoute une treizième mesure dans [Li et al. 05]. Nous allons détailler cette propriété en nous plaçant dans le même contexte de règles de classe que précédemment. Cependant, les attributs ne sont pas nécessairement catégoriels, nous noterons donc simplement P de manière générale, car même dans le cas catégoriel, l'instanciation n'est pas si importante que dans le cadre de la propriété UEUC.

7.4.1 Une propriété d'anti-monotonie

Nous présentons ici le théorème et les corollaires proposés par [Li 06]. Nous observons qu'ils concernent 13 mesures parmi les 42 que nous étudions dans cette thèse.

> **Théorème 1 – anti-monotonie – :** Si $supp(PX\neg c) = supp(P\neg c)$ alors ni la règle $PX \to c$ ni aucune de ses spécifications n'apparaissent dans un ensemble de règles optimales défini par les mesures suivantes : **confiance**, **odds ratio**, **lift**, **gain**, **confiance centrée**, **Klosgen**, **conviction**, **levier**, **Laplace**, **cosine**, **Loevinger**, **risque relatif** ou **Jaccard**.

Il est très naturel de se demander si cet ensemble de mesures pourrait être étendu, mais la question est loin d'être triviale. En effet, certaines mesures ne conviennent pas, ainsi que le démontre le détail de la table 7.2 qui s'appuie sur la base de données jouet de la table 7.1. Cet exemple jouet a été construit de telle manière que $supp(PX\neg c) = supp(P\neg c)$. Nous y présentons les cas de 8 mesures dont certaines n'ont pas été étudiées dans [Li 06]. Pour les cinq premières, l'inégalité $\mu(PX \to c) \leq \mu(P \to c)$ est respectée, ainsi que $\mu(PXY \to c) \leq \mu(PY \to c)$. La règle $PX \to c$ a donc un intérêt moindre que sa règle plus générale, et il en est de même pour toutes ses spécialisations. Attention, cela n'est pas une preuve que toutes ces mesures sont compatibles avec le théorème 1. Par contre, ce n'est pas le cas des trois dernières mesures, ce qui est en revanche une preuve que

ces mesures ne sont pas compatibles avec ce même théorème. Ce théorème possède deux corollaires directs.

> **Corollaire 1 – Propriété de clôture – :** Si $supp(\mathtt{P}) = supp(\mathtt{PX})$ alors pour toute valeur de `c`, ni la règle $\mathtt{PX} \rightarrow \mathtt{c}$ ni aucune de ses spécifications n'apparaissent dans un ensemble de règles optimales défini par les mesures suivantes : **`confiance`**, **`odds ratio`**, **`lift`**, **`gain`**, **`confiance centrée`**, **`Klosgen`**, **`conviction`**, **`levier`**, **`Laplace`**, **`cosine`**, **`Loevinger`**, ou **`Jaccard`**.

> **Corollaire 2 – Propriété de terminaison – :** Si $supp(\mathtt{P}\neg\mathtt{c}) = 0$ alors aucune règle plus spécifique que $\mathtt{P} \rightarrow \mathtt{c}$ n'apparait dans un ensemble de règles optimales défini par les mesures suivantes : by **`confiance`**, **`odds ratio`**, **`lift`**, **`gain`**, **`confiance centrée`**, **`Klosgen`**, **`conviction`**, **`levier`**, **`Laplace`**, **`cosine`**, **`Loevinger`**, ou **`Jaccard`**.

Ce théorème et ses deux corollaires inspirent une méthode d'élagage facilement implémentable. L'article propose une adaptation de la version de Apriori utilisée dans l'algorithme de classification CBA en insérant les propriétés d'anti-monotonie citées précédemment en plus de la propriété d'anti-monotonie du **`support`**.

7.4.2 Un algorithme de recherche efficace.

Dans l'algorithme suivant, nous appellerons candidat un couple $(\mathtt{P}, C)$ formé d'un antécédent $\mathtt{P}$ et d'un ensemble de conséquents C, c'est-à-dire un ensemble d'instances de l'attribut de classe : $C \subset \mathcal{C}$. Si la taille de $\mathtt{P}$ est l, nous appellerons le candidat un l-candidat. Un parent d'un l-candidat $(\mathtt{P}, C)$ est un couple $(\mathtt{P}', C')$ tel que $\mathtt{P}' \subset \mathtt{P}$ et $C \subset C'$.

L'algorithme OCGA (pour *Optimal Candidates Generation Algorithm*) génère des candidats optimaux, qui sont des règles candidates à l'optimalité. Nous allons voir qu'il est indépendant de la mesure choisie : celle-ci intervient dans une dernière phase pour vérifier que les candidats optimaux sont, ou non, réellement des règles optimales. L'ensemble de ces deux phases forme l'algorithme ORD (pour *Optimal Rule Discovery*). OCGA est composé de deux étapes principales. La première étape génère une base de $l+1$-candidats par combinaisons des candidats de taille l par le même principe que dans Apriori, et vérifie leur hérédité (tous les parents doivent être des candidats optimaux). A l'intérieur de OCGA, la deuxième étape consiste en une phase d'élagage, qui sera notre principal intérêt. Cet élagage s'appuie sur le théorème 1 et les corollaires 1 et 2. Nous insistons sur le fait que cette phase d'élagage est indépendante de la mesure : on récupère en sortie de l'algorithme un ensemble de règles qui contient les règles optimales pour n'importe laquelle des mesures d'intérêt du théorème 1. Cet ensemble est donc complet, mais, pour une mesure donnée, il n'est pas minimal. Nous décrivons ici ces deux algorithmes : la partie générale est donnée dans l'algorithme 4 et la fonction d'élagage optimal dans l'algorithme 5.

Après avoir généré l'ensemble des candidats optimaux sans avoir recours à aucune mesure d'intérêt, il est facile d'obtenir, étant donnée une mesure d'intérêt compatible et un seuil, l'ensemble des règles optimales, ainsi que l'ensemble des règles optimales et intéressantes. Ces opérations rendent l'ensemble minimal.

Remarquons cependant que l'ensemble de règles que nous obtenons n'est pas l'ensemble des règles intéressantes, car il manque les règles intéressantes mais non optimales. Cependant, devant la grande efficacité de cet algorithme, notre démarche va être d'essayer d'étendre l'ensemble des mesures qui peuvent être utilisées.

Données : base $\mathcal{DB}$, seuil de `rappel`
Résultat : un ensemble complet de candidats optimaux
construire les 1-candidats;
générer les 2-candidats;
si *l'ensemble des l-candidats n'est pas vide* **alors**
  pour $(P, C) \subset$ *l-candidats* **faire**
    appliquer l'élagage optimal à $(\mathtt{P}, C)$;
    si C *est vide* **alors** retirer $(\mathtt{P}, C)$;
  fin
  générer les $l+1$-candidats;
fin

ALGORITHME 4: OCGA [Li 06]

Données : un l-candidat $(\mathtt{P}, C)$
Résultat : le conséquent C est élagué
pour $\mathtt{c} \in C$ **faire**
  élagage sur le `rappel` pour $\mathtt{P} \rightarrow \mathtt{c}$;
  élagage sur le corollaire 2 $\mathtt{P} \rightarrow \mathtt{c}$;
fin
si C *est vide* **alors** sortir;
pour $P' \subset P$ **faire**
  élagage sur le corollaire 1;
  élagage sur le théorème 1;
fin
si C *est vide* **alors** sortir

ALGORITHME 5: Fonction d'élagage optimal [Li 06]

CONCLUSION

Nous avons mis en évidence trois propriétés d'élagage qui permettent de découvrir trois ensembles de règles distincts. La propriété UEUC permet d'obtenir l'ensemble des règles de mesure de `confiance` dépassant un certain seuil. Les autres propriétés, liées à la `all-confidence` et à la recherche de règles optimales, ne renvoient pas toutes les règles intéressantes au sens d'une mesure, mais offrent la possibilité d'utiliser des algorithmes efficaces pour obtenir des ensembles intéressants de règles sous certains critères. Les trois chapitres suivants sont consacrés à la généralisation de ces travaux, et à l'établissement de conditions nécessaires et suffisantes permettant d'établir la compatibilité des mesures avec ces généralisations.

8 Généralisation de la all-confidence

Dans ce chapitre, nous nous concentrons sur le système descripteur S_{ex} associé au domaine adapté D_{ex}. Lorsque nous étudions les variations d'une fonction f de x, y et z définie sur un domaine D "par rapport à sa 1ère variable", nous étudions en fait les variations de la fonction définie par $x \mapsto f(x, y, z)$ pour tout couple (y, z) sur le domaine $\{x|(x, y, z) \in D\}$.

8.1 Une transformation des mesures

Nous commençons par définir un procédé de transformation des mesures identique au passage de la **confiance** à la **all-confidence** [Le Bras et al. 09c]. Il s'agit de transférer les mesures des règles dans les mesures des motifs. Nous généralisons donc la définition de la **all-confidence**.

> **Définition 12 – omni-mesure – :** Soit m une mesure d'intérêt des règles d'association. Nous définissons l'omni-mesure $o\text{-}m$ sur un motif $\mathtt{I}$ à partir de m de la façon suivante :
>
> $$o\text{-}m(\mathtt{I}) \quad = \quad \min_{\mathtt{AB}=\mathtt{I},\mathtt{A}\neq\emptyset,\mathtt{B}\neq\emptyset}(m(\mathtt{A} \rightarrow \mathtt{B})).$$

L'omni-mesure d'un motif $\mathtt{I}$ est donc la plus petite valeur que puisse prendre la mesure considérée sur une règle extraite du motif $\mathtt{I}$. Nous remarquons donc que la **all-confidence** est bien l'omni-mesure de la **confiance**.

> **Définition 13 – omni-monotonie – :** m est dite omni-monotone si la mesure $o\text{-}m$ est anti-monotone.

Cette propriété très intéressante du point de vue algorithmique puisqu'elle permettrait d'utiliser des mesures directement au cœur d'Apriori, a été démontrée dans sa version non générale par [Omiecinski 03] pour la **confiance**. Nous montrons que ce résultat est vérifié pour les transformées monotones de la **confiance** :

> **Propriété 3 :** La **confiance**, et toutes les mesures fonctions croissantes de la **confiance** seule sont omni-monotones.

Démonstration. En effet, soit m une telle mesure, et $\mathtt{I} \subset \mathtt{I}'$ deux motifs. Puisque m est croissante en fonction de la **confiance**, $\min_{\mathtt{AB}=\mathtt{I}} m(conf(\mathtt{A} \rightarrow \mathtt{B})) = m\left(\min_{\mathtt{AB}=\mathtt{I}}(conf(\mathtt{A} \rightarrow \mathtt{B}))\right)$. Et comme la **confiance** est elle même omni-monotone, $m\left(\min_{\mathtt{AB}=\mathtt{I}}(conf(\mathtt{A} \rightarrow \mathtt{B}))\right) \geq m\left(\min_{\mathtt{AB}=\mathtt{I}'}(conf(\mathtt{A} \rightarrow \mathtt{B}))\right)$. Finalement, m est omni-monotone :

$$\min_{\mathtt{AB}=\mathtt{I}}(m(conf(\mathtt{A} \rightarrow \mathtt{B}))) = m\left(\min_{\mathtt{AB}=\mathtt{I}}(conf(\mathtt{A} \rightarrow \mathtt{B}))\right) \geq m\left(\min_{\mathtt{AB}=\mathtt{I}'}(conf(\mathtt{A} \rightarrow \mathtt{B}))\right) = \min_{\mathtt{AB}=\mathtt{I}'} m(conf(\mathtt{A} \rightarrow \mathtt{B}))$$

□

Ainsi, la **confiance** n'est pas la seule mesure à posséder cette propriété et donc à être *anti-monotonisable* par le même procédé. Cette propriété a été publiée dans [Le Bras et al. 09c], mais nous donnons ici un résultat un peu plus général, mais compatible, et qui nous permettra de traiter plus de mesures.

> **Propriété 4 :** Soit m une mesure d'intérêt et Φ_m sa fonction de mesure adaptée au domaine des exemples. Si Φ_m est constante par rapport à sa troisième variable, croissante par rapport à la première et décroissante par rapport à la deuxième, alors m est omni-monotone.

Démonstration. Si $\Phi_m^{S_{ex}}$ remplit ces conditions, alors il est possible de donner explicitement la valeur de $o\text{-}m(\mathtt{I})$ pour un motif $\mathtt{I}$ quelconque. On a en effet pour tout sous-motif $\mathtt{X}$ de $\mathtt{I}$: $supp(\mathtt{X}) \leq \max_{\mathtt{A}\in\mathtt{I}}(supp(\mathtt{A}))$. On note $\mathtt{A}_0$ l'attribut réalisant ce maximum, et $\mathtt{B}_0$ son complément dans $\mathtt{I}$, on a donc :

$$\begin{aligned} o\text{-}m(\mathtt{I}) &= \min_{\mathtt{AB}=\mathtt{I}}(m(\mathtt{A}\rightarrow\mathtt{B})) \\ &= \min_{\mathtt{AB}=\mathtt{I}}(\Phi_m(p_{\mathtt{AB}}, p_{\mathtt{A}}, p_{\mathtt{B}})) \\ &= \min_{\mathtt{AB}=\mathtt{I}}(\Phi_m(p_{\mathtt{I}}, p_{\mathtt{A}}, p_{\mathtt{B}})) \\ &= \min_{\mathtt{AB}=\mathtt{I}}(\Phi_m(p_{\mathtt{I}}, p_{\mathtt{A}}, p_{\mathtt{B}_0})) \\ &\geq \Phi_m(p_{\mathtt{I}}, p_{\mathtt{A}_0}, p_{\mathtt{B}_0}) \end{aligned}$$

Puisque la règle $\mathtt{A}_0 \rightarrow \mathtt{B}_0$ est une règle particulière issue de $\mathtt{I}$, on a en fait égalité. Nous allons nous servir de ceci pour montrer l'anti-monotonie. Donnons-nous un motif $\mathtt{I}'$ tel que $\mathtt{I} \subset \mathtt{I}'$. En particulier, $\mathtt{A}_0 \in \mathtt{I}'$ et $\mathtt{B}_0 \in \mathtt{I}'$. Ainsi nous avons l'inégalité

$$supp(\mathtt{A}'_0) = \max_{\mathtt{A}\in\mathtt{I}'}(supp(\mathtt{A})) \geq supp(\mathtt{A}_0)$$

qui nous permet d'écrire :

$$\begin{aligned} o\text{-}m(\mathtt{I}') &= \Phi_m(p_{\mathtt{I}'}, p_{\mathtt{A}'_0}, p_{\mathtt{B}'_0}) \\ &\leq \Phi_m(p_{\mathtt{I}}, p_{\mathtt{A}'_0}, p_{\mathtt{B}'_0}) \\ &\leq \Phi_m(p_{\mathtt{I}}, p_{\mathtt{A}'_0}, p_{\mathtt{B}_0}) \\ &\leq \Phi_m(p_{\mathtt{I}}, p_{\mathtt{A}_0}, p_{\mathtt{B}_0}) \\ &\leq o\text{-}m(\mathtt{I}) \end{aligned}$$

Nous avons ainsi montré la condition d'anti-monotonie. □

La constance par rapport à la troisième variable est indispensable, nous allons voir que lorsqu'elle n'est pas présente, nos résultats (théorèmes 2 et 3) montrent qu'un grand nombre de mesures classiquement étudiées dans la littérature ne vérifient pas cette propriété.

8.2 Théorèmes d'exclusion

Nous allons donc énoncer ici deux théorèmes dont nous verrons qu'ils excluent un grand nombre de mesures de la propriété d'omni-monotonie.

> **Théorème 2 :** Soit m une mesure d'intérêt des règles d'association, et (Φ_m, D_{ex}) la fonction de mesure adaptée à m. Si Φ_m est strictement décroissante en la deuxième (antécédents) et troisième (conséquents) variable, alors m n'est pas omni-monotone.

Démonstration. Supposons que cette fonction de mesure soit décroissante strictement en la deuxième et la troisième variable. Nous allons montrer qu'il est possible de construire une base de données qui sert de contre-exemple à l'omni-monotonie. Choisissons de travailler sur le nombre minimal d'attributs dont nous avons besoin pour étudier une situation simple d'omni-monotonie, c'est-à-dire 3. Nommons les A, B et C. Nous voulons donc comparer dans une certaine base de données les règles issues du motif AB et celles issues du motif ABC. Nous allons donner la définition de valeurs rationnelles $\{x, y, z, t, y', z'\}$ en fonction de diverses contraintes, avec pour but d'obtenir dans notre base de données les relations :

$$supp(\mathtt{AB}) = x,\ supp(\mathtt{A}) = y,\ supp(\mathtt{B}) = z$$
$$supp(\mathtt{ABC}) = x,\ supp(\mathtt{C}) = t,\ supp(\mathtt{AC}) = y',\ supp(\mathtt{BC}) = z'$$

Les propriétés des fréquences impliquent :
- $0 < supp(\mathtt{A}) < 1$, c'est-à-dire $0 < y < 1$;
- $0 < supp(\mathtt{B}) < 1$, c'est-à-dire $0 < z < 1$;
- $0 < supp(\mathtt{ABC}) \leq \min(supp(\mathtt{AC}), supp(\mathtt{BC}))$, c'est-à-dire $0 < x \leq \min(y', z')$;
- $y' = supp(\mathtt{AC}) < supp(\mathtt{A}) = y$;
- $z' = supp(\mathtt{BC}) < supp(\mathtt{B}) = z$;
- $t = supp(\mathtt{C}) \geq supp(\mathtt{AC}) + supp(\mathtt{BC}) - supp(\mathtt{ABC}) = y' + z' - x$.

Nous posons des inégalités strictes pour pouvoir exploiter pleinement la stricte décroissance de la mesure. Supposons de plus que l'attribut C apparait moins fréquemment que B, c'est-à-dire $t \leq z$.

Résumons donc les contraintes qui permettent de définir nos rationnels :

$$y \in]0, 1[\tag{8.1}$$

$$z \in]0, 1[\tag{8.2}$$

$$0 < y' < y \tag{8.3}$$

$$0 < z' < z \tag{8.4}$$

$$0 < x \leq \min\{z', y'\} \tag{8.5}$$

$$y' + z' - x < t \tag{8.6}$$

$$t \leq z \tag{8.7}$$

$$t + (y - y') + (z - z') \leq 1 \tag{8.8}$$

Remarque 1 : Notons immédiatement que les valeurs $y = \frac{4}{16}$, $z = \frac{8}{16}$, $y' = \frac{2}{16}$, $z' = \frac{3}{16}$, $x = \frac{1}{16}$, $t = \frac{7}{16}$ vérifient ces contraintes et permettent d'engendrer la base de données décrite table 8.1.

Remarque 2 : D'une manière plus générale, on s'aperçoit que pour un entier n donné, toute base de données à trois attributs $\{\mathtt{A}, \mathtt{B}, \mathtt{C}\}$ et $n + 3$ transactions $\mathtt{T}_1, \cdots, \mathtt{T}_{n+3}$ réparties comme suit : $\mathtt{T}_1 = \{\mathtt{A}\}, \mathtt{T}_2 = \{\mathtt{B}\}, \mathtt{T}_3 = \{\mathtt{C}\}, \mathtt{T}_4 = \mathtt{T}_5 = \cdots = \mathtt{T}_{n+3} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}\}$, vérifie les contraintes.

A	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
B	1	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0
C	1	1	0	0	1	1	0	0	0	0	0	1	1	1	0	0

TABLE 8.1 : Base de données à 16 transactions

Sur chacune de ces bases de données, les contraintes nous permettent d'écrire les inégalités suivantes (où on trouve au-dessus du signe d'inégalité la contrainte qui permet le passage), en prenant en compte les hypothèses de variations :

$$\begin{array}{lcl}
m(\mathtt{A}\to\mathtt{B}) = \Phi_m(x,y,z) & \overset{8.3}{<} & \Phi_m(x,y',z) = m(\mathtt{AC}\to\mathtt{B}) \\
& \overset{8.4}{<} & \Phi_m(x,y,z') = m(\mathtt{A}\to\mathtt{BC}) \\
& \overset{8.5+8.3}{<} & \Phi_m(x,x,z) \overset{8.7}{\leq} \Phi_m(x,x,t) = m(\mathtt{AB}\to\mathtt{C}) \\
m(\mathtt{B}\to\mathtt{A}) = \Phi_m(x,z,y) & \overset{8.4}{<} & \Phi_m(x,z',y) = m(\mathtt{BC}\to\mathtt{A}) \\
& \overset{8.3}{<} & \Phi_m(x,z,y') = m(\mathtt{B}\to\mathtt{AC}) \\
& \overset{8.5+8.4}{<} & \Phi_m(x,z,x) \overset{8.7}{\leq} \Phi_m(x,t,x) = m(\mathtt{C}\to\mathtt{AB})
\end{array}$$

Ces inégalités nous donnent la conclusion : $o\text{-}m(\{\mathtt{A},\mathtt{B}\}) < o\text{-}m(\{\mathtt{A},\mathtt{B},\mathtt{C}\})$, ce qui contredit l'anti-monotonie. □

L'une des conséquences de ce théorème est en particulier que les mesures qui respectent strictement les principes de Piatetsky-Shapiro (strictement signifie que les croissances sont entendues au sens strict) ne seront pas omni-monotones ! La propriété d'omni-monotonie est donc très contraignante, mais nous relèverons néanmoins 10 mesures la vérifiant.

Exemple 5 – `facteur bayésien` – : La fonction de mesure adaptée sur D_{ex} à la mesure du **`facteur bayésien`** a la forme suivante : $\Phi_{BF}(x,y,z) = \frac{x}{y-x} \times \frac{1-z}{z}$. Elle possède les propriétés de décroissance énoncées précédemment. Elle doit donc ne pas être omni-monotone. Calculons BF sur la base de la table 8.1 :

$$\begin{array}{llll}
BF(\mathtt{A}\to\mathtt{B}) = \frac{1}{3} & BF(\mathtt{AC}\to\mathtt{B}) = 1 & BF(\mathtt{B}\to\mathtt{AC}) = 1 & BF(\mathtt{BC}\to\mathtt{A}) = \frac{3}{2} \\
BF(\mathtt{B}\to\mathtt{A}) = \frac{3}{7} & BF(\mathtt{A}\to\mathtt{BC}) = \frac{13}{9} & BF(\mathtt{AB}\to\mathtt{C}) = \infty & BF(\mathtt{C}\to\mathtt{AB}) = \infty
\end{array}$$

Ces calculs mettent en évidence l'inexistence de l'omni-monotonie, puisque $o\text{-}BF(\{\mathtt{A},\mathtt{B}\}) < o\text{-}BF(\{\mathtt{A},\mathtt{B},\mathtt{C}\})$.

Remarque 3 : L'hypothèse de stricte monotonie est trop forte. Voyons la mesure de **`Loevinger`** : $L(\mathtt{A}\to\mathtt{B}) = 1 - \frac{supp(\mathtt{A}\bar{\mathtt{B}})}{supp(\mathtt{A})supp(\bar{\mathtt{B}})}$. Sa fonction de mesure adaptée sur D_{ex} est $\Phi_L(x,y,z) = 1 - \frac{y-x}{y\cdot(1-z)}$. Et sur le plan défini par $x-y=0$, sa dérivée par rapport à la troisième variable est nulle. Il n'y a donc pas stricte monotonie par rapport à la troisième variable. Pourtant, sur la base de données définie précédemment :

$$\begin{array}{llll}
L(\mathtt{A}\to\mathtt{B}) = -\frac{1}{2} & L(\mathtt{AC}\to\mathtt{B}) = 0 & L(\mathtt{B}\to\mathtt{AC}) = 0 & L(\mathtt{BC}\to\mathtt{A}) = \frac{1}{9} \\
L(\mathtt{B}\to\mathtt{A}) = -\frac{1}{6} & L(\mathtt{A}\to\mathtt{BC}) = \frac{1}{13} & L(\mathtt{AB}\to\mathtt{C}) = 1 & L(\mathtt{C}\to\mathtt{AB}) = \frac{3}{35}
\end{array}$$

La mesure de **`Loevinger`** n'est donc pas omni-monotone. D'autres mesures présentent la même particularité en la deuxième variable sur le plan $x-z=0$.

Puisque l'hypothèse de stricte monotonie est trop forte, nous allons l'affaiblir un peu en donnant le théorème suivant.

Théorème 3 : Soit m une mesure d'intérêt des règles d'association, et (Φ_m, D_{ex}) la fonction de mesure adaptée à m. Si $\Phi_m(x, y, z)$ est strictement décroissante en la deuxième et troisième variable, ailleurs que sur les plans $x - z = 0$ et $x - y = 0$, où elle peut éventuellement être constante respectivement en la deuxième et la troisième variable, alors m n'est pas omni-monotone.

Démonstration. La démonstration est la même que pour le théorème 2, les inégalités $\overset{8.7}{\leq}$ devenant des égalités. □

La réciproque du théorème 3 n'est pas vérifiée et nous n'avons donc pas de condition nécessaire et suffisante. Le cas de la mesure **moindre contradiction** $contr(\mathtt{AB}) = \frac{supp(\mathtt{AB}) - supp(\mathtt{A\bar{B}})}{supp(\mathtt{B})}$ [Azé et Kodratoff 04] justifie cette affirmation. On l'applique à une base de données de taille n décrite remarque 2 : on mesure alors : $o\text{-}contr(\{\mathtt{A}, \mathtt{B}\}) = \frac{n-1}{n+1}$ et $o\text{-}contr(\{\mathtt{A}, \mathtt{B}, \mathtt{C}\}) = \frac{n-1}{n}$. C'est-à-dire $o\text{-}contr(\{\mathtt{A}, \mathtt{B}\}) < o\text{-}contr(\{\mathtt{A}, \mathtt{B}, \mathtt{C}\})$ dès que $n > 1$. Ainsi la mesure de **moindre contradiction** n'est pas omni-monotone. Sa fonction de mesure adaptée est $\frac{2x-y}{z}$, dont le sens de variation par rapport à z dépend du signe de $2x - y$: *contr* n'est pas décroissante en fonction de sa troisième variable, pas même au sens large.

8.3 Classification des mesures

Ces travaux nous permettent d'apporter trois sortes de conclusion sur les 42 mesures que nous avons relevées, la première s'appuyant sur le théorème 3, la deuxième sur la propriété 3 et la propriété 4, et la dernière mettant en avant le lien entre **rappel** et **confiance** :

- En combinant le résultat du théorème 3 et, pour certains cas, l'utilisation des différents contre-exemples des remarques 1 et 2, nous observons qu'un certain nombre de mesures ne vérifient pas la propriété de omni-monotonie : **spécificité**, **précision**, **lift**, **levier**, **confiance centrée**, **Jaccard**, **confiance positive**, **odds ratio**, **Klosgen**, **moindre contradiction**, **valeur ajoutée symétrique**, **conviction**, **One Way Support**, **J1-mesure**, **Piatetsky-Shapiro**, **cosine**, **Loevinger**, **gain informationnel**, **facteur bayésien**, **Zhang**, **indice d'implication**, **kappa**... ;
- A l'opposé, les mesures de **confiance**, **Sebag-Shoenauer**, **Ganascia** et le **TEC** vérifient cette propriété, car elles sont des fonctions croissantes de la **confiance** (elles classent les règles de la même façon mais ces mesures se différencient selon des objectifs utilisateurs, [Lenca et al. 08]), il est donc possible de les utiliser dans un algorithme de type Apriori pour réaliser un élagage du treillis des motifs, au même titre que les mesures de **Laplace**, **couverture**, **gain**, **prevalence** qui relève de la propriété 4, plus générale ;
- Enfin, concernant la mesure de **rappel**, on remarque que pour deux itemsets $\mathtt{A}$ et $\mathtt{B}$, $rec(\mathtt{A} \rightarrow \mathtt{B}) = \frac{supp(\mathtt{AB})}{supp(\mathtt{B})} = conf(\mathtt{B} \rightarrow \mathtt{A})$. Ainsi, pour un motif $\mathtt{I}$, $\{rec(\mathtt{A} \rightarrow \mathtt{B}) | \mathtt{AB} = \mathtt{I}\} = \{conf(\mathtt{A} \rightarrow \mathtt{B}) | \mathtt{AB} = \mathtt{I}\}$. Donc $o\text{-}rec(\mathtt{I}) = o\text{-}conf(\mathtt{I})$, et la mesure de **rappel** est omni-montone.

Ces résultats sont résumés dans le tableau 8.2 qui montre que les mesures **indice de Gini**, **intérêt** et **J-mesure** ne peuvent pas être classées car elles n'entrent dans aucune des propriétés énoncées. Souvenons-nous que ce résultat est essentiellement négatif, car il exclue un grand nombre de mesures classiques, contrairement aux résultats que nous allons proposer sur les deux autres propriétés.

Conclusion

La mesure de **all-confidence** est une version de la **confiance** adaptée aux motifs, et qui possède une propriété d'anti-monotonie. Nous avons montré comment adapter toute mesure d'une

mesure	OMNI	mesure	OMNI
confiance	✓	confiance centrée	✗
moindre contradiction	✗	conviction	✗
cosine	✗	couverture	✓
Czekanowski	✗	facteur bayésien	✗
force collective	✗	gain	✓
gain informationnel	✗	Ganascia	✓
indice de Gini	?	indice d'implication	✗
intérêt	?	J1-mesure	✗
Jaccard	✗	J-mesure	?
Kappa	✗	Klosgen	✗
Kulczynski	✗	Laplace	✓
levier	✗	lift	✗
Loevinger	✗	odds ratio	✗
one way support	✗	coefficient de Pearson	✗
Piatetsky-Shapiro	✗	précision	✗
prevalence	✓	Q de Yule	✗
rappel	✓	risque relatif	✗
Sebag-Shoenauer	✓	spécificité	✗
spécificité relative	✗	support	✓
taux exemples contre-exemples	✓	valeur ajoutée	✗
Y de Yule	✗	Zhang	✗

TABLE 8.2 : Existence de l'anti-monotonie pour l'omni-mesure correspondante. Dans cette table et les suivantes, le symbole ✓ signifie que la propriété est vérifiée, ✗ que la propriété n'est pas vérifiée, et un ? indique que nos conditions ne permettent pas de répondre.

façon identique, et nous avons proposé deux conditions suffisantes d'existence d'une propriété d'anti-monotonie pour cette adaptation. Nous avons ainsi pu mettre en évidence 10 mesures possédant cette propriété. La condition nécessaire que nous avons énoncée permet d'en exclure 29 autres. Trois autres mesures restent non classées, par manque d'une condition nécessaire *et* suffisante. Le chapitre suivant établit des résultats similaires pour généraliser la propriété UEUC, qui est aussi une propriété de la `confiance`.

9 Généralisation de la propriété UEUC

Nous nous intéressons ici à la propriété UEUC [Wang et al. 01] qui a été définie à l'origine pour la mesure de **confiance**, tout comme la notion de **all-confidence**. Nous proposons ici d'énoncer une condition nécessaire, ainsi qu'une condition suffisante à l'existence d'une propriété de monotonie liée à la propriété UEUC pour n'importe quelle mesure. Ce sera notamment l'occasion d'utiliser un autre domaine adapté, celui associé à la confiance, et donc de justifier encore un peu plus la définition de notre cadre formel.

9.1 Généralisation de la propriété UEUC

L'une des limites de la propriété UEUC est sa définition dépendante de la **confiance**. D'après l'étude que nous en avons faite, cette propriété possède cependant des avantages certains, notamment algorithmiques, que nous aimerions pouvoir transférer à d'autres mesures.

Nous proposons une propriété UEUC générale pour toute mesure [Le Bras et al. 10a], puis nous étudierons les mesures vérifiant effectivement cette propriété.

> **Définition 14 – Propriété générale UEUC – :** Une mesure d'intérêt m vérifie la propriété générale UEUC (GUEUC) si et seulement si pour tout attribut $\mathtt{A}_i$ n'apparaissant pas dans une règle r : $\mathtt{P} = p \rightarrow \mathtt{c}$, au moins l'une des $\mathtt{A}_i$-spécialisations de r prend une valeur de mesure m au moins égale à celle de r. Une conséquence directe est que si r est une règle intéressante pour un seuil donné, alors au moins l'une de ses $\mathtt{A}_i$-spécialisations l'est aussi.

La **confiance** vérifie clairement cette propriété GUEUC et nous voulons, comme cela a été fait dans le chapitre précédent, pouvoir établir l'ensemble des mesures la vérifiant de manière automatique, c'est-à-dire à l'aide de conditions nécessaires et/ou suffisantes. Nous allons commencer par étudier la propriété GUEUC sur deux mesures particulières dont nous allons voir que l'une respecte cette propriété mais pas l'autre et qui vont nous permettre de faire apparaître des propriétés intéressantes.

Considérons la mesure de **Sebag-Shoenauer** qui peut s'écrire sous la forme

$$seb(\mathtt{P} = p \rightarrow \mathtt{c}) = \frac{supp(\mathtt{P} = p, \mathtt{c})}{supp(\mathtt{P} = p, \neg\mathtt{c})}.$$

Soit $\mathtt{A}$ un attribut ne se trouvant pas dans $\mathtt{P}$, les égalités suivantes sont vérifiées :

$$
\begin{aligned}
seb(\mathtt{P}=p\rightarrow\mathtt{c}) &= \frac{\sum_{a\in\mathcal{A}} supp(\mathtt{P}=p,\mathtt{A}=a,\mathtt{c})}{supp(\mathtt{P}=p,\neg\mathtt{c})}\\
&= \sum_{a\in\mathcal{A}} \frac{supp(\mathtt{P}=p,\mathtt{A}=a,\mathtt{c})}{supp(\mathtt{P}=p,\neg\mathtt{c})}\\
&= \sum_{a\in\mathcal{A}} \frac{supp(\mathtt{P}=p,\mathtt{A}=a,\neg\mathtt{c})}{supp(\mathtt{P}=p,\neg\mathtt{c})} \times \frac{supp(\mathtt{P}=p,\mathtt{A}=a,\mathtt{c})}{supp(\mathtt{P}=p,\mathtt{A}=a,\neg\mathtt{c})}\\
&= \sum_{a\in\mathcal{A}} \alpha_a \times \frac{supp(\mathtt{P}=p,\mathtt{A}=a,\mathtt{c})}{supp(\mathtt{P}=p,\mathtt{A}=a,\neg\mathtt{c})}\\
&= \sum_{a\in\mathcal{A}} \alpha_a \times seb(\mathtt{P}=p,\mathtt{A}=a\rightarrow\mathtt{c})
\end{aligned}
$$

où les α_a sont des nombres positifs tels que $\sum_{a\in\mathcal{A}} \alpha_a = 1$. La mesure de **Sebag-Shoenauer** présente donc la même propriété barycentrique que la mesure de **confiance** et nous pouvons en déduire aisément qu'elle respecte donc la propriété GUEUC.

Au contraire, considérons la mesure de **levier** définie par

$$lev(\mathtt{P}=p\rightarrow\mathtt{c}) = supp(\mathtt{P}=p,\mathtt{c}) - supp(\mathtt{P}=p)\times supp(\mathtt{c}).$$

Pour cette mesure, nous pouvons écrire les égalités suivantes :

$$
\begin{aligned}
lev(\mathtt{X}=x\rightarrow\mathtt{c}) &= \sum_{a\in\mathcal{A}} (supp(\mathtt{P}=p,\mathtt{A}=a,\mathtt{c}) - supp(\mathtt{P}=p,\mathtt{A}=a)\times supp(\mathtt{c}))\\
&= \sum_{a\in\mathcal{A}} lev(\mathtt{P}=p,\mathtt{A}=a\rightarrow\mathtt{c})
\end{aligned}
$$

La mesure de **levier** d'une règle est donc la somme des mesures de ses spécifications, et ne vérifie pas la propriété GUEUC. Une question arrive alors naturellement : comment savoir quelles mesures vérifient la propriété GUEUC, et quelles mesures ne la vérifient pas ?

9.2 Conditions d'existence de GUEUC

Pour étudier les conditions d'existence de la propriété GUEUC nous allons nous concentrer sur le domaine adapté lié à la confiance D_{conf}. Comme nous l'avons vu précédemment nous étudierons les variations des fonctions de mesure adaptées sur ce domaine.

9.2.1 Une condition suffisante

Nous allons commencer par étudier une condition suffisante d'existence de la propriété GUEUC pour une mesure d'intérêt m. Remarquons que nous en avons déjà découvert une avec la propriété de barycentre énoncée plus haut :

Proposition 1 – Condition suffisante triviale – : Soit m une mesure d'intérêt des règles d'association, transformation affine de la **confiance** (où $p_{\mathtt{B}}$ peut éventuellement apparaître mais est fixé dans le cadre des règles de classe). Alors m vérifie la même propriété de barycentre que la **confiance**, et possède donc la propriété GUEUC.

Cette condition regroupe 6 mesures parmi les 42 que nous avons relevées : **confiance**, **prevalence**, **confiance centrée**, **lift**, **Ganascia** et **Loevinger**. Le problème de cette condition est qu'elle n'est pas automatiquement applicable, comme pourrait l'être une condition sur les variations des fonctions. Nous proposons donc la condition suivante :

> **Proposition 2 – Condition suffisante – :** Soit m une mesure d'intérêt des règles d'association et (Φ_m, D_{conf}) sa fonction de mesure adaptée. Supposons que Φ_m vérifie les propriétés suivantes sur D_{conf} :
>
> **(a)** $\forall (y,z) \in [0,1]^2$, la fonction $c \mapsto \Phi_m(c,y,z)$, où c est tel que $(c,y,z) \in D_{conf}$, est monotone par rapport à sa première variable (croissante ou décroissante) ;
>
> **(b)** $\forall (c,z) \in [0,1]^2$, la fonction $y \mapsto \Phi_m(c,y,z)$, où y est tel que $(c,y,z) \in D_{conf}$, est décroissante par rapport à sa deuxième variable (au sens large).
>
> Alors m possède la propriété GUEUC.

Démonstration. Commençons par faire une remarque qui nous sera utile afin de s'assurer que dans les calculs suivants, tout se passe bien dans D_{conf}. Nous remarquons que si $(c,y,z) \in D_{conf}$ et que l'on se donne un nombre y' positif tel que $y' \leq y$, alors le triplet (c,y',z) appartient lui aussi à D_{conf}.

Considérons donc une règle r : $\mathtt{P} = p \rightarrow \mathtt{c}$ et un attribut $\mathtt{A}$ ne faisant pas partie des attributs composant le motif $\mathtt{P}$. Puisque la **confiance** possède la propriété de barycentre, il existe deux instances de l'attribut $\mathtt{A}$, que nous notons $a^{\uparrow}$ et $a^{\downarrow}$ telles que

$$conf(\mathtt{P} = p, \mathtt{A} = a^{\downarrow} \rightarrow \mathtt{c}) \leq conf(\mathtt{P} = p \rightarrow \mathtt{c}) \leq conf(\mathtt{P} = p, \mathtt{A} = a^{\uparrow} \rightarrow \mathtt{c}).$$

Supposons alors que, pour $y = supp(\mathtt{X} = x, \mathtt{A} = a^{\uparrow})$ (respectivement $y = supp(\mathtt{X} = x, \mathtt{A} = a^{\downarrow})$) la propriété (a) soit une propriété de croissance (resp. décroissance). Nous posons $a = a^{\uparrow}$ (resp. $a = a^{\downarrow}$) et les inégalités suivantes sont vérifiées (les lettres au-dessus des inégalités indiquent la propriété utilisée) :

$$\begin{aligned} m(\mathtt{P} = p \rightarrow \mathtt{c}) &= \Phi_m(conf(\mathtt{P} = p \rightarrow \mathtt{c}), supp(\mathtt{P} = p), supp(\mathtt{c})) \\ &\overset{(b)}{\leq} \Phi_m(conf(\mathtt{P} = p \rightarrow \mathtt{c}), supp(\mathtt{P} = p, \mathtt{A} = a), supp(\mathtt{c})) \\ &\overset{(a)}{\leq} \Phi_m(conf(\mathtt{P} = p, \mathtt{A} = a \rightarrow \mathtt{c}), supp(\mathtt{P} = p, \mathtt{A} = a), supp(\mathtt{c})) \\ &\leq m(\mathtt{P} = p, \mathtt{A} = a \rightarrow \mathtt{c}) \end{aligned}$$

Ainsi la règle $\mathtt{P} = p \rightarrow \mathtt{c}$ admet une $\mathtt{A}$-spécialisation $\mathtt{P} = p, \mathtt{A} = a \rightarrow \mathtt{c}$ plus intéressante et m possède donc la propriété GUEUC. □

Cette condition nous permet donc d'ajouter à la liste des mesures qui vérifient GUEUC deux mesures fonctions croissantes de la **confiance** (**Sebag-Shoenauer**, **TEC**) ainsi que cinq autres mesures (**conviction**, **gain informationnel**, **facteur bayésien**, **Zhang**, **Kulczynski**). On peut donc substituer l'une de ces mesures à la mesure de **confiance** dans l'algorithme, ou encore les agréger puisqu'elles vérifient la même condition d'élagage. Cette condition ne s'applique cependant pas à toutes les mesures que nous avons énoncées. Et en effet, nous allons voir que certaines mesures ne satisfont pas la propriété GUEUC.

9.2.2 Une condition nécessaire

Nous donnons à présent une condition nécessaire pour l'existence de la propriété GUEUC pour une mesure m. Rappelons que si (c,y,z) est un élément de D_{conf} alors $(c, \frac{y}{2}, z)$ aussi (c'est un cas

particulier de la remarque que nous avons faite plus haut).

Proposition 3 – Condition nécessaire pour GUEUC – : Si m est une mesure d'intérêt qui vérifie la condition GUEUC, de fonction de mesure adaptée (Φ_m, D_{conf}), alors pour tout triplet $(c, y, z) \in D_{conf}$, on a

$$\Phi_m(c, y, z) \quad \leq \quad \Phi_m(c, \frac{y}{2}, z). \tag{9.1}$$

Supposons qu'il existe deux valeurs de c et z telles que la fonction $y \mapsto \Phi_m(c, y, z)$ de $\{y|(c, y, z) \in D_{conf}\}$ dans $\mathbb{R}$ soit une fonction croissante. Alors l'inégalité 9.1 n'est pas satisfaite et m ne peut pas avoir la propriété GUEUC. Cela concerne en particulier toutes les mesures dont la variation de la fonction de mesure par rapport à la seconde variable dépend de la situation par rapport à l'indépendance. Nous proposons une démonstration de ce résultat.

Démonstration. Nous utilisons pour cette preuve une version graphique des motifs, sous la forme d'ensembles. Nous considérons donc la figure 9.1 : sur cette figure, on suppose que $\mathtt{P}$, $\mathtt{C}$, et $\mathtt{A}$ sont des attributs catégoriels pouvant prendre la valeur 0 ou 1. Pour $\mathtt{P}$ et $\mathtt{C}$ seule la valeur 1 est indiquée. Supposons que les proportions soient les suivantes :

$$supp(\mathtt{P} = 1) = y, \ supp(\mathtt{C} = 1) = z, \ supp(\mathtt{P} = 1, \mathtt{C} = 1) = cy.$$

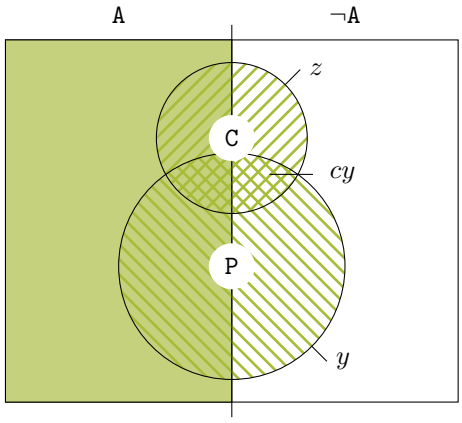


FIGURE 9.1 : Éléments de preuve pour la condition nécessaire

Il est clair, d'après la figure et les proportions, que :

$$supp(\mathtt{P} = 1, \mathtt{A} = 1) = \frac{y}{2}, \ supp(\mathtt{P} = 1, \mathtt{A} = 0) = \frac{y}{2}$$

$$conf(\mathtt{P} = 1, \mathtt{A} = 1 \rightarrow \mathtt{C} = 1) = c, \ conf(\mathtt{P} = 1, \mathtt{A} = 0 \rightarrow \mathtt{C} = 1) = c$$

Alors, puisque m vérifie la propriété GUEUC, au moins l'une des $\mathtt{A}$-spécialisations de $\mathtt{P} = 1 \rightarrow \mathtt{C} = 1$ possède une valeur de mesure supérieure. Mais puisque les deux $\mathtt{A}$-spécialisations sont projetées sur le même point de D_{conf}, à savoir $(c, \frac{y}{2}, z)$, nous avons $m(\mathtt{P} = 1, \mathtt{A} = 1 \rightarrow \mathtt{C} = 1) = m(\mathtt{P} = 1, \mathtt{A} = 0 \rightarrow \mathtt{C} = 1)$. Par conséquent, nous obtenons l'inégalité 9.1 □

Nous pouvons ainsi exclure toutes les mesures dont la fonction de mesure adaptée est croissante avec la seconde variable (la proportion d'antécédents), comme par exemple le **`support`**, **`Jaccard`** ou la **`couverture`**. De plus toutes les mesures dont la variation de la fonction de mesure

adaptée par rapport à la seconde variable dépend des valeurs de la première et de la troisième variable ne vérifieront pas la propriété GUEUC. C'est le cas de la **spécificité**, **risque relatif** ou **Piatetsky-Shapiro**.

Nous résumons dans la prochaine section l'ensemble des résultats.

9.3 Classification des mesures

Dans la section précédente, nous avons démontré deux conditions nous permettant de classer un grand nombre de mesures par rapport à la propriété GUEUC. Cette propriété permet de déduire un algorithme efficace s'appuyant sur une stratégie d'élagage du haut vers le bas, contrairement à Apriori, c'est-à-dire partant des motifs formés par les transactions vers les motifs formés par les attributs. La table 9.1 détaille pour chaque mesure les variations correspondant aux différentes conditions, ainsi que la présence, ou non, de la propriété GUEUC pour la mesure. Pour l'expression des mesures dans le domaine adapté à la **confiance**, le lecteur pourra se référer à la table 3.7 de la page 40.

mesure	c	y	GUEUC	mesure	c	y	GUEUC
confiance	↗	→	✓	confiance centrée	↗	→	✓
moindre contradiction	↗	↗	✗	conviction	↗	→	✓
cosine	↗	↗	✗	couverture	↗	→	✓
Czekanowski	↗	↗	✗	facteur bayésien	↗	→	✓
force collective	↗	⤨	✗	gain	↗	⤨	✗
gain informationnel	↗	→	✓	Ganascia	↗	→	✓
indice de Gini	⤨	↗	✗	indice d'implication	↗	⤨	✗
intérêt	⤨	↗	✗	J1-mesure	⤨	⤨	✗
Jaccard	↗	↗	✗	J-mesure	⤨	↗	✗
Kappa	↗	⤨	✗	Klosgen	⤨	⤨	✗
Kulczynski	↗	↘	✓	Laplace	↗	⤨	✗
levier	↗	⤨	✗	lift	↗	→	✓
Loevinger	↗	→	✓	odds ratio	↗	⤨	✗
one way support	⤨	→	?	coefficient de Pearson	↗	⤨	✗
Piatetsky-Shapiro	↗	⤨	✗	précision	↗	⤨	✗
prevalence	→	→	✓	Q de Yule	?	⤨	✗
rappel	↗	↗	✗	risque relatif	↗	⤨	✗
Sebag-Shoenauer	↗	→	✓	spécificité	↗	↘	✓
spécificité relative	↗	⤨	✗	support	↗	↗	✗
taux exemples contre-exemples	↗	→	✓	valeur ajoutée	↗	⤨	✗
Y de Yule	?	⤨	✗	Zhang	↗	→	✓

Table 9.1 : Existence de la propriété GUEUC pour l'ensemble de nos mesures. Le tableau précise la variation des fonctions de mesure par rapport à la première et la deuxième variable. Le symbole ⤨ indique que la variation n'est pas fixe.

Un point important est que parmi les mesures vérifiant la propriété, un utilisateur peut caractériser un grand nombre de situations, comme par exemple le rapport à l'indépendance avec la mesure de **lift**, ou le **gain informationnel**, ou prendre en compte les contre-exemples avec les mesure de **Sebag-Shoenauer**, **TEC** ou **facteur bayésien**, ou encore la taille du conséquent avec la **prevalence**. Ces mesures peuvent être agrégées dans l'algorithme et éventuellement combinées avec le **support**. Sans contrainte de **support**, on a notamment accès aux pépites, qui sont très présentes comme le montre la figure 9.2.

Nous remarquerons de plus que la mesure de **One Way Support** ne peut pas être classée, car elle ne respecte aucune des hypothèses énoncées par les conditions établies. En particulier, la

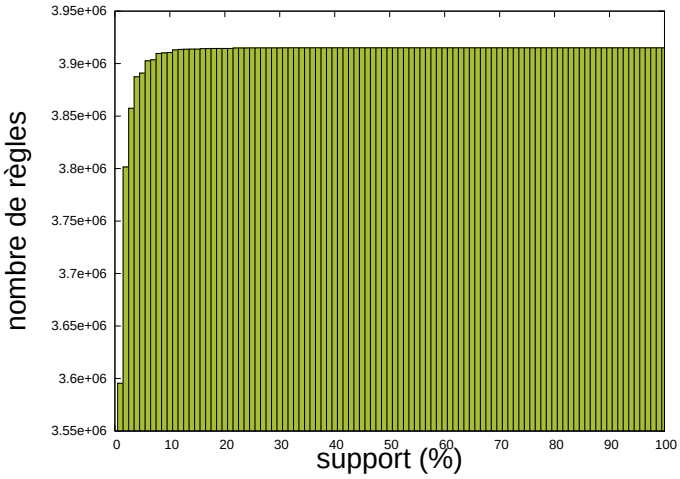


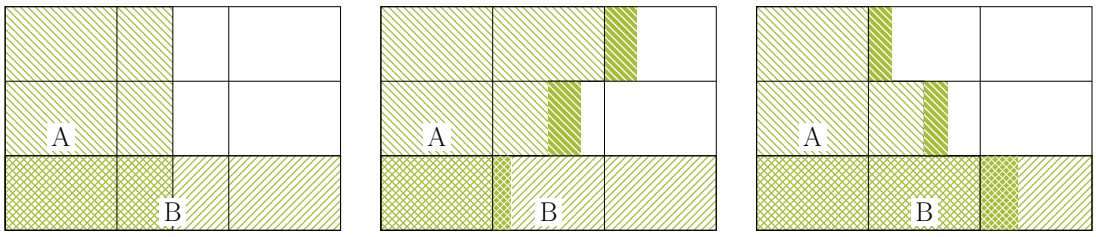
FIGURE 9.2 : Pépites de connaissance. La figure montre la répartition des valeurs de **support** pour toutes les règles de **confiance** supérieure à 0.8 dans la base *mushroom*. Pour la plupart, les règles ont un **support** inférieur à 1% : un seuil trop élevé dans APRIORI les aurait éliminées.

variation de sa fonction de mesure par rapport à la **confiance** dépend de la position par rapport à l'indépendance. C'est aussi le cas de l'**intérêt**, de la **J1-mesure**. Toutes les autres sont croissantes avec la **confiance**. La variation avec la proportion d'antécédents, c'est-à-dire la seconde variable de la fonction de mesure, est plus diversifiée. Souvent, elle varie en fonction de la position par rapport à l'indépendance, soit la configuration de variables $c = z$.

Nous avons essayé de comprendre ce qui distinguait les mesures croissantes en fonction de la proportion d'antécédents et les mesures dont la variation change par rapport à l'indépendance. D'un côté, on peut se dire que si le nombre d'antécédents croît, alors la règle est plus intéressante, car en effet si la proportion du conséquent est fixée, la proportion d'exemple croît elle aussi. Par contre, il en est de même de la proportion de contre-exemples... ce qui devrait rendre la règle moins intéressante.

On pourrait donc tenter de combiner les deux. Donnons-nous deux règles r : $\mathtt{A} \rightarrow \mathtt{B}$ et r' : $\mathtt{A}' \rightarrow \mathtt{B}$ de même valeur de **confiance**, mais telles ques $\mathbb{P}(\mathtt{A}') = (1+\delta)\mathbb{P}(\mathtt{A})$, avec $\delta \geq 0$. La règle r' a donc une proportion d'antécédents plus grande. Est-elle pour autant plus intéressante ? Nous allons nous intéresser à deux quantités : la proportion de vrais-positifs parmi les cas positifs (pour r, elle est représentée par $\mathbb{P}(\mathtt{A}|\mathtt{B})$) et la proportion de faux-négatifs parmi les cas négatifs (représentée par $\mathbb{P}(\mathtt{A}|\neg\mathtt{B})$) ; et plus particulièrement à la variation de ces quantités entre r et r'. La figure 9.3 montre comment évoluent ces proportions. Notons donc $\Delta = \mathbb{P}(\mathtt{A}'|\mathtt{B}) - \mathbb{P}(\mathtt{A}|\mathtt{B})$ et $\bar{\Delta} = \mathbb{P}(\mathtt{A}'|\neg\mathtt{B}) - \mathbb{P}(\mathtt{A}|\neg\mathtt{B})$: r' sera plus intéressante si la variation de faux-négatifs parmi les cas négatifs est plus faible que la variation de vrais-positifs parmi les cas positifs, c'est-à-dire si $\Delta - \bar{\Delta} \geq 0$. Étudions donc ces quantités, en commençant par leur numérateur.

$$\begin{aligned}
\mathbb{P}(\mathtt{A}'\mathtt{B}) - \mathbb{P}(\mathtt{AB}) &= \mathbb{P}(\mathtt{A}') \times conf(\mathtt{A}' \rightarrow \mathtt{B}) - \mathbb{P}(\mathtt{AB}) \\
&= (1+\delta)\mathbb{P}(\mathtt{A}) \times conf(\mathtt{AB}) - \mathbb{P}(\mathtt{AB}) \\
&= (1+\delta)\mathbb{P}(\mathtt{AB}) - \mathbb{P}(\mathtt{AB}) \\
&= \delta \times \mathbb{P}(\mathtt{AB})
\end{aligned}$$



(a) Situation à l'indépendance (b) Corrélation négative : $\bar{\Delta} \geq \Delta$ (c) Corrélation positive : $\bar{\Delta} \leq \Delta$

FIGURE 9.3 : Différentes situations par rapport à l'indépendance. Les parties foncées représentent l'apport d'une règle $\mathtt{A'} \rightarrow \mathtt{B}$ telle que $\mathbb{P}(\mathtt{A'}) > \mathbb{P}(\mathtt{A})$, mais possédant la même valeur de **confiance**, et le même conséquent. Ici, $\mathbb{P}(\mathtt{A}) = \frac{1}{2}$, $\mathbb{P}(\mathtt{B}) = \frac{1}{3}$ et $\mathbb{P}(\mathtt{A'}) = \frac{5}{8}$. On voit que dans le cas de la corrélation positive, la proportion de vrais-positifs est plus importante que la proportion de faux-négatifs.

De la même façon, nous pouvons montrer que $\mathbb{P}(\mathtt{A'}\neg\mathtt{B}) - \mathbb{P}(\mathtt{A}\neg\mathtt{B}) = \delta \times \mathbb{P}(\mathtt{A}\neg\mathtt{B})$. Finalement,

$$\Delta = \frac{\mathbb{P}(\mathtt{A'B}) - \mathbb{P}(\mathtt{AB})}{\mathbb{P}(\mathtt{B})} = \delta\mathbb{P}(\mathtt{A}|\mathtt{B})$$

et

$$\bar{\Delta} = \frac{\mathbb{P}(\mathtt{A'}\neg\mathtt{B}) - \mathbb{P}(\mathtt{A}\neg\mathtt{B})}{\mathbb{P}(\neg\mathtt{B})} = \delta\mathbb{P}(\mathtt{A}|\neg\mathtt{B}).$$

Déterminer le signe de $\Delta - \bar{\Delta}$ revient donc à déterminer le signe de $\mathbb{P}(\mathtt{A}|\mathtt{B}) - \mathbb{P}(\mathtt{A}|\neg\mathtt{B})$:

$$\begin{aligned}
\mathbb{P}(\mathtt{A}|\mathtt{B}) - \mathbb{P}(\mathtt{A}|\neg\mathtt{B}) &= \frac{\mathbb{P}(\neg\mathtt{B})\mathbb{P}(\mathtt{AB}) - \mathbb{P}(\mathtt{B})\mathbb{P}(\mathtt{A}\neg\mathtt{B})}{\mathbb{P}(\mathtt{B})\mathbb{P}(\neg\mathtt{B})} \\
&= \frac{\mathbb{P}(\neg\mathtt{B})\mathbb{P}(\mathtt{AB}) - \mathbb{P}(\mathtt{B})\mathbb{P}(\mathtt{A}) + \mathbb{P}(\mathtt{B})\mathbb{P}(\mathtt{AB})}{\mathbb{P}(\mathtt{B})\mathbb{P}(\neg\mathtt{B})} \\
&= \frac{\mathbb{P}(\mathtt{AB}) - \mathbb{P}(\mathtt{B})\mathbb{P}(\mathtt{A})}{\mathbb{P}(\mathtt{B})\mathbb{P}(\neg\mathtt{B})}
\end{aligned}$$

Le signe de $\Delta - \bar{\Delta}$ dépend donc de la situation par rapport à l'indépendance. Puisque r et r' ont la même valeur de **confiance** et la même proportion de conséquents, les deux règles sont du même côté de l'indépendance. Supposons donc que r se trouve *à gauche*, c'est-à-dire que $\mathbb{P}(\mathtt{AB}) - \mathbb{P}(\mathtt{A})\mathbb{P}(\mathtt{B}) \leq 0$, alors $\Delta - \bar{\Delta} \leq 0$, et donc en construisant la règle r', on a ajouté plus de faux-négatifs parmi les cas négatifs que de vrais-positifs parmi les cas positifs. Si les coûts sont équivalents, on peut donc juger que r' est moins intéressante. Dans le cas opposé, si $\Delta - \bar{\Delta} \geq 0$, r' sera meilleure puisqu'elle aura ajouté plus de vrais-positifs parmi les positifs que de faux-négatifs parmi les négatifs.

Des mesures faisant cette distinction pourraient naturellement être désirées par un utilisateur. On voit cependant qu'elles n'ont pas cette propriété de monotonie... Il faudra donc faire un choix entre qualité de la mesure et performances algorithmiques.

CONCLUSION

La propriété d'*Universal Existential Upward Closure* de la mesure de **confiance** permet, grâce à une stratégie d'élagage par le haut, d'extraire l'ensemble des règles dont la mesure de **confiance** se situe au dessus d'un seuil donné. Nous avons proposé une généralisation de cette propriété, et donné une condition nécessaire ainsi qu'une condition suffisante de l'existence de cette généralisation pour

une mesure quelconque. Cela nous a permis de mettre en évidence que 15 mesures peuvent être utilisées dans un algorithme identique à celui utilisé pour la `confiance`, alors que 26 sont à exclure. Une mesure, celle de `One Way Support`, ne peut être classée. La propriété liée à la recherche de règles otpimales du chapitre suivant ne présentera pas ce défaut puisque nous allons proposer une condition nécessaire *ET* suffisante.

10 Généralisation de la recherche de règles optimales

Les deux précédentes propriétés d'élagage ont été originalement définies pour la mesure de **confiance**, et nous avons vu que nous pouvions les étendre à un certain nombre de mesures : 10 pour l'omni-monotonie, et 15 pour la propriété de monotonie GUEUC. Nous nous intéressons maintenant à la recherche de règles optimales, qui est dirigée par une propriété d'anti-monotonie que nous avons détaillée précédemment. Cette propriété permet de renvoyer un ensemble complet non minimal pour 13 mesures (en réalité 12 [Li 06] + 1 [Li et al. 05]). Nous allons ici obtenir un résultat plus fort que pour les deux propriétés précédentes, puisque nous proposons une condition nécessaire *et* suffisante s'appuyant sur notre troisième domaine adapté : D_{cex}.

10.1 La recherche de règles optimales

Comme nous l'avons déjà fait précédemment, nous considérons ici que P est un motif, X est un attribut ne se trouvant pas dans P et c est une instance de l'attribut de classe. Nous considérons une mesure d'intérêt m, et nous allons définir une propriété d'opti-monotonie sur les mesures.

10.1.1 Une propriété d'opti-monotonie

Nous nous sommes inspirés des travaux de [Li 06], dans lesquels l'auteur donne la preuve au cas par cas de la compatibilité des mesures avec la recherche de règles optimales, afin de définir une propriété d'*opti-monotonie* générale [Le Bras et al. 09a]. Nous chercherons donc par la suite à savoir si une mesure est, ou n'est pas, *opti-monotone*.

Définition 15 – Opti-monotonie – : Une mesure d'intérêt m est opti-monotone si, étant donnée une règle $\mathtt{P} \rightarrow \mathtt{c}$ et l'une de ses spécifications $\mathtt{PX} \rightarrow \mathtt{c}$, on a :

$$supp(\mathtt{P\bar{c}}) = supp(\mathtt{PX\bar{c}}) \implies m(\mathtt{PX} \rightarrow \mathtt{c}) \leq m(\mathtt{P} \rightarrow \mathtt{c}).$$

Remarquons qu'il n'est pour l'instant pas question d'ensemble de règles optimales. Cependant, notre théorème 4 va faire le lien entre l'opti-monotonie et la recherche d'ensembles de règles optimales sous la forme proposée par [Li 06]. Nous allons voir que toutes les mesures opti-monotones sont compatibles avec ce processus, c'est-à-dire que l'ensemble retourné par la procédure ORD renvoie un ensemble complet pour toute mesure opti-monotone. De plus, toutes les mesures citées dans le théorème original 1 sont opti-monotones... à l'exception de la mesure de **Klosgen**

(contrairement au résultat énoncé dans [Li 06]) comme nous allons le voir sur l'exemple suivant.

Exemple 6 – Incompatibilité de `Klosgen` – : La mesure de **`Klosgen`** n'est pas opti-monotone. En effet, considérons la base de données de la table 10.1 (pour des raisons de présentation, les attributs sont représentés en ligne et les transactions en colonne). Dans celle-ci, nous pouvons constater que $supp(\mathtt{PX\bar{c}}) = supp(\mathtt{P\bar{c}})$, mais le calcul de la mesure de **`Klosgen`** donne $Kl(\mathtt{P} \rightarrow \mathtt{c}) = -0.041 < 0$ et $Kl(\mathtt{PX} \rightarrow \mathtt{c}) = 0$ i.e. $Kl(\mathtt{PX} \rightarrow \mathtt{c}) \not\leq Kl(\mathtt{P} \rightarrow \mathtt{c})$. Ceci contredit l'opti-monotonie. Mais cela contredit aussi les résultats des travaux originaux [Li 06] : $\mathtt{PX} \rightarrow \mathtt{c}$ est optimale dans cette base pour la mesure de **`Klosgen`** (aucune de ses règles plus générales n'a de valeur de mesure plus élevée) alors même que l'on avait l'égalité $supp(\mathtt{P\bar{c}}) = supp(\mathtt{PX\bar{c}})$.

P	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
X	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
c	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

TABLE 10.1 : Contre-exemple pour **`Klosgen`**

Nous établissons, en parallèle avec le théorème 1 le théorème suivant :

Théorème 4 : Si $supp(\mathtt{PX\bar{c}}) = supp(\mathtt{P\bar{c}})$ alors ni la règle $\mathtt{PX} \rightarrow \mathtt{c}$ ni aucune de ses règles plus spécifique n'apparaîtra dans un ensemble de règles optimales défini par une mesure opti-monotone.

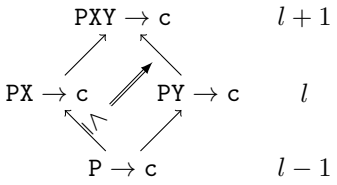


FIGURE 10.1 : Une information au niveau $(l-1)$ renseigne sur les niveaux supérieurs.

Démonstration. La démonstration repose sur le fait que si $supp(\mathtt{PX\bar{c}}) = supp(\mathtt{P\bar{c}})$, alors pour tout motif `Y` n'apparaissant pas dans `PX`, on a aussi $supp(\mathtt{PXY\bar{c}}) = supp(\mathtt{PY\bar{c}})$. Et donc par opti-monotonie, $m(\mathtt{PXY} \rightarrow \mathtt{c}) \leq m(\mathtt{PY} \rightarrow \mathtt{c})$. En fait, l'information d'égalité sur les valeurs de **`support`** entre les niveaux $l-1$ et l va nous donner une information d'égalité sur les valeurs de **`support`** aux niveaux supérieurs, et par conséquent une information d'inégalité sur les valeurs de mesure. Ceci est résumé sur la figure 10.1, où le transfert d'information est illustré par le signe $\Longrightarrow$. Ainsi donc, toute la partie du treillis dans le sillage de `PX` n'a pas besoin d'être parcourue. □

Le théorème 4 est une généralisation du théorème 1 et présente de manière similaire les deux corollaires suivants.

Corollaire 3 – Propriété de clôture – : Si $supp(\mathtt{P}) = supp(\mathtt{PX})$ alors pour toute valeur de `c`, ni la règle $\mathtt{PX} \rightarrow \mathtt{c}$ ni aucune de ses spécifications n'apparaissent dans un ensemble de règles optimales défini par une mesure opti-monotone.

Corollaire 4 – Propriété de terminaison – : Si $supp(\mathtt{P\bar{c}}) = 0$ alors aucune règle plus spécifique que $\mathtt{P} \rightarrow \mathtt{c}$ n'apparait dans un ensemble de règles optimales défini par une mesure opti-monotone.

Nous l'avons vu notamment avec la mesure de **Klosgen**, savoir quelles sont les mesures opti-monotones est une question non triviale. On pourrait faire des études au cas par cas, mais nous proposons une approche plus formelle permettant éventuellement d'automatiser le processus, et en tous cas de ramener encore une fois cette étude à de simples calculs de dérivées. Le théorème suivant fournit une condition nécessaire et suffisante pour une mesure afin qu'elle soit opti-monotone.

Théorème 5 – CNS d'opti-monotonie – : Soit m une mesure d'intérêt des règles d'association, et (Φ_m, D_{cex}) sa fonction de mesure adaptée dans le domaine lié aux contre-exemples. m est opti-monotone *si et seulement si* Φ_m est croissante par rapport à sa seconde variable (antécédent).

	1		$(z-(y-x))\times n$			$(z-(y'-x))\times n$			$z\times n$			$(z+x)\times n$			n
P	0	···	0	1	···	···	···	···	···	···	···	1	0	···	0
X	0	···	···	···	···	0	1	···	···	···	···	1	0	···	0
C	1	···	···	···	···	···	···	···	1	0	···	···	···	···	0

P, PC̄=PXC̄, PX, C

TABLE 10.2 : Base de données pour la condition nécessaire et suffisante

Démonstration. Commençons par prouver l'implication réciproque. Supposons donc que Φ_m soit croissante par rapport à sa seconde variable, donnons-nous P un motif, c une instance de l'attribut de classe, et X un attribut n'apparaissant pas dans P. Puisque nous voulons montrer l'opti-monotonie, considérons de plus que $supp(\mathtt{P}\bar{\mathtt{c}}) = supp(\mathtt{PX}\bar{\mathtt{c}})$, et montrons qu'alors $m(\mathtt{PX} \to \mathtt{c}) \leq m(\mathtt{P} \to \mathtt{c})$. D'après la définition du **support**, nous pouvons nous appuyer sur l'inégalité $supp(\mathtt{P}) \geq supp(\mathtt{PX})$ et écrire les équations suivantes :

$$\begin{aligned} m(\mathtt{PX} \to \mathtt{c}) &= \Phi_m(supp(\mathtt{PX}\bar{\mathtt{c}}), supp(\mathtt{PX}), supp(\mathtt{c})) \\ &\overset{\text{hypothèse}}{=} \Phi_m(supp(\mathtt{P}\bar{\mathtt{c}}), supp(\mathtt{PX}), supp(\mathtt{c})) \\ &\overset{\text{monotonie}}{\leq} \Phi_m(supp(\mathtt{P}\bar{\mathtt{c}}), supp(\mathtt{P}), supp(\mathtt{C})) \\ &\leq m(\mathtt{P} \to \mathtt{c}) \end{aligned}$$

C'est ce que nous recherchions.

Pour le sens direct, nous voulons donc prouver qu'étant donnés deux éléments (x, y, z) et (x, y', z) de D_{cex} tels que $y' \leq y$, l'inégalité $\Phi_m(x, y', z) \leq \Phi_m(x, y, z)$ est vérifiée. Puisque x, y, y' et z sont des nombres rationnels, il existe un entier n tel que $n \times x$, $n \times y$, $n \times y'$ et $n \times z$ soient des entiers, et construisons la base de données de la table 10.2 dans laquelle on peut vérifier que les valeurs de **support** sont telles que

$$supp(\mathtt{P}\bar{\mathtt{c}}) = x = supp(\mathtt{PX}\bar{\mathtt{c}})$$

$$supp(\mathtt{c}) = z,\ supp(\mathtt{P}) = y,\ supp(\mathtt{PX}) = y'$$

L'important étant que l'on a l'égalité $supp(\mathtt{PX}\bar{\mathtt{c}}) = supp(\mathtt{P}\bar{\mathtt{c}})$. D'après l'hypothèse, nous avons donc $m(\mathtt{PX} \to \mathtt{c}) \leq m(\mathtt{P} \to \mathtt{c})$ que l'on peut écrire, par le biais de la fonction de mesure adaptée : $\Phi_m(x, y', z) \leq \Phi_m(x, y, z)$. Puisque $y' \leq y$, cela se traduit par : Φ_m est croissante par rapport à sa seconde variable. □

Ce théorème permet donc une classification complète des mesures d'intérêt objectives, à condition qu'elles puissent s'écrire en fonction de la table de contingence : ici, la taille de la base est considérée comme une constante, cela concerne une grande majorité de mesures objectifs, à l'exception par exemple de la mesure introduite dans [Gay et Boullé 11] qui prend compte la taille de la règle. Cela permet d'obtenir deux familles de mesures : les mesure opti-monotones, pour lesquelles la procédure ORD renverra un ensemble complet, et les autres, pour lesquelles l'ensemble ne sera pas nécessairement complet.

10.1.2 Variation des mesures

Nous avons mené l'étude des dérivées sur les 42 mesures sur lesquelles nous nous concentrons dans cette thèse. Parmi ces mesures, quelques-unes demandent plus d'attention dans le calcul de leurs dérivées. Nous proposons ici de nous attarder sur ces cas pathologiques.

Intéressons-nous tout d'abord à la mesure de **`force collective`**, dont nous rappelons l'écriture

$$fc(\mathtt{A} \to \mathtt{B}) = \frac{p_{\mathtt{AB}} + p_{\neg\mathtt{A}\neg\mathtt{B}}}{p_{\mathtt{A}} \times p_{\mathtt{B}} + p_{\neg\mathtt{A}} \times p_{\neg\mathtt{B}}} \times \frac{p_{\neg\mathtt{A}} \times p_{\mathtt{B}} + p_{\mathtt{A}} \times p_{\neg\mathtt{B}}}{p_{\neg\mathtt{AB}} + p_{\mathtt{A}\neg\mathtt{B}}}$$

$$\Phi_{fc}^{S_{cex}} = \frac{1 + y - z - 2x}{y \times z + (1-y) \times (1-z)} \times \frac{y \times (1-z) + z \times (1-y)}{2x + z - y}$$

Cette fonction est le produit de deux fonctions f_1 et f_2 positives et croissantes par rapport à leur seconde variable :

$$f_1(x,y,z) = \frac{1 + y - z - 2x}{y \times z + (1-y) \times (1-z)}$$

$$f_2(x,y,z) = \frac{y \times (1-z) + z \times (1-y)}{2x + z - y}.$$

f_1 est positive car dans le domaine adapté aux contre-exemples, on a $x < 1 - z$ et $x < y$. De même f_2 est positive puisque sur ce même domaine, $x > y - z$. Ces deux fonctions, en tant que fonctions de trois variables, sont continues sur D_{cex}. Étudions par exemple le comportement de f_1 par rapport à sa deuxième variable :

$$\begin{aligned} \partial_2 f_1(x,y,z) &= 2\,\frac{1 - 2\,z + z^2 + 2\,zx - x}{(2\,zy + 1 - z - y)^2} \\ &= 2\,\frac{(1-z)(1-z-x) + zx}{(2\,zy + 1 - z - y)^2} \end{aligned}$$

Puisque $x < 1 - z$, cette dérivée est positive, f_1 est donc croissante par rapport à sa deuxième variable. On peut le prouver aussi pour f_2, ce qui nous permet de dire que $\Phi_{fc}^{S_{cex}}$ est une fonction positive croissante par rapport à sa deuxième variable comme produit de deux fonctions positives et croissantes. La **`force collective`** est donc opti-monotone.

On trouve dans [Li et al. 05] l'affirmation que la mesure de **`risque relatif`** peut aussi être utilisée dans l'algorithme ORD, mais aucune preuve n'est donnée. Nous allons donc montrer que la fonction de mesure du **`risque relatif`** adaptée au domaine des contre-exemples est une fonction croissante par rapport à sa deuxième variable. La dérivée par rapport à cette variable est donnée par :

$$\frac{y^2 \times (1-z) - 2\,y \times x + x \times (z+x)}{(z - y + x)^2 \times y^2}.$$

Son signe est donné par la quantité $y^2 \times (1-z) - 2\,y \times x + x \times (z+x)$ qui est un polynôme du second degré en y. Son discriminant vaut $4\,x \times z \times (x - (1-z))$ qui est négatif puisque sur le

measure	y	OPTI	measure	y	OPTI
confiance	↗	✓	confiance centrée	↗	✓
moindre contradiction	↗	✓	conviction	↗	✓
cosine	↗	✓	couverture	↗	✓
Czekanowski	↗	✓	facteur bayésien	↗	✓
force collective	↗	✓	gain	⤨	✗
gain informationnel	↗	✓	Ganascia	↗	✓
indice de Gini	⤨	✗	indice d'implication	↗	✓
intérêt	⤨	✗	J1-mesure	⤨	✗
Jaccard	↗	✓	J-mesure	⤨	✗
Kappa	↗	✓	Klosgen	⤨	✗
Kulczynski	↗	✓	Laplace	↗	✓
levier	↗	✓	lift	↗	✓
Loevinger	↗	✓	odds ratio	↗	✓
one way support	⤨	✗	coefficient de Pearson	↗	✓
Piatetsky-Shapiro	↗	✓	précision	↗	✓
prevalence	–	✗	Q de Yule	↗	✓
rappel	↗	✓	risque relatif	↗	✓
Sebag-Shoenauer	↗	✓	spécificité	↗	✓
spécificité relative	↗	✓	support	↗	✓
taux exemples contre-exemples	↗	✓	valeur ajoutée	↗	✓
Y de Yule	↗	✓	Zhang	↗	✓

TABLE 10.3 : Variation et opti-monotonie des mesures d'intérêt.

domaine, $x < 1 - z$. Ce polynôme n'a donc aucune racine réelle et est toujours positif : la dérivée est donc positive et la fonction est croissante, ce dont nous pouvons déduire que la mesure de **risque relatif** est bien opti-monotone.

Même si cette propriété concerne une grande partie des mesures d'intérêt, certaines mesures ne sont pas décroissantes avec la proportion d'antécédents, comme par exemple l'**indice de Gini** dont la fonction de mesure adaptée est donnée par

$$\Phi_g^{S_{cex}}(x, y, z) = \frac{1}{y} \times ((y - x)^2 + x^2) + \frac{1}{1 - y} \times ((z - y + x)^2 + (1 - z - x)^2) - z^2 - (1 - z)^2.$$

Sa dérivée par rapport à y se simplifie sous la forme :

$$2 * \frac{((1 - z - x) \cdot y + x \cdot (1 - y)) \cdot ((1 - z) \cdot y - x)}{y^2 \cdot (1 - y)^2}.$$

La partie gauche du numérateur $((1 - z - x) \cdot y + x \cdot (1 - y))$ est positive d'après les contraintes du domaine D_{cex}, le signe de cette dérivée est donc donné par le signe de $(1 - z) \cdot y - x$. Il dépend donc de la situation par rapport à l'indépendance. On ne peut ainsi pas assurer la décroissance avec y, et l'**indice de Gini** n'est donc pas opti-monotone.

Enfin, nous pouvons donner une preuve formelle de l'exclusion de la mesure de **Klosgen**, bien que le contre-exemple exhibé soit suffisant. La mesure de **Klosgen** s'écrit dans le domaine adapté aux contre-exemples sous la forme $\sqrt{y - x} \times (1 - z - \frac{x}{y})$. C'est le produit de deux fonctions croissantes par rapport à y, comme dans le cas de la **force collective**, mais cette fois-ci, les deux fonctions ne sont pas toutes les deux positives : la partie droite peut prendre des valeurs négatives en fonction de la corrélation des motifs. Ainsi, la mesure de **Klosgen** n'est pas opti-monotone.

Nos résultats permettent donc de savoir quelles mesures peuvent être utilisées dans l'algorithme ORD, c'est-à-dire de connaître les mesures pour lesquelles l'ensemble de règles généré par l'algorithme OCGA est complet dans le sens où il contient toutes les règles optimales. Rappelons qu'il ne reste plus après qu'à effectuer un filtrage exhaustif pour extraire les règles effectivement optimales. Le tableau 10.3 détaille l'application de notre condition nécessaire et suffisante sur l'ensemble des mesures.

10.2 Expériences

Forts de cette généralisation, nous avons mené des expériences afin d'évaluer l'efficacité de l'heuristique d'élagage liée à la recherche de règles optimales. Nous étudions en particulier les paramètres suivants : nombre de candidats générés par rapport à Apriori, efficacité des différents élagages (théorème et corollaires), et nombre de règles finalement générées (importance du filtrage *a posteriori* des règles optimales). En effet ces trois étapes sont les étapes significatives pour l'évaluation de cette stratégie d'élagage. Les candidats générés sont source de calculs de `support` coûteux, l'efficacité des stratégies montrera l'utilité de les implémenter, et finalement, la proportion entre les règles finales et les candidats témoigne de la qualité des stratégies : moins on supprimera de règles *a posteriori*, plus les stratégies d'élagages sont bonnes, car les ensembles obtenus sont alors quasiment minimaux.

10.2.1 Protocole

Pour notre étude nous nous concentrons sur 5 bases de données issues de l'archive UCI. Notre choix a été motivé par leur relative popularité. Parmi elles, la base *mushroom* a été étudiée dans [Li 06], mais nous n'avons pas pris en compte les autres bases de l'article original. En effet, ces bases contenaient des attributs continus qui ont été discrétisés sans préciser la méthode employée, ce qui aurait entrainé des différences de résultats. Nous avons préféré utiliser d'autres bases de données de dimensions comparables telles que *tic-tac-toe*, *soybean*, *car* ou encore *house votes 84*. La table 10.4 résume les caractéristiques de chaque base, la colonne *attributs bin.* indiquant le nombre d'attributs binaires obtenus par binarisation des attributs de la base catégorielle (*attributs cat.*). Dans toute cette section, nous nous concentrons sur des règles de taille 8 au plus. Nous avons donc effectué différentes expériences sur ces bases afin d'évaluer correctement l'effi-

nom	transactions	attributs cat.	attributs bin.	classes
mushroom	8124	22	124	2
tic-tac-toe	858	9	27	2
soybean	307	35	134	19
car	1728	6	21	4
house votes 84	435	16	32	2

Table 10.4 : Description des bases.

cacité de la stratégie d'élagage sous-jacente à la notion d'ensemble de règles optimales. Les deux premières expériences ne prennent pas en compte la mesure car elles se concentrent exclusivement sur l'élagage. La troisième série d'expériences évalue le résultat de cette phase d'élagage en faisant entrer en scène les mesures d'intérêt.

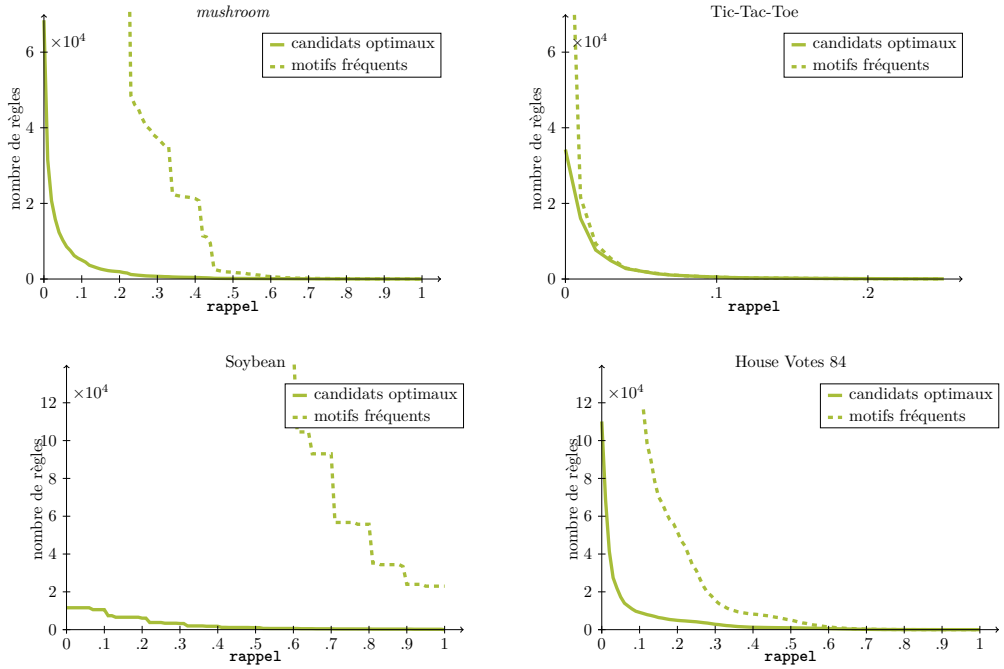


FIGURE 10.2 : Comparaison entre les candidats de l'algorithme APRIORI (pointillés) et de l'algorithme OCGA (plein)

10.2.2 Comparaison avec APRIORI

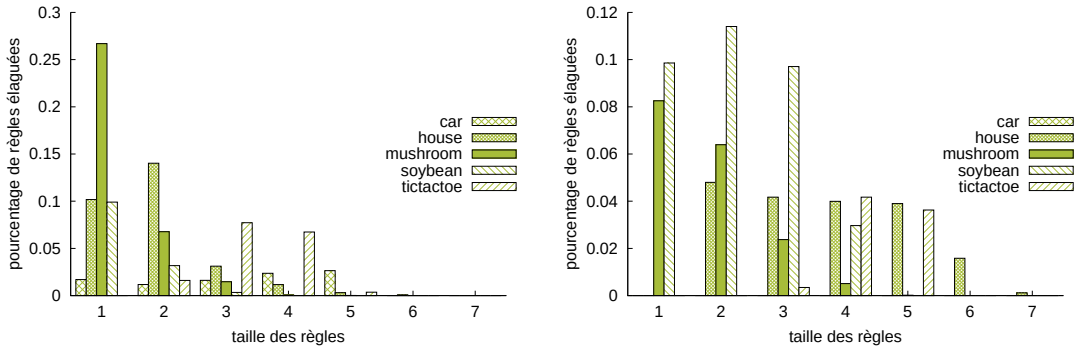
Afin d'examiner les avantages offerts par l'algorithme OCGA par rapport à APRIORI, nous décidons de comparer les règles générées par OCGA et les motifs fréquents générés par une version de l'algorithme APRIORI adaptée aux règles de classe. En effet, les règles générées par OCGA sont des candidats à l'optimalité, sur lesquels seront évaluées les mesures. Pour APRIORI, nous nous sommes appuyés sur notre implémentation de OCGA dans laquelle nous avons supprimé les élagages des lignes 2, 6 et 7 de la fonction d'élagage. Les règles générées sont donc toutes les règles possédant une valeur de **rappel** (support local) supérieure au seuil préfixé : ce sont les candidats à l'intérêt. Nous comparons donc ces deux types de candidats, qui sont comparables car ils vont engendrer un calcul de mesure.

Nous avons donc exécuté l'algorithme OCGA sur chaque base avec 100 seuils de **rappel** différents équirépartis entre 0 et 1. Nous avons effectué les mêmes expériences avec notre version d'APRIORI et enregistré à chaque fois le nombre de candidats générés (figure 10.2). Nous ne présenterons pas les résultats de la base de données *car* dont les résultats sont similaires à la base de données *tic-tac-toe* à une moindre échelle.

Les avantages de l'algorithme OCGA sont clairs pour toutes les bases, mais sont particulièrement marqués pour la base de données *soybean*. Cette base présente un grand nombre d'attributs ayant la même valeur dans toutes les transactions. Ces attributs sont immédiatement élagués par OCGA alors qu'ils entraînent une explosion de complexité dans le cas d'APRIORI. Une remarque générale est la faible proportion de motifs fréquents générés par l'algorithme OCGA : il en résulte un gain important en temps d'exécution ainsi qu'en taille des bases exploitables.

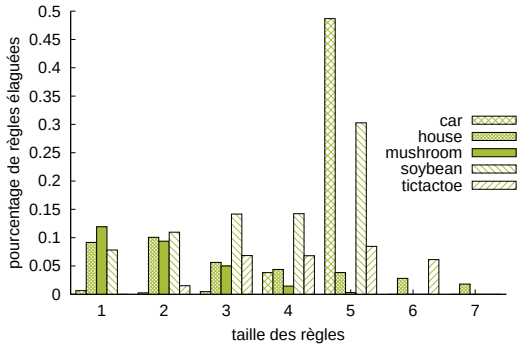
Nous allons maintenant nous concentrer sur l'efficacité de l'élagage.

10.2.3 Efficacité de l'élagage

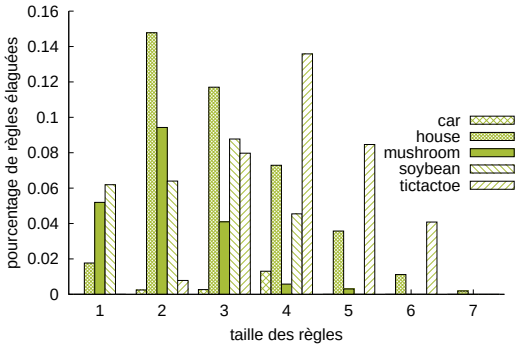


(a) Élagage par le **rappel**

(b) Élagage sur le Corollaire 3



(c) Élagage sur le Corollaire 4



(d) Élagage sur le Théorème 4

10

FIGURE 10.3 : Efficacité de l'élagage : sur l'axe x est représentée la taille des règles, et sur l'axe y, la proportion de règles élaguées par rapport aux règles générées (nombre de règles présentes après la ligne 7 de l'algorithme OCGA)

Comme on a pu le voir dans l'algorithme OCGA, la recherche de règles optimales s'appuie sur 4 stratégies d'élagage différentes : la première est la phase classique d'élagage par le **support**, tandis que les 3 autres sont des stratégies spécifiques à cet algorithme. Pour cette expérience, nous avons fixé le seuil de **support** à 0, pour avoir accès aux pépites, mais l'élagage reste actif pour éliminer les règles de **support** nul. Nous avons ensuite placé des compteurs pour récupérer le nombre de règles élaguées dans chaque étape. Les résultats sont présentés sur la figure 10.3. Nous pouvons nous rendre compte que l'efficacité des différentes heuristiques dépend beaucoup de la base de données :

- l'élagage par le corollaire 4 est presque inexistant pour la base de données *car* jusqu'au rang 5 où il concerne alors près de 50% des règles ;
- le théorème 4 permet d'élaguer un nombre de règles croissant pour la base *tic-tac-toe*, mais décroissant pour la base *house votes 84*.

La figure 10.4 montre la proportion de candidats optimaux pour chaque taille de règle, c'est-à-dire le nombre de règles qui ont survécu à l'élagage divisé par le nombre de règles générées par combinaison (toujours ligne 7 de OCGA). Cet ensemble contient toutes les règles optimales et est donc complet, mais il doit encore être filtré pour devenir minimal. On voit que l'élagage est par exemple très efficace pour la base *mushroom* sur les règles de petite taille (près de 50% de règles élaguées au rang 1), puis de moins en moins efficace pour les tailles supérieures. À l'inverse, la base *car* présente un très faible élagage pour les petites tailles, et l'efficacité croît avec la taille des

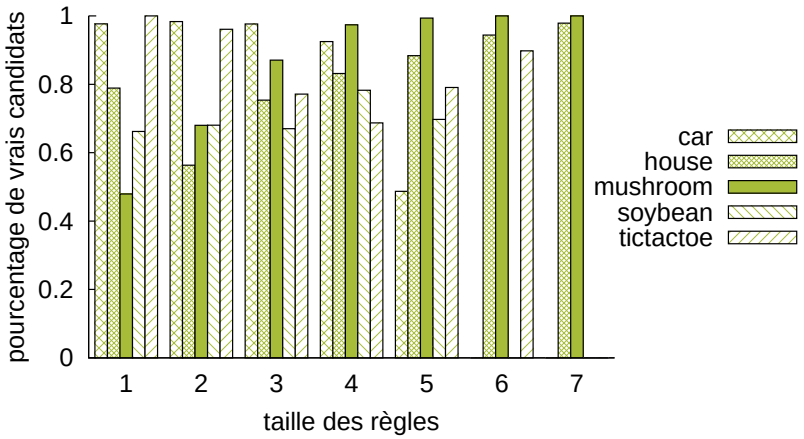


FIGURE 10.4 : Proportion de candidats optimaux pour chaque taille de règle.

règles. Cette base ne contient que 5+1 attributs, c'est pourquoi il n'y a pas de résultats pour les tailles supérieures à 6.

Dans la suite, nous évaluons la qualité de l'élagage en analysant la dernière partie de l'algorithme ORD pour les mesures opti-monotones de la table 10.3. Il s'agit d'évaluer la quantité de travail effectué par la phase de filtrage.

10.2.4 Qualité de l'élagage

Nous nous intéressons à la phase d'évaluation de l'algorithme ORD, semblable à la phase d'évaluation de l'algorithme APRIORI. Cette étape est exécutée exhaustivement pour tous les candidats optimaux : pour chacun, nous vérifions s'il possède une valeur de mesure plus élevée que toutes ses généralisations. Le seuil de **`support`** est fixé à 0, et nous ne nous intéressons pas ici aux règles intéressantes : aucun seuil de mesure n'est fixé. Nous ne traitons pas ici le cas du **`support`**, ni de la mesure de **`gain`** à cause du paramètre θ qu'elle utilise. Les résultats sont donnés dans la table 10.5. La première ligne indique le nombre de candidats optimaux, puis pour chaque mesure nous indiquons un couple de valeurs : la seconde représente le nombre de règles effectivement optimales, et la première représente le nombre de ces règles optimales qui sont rares (**`rappel`** inférieur à 1%) et auraient donc été ratées par une approche type APRIORI. Par exemple, dans la base de données *house votes 84*, il y a 110116 candidats optimaux. Parmi eux, on trouve 10859 règles optimales pour la mesure de **`confiance centrée`**, c'est-à-dire moins de 10%. Si l'on prend en compte toutes les mesures, ce rapport moyen est de 5.51% (dernière ligne). Enfin, dans le cas de la mesure de **`confiance centrée`**, 3124 règles optimales ont une mesure de **`rappel`** inférieure à 1%. Cela montre qu'une approche de type APRIORI classique aurait laissé passer un grand nombre de ces règles. Nous remarquons que la proportion de règles réellement optimales indiquée dans la dernière ligne varie entre 5% pour la base *house votes 84* et 25% pour la base *soybean*. Un élagage parfait donnerait une proportion proche de 100% : c'est le prix de la généralité de l'approche.

Observons la distribution de la valeur de mesure parmi les règles optimales rares pour la mesure **`facteur bayésien`**. Nous nous intéressons en particulier à cette mesure car elle apparaît souvent en tête des classements effectués par [Lenca et al. 08] sur différents systèmes de préférences utilisateur. La figure 10.5 montre cette distribution sur la base de données *mushroom* pour les règles optimales rares (**`rappel`** inférieur à 1%). Nous voyons que la majorité des règles optimales rares dans cette base ont une mesure de **`facteur bayésien`** très élevée. Ces règles, potentiellement intéressantes,

mesure	*car*	*house votes 84*	*mushroom*	*soybean*	*tic-tac-toe*
précision	325/3174	4106/6151	3753/5637	0/5178	4393/4393
confiance centrée	279/1841	3124/10859	1827/7855	0/5383	2968/7382
facteur bayésien	281/1870	3124/10864	1829/7874	0/5396	2968/7382
force collective	294/535	934/1818	2915/4528	0/788	376/903
confiance	279/1841	3124/10859	1827/7855	0/5383	2968/7382
moindre contradiction	291/3083	3597/5539	3794/5698	0/5183	2330/4553
conviction	281/1853	3126/10871	1829/7880	0/5362	2968/7390
cosine	0/75	1/127	27/225	0/494	0/54
couverture	0/67	0/96	8/144	0/129	0/54
TEC	281/1878	3124/10861	1830/7871	0/5387	2968/7390
indice d'implication	17/208	653/1175	1412/3005	0/1878	245/698
gain informationnel	282/1871	3124/10863	1834/7856	0/5395	2968/7390
Jaccard	0/130	0/115	9/184	0/530	0/54
kappa	299/618	994/2282	2895/4506	0/1100	372/863
Laplace	170/2543	420/3983	1605/7436	0/5065	1301/4688
levier	299/464	947/1869	2915/4527	0/631	380/939
lift	282/1876	3124/10863	1834/7856	0/5395	2968/7390
Loevinger	282/1905	3134/10886	1834/7868	0/5450	2976/7398
odds ratio	279/1476	3127/7986	1815/6441	0/2489	2864/6736
coefficient de Pearson	17/180	737/1416	1513/2989	0/1097	229/660
rappel	0/67	0/96	8/144	0/129	0/54
risque relatif	279/1353	514/928	1135/3156	0/974	2734/5980
spécificité relative	0/67	0/96	8/145	0/129	0/54
Sebag-Shoenauer	281/1874	3124/10864	1829/7876	0/5396	2968/7382
spécificité	318/538	2380/5967	3051/4777	0/269	388/951
valeur ajoutée symétrique	578/1553	3862/11881	4509/11356	0/3488	3279/8055
Q de Yule	279/1441	3127/7986	1815/6441	0/943	2864/6736
Y de Yule	279/1441	3127/7986	1815/6441	0/943	2864/6736
Zhang	312/1925	3134/10893	1834/7875	0/5407	2976/7390
optimal candidates	7427	110116	68395	11570	34242
average ratio	0.175	0.0551	0.079	0.254	0.128

TABLE 10.5 : Proportion de règles rares (**rappel** $< 1\%$) parmi les règles optimales, et moyenne des règles optimales pour chaque base.

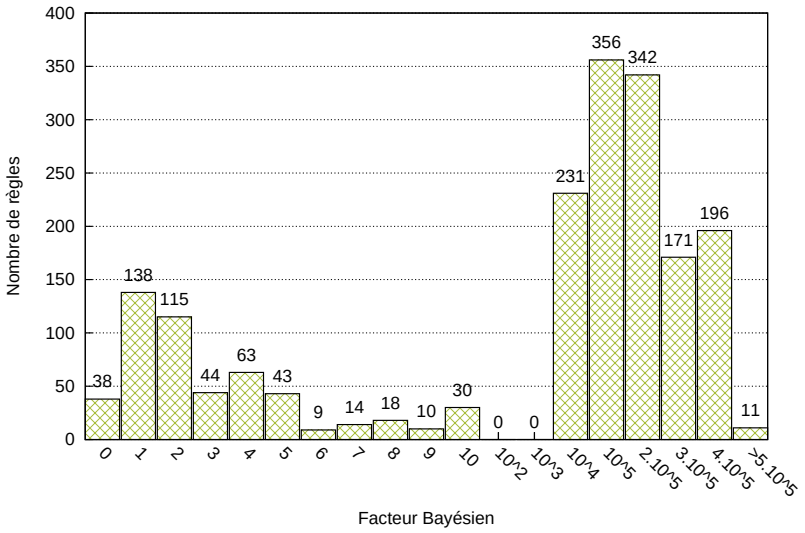


FIGURE 10.5 : Distribution de la mesure de **`facteur bayésien`** dans l'ensemble des règles optimales rares sur la base *mushroom*.

ne seraient pas apparues dans une approche classique : il est donc important de choisir la bonne mesure (d'un point de vue utilisateur), mais aussi la bonne propriété algorithmique.

CONCLUSION

La recherche de règles optimale est dirigée par un algorithme indépendant des mesures mais dont le résultat est cohérent pour certaines mesures : pour chaque mesure compatible avec cet algorithme, l'ensemble retourné est complet au sens de l'optimalité. Les mesures compatibles sont utilisées *a posteriori* pour rendre l'ensemble minimal. Nous avons proposé une condition nécessaire et suffisante de compatibilité des mesures avec cet algorithme. Cette condition nous a permis de montrer que 34 des 42 mesures que nous étudions sont compatibles avec cette démarche. Les expériences que nous avons menées ont montré l'efficacité de l'algorithme sous-jacent. Cependant, l'ensemble de règles obtenu n'est pas complet au sens de l'intérêt puisque les règles intéressantes mais non-optimales sont absentes. Dans la partie suivante, nous énonçons des propriétés d'anti-monotonie de différentes mesures permettant d'extraire un ensemble minimal et complet au sens de l'intérêt.

Conclusion sur la généralisation de propriétés

mesure	OMNI	GUEUX	OPTI	mesure	OMNI	GUEUC	OPTI
confiance	✓	✓	✓	confiance centrée	✗	✓	✓
moindre contradiction	✗	✗	✓	conviction	✗	✓	✓
cosine	✗	✗	✓	couverture	✓	✓	✓
Czekanowski	✗	✗	✓	facteur bayésien	✗	✓	✓
force collective	✗	✗	✓	gain	✓	✗	✗
gain informationnel	✗	✓	✓	Ganascia	✓	✓	✓
indice de Gini	?	✗	✗	indice d'implication	✗	✗	✓
intérêt	?	✗	✗	J1-mesure	✗	✗	✗
Jaccard	✗	✗	✓	J-mesure	?	✗	✗
Kappa	✗	✗	✓	Klosgen	✗	✗	✗
Kulczynski	✗	✓	✓	Laplace	✓	✗	✓
levier	✗	✗	✓	lift	✗	✓	✓
Loevinger	✗	✓	✓	odds ratio	✗	✗	✓
one way support	✗	?	✗	coefficient de Pearson	✗	✗	✓
Piatetsky-Shapiro	✗	✗	✓	précision	✗	✗	✓
prevalence	✓	✓	✗	Q de Yule	✗	✗	✓
rappel	✓	✗	✓	risque relatif	✗	✗	✓
Sebag-Shoenauer	✓	✓	✓	spécificité	✗	✓	✓
spécificité relative	✗	✗	✓	support	✓	✗	✓
taux exemples contre-exemples	✓	✓	✓	valeur ajoutée	✗	✗	✓
Y de Yule	✗	✗	✓	Zhang	✗	✓	✓

TABLE 10.6 : Variation et opti-monotonie des mesures d'intérêt.

Nos travaux sur la généralisation de propriétés algorithmiques nous ont conduit à étudier 3 domaines particuliers : la **all-confidence**, conversion de la **confiance** qui la rend anti-monotone sur les motifs, la propriété UEUC, propriété de monotonie descendante de la **confiance**, et enfin la recherche de règles optimales, qui s'appuie sur une propriété d'anti-monotonie pour la génération des candidats. Pour chacune de ces propriétés, nous avons fourni une généralisation (omni-monotonie, GUEUC et opti-monotonie), et proposé des conditions d'existence de ces généralisations. Dans les deux premiers cas, nous avons pu énoncer une condition nécessaire et une condition suffisante sans malheureusement les faire se rejoindre. Dans le troisième cas, nous avons pu mettre en évidence une condition à la fois nécessaire et suffisante. Nous avons appliqué ces conditions à notre ensemble de 42 mesures, et nous avons retranscrit les résultats dans le tableau 10.6. Nous en déduisons que pour la propriété d'omni-monotonie, 29 mesures sont exclues, 10 sont compatibles, et 3 ne peuvent être classées. Concernant la propriété GUEUC, nous avons 15 mesures la vérifiant, 26 ne la vérifiant pas et 1 qui ne peut être classée. Puisqu'elle propose une condition nécessaire et suffisante, la propriété d'opti-monotonie permet de classer toutes les mesures : 34 la vérifient, et 8 ne la vérifient pas. Nous proposons dans la suite d'étudier les mesures d'intérêt sous un angle nouveau et nous verrons que cette étude nous permettra de définir des propriétés d'anti-monotonie dans le cadre de la recherche de règles de classe.

QUATRIÈME PARTIE : RECHERCHE DE PÉPITES

Nous avons proposé dans les parties précédentes la généralisation de propriétés algorithmiques. Pour chacune d'entre elles, nous avons établi des conditions analytiques d'existence. Il nous reste à faire le chemin inverse : partir des mesures, de leurs caractéristiques analytiques propres, et essayer d'en déduire des propriétés algorithmiques. Dans cette partie, nous nous concentrons en premier lieu sur la mesure de `Jaccard` (chapitre 11) pour laquelle nous définissons une propriété d'anti-monotonie dans le cadre des règles de classe. Puis nous généraliserons cette démarche à l'ensemble des mesures, afin de définir des propriétés d'anti-monotonie pour 10 d'entre elles (chapitre 12). Cela nous permet d'éviter la phase de recherche de motifs lors de la recherche de pépites de connaissance. Ces travaux ont été publiés dans [Le Bras et al. 11].

TELECOM Bretagne

11 Anti-monotonie de la mesure de Jaccard

Comme nous l'avons vu dans les parties précédentes, seuls le **support** et la **all-confidence**, ainsi que les quelques mesures compatibles avec la généralisation, possèdent une propriété d'anti-monotonie. Même en se limitant aux règles de classe, comme c'est le cas pour la propriété UEUC et la recherche de règles optimales, il est difficile d'avoir un ensemble minimal et complet de règles intéressantes pour une mesure donnée. Cela pose notamment des problèmes dans la recherche de pépites de connaissance qui s'appuient sur des motifs non-fréquents. Si [Surana et al. 10] s'intéresse aux mesures qui évaluent correctement ces règles, ce travail s'appuie toujours sur la recherche de motifs rares, quand nous pensons qu'il faudrait pouvoir, sinon s'en affranchir, au moins la compléter par des propriétés algorithmiques sur les mesures. Pour la recherche de pépites de connaissance au sens de la **confiance**, [Szathmary et al. 10] propose d'agréger plusieurs domaines (motifs rares minimaux, clôture...) pour découvrir un ensemble de règles d'association rares. Mais la contrainte de **support** subsiste. Nous proposons ici une approche novatrice s'appuyant sur notre vision projetée des règles d'association afin de déterminer l'existence d'une propriété d'anti-monotonie sur les mesures. L'état de l'art sur ce domaine a été présenté dans le chapitre précédent.

11.1 Contexte et premières remarques

Dans toute cette partie, nous nous restreignons aux règles de classe, les règles dont le conséquent est fixé. Si nous nous plaçons dans le domaine adapté aux exemples D_{ex}, toutes les règles ayant le même conséquent sont projetées dans un même plan, dans une zone délimitée par le domaine adapté. La figure 11.1(a) montre l'apparence de cette zone pour une proportion de conséquents $p_{\mathtt{c}}$ donnée. Si l'on considère deux règles r : $\mathtt{P} = p \rightarrow \mathtt{c}$ et r' : $\mathtt{P}' = p' \rightarrow \mathtt{c}$ telles que r soit une généralisation de r', c'est à dire que $\mathtt{P} \subset \mathtt{P}'$ (on notera r'$\preccurlyeq$r, r' est plus spécifique que r), la propriété d'anti-monotonie du **support** nous rappelle que

$$supp(\mathtt{P}' = p') \leq supp(\mathtt{P} = p)$$

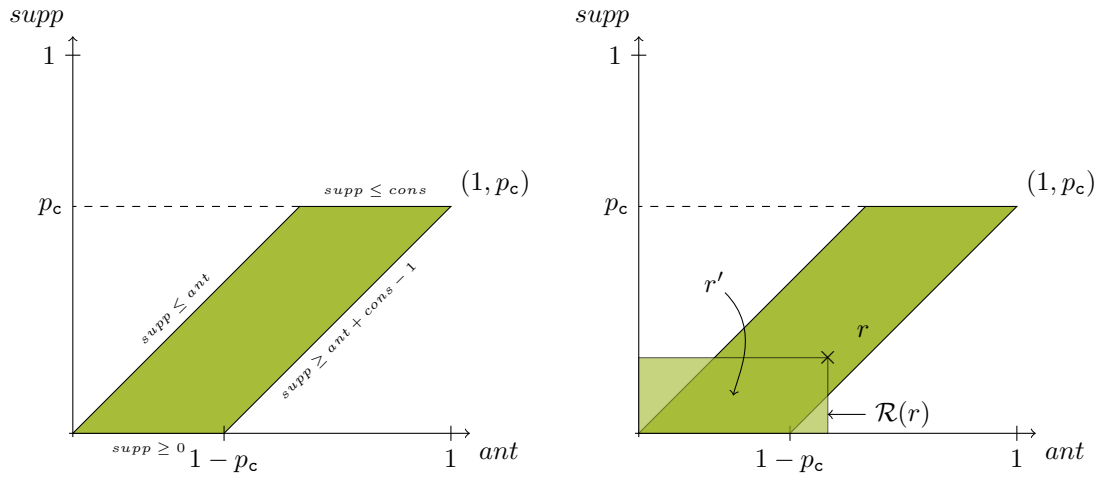
$$supp(\mathtt{P}' = p', \mathtt{c}) \leq supp(\mathtt{P} = p, \mathtt{c})$$

et nous permet donc de localiser la règle r' par rapport à la règle r : r' se trouve dans un rectangle délimité par l'origine et la projection de r noté $\mathcal{R}(r)$ comme cela est illustré dans la figure 11.1(b).

Soit maintenant m une mesure d'intérêt des règles d'association, définissons par $m_{\downarrow}$ un seuil de mesure pré-fixé. Nous dirons que m présente une propriété d'anti-monotonie s'il existe un prédicat $\mathcal{P}$ sur les règles de classe tel que l'implication suivante soit vérifiée :

$$\text{si } \mathcal{P}(r) \text{ alors } \forall r' \preccurlyeq r, m(r') < m_{\downarrow}.$$

Nous appellerons un tel prédicat $\mathcal{P}$ un prédicat d'élagage. Ce prédicat induit naturellement une stratégie d'élagage du bas vers le haut : dans le cas du **support**, on aurait par exemple $\mathcal{P}(r) = (supp(r) < m_{\downarrow})$. Le prédicat dépend de la mesure utilisée, ainsi que du seuil fixé. Jusqu'à présent, et



(a) Domaine adapté aux exemples pour $cons = p_c$. Chaque contrainte est rappelée sur la droite concernée.

(b) Localisation de la règle r' plus spécifique que la règle r.

FIGURE 11.1 : Projections de règles et dépendance d'une règle plus spécifique.

à notre connaissance, aucune des mesures étudiées dans cette thèse ne présentait de telle propriété d'anti-monotonie.

Nous allons dans un premier temps nous concentrer sur la mesure de **Jaccard** et nous fixer la tâche de découvrir, dans une base de données, l'ensemble des règles de classe dont la valeur de la mesure de **Jaccard** se trouve au dessus de $m_\downarrow$. Dans nos travaux sur la robustesse, nous avons mis en évidence que la mesure de **Jaccard** était une mesure plane, c'est-à-dire que la surface définie par $jacc(r) = m_\downarrow$ est un plan. Sa restriction au plan $(z = p_c)$ est donc une droite que nous avons représentée sur la figure 11.2(a). Cette droite partage le domaine adapté en deux régions distinctes : au dessus de la droite, $S^+_{m_\downarrow}$, dans laquelle toutes les règles projetées sont intéressantes, et sous la droite, $S^-_{m_\downarrow}$, dans laquelle les règles projetées ne sont pas intéressantes.

Pour trouver une propriété d'anti-monotonie, il faut trouver un point r du domaine adapté (r pourra être, ou non, la projection d'une règle) tel que le rectangle associé $\mathcal{R}(r)$ se trouve entièrement inclus dans la zone $S^-_{m_\downarrow}$.

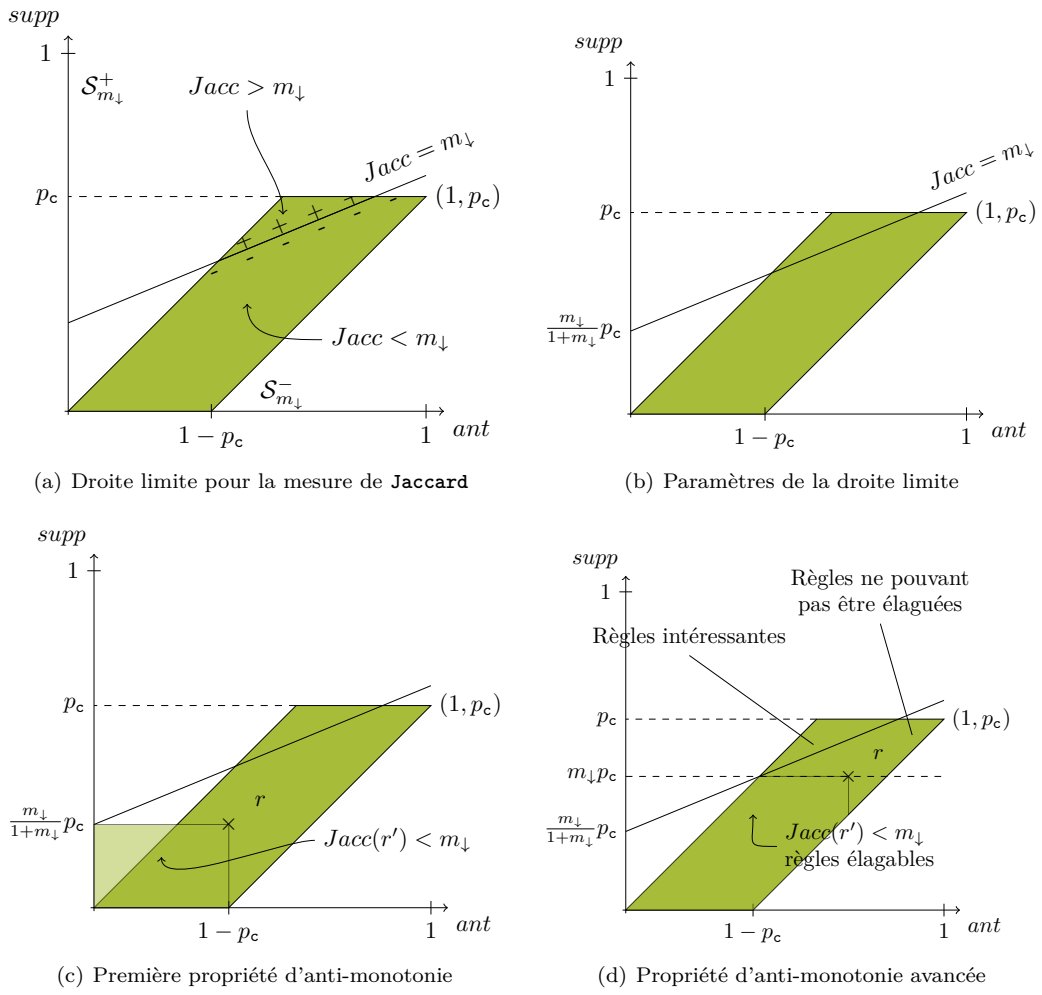
11.2 PROPRIÉTÉ D'ANTI-MONOTONIE DE Jaccard

Notre approche est essentiellement graphique. Considérons la figure 11.2(b) qui montre que la droite définie par l'ensemble des points tels que $jacc = m_\downarrow$ a pour équation

$$supp = \frac{m_\downarrow}{1 + m_\downarrow} ant + \frac{m_\downarrow}{1 + m_\downarrow} p_c$$

et coupe l'axe des abscisses au point de coordonnées $(0, \frac{m_\downarrow}{1+m_\downarrow} p_c)$. Puisque la mesure de **Jaccard** prend ses valeurs entre 0 et 1, la quantité $\frac{m_\downarrow}{1+m_\downarrow} p_c$ est un réel positif. Ainsi, toutes les règles r ayant une valeur de **support** inférieure à cette quantité sont telles que $\mathcal{R}(r) \in (S^-_{m_\downarrow})$ (figure 11.2(c)). Posons donc le prédicat

$$\mathcal{P}_0(r) = (supp(r) < \frac{m_\downarrow}{1 + m\downarrow} p_c).$$



(a) Droite limite pour la mesure de `Jaccard`

(b) Paramètres de la droite limite

(c) Première propriété d'anti-monotonie

(d) Propriété d'anti-monotonie avancée

FIGURE 11.2 : Construction d'une propriété d'anti-monotonie pour la mesure de `Jaccard`.

Naturellement, la mesure de **Jaccard** possède la propriété d'anti-monotonie suivante :

$$\text{si } \mathcal{P}_0(r) \text{ alors } \forall r' \preccurlyeq r, jacc(r') < m_{\downarrow}.$$

Nous pouvons affiner cette propriété en nous rappelant que seuls les points du domaine adapté sont d'intérêt pour nous. Considérons le point d'intersection supérieure gauche entre D_{ex} et la droite limite. Ses coordonnées sont $(m_{\downarrow}p_c, m_{\downarrow}p_c)$. Finalement, aucune règle ayant une valeur de **support** inférieure à $m_{\downarrow}p_c$ ni aucune de ses spécifications ne possèdent une valeur de mesure de **Jaccard** supérieure à $m_{\downarrow}$, comme le montre la figure 11.2(d). Nous pouvons donc définir le prédicat

$$\mathcal{P}_J(r) = (supp(r) < m_{\downarrow}p_c),$$

et affirmer que la mesure de **Jaccard** possède la propriété d'anti-monotonie suivante :

$$\text{si } \mathcal{P}_J(r) \text{ alors } \forall r' \preccurlyeq r, jacc(r') < m_{\downarrow}.$$

Cela peut se vérifier par le calcul : considérons deux règles de classe r'≼r, et supposons que $\mathcal{P}_J(r)$ est vérifiée, c'est-à-dire que $supp(r) < m_{\downarrow}p_c$. Nous pouvons écrire les équations suivantes :

$$\begin{aligned}
jacc(r') &= \frac{supp(r')}{ant(r') + cons(r') - supp(r')} \\
&= \frac{supp(r')}{ant(r') + cons(r) - supp(r')} \\
&\leq \frac{supp(r')}{cons(r)} \\
&\leq \frac{supp(r)}{cons(r)} \\
&< \frac{m_{\downarrow}cons(r)}{cons(r)} \\
&< m_{\downarrow}
\end{aligned}$$

Nous avons donc bien décrit une propriété d'anti-monotonie de la mesure de **Jaccard** dans le cadre de la recherche de règles de classe.

Propriété 5 – Anti-monotonie de Jaccard – : La mesure de **Jaccard** vérifie la propriété d'anti-monotonie suivante :

$$\text{si } (supp(r) < m_{\downarrow}p_c) \text{ alors } \begin{cases} Jacc(r) < m_{\downarrow} \\ \text{et } \forall r' \preccurlyeq r,\ Jacc(r') < m_{\downarrow} \end{cases}.$$

Nous allons maintenant étudier cette propriété de manière expérimentale.

11.3 L'ALGORITHME

Pour évaluer la réelle efficacité de cette propriété d'anti-monotonie de la mesure de Jaccard, nous utilisons un algorithme s'appuyant sur l'algorithme de recherche de règles de classe utilisé dans CBA [Liu et al. 98] et adapté ensuite dans l'algorithme ORD [Li 06] pour la recherche de règles optimales. Cet algorithme se divise en trois parties : la fonction d'élagage, la génération de

candidats et la fonction globale permettant la vérification des candidats. Nous détaillons ci-dessous ces trois parties.

Donnons nous une base de données $\mathcal{DB}$ possédant un attribut de classe $\mathtt{C}$ dont les valeurs sont $\mathcal{C} = \{\mathtt{c}_1, \ldots, \mathtt{c}_k\}$. Un l-candidat est un couple $(\mathtt{P} = p, C)$ où $|\mathtt{P}| = l$ et C est un sous-ensemble de $\mathcal{C}$ indiquant que pour tout $\mathtt{c} \in C$, la règle $\mathtt{P} = p \rightarrow \mathtt{c}$ est potentiellement intéressante pour la mesure de **Jaccard** par rapport au seuil $m_\downarrow$. Nous nous fixons comme but de découvrir toutes les règles r, et seulement celles-ci, telles que $jacc(r) \geq m_\downarrow$. Pour un confort d'écriture, nous noterons simplement $(\mathtt{P}, C)$ au lieu de $(\mathtt{P} = p, C)$.

La procédure de génération des candidats est décrite dans l'algorithme 6. C'est une procédure classique de génération de candidats prenant en compte le fait que l'attribut conséquent est préfixé. On pourra trouver plus de détails dans [Li 06]. Notre contribution se situe surtout dans la fonction

```
   Données : Ensemble de l-candidats CS_l
   Résultat : Ensemble de l+1-candidats CS_{l+1}
 1 pour (P_{l-1}s, C_s) et (P_{l-1}t, C_t) ∈ CS_l faire
 2 |   C ← C_s ∩ C_t;
 3 |   P_{l+1} ← P_{l-1}st;
 4 |   CS_{l+1} ← CS_{l+1} ∪ {(P_{l+1}, C)};
 5 |   pour P_l ⊂ P_{l+1} faire
 6 |   |   si ¬(∃C_l s.t. (P_l, C_l) ∈ CS_l) alors
 7 |   |   |   C ← ∅;
 8 |   |   |   stop
 9 |   |   sinon
10 |   |   |   C ← C ∩ C_l
11 |   |   fin
12 |   fin
13 |   si C == ∅ alors
14 |   |   CS_{l+1} ← CS_{l+1} − {(P_{l+1}, C)}
15 |   fin
16 fin
17 retourner CS_{l+1}
```

ALGORITHME 6: Génération de candidats

d'élagage des candidats à chaque niveau. Elle consiste à introduire le prédicat d'élagage dans l'algorithme. Nous notons ce prédicat $\mathcal{P}$ (dans le cas de la mesure de **Jaccard**, ce prédicat est $\mathcal{P}_J$), et nous détaillons cette procédure dans l'algorithme 7. Il est classique dans le sens où il consiste en la construction d'un ensemble de candidats étoilé constitué des candidats qui auront passé le prédicat d'élagage. Les lignes 2 à 4 parcourent l'ensemble des candidats et la ligne 5 se charge de l'élagage : puisque $\mathcal{P}$ est une propriété d'élagage, les conséquents sont conservés si le prédicat n'est pas vérifié. Si au moins une règle dérivée d'un candidat a passé le prédicat, la ligne 8 la sauvegarde dans l'ensemble étoilé. Finalement, nous décrivons la fonction principale dans l'algorithme 8 appelé *Découverte Anti-monotone de Règles de Classe* (DARC).

Cet algorithme peut être utilisé dans le but d'obtenir une implémentation classique de APRIORI pour les règles de classe : il faudrait simplement prendre $m =$**confiance**, $m_\downarrow$ comme seuil de confiance, et un prédicat d'élagage $supp < \sigma$ où σ serait le seuil de **support**. Une autre mesure pourrait évidemment aussi être utilisée à la place de la **confiance**. Nous allons avoir besoin par la suite d'une telle version d'APRIORI, que nous noterons APRIORI-RC pour nous y référer (RC : Règles de Classe). Afin de rechercher toutes les règles intéressantes au sens de la mesure de **Jaccard**, m sera la mesure de **Jaccard**, $m_\downarrow$ sera le seuil fixé, et nous choisirons comme prédicat d'élagage $\mathcal{P}_J$ défini précédemment.

Données : Ensemble de l-candidats CS_l, prédicat d'élagage $\mathcal{P}$
Résultat : Ensemble élagué de l-candidats CS_l^*
$CS_l^* \leftarrow \emptyset$;
pour $(P_l, C) \in CS_l$ **faire**
  $C^* \leftarrow \emptyset$;
  pour $c \in C$ **faire**
    si $\neg\mathcal{P}(P_l \rightarrow c)$ **alors**
      $C^* \leftarrow C^* \cup \{c\}$
    fin
    si $C^* \neq \emptyset$ **alors**
      $CS_l^* \leftarrow CS_l^* \cup \{(P_l, C^*)\}$
    fin
  fin
fin
retourner CS_l^*

ALGORITHME 7: Algorithme d'élagage par les mesures

Données : Base $\mathcal{DB}$, mesure m, seuil $m_\downarrow$, prédicat d'élagage $\mathcal{P}$
Résultat : Toutes les règles intéressantes R
$R \leftarrow \emptyset$;
Générer l'ensemble des 1-candidats CS_1;
Élaguer l'ensemble des 1-candidats CS_1;
Mettre les règles intéressantes dans R;
tant que $CS_l \neq \emptyset$ **faire**
  $CS_{l+1} \leftarrow Générer(CS_l)$;
  $CS_{l+1} \leftarrow Élaguer(CS_{l+1})$;
  Mettre les règles intéressantes dans R;
fin
retourner R

ALGORITHME 8: DARC : Découverte Anti-monotone de Règles de Classe

11.4 ÉTUDE SUR LA BASE *mushroom*

Nous nous intéressons ici à la base de données *mushroom* de l'archive UCI. Rappelons que cette base de données contient 24 attributs dont un attribut de classe pouvant prendre deux modalités et 8124 transactions. Trouver toutes les règles de **`support`** positif et de **`confiance`** positive est une tâche difficile, et même le recours à un puissant serveur (Xeon 2.33GHz (4 cœurs) - 4Go RAM) ne nous a pas permis d'en venir à bout après plus d'une heure d'exécution malgré l'utilisation de l'implémentation performante de Christian Borgelt. Notre version de l'algorithme DARC, en Python, nous a permis de trouver toutes les règles de mesure de **`Jaccard`** supérieure à 0.6 en moins d'une minute sur un ordinateur de bureau classique (Pentium 4 2.8GHz - 1Go RAM). Nous avons obtenu les 218 règles intéressantes en n'évaluant la mesure que 654 fois, et en ne testant le prédicat que 889 fois. Rappelons que les tests du prédicat correspondent aux calculs de **`support`** : ainsi, seuls 889 calculs de **`support`** ont été nécessaires. Mais revenons à des critères de comparaison plus objectifs.

Nous limitons maintenant la base *mushroom* à ses 10 premiers attributs (attribut de classe compris), et nous comparons le nombre de règles générées avec APRIORI-RC (sur lesquelles il faudra évaluer la mesure) et le nombre de **`support`** calculés avec l'algorithme DARC. En effet,

dans Apriori-RC, avec un seuil de **support** de 0, les valeurs de **support** sont calculées au moment de l'évaluation de la mesure. Cette tâche est la plus coûteuse et permet donc une comparaison objective.

Avec cette approche, Apriori-RC trouve 76016 règles et effectue donc 76016 calculs de la mesure de **support** pour l'évaluation de la mesure. En comparaison, l'algorithme DARC trouve les dix règles intéressantes en calculant 26 fois la mesure et 101 fois une valeur de **support**. Le gain est très significatif.

Dans le même contexte, nous nous intéressons aux deux prédicats d'élagage établis pour la mesure de **Jaccard**, $\mathcal{P}_0$ et $\mathcal{P}_J$, afin d'évaluer le gain dû au raffinement de ce prédicat. Nous les comparons dans la base *mushroom* complète. Rappelons-nous que pour la version affinée, nous avions 654 évaluations de mesure, 889 tests du prédicat et 218 règles intéressantes au final. Avec le premier prédicat d'élagage, nous obtenons aussi 218 règles intéressantes (à titre de contrôle), mais 26260 évaluations de la mesure et 26696 tests du prédicat. Le gain dû au raffinement est donc notable, avec un nombre de calculs de **support** divisé par 30.

Conclusion

Nous avons montré que dans le cadre des règles de classe, il était possible de définir une propriété d'anti-monotonie pour la mesure de **Jaccard**. Cette mesure est couramment utilisée et pouvoir l'introduire au cœur d'un algorithme comme mesure pour l'élagage est un grand avantage. Dans le chapitre suivant, nous allons présenter un résultat théorique permettant d'établir l'existence d'une propriété d'anti-monotonie pour une mesure d'intérêt quelconque. Nous allons ainsi pouvoir définir d'autres prédicats d'élagage, s'appuyant sur d'autres mesures d'intérêt.

12 Extension de l'ensemble des mesures anti-monotones

Dans ce chapitre, nous donnons une condition suffisante d'existence d'une propriété d'anti-monotonie pour une mesure quelconque dans le cadre des règles de classe. Nous étendons ensuite l'ensemble des mesures anti-monotones en précisant les prédicats d'élagage de chacune. Nous clôturerons ce chapitre par des résultats expérimentaux.

12.1 Qui est anti-monotone ?

Soit une mesure d'intérêt m et $m_{\downarrow}$ un seuil d'intérêt fixé. Nous restons concentrés sur le domaine D_{ex} des exemples et notons Φ_m la fonction de mesure adaptée à m. Puisque nous sommes dans un contexte de règles de classe, soit c une instance de l'attribut de classe, nous notons $\phi^{\mathsf{c}}_{m_{\downarrow}}(x,y) = \Phi_m(x,y,p_{\mathsf{c}}) - m_{\downarrow}$.

Une règle intéressante r de projection $(supp(r), ant(r), p_{\mathsf{c}})$ sera donc telle que $\phi^{\mathsf{c}}_{m_{\downarrow}}(supp(r), ant(r)) > 0$. En reprenant les notations précédentes, rappelons que pour établir une propriété d'anti-monotonie, il suffit de montrer l'existence d'un point r dans le domaine adapté tel que $\mathcal{R}(r) \subset S^{-}_{m_{\downarrow}}$. Nous énonçons par conséquent la propriété suivante :

> **Propriété 6 :** Si $\phi^{\mathsf{c}}_{m_{\downarrow}}(0,0) < 0$ ou si $\phi^{\mathsf{c}}_{m_{\downarrow}}$ admet un prolongement par continuité en $(0,0)$ strictement négatif, alors m admet une propriété d'anti-monotonie.

Cette propriété n'est pas une propriété constructive, mais une propriété d'existence. Elle indique s'il est intéressant de se pencher sur une mesure particulière. Elle assure l'existence d'un algorithme d'élagage sous-jacent, mais pas nécessairement son efficacité : l'heuristique pourrait très bien n'effectuer qu'un élagage superficiel. Nous donnons ici une preuve de ce résultat.

Démonstration. Si $\phi^{\mathsf{c}}_{m_{\downarrow}}(0,0) < 0$ ou s'il existe un prolongement continu strictement négatif en $(0,0)$, alors il existe un voisinage $\mathcal{V}$ de $(0,0)$ sur lequel la fonction $\phi^{\mathsf{c}}_{m_{\downarrow}}$ est strictement négative. Dans un tel voisinage, nous pouvons inclure une boule $\mathcal{B}$ de centre $(0,0)$ et de rayon x. Soit r un point de la sphère associée, inclu dans le domaine adapté, alors

$$\mathcal{R}(r) \subset \mathcal{B} \subset \mathcal{V} \subset S^{-}_{m_{\downarrow}}.$$

Par conséquent, si, durant la recherche de règles intéressantes pour la mesure m et le seuil $m_{\downarrow}$, on tombe sur une règle dont la projection se trouve dans $\mathcal{R}(r)$, alors cette règle n'est pas intéressante, ni aucune de ses règles plus spécifiques. □

Cette propriété est très prometteuse car elle nous oriente dans l'obscurité du monde des mesures : nous allons facilement trouver des mesures présentant une propriété d'anti-monotonie en nous concentrant simplement sur la valeur en $(0,0)$. Une question légitime et immédiate serait de

se demander ce qu'il en est pour la mesure de **confiance**. Celle-ci n'admet aucun prolongement continu en $(0,0)$, et ne peut donc pas passer le filtre. Cela ne signifie pas pour autant que la mesure de **confiance** ne possède pas de bonnes propriétés algorithmiques, mais notre méthode ne permet pas de le déduire. Par contre, la mesure de Jaccard donnera une fonction $\phi^{\mathsf{c}}_{m_\downarrow}(x,y) = \frac{x}{y+p_{\mathsf{c}}-x} - m_\downarrow$ qui prend la valeur $-m_\downarrow$ en $(0,0)$. Donc si l'on fixe un seuil positif, ce qui devrait être le cas puisque la mesure de **Jaccard** prend ses valeurs entre 0 et 1, elle présente une propriété d'anti-monotonie.

12.2 Mesures anti-monotones

Grâce à cette propriété, nous pouvons mettre en évidence 10 mesures anti-monotones associées à leur prédicat. La table 12.1 en donne les détails.

mesure	$\phi^{\mathsf{c}}_{m_\downarrow}$ en $(0,0)$	condition d'anti-monotonie	prédicat
support	$-m_\downarrow$	$m_\downarrow > 0$	$supp(r) < m_\downarrow$
Jaccard	$-m_\downarrow$	$m_\downarrow > 0$	$supp(r) < m_\downarrow p_{\mathsf{c}}$
précision	$1 - p_{\mathsf{c}} - m_\downarrow$	$m_\downarrow > 1 - p_{\mathsf{c}}$	$supp(r) < p_{\mathsf{c}} + m_\downarrow - 1$
contramin	$-m_\downarrow$	$m_\downarrow > 0$	$supp(r) < m_\downarrow p_{\mathsf{c}}$
kappa	$-m_\downarrow$	$m_\downarrow > 0$	$supp(r) < \frac{m_\downarrow p_{\mathsf{c}}}{2-m_\downarrow-2p_{\mathsf{c}}(1-m_\downarrow)}$
levier	$-m_\downarrow$	$m_\downarrow > 0$	$supp(r) < \frac{m_\downarrow}{1-p_{\mathsf{c}}}$
spécificité	$1 - p_{\mathsf{c}} - m_\downarrow$	$m_\downarrow > 1 - p_{\mathsf{c}}$	$supp(r) < 1 - \frac{1-p_{\mathsf{c}}}{m_\downarrow}$
spécificité relative	$-m_\downarrow$	$m_\downarrow > 0$	$supp(r) < m_\downarrow$
Kulczynski	$-m_\downarrow$	$m_\downarrow > 0$	$supp(r) < \frac{m_\downarrow}{1+m_\downarrow} p_{\mathsf{c}}$
Czekanowski	$-m_\downarrow$	$m_\downarrow > 0$	$supp(r) < \frac{m_\downarrow}{2-m_\downarrow} p_{\mathsf{c}}$

Table 12.1 : Un ensemble de mesures anti-monotones

Une première remarque que l'on peut faire est que tous les prédicats sont croissants en fonction du seuil de mesure, ainsi plus on cherchera des règles intéressantes, plus ils seront efficaces. Ensuite, nous voyons que la condition d'existence d'un propriété d'anti-monotonie n'est pas toujours triviale. Par exemple, pour les mesures de **précision** et **spécificité**, il faut choisir le seuil en fonction de la base car l'existence de l'anti-monotonie dépend de la valeur de p_{c}. Cela n'est pas une limitation car dans les premières phases du processus de fouille de données, nous prenons normalement connaissance des fréquences des classes. Pour la mesure **kappa**, il faut choisir un seuil positif alors que cette mesure pourrait bien avoir un seuil négatif. Pour chaque cas, le prédicat d'élagage est établi suivant la méthode utilisée pour la mesure de **Jaccard**.

On peut aussi remarquer que chacune de ces propriétés est de la forme $m(r) > m_\downarrow \implies supp(r) > f(m_\downarrow, p_{\mathsf{c}})$, et l'on peut donc visionner le problème sous un angle différent. En effet, on pourrait prendre ce problème comme la résolution de la question suivante : afin d'extraire les règles de mesure supérieure à un seuil $m_\downarrow$, à combien puis-je fixer le seuil de **support** dans l'algorithme Apriori pour être sûr de toutes les obtenir ? Par exemple, pour la mesure de **Jaccard**, fixer le seuil de **support** dans Apriori à $m_\downarrow p_{\mathsf{c}}$ assure que les motifs retournés engendreront l'ensemble des règles intéressantes.

Il est clair que pour chacune de ces mesures, l'algorithme DARC peut être utilisé en changeant uniquement la condition d'élagage et la fonction d'évaluation. De plus on peut imaginer utiliser plusieurs prédicats en même temps et agréger les mesures, et n'importe quelle combinaison linéaire positive de mesures anti-monotones peut être exploitée en utilisant la même combinaison linéaire de seuils et une conjonction de prédicats : si $(m_1, \ldots, m_k)$ sont des mesures anti-monotones pour les seuils $(m^1_\downarrow, \ldots, m^k_\downarrow)$ et les prédicats $(\mathcal{P}_1, \ldots, \mathcal{P}_k)$, et si $(\lambda_1, \ldots, \lambda_k)$ sont des nombres réels positifs

tels que $\lambda_1 + \cdots + \lambda_k = 1$, alors la mesure définie par

$$m = \sum_{i=1}^{i=k} \lambda_i m_i$$

est anti-monotone pour le seuil

$$m_\downarrow = \sum_{i=1}^{i=k} \lambda_i m_\downarrow^i$$

et le prédicat d'élagage

$$\mathcal{P} = \bigcap_{i=1}^{i=k} \mathcal{P}_i.$$

Comme dans le cas de la mesure de **Jaccard**, il peut exister plusieurs prédicats et celui obtenu par agrégation n'est pas nécessairement le prédicat le plus fin possible, mais c'est un prédicat fonctionnel.

Par exemple, considérons l'agrégation de la mesure de **Jaccard** avec la mesure de **levier** avec des coefficients respectifs de 0.7 et 0.3 et les seuils respectifs $m_\downarrow^J$ et $m_\downarrow^l$. La mesure m est alors définie par

$$m(r) = 0.7\frac{supp(r)}{ant(r) + cons(r) - p_c} + 0.3(supp(r) - ant(r) \times p_c).$$

Nous affirmons que m est une mesure présentant une propriété d'antimonotonie pour le seuil

$$m_\downarrow = 0.7m_\downarrow^J + 0.3m_\downarrow^l$$

et le prédicat

$$\mathcal{P}(r) = (supp(r) < m_\downarrow^J p_c) \cap (supp(r) < \frac{m_\downarrow}{1 - p_c}).$$

En effet, si r' est plus spécifique que r et que le prédicat est vérifié sur r alors :

$$\begin{aligned} m(r') &= 0.7\frac{supp(r')}{ant(r') + p_c - supp(r')} + 0.3(supp(r') - ant(r') \times p_c) \\ &\leq 0.7\frac{supp(r)}{p_c} + 0.3supp(r)(1 - p_c) \\ &\leq O.7m_\downarrow^J + 0.3m_\downarrow^l \\ &\leq m_\downarrow \end{aligned}$$

Ainsi la mesure m présente bien une propriété d'anti-monotonie.

12.3 RÉSULTATS D'EXPÉRIENCES

Nous présentons dans cette section quelques résultats d'expériences. Nous considérons 10 bases de données classiques de l'UCI : *mushroom*, *audiology*, *house votes 84*, *sick* (discrétisée par Weka [1]), *soybean* (large), *car*, *connect*, *kr-vs-kp*, *nursery* et *tic-tac-toe*. Pour des raisons de complexité de calculs, notamment dans la recherche de l'ensemble des règles de **support** positif, nous nous limitons ici aux 11 premiers attributs des bases de données (classe incluse). Pour chaque base, nous utilisons tout d'abord DARC en configuration APRIORI-RC afin d'extraire toutes les règles de **support** positif, puis nous effectuons la recherche avec les mesures anti-monotones et un seuil fixé

1. http://www.cs.waikato.ac.nz/ml/weka/

	nursery	*tic-tac-toe*	*car*	*connect*	*kr-vs-kp*	*house votes 84*	*mushroom*	*sick*	*soybean*	*audiology*
support	192789	147061	12506	254282	44705	80234	159272	31164	211181	43687
kappa	1.01	0.45	5.68	1.48	5.79	2.83	0.59	6.44	14.49	4.66
levier	0.06	0.04	0.55	0.04	0.15	0.09	0.06	0.14	0.21	0.21
Jaccard	0.67	0.40	4.48	1.04	5.72	3.03	0.67	10.74	6.24	4.60
spécificité	0.25	0.09	1.43	0.23	1.17	0.71	0.15	4.83	0.82	1.24
précision	0.25	0.13	1.43	0.26	3.32	1.65	0.28	4.98	0.82	2.37
Czekanowski	1.23	0.96	7.28	2.29	11.34	5.81	1.61	12.78	15.25	10.08
Kulczynski	0.81	0.56	5.30	1.47	7.40	3.70	0.83	11.11	13.31	5.99
contramin	0.73	0.40	4.48	1.04	5.72	3.03	0.67	10.74	10.80	4.60
spé.relative	0.06	0.09	0.74	0.19	1.59	0.87	0.16	5.77	0.21	1.21

TABLE 12.2 : Pourcentage de valeurs de **support** calculées pour chaque mesure par rapport à un APRIORI-RC classique, seuils à 0. La première ligne donne le nombre de valeurs de **support** calculées par APRIORI-RC. Les autres représentent le pourcentage de valeurs calculées par DARC.

	nursery	*tic-tac-toe*	*car*	*connect*	*kr-vs-kp*	*house votes 84*	*mushroom*	*sick*	*soybean*	*audiology*
kappa	111	6	51	0	183	1928	549	0	20989	360
levier	1	0	0	0	0	3	0	0	0	0
Jaccard	67	46	24	936	1749	2195	770	1789	8363	1522
spécificité	383	1	24	0	20	399	92	353	1326	127
précision	4	0	0	0	33	992	190	0	974	32
Czekanowski	362	299	173	3036	4289	4256	2140	2165	23617	3759
Kulczynski	121	109	39	1436	2467	2673	983	1873	16800	2029
contramin	25	9	5	223	242	2049	647	1786	6817	229
spé.relative	1	1	2	0	0	520	106	193	0	0

TABLE 12.3 : Nombre de règles intéressantes pour chaque base de donnée et chaque mesure.

arbitrairement à 0.2 pour les mesures dont la condition d'anti-monotonie ne dépend pas de la probabilité du conséquent. Pour les autres, c'est-à-dire la **spécificité** et la **précision**, nous fixons le seuil à $1 - \frac{3}{4}\min_{\mathtt{c}}(p_{\mathtt{c}})$. On s'assure ainsi que la condition d'anti-monotonie est respectée.

La table 12.2 montre le pourcentage de valeurs de **support** calculées dans le cadre de l'algorithme DARC par rapport au nombre de valeurs de **support** calculées dans une approche type APRIORI. Cela montre que dans la plupart des cas, cette nouvelle propriété d'anti-monotonie calcule moins de 1% des valeurs calculées par l'approche classique.

De plus, un seuil fixé à 0.2 ne nous place pas vraiment dans le cadre d'une recherche de pépites (faible **support** et mesure élevée). Dans ce cadre là, la propriété serait encore plus efficace, comme l'a montré l'exemple de la mesure de **Jaccard** avec un seuil fixé à 0.6 dans le chapitre 11. Nous remarquons que la valeur la plus élevée pour ce pourcentage est atteinte pour la mesure de **Czekanowski** et la base de données *soybean* et vaut 15.25%. De plus, cette mesure génère toujours le plus grand nombre de calculs de valeurs de **support**, mais la proportion reste toujours sous la barre des 16% ce qui permet d'affirmer que l'approche reste efficace.

La table 12.3 montre le nombre de règles découvertes pour chaque mesure et chaque base de données. Nous remarquons que la mesure de **Czekanowski** est aussi la mesure qui génère le plus grand nombre de règles intéressantes. Le cas de la mesure de **levier** qui ne génère presque aucune règle intéressante peut facilement être expliqué. En effet, les valeurs de la mesure de **levier** peuvent varier entre -0.25 et 0.25, et un seuil de 0.2 est donc déjà très élevé pour cette mesure. On remarque que le nombre de calculs de valeurs de **support** pour la mesure de **levier** est toujours le plus faible, et est d'ailleurs très bas : aucune règle n'est intéressante, et très peu de travail inutile est fait. Cette remarque confirme l'efficacité de cette propriété d'anti-monotonie.

La table 12.4 représente la proportion de règles intéressantes par rapport au nombre de fois où la mesure a été évaluée. Dans le cas général, la stratégie d'élagage est une implication *si* $\mathcal{P}(r)$ *alors* r *n'est pas intéressante*. Si cette proportion est proche de 1, alors toutes les règles non intéressantes ont été élaguées et la stratégie est donc proche d'une équivalence $\mathcal{P}(r)$ *ssi* r *n'est pas intéressante*.

	nursery	*tic-tac-toe*	*car*	*connect*	*kr-vs-kp*	*house votes 84*	*mushroom*	*sick*	*soybean*	*audiology*	moyenne
kappa	0.12	0.05	0.15	×	0.08	0.91	0.75	×	0.75	0.19	0.37
levier	1.00	×	×	×	×	0.21	×	×	×	×	0.61
Jaccard	0.11	0.45	0.18	0.38	0.74	0.98	0.91	0.54	0.73	0.80	0.58
spécificité	0.99	0.08	0.44	×	0.05	0.84	0.70	0.24	1.00	0.26	0.51
précision	0.01	×	×	×	0.03	0.78	0.62	×	0.73	0.03	0.37
Czekanowski	0.32	0.77	0.41	0.55	0.89	0.98	0.98	0.55	0.81	0.89	0.71
Kulczynski	0.18	0.70	0.19	0.41	0.79	0.98	0.92	0.55	0.66	0.82	0.62
contramin	0.04	0.09	0.04	0.09	0.10	0.92	0.76	0.54	0.33	0.12	0.30
spé.relative	1.00	0.07	0.25	×	×	0.89	0.70	0.11	×	×	0.50

TABLE 12.4 : Proportion de calculs de mesure utiles. Plus ce rapport est proche de un, plus la propriété d'anti-monotonie est proche de l'équivalence.

La valeur 1 apparaît dans trois cas. Dans la base *nursery*, le **levier** et la **spécificité relative** atteignent cette valeurs, mais cela ne semble pas significatif car seulement 1 règle est intéressante pour ces mesures dans cette base. Le cas de la mesure de **spécificité** dans la base *soybean* est plus intéressant car il concerne 1326 règles. Cependant, l'ajustement du seuil pour cette mesure la rend très contraignante avec un seuil très élevé dans cette base *soybean* présentant un attribut de classe très rare, ce qui peut expliquer cette situation. La mesure de **Czekanowski** est celle qui propose la meilleure moyenne, et donc la stratégie d'élagage la plus performante. À l'opposé, la mesure de **moindre contradiction** (**contramin** dans les tableaux) semble générer beaucoup de calculs de mesure inutiles par rapport au nombre de règles intéressantes découvertes. Cependant, pour cette mesure, le nombre de règles se trouve souvent parmi les plus bas, ce qui peut compenser l'inefficacité de la stratégie.

12.4 UTILITÉ DE DARC

Finalement, nous montrons pour conclure l'intérêt de DARC en comparaison de l'approche classique APRIORI dont l'élagage élimine des règles potentiellement intéressantes. Si l'utilisateur est intéressé par des pépites de connaissance, des règles de faible **support** mais de valeur de mesure élevée, un algorithme de type APRIORI effectuera beaucoup de travail inutile pour rechercher les règles et les évaluer. Dans cette section, nous allons souligner ceci en montrant que dans l'ensemble de règles découvert par l'algorithme DARC, un large sous-ensemble possède un **support** très bas et aurait donc été manqué par APRIORI. Par **support** bas, nous entendons ici inférieur à 5%.

Pour cette expérience, nous allons nous intéresser à la base de données *nursery*, et fixer le seuil de mesure arbitrairement à 0.1 pour les mesures **kappa**, **Czekanowski**, **Kulczynski** et **moindre contradiction**, et à 0.01 pour les mesures de **précision**, **spécificité**, **levier** et **spécificité relative** (afin d'obtenir un nombre significatif de règles). La correction pour les mesures dont la condition d'anti-monotonie dépend du conséquent est toujours active. La figure 12.1 montre la proportion de règles rares dans l'ensemble des règles découvertes par l'algorithme DARC, en comparaison des règles qui auraient été découvertes par une approche APRIORI classique.

On se rend compte que la part de règles rares est presque toujours au-dessus de 50% des règles intéressantes sauf pour la mesure de **moindre contradiction**. Cela est particulier à la base *nursery*, et très marqué pour certaines mesures comme **Czekanowski** ou **kappa**, pour lesquelles cette part dépasse les 80%. Le même travail effectué sur la base *house votes 84* (figure 12.2) montre un résultat différent puisque la proportion de règles fréquentes prédomine pour toutes les mesures, excepté **Czekanowski**, la **spécificité relative** et la mesure de **levier**. Ces mesures semblent produire plus de pépites de connaissance. Bien que ces proportions dépendent largement de la mesure utilisée et de la base étudiée, ces graphiques nous donnent une information importante : la contrainte de **support** utilisée dans APRIORI élimine systématiquement des règles intéressantes. La possibilité de découvrir des règles à l'aide d'une nouvelle stratégie d'élagage est un avantage incontestable.

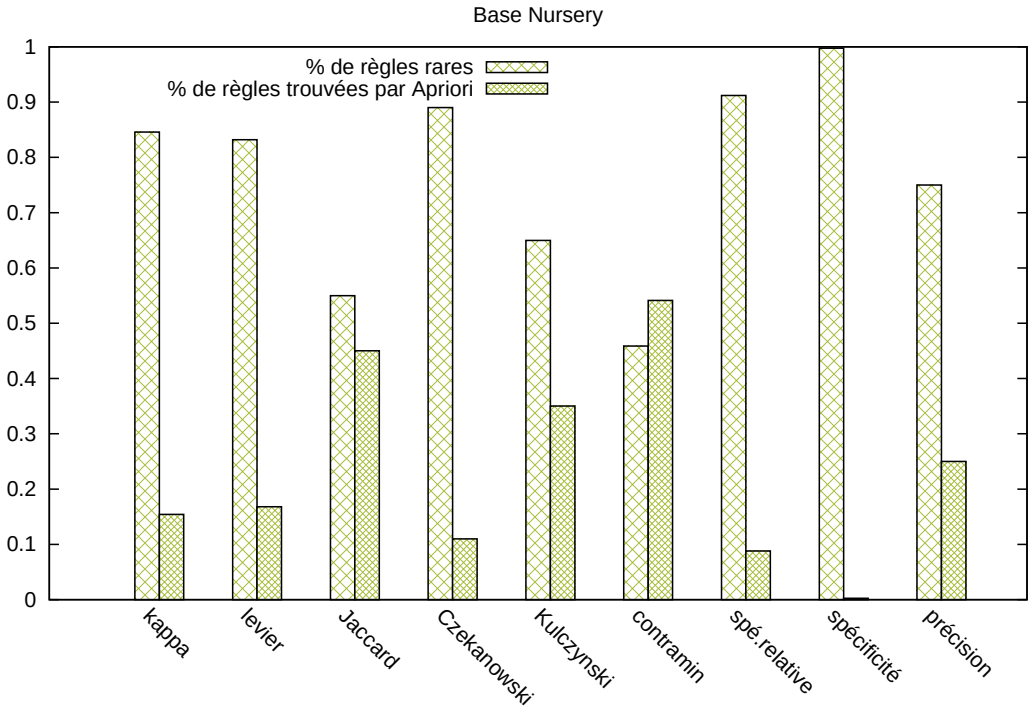


FIGURE 12.1 : Comparaison entre les règles rares de DARC et les règles obtenues par APRIORI-RC dans la base *nursery*.

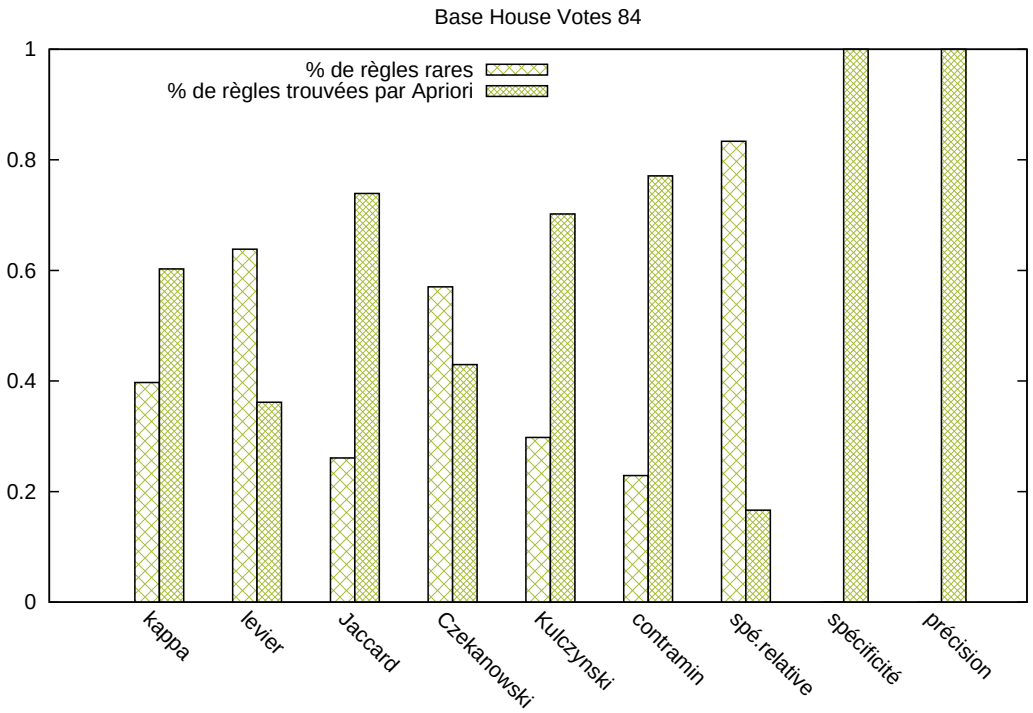


FIGURE 12.2 : Comparaison entre les règles rares de DARC et les règles obtenues par APRIORI-RC dans la base *house votes 84*.

Conclusion

Nous venons d'énoncer un ensemble de propriétés d'élagage pour différentes mesure d'intérêt. Ces propriétés se présentent sous la forme de prédicats et sont intégrables dans un algorithme générique d'élagage du bas vers le haut (le prédicat est un paramètre de l'algorithme), dans la même veine que l'algorithme Apriori. Ces prédicats d'élagage nous permettent principalement d'extraire des bases de données les règles intéressantes sans aucune contrainte de `support`. L'ensemble de règles obtenu est complet et suffisamment fin pour que le travail nécessaire à le rendre minimal soit raisonnable, contrairement aux approches actuelles pour lesquelles il est nécessaire de rechercher l'ensemble des motifs pour avoir accès à l'ensemble des règles. Les expériences menées confirment l'efficacité et la qualité des stratégies d'élagage.

Conclusion sur les propriétés d'anti-monotonie des mesures

La propriété d'anti-monotonie du **support** a permis un grand bond en avant dans le domaine de la recherche de motifs et de la recherche de règles d'association. Mais depuis son avènement au tout début des années 90, très peu de travaux avaient mis en évidence des propriétés algorithmiques de mesures d'intérêt. Nous avons montré tout au long de cette Thèse qu'il s'agissait pourtant là du cœur du problème. Être capable d'extraire directement les règles intéressantes sans passer par le filtre du **support** était un défi nécessaire dans un domaine où les limites du calcul ont essentiellement été repoussées par des avancées techniques.

Nous avons proposé des propriétés d'anti-monotonie pour un ensemble de 10 mesures. Ces propriétés sont aisément compréhensibles en utilisant notre cadre d'étude des règles d'association et leur établissement est essentiellement visuel. Nous avons proposé deux façons d'établir une propriété d'anti-monotonie pour des mesures dont la valeur à l'origine est inférieure au seuil fixé. L'étude de ces propriétés d'anti-monotonie et de leur efficacité sur un ensemble de 5 bases classiques a montré d'excellents résultats par rapport à l'approche Apriori, notamment dans la recherche de pépites de connaissances.

Parmi les 42 mesures que nous étudions, 10 présentent une propriété d'anti-monotonie, mais notre travail n'est pas exhaustif, et d'autres mesures pourraient présenter le même type de propriété. Notre condition d'existence permet notamment d'avoir une première idée rapide sur la présence d'une propriété d'anti-monotonie. La restriction de ce résultat aux règles de classe doit nous inciter à étudier l'efficacité de telles mesures dans une tâche de prédiction notamment en parallèle aux travaux de [Jalali-Heravi et Zaïane 10] : par exemple, la mesure de **levier** ou de **précision** sont deux mesures qui arrivent en tête de classement pour certaines bases de données, et qui possèdent, d'après nos travaux, une propriété d'anti-monotonie.

Il reste aussi à se poser la question des pépites de connaissance dans ce contexte, car si l'on est capable, grâce à l'algorithme DARC d'extraire directement les règles intéressantes, la remarque que nous avons formulée quant à la fixation du seuil dans Apriori pose un autre problème. En effet, les mesures possédant une telle propriété d'anti-monotonie sont-elles vraiment aptes à la recherche de pépites au sens premier du terme, c'est-à-dire des règles de faible **support** mais de mesure élevée ? La continuité en 0 ou la continuité par prolongement en 0 assure en fait que les règles de faible **support** aient une valeur de mesure inférieure au seuil, contrairement à la **confiance**, non continue, pour laquelle des mesures de faible **support** peuvent avoir des valeurs de mesure élevées. Notre méthode permet cependant d'extraire toutes les règles intéressantes et d'abaisser le seuil de **support** en diminuant considérablement les temps de calcul.

Conclusion générale

Trois années de thèse ne peuvent se conclure en une page, d'ailleurs, une thèse se conclue-t-elle vraiment un jour ? Ces travaux ne sont qu'une ouverture, un ensemble d'idées qui ne demandent qu'à survivre à leur auteur qui espère qu'elles ne s'endormiront pas au fond d'un vieux placard. Raymond Devos disait :

> *“ Vous savez, les idées elles sont dans l'air. Il suffit que quelqu'un vous en parle de trop près, pour que vous les attrapiez ! ”*

Tout au long de ma thèse, j'ai attrapé des idées au vol, comme elles se présentaient, et leur ai insuflé l'énergie que je pouvais puiser au fond de moi. Certaines m'ont été soufflées, d'autres sont arrivées d'elles-mêmes sans que l'on ne s'explique encore aujourd'hui comment. Certaines me survivront, d'autres sont déjà peut-être enterrées, jusqu'à ce qu'un autre ne les réanime et ne les exploite mieux que je ne l'ai fait. Ainsi vont les idées, si je ne les avais attrapées, d'autres l'auraient fait, mais je suis fier d'avoir parfois été le premier.

Ainsi, notre définition de la robustesse, s'appuyant sur une notion de distance à une surface définie par la mesure et le seuil, est cohérente, naturelle, trouve un écho dans le domaine des statistiques, et a été validée par l'expérience. Nous avons pu distinguer deux familles de mesures, d'un côté les mesures planes pour lesquelles il existe un calcul exact de la robustesse, et de l'autre côté, les mesures quadratiques qui nécessitent de faire appel aux méthodes de Newton. Nous n'avons cependant pas eu le temps d'approfondir les applications de cette robustesse, et des idées restent en suspens, notamment au sujet de son implication dans les tâches de classification, et de la pertinence de la distance euclidienne dans ce contexte.

Les propriétés algorithmiques que nous avons généralisées doivent être étudiées encore, car seule l'opti-monotonie a fait l'objet d'une condition nécessaire et suffisante, alors que l'omni-monotonie et la propriété GUEUC ne nous ont laissé entrevoir que des conditions à sens unique. Nous avons l'intuition qu'il manque peu de chose pour faire se rejoindre ces deux bouts. Et puis, nous n'avons étudié qu'un ensemble de 42 mesures, qu'en est-il de toutes les autres ? L'auteur de ces lignes avait-il raison de laisser entendre que peu importent les applications, tant que l'on a le résultat théorique...

Enfin, nous croyons que l'étude des propriétés d'anti-monotonie des mesures d'intérêt *via* l'aspect visuel offert par notre cadre formel d'étude est un grand pas en avant qui offre une contribution théorique, et non pas technique, au problème de la recherche de règles. Seulement, cette étude est restreinte à la recherche de règles de classe, ne pourrait-on pas la généraliser, ou l'implémenter dans des algorithmes plus avancés ? Il reste de plus encore à étudier les applications pratiques que peuvent avoir de telles propriétés.

Parmi les applications possibles, toutes ces propriétés pourraient faire partie des critères de décision lors du choix d'une mesure. Des travaux tels que ceux de [Tan et al. 04] ou [Lenca et al. 08] recensent un grand nombre de propriétés souhaitables des mesures d'intérêt (10 propriétés pour le premier, 8 pour le second). L'article [Lenca et al. 08] s'appuie sur une technnique d'aide à la décision multicritère pour proposer, à partir de préférences de l'utilisateur, un classement des mesures. Les propriétés étudiées sont pour certaines totalement objectives et décrivent des propriétés descriptives des règles (valeur à l'indépendance, variations...) mais d'autres sont subjectives en décrivant par exemple l'intelligibilité ou la facilité de fixer un seuil. Dans [Tan et al. 04], toutes les propriétés sont descriptives. Cependant ces travaux se focalisaient principalement sur l'aspect qualitatif des règles d'associations. Plus récemment des travaux tels que ceux de [Jalali-Heravi et Zaïane 10] sur

mesure	PLANE	QUAD	OMNI	GUEUC	OPTI	ANTI
confiance	✓	✗	✓	✓	✓	✗
confiance centrée	✗	✓	✗	✓	✓	✗
moindre contradiction	✓	✗	✗	✗	✓	✓
conviction	✗	✓	✗	✓	✓	✗
cosine	✗	✓	✗	✗	✓	✗
couverture	✓	✗	✓	✓	✓	✗
Czekanowski	✓	✗	✗	✗	✓	✓
facteur bayésien	✗	✓	✗	✓	✓	✗
force collective	✗	✗	✗	✗	✓	✗
gain	✓	✗	✓	✗	✗	✗
gain informationnel	✗	✓	✗	✓	✓	✗
Ganascia	✓	✗	✓	✓	✓	✗
indice de Gini	✗	✗	?	✗	✗	✗
indice d'implication	✗	✗	✗	✗	✓	✗
intérêt	✗	✗	?	✗	✗	✗
J1-mesure	✗	✗	✗	✗	✗	✗
Jaccard	✓	✗	✗	✗	✓	✓
J-mesure	✗	✗	?	✗	✗	✗
Kappa	✗	✓	✗	✗	✓	✓
Klosgen	✗	✗	✗	✗	✗	✗
Kulczynski	✓	✗	✗	✓	✓	✓
Laplace	✗	✗	✓	✗	✓	✓
levier	✗	✓	✗	✗	✓	✓
lift	✗	✓	✗	✓	✓	✗
Loevinger	✗	✓	✗	✓	✓	✗
odds ratio	✗	✓	✗	✗	✓	✗
one way support	✗	✗	✗	?	✗	✗
coefficient de Pearson	✗	✗	✗	✗	✓	✗
Piatetsky-Shapiro	✗	✗	✗	✗	✓	✗
précision	✓	✗	✗	✗	✓	✓
prevalence	✓	✗	✓	✓	✗	✗
Q de Yule	✗	✓	✗	✗	✓	✗
rappel	✓	✗	✓	✗	✓	✗
risque relatif	✗	✓	✗	✗	✓	✗
Sebag-Shoenauer	✓	✗	✓	✓	✓	✗
spécificité	✓	✗	✗	✓	✓	✓
spécificité relative	✗	✓	✗	✗	✓	✓
support	✓	✗	✓	✗	✓	✓
taux exemples contre-exemples	✓	✗	✓	✓	✓	✗
valeur ajoutée	✗	✗	✗	✗	✓	✗
Y de Yule	✗	✗	✗	✗	✓	✗
Zhang	✗	✗	✗	✓	✓	✗

TABLE 12.5 : Résumé de l'ensemble des propriétés que nous avons étudiées au cours de cette thèse.

l'efficacité des mesures dans des tâches de classification ou de [Surana et al. 10] concernant le choix d'une mesure pour les règles rares se sont intéressés aux liens avec les algorithmes. Notre travail apporte lui aussi une forte contribution dans cette direction, puisqu'il propose 6 critères opérationnels pour le choix d'une mesure. Ces critères concernent d'une part la possibilité de calculer la

robustesse, et d'autre part d'utiliser des algorithmes efficaces, car dans un contexte de croissance des données omniprésente, l'efficacité des algorithmes est une question clé, à laquelle les parties III et IV de ce mémoire répondent.

À titre d'exemple, dans [Lenca et al. 08], les mesures du **facteur bayésien** et la **conviction** arrivent en tête suivant l'un des scénarii proposés. Or d'après nos travaux, chacune est quadratique (on peut donc évaluer convenablement la robustesse) et possède les propriétés GUEUC et d'opti-monotonie : elles ont donc des propriétés opérationnelles fortes. À l'opposé, les mesures de **kappa** et **Laplace** arrivent en fin de classement, mais d'après nos travaux, elles possèdent toutes les deux une bonne propriété d'anti-monotonie dans le cadre des règles de classe : cette information opérationnelle pourrait bien changer leur position, relativement aux mesures n'en possédant pas. C'est aussi le cas de la **confiance centrée** qui arrive, dans les deux scénarii, en quatrième et cinquième positions et qui, d'après nos travaux, possède une propriété d'anti-monotonie : une telle mesure d'intérêt est donc à la fois adaptée *et* efficace. Dans les travaux de [Jalali-Heravi et Zaïane 10], les mesures **levier**, **précision**, **Laplace** ou **confiance centrée** reviennent plusieurs fois comme de *bonnes* mesures pour la classification, soit on niveau du filtrage, soit au niveau de l'évaluation. Nos travaux montrent que ces mesures possèdent une propriété d'anti-monotonie dans le cadre des règles de classe, qui pourrait donc être utilisée dans la construction du classifieur.

Trois années de travaux tiennent malgré tout en quelques lignes, et sont résumées dans le tableau 12.5. Nous aurons au final mis en évidence 15 mesures planes, 13 mesures quadratiques, permettant un calcul de la robustesse ; 10 mesures sont omni-monotones, 15 possèdent la propriété GUEUC et 34 sont opti-monotones, et sont ainsi utilisables dans des algorithmes connus et efficaces ; enfin, nous montrons que 11 mesures possèdent une propriété d'anti-monotonie dans le cadre des règles de classe, et peuvent ainsi élaguer l'espace de recherche indépendemment de toute contrainte de **support**. Ce travail ouvre de nombreuses voies pour des travaux futurs, et je les laisse, si cela n'est pas trop paradoxal, avec regret et soulagement, espérant sincèrement que quelqu'un, un jour, s'en approchera suffisamment près pour pouvoir attraper au vol toutes les idées que je laisse en suspension autour de cette thèse.

. . . puis heureux et satisfait, il prit une position commode pour méditer sur les événements et les attendre.

Figure ⊤ : D'après *L'idée fixe du Savant Cosinus*, 11è Chant, Christophe, 1893

Bibliographie

[Aggarwal et Yu 98] C. C. Aggarwal & P. S. Yu. *A new framework for itemset generation.* In 98 ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Seattle, Washington, United States, pages 18–24, New York, NY, USA, 1998. ACM. 27

[Agrawal et Srikant 94] R. Agrawal & R. Srikant. *Fast Algorithms for Mining Association Rules in Large Databases.* In J. B. Bocca, M. Jarke & C. Zaniolo, editeurs, 20th International Conference on Very Large Data Bases, Santiago de Chile, Chile, pages 478–499. Morgan Kaufmann, December 1994. 7, 9, 10

[Agrawal et al. 93] R. Agrawal, T. Imieliski & A. Swami. *Mining association rules between sets of items in large databases.* In P. Buneman & S. Jajodia, editeurs, ACM SIGMOD International Conference on Management of Data, Washington, D.C., United States, pages 207–216, New York, NY, USA, May 1993. ACM Press. 7, 21, 28

[Asuncion et Newman 07] A. Asuncion & D. Newman. *UCI Machine Learning Repository*, 2007. 53

[Azé et Kodratoff 02] J. Azé & Y. Kodratoff. *Evaluation de la résistance au bruit de quelques mesures d'extraction de règles d'association.* In D. Hérin & D. A. Zighed, editeurs, 2nd Extraction et Gestion des Connaissances conference, Montpellier, France, volume 1-4 of *Extraction des Connaissances et Apprentissage*, pages 143–154. Hermes Science Publications, Januar 2002. 27, 49

[Azé et Kodratoff 04] J. Azé & Y. Kodratoff. *Extraction de "pépites" de connaissance dans les données : une nouvelle approche et une étude de la sensibilité au bruit.* numéro spécial RNTI-2 Mesures de qualité pour la fouille de données, vol. 1, pages 247–270, 2004. 87

[Azé et al. 03] J. Azé, S. Guillaume & P. Castagliola. *Evaluation de la résistance au bruit de quelques mesures quantitatives.* n^o spécial RNTI Entreposage et fouille de données, vol. 1, pages 159–170, 2003. 49

[Azé et al. 07] J. Azé, P. Lenca, S. Lallich & B. Vaillant. *A study of the robustness of association rules.* In R. Stahlbock, S. F. Crone & S. Lessmann, editeurs, The 2007 International Conference on Data Mining, Las Vegas, Nevada, USA, pages 163–169, Las Vegas, Nevada, USA, 2007. CSREA Press. 47

[Bahri et Lallich 09] E. Bahri & S. Lallich. *FCP-Growth : Class Itemsets for Class Association Rules.* In Proceedings of the Twenty-Second International Florida Artificial Intelligence Research Society Conference, May 19-21, 2009, Sanibel Island, Florida, USA. AAAI Press, 2009. 23

[Balcázar 09] J. L. Balcázar. *Confidence Width : an Objective Measure of Novelty for Association Rules.* In P. Lenca & S. Lallich, editeurs, Workshop on Quality Issues, Measures of Interestingness and Evaluation of Data Mining Models, in conjunction with the 13th Pacific-Asia

Conference on Knowledge Discovery and Data Mining, Bangkok, Thailand, April 2009. 77

[Balcázar et al. 10] J. L. Balcázar, C. Tîrnăucă & M. E. Zorrilla. *Mining Educational Data for Patterns With Negations and High Confidence Boost*. In A. Troncoso & J. C. Riquelme, editeurs, In Proceedings of the III Congreso Español de Informática (CEDI 2010). Simposio de Teoría y Aplicaciones de Minería de Datos (TAMIDA), pages 329–338. Ibergarceta Publicaciones, S.L. Madrid, 2010. 77

[Barthélemy et al. 06] J.-P. Barthélemy, A. Legrain, P. Lenca & B. Vaillant. *Aggregation of Valued Relations Applied to Association Rule Interestingness Measures*. In V. Torra, Y. Narukawa, A. Valls & J. Domingo-Ferrer, editeurs, 3rd International Conference on Modeling Decisions for Artificial Intelligence, Tarragona, Spain, volume 3885 of *Lecture Notes in Computer Science*, pages 203–214. Springer, April 2006. 30

[Bayardo et Agrawal 99] R. J. Bayardo & R. Agrawal. *Mining the Most Interesting Rules*. In S. Chaudhuri & D. Madigan, editeurs, 1999 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, pages 145–154. ACM, 1999. 32, 33

[Bayardo et al. 99] R. J. Bayardo, R. Agrawal & D. Gunopulos. *Constraint-Based Rule Mining in Large, Dense Databases*. In 15th International Conference on Data Engineering, Sydney, Australia, pages 188–197, Washington, DC, USA, 1999. IEEE Computer Society. 79

[Benjamini et Liu 99] Y. Benjamini & W. Liu. *A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence*. Journal of Statistical Planning and Inference, vol. 82, no. 1-2, pages 163–170, 1999. 57

[Berti-Equille 07] L. Berti-Equille. *Measuring and Modelling Data Quality for Quality-Awareness in Data Mining*. In F. Guillet & H. J. Hamilton, editeurs, Quality Measures in Data Mining, volume 43 of *Studies in Computational Intelligence*, pages 101–126. Springer, 2007. 49

[Bodon 06] F. Bodon. *A Survey on Frequent Itemset Mining*. Rapport technique, Budapest University of Technology and Economics, 2006. 12

[Bonchi et Lucchese 05] F. Bonchi & C. Lucchese. *Pushing Tougher Constraints in Frequent Pattern Mining*. In T. B. Ho, D. W.-L. Cheung & H. Liu, editeurs, 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Hanoi, Vietnam, volume 3518, pages 114–124. Springer, May 2005. 76

[Borgelt et Kruse 02] C. Borgelt & R. Kruse. *Induction of Association Rules :* Apriori *Implementation*. In 15th Conference on Computational Statistics, Berlin, Germany, pages 395–400, Heidelberg, Germany, 2002. Physika Verlag. 53

[Borgelt 03] C. Borgelt. *Efficient Implementations of* Apriori *and* Eclat. In 1st Workshop on Frequent Item Set Mining Implementations, Aachen, Germany, 2003. CEUR Workshop Proceedings 90. 17, 53

[Boulicaut et al. 03] J.-F. Boulicaut, A. Bykowski & C. Rigotti. *Free-Sets : A Condensed Representation of Boolean Data for the Approximation of Frequency Queries*. Data Mining and Knowledge Discovery, vol. 7, pages 5–22, 2003. 17

BIBLIOGRAPHIE

[Brin et al. 97a] S. Brin, R. Motwani & C. Silverstein. *Beyond Market Baskets : Generalizing Association Rules to Correlations.* In J. Peckham, editeur, ACM SIGMOD International Conference on Management of Data, Tucson, Arizona, USA, pages 265–276, New York, NY, USA, May 1997. ACM Press. 26, 28

[Brin et al. 97b] S. Brin, R. Motwani, J. D. Ullman & S. Tsur. *Dynamic itemset counting and implication rules for market basket data.* In J. Peckham, editeur, ACM SIGMOD International Conference on Management of Data, Tucson, Arizona, USA, pages 255–264, New York, NY, USA, May 1997. ACM Press. 27

[Cadot 05] M. Cadot. *A Randomization Test for extracting Robust Association Rules.* In Computational Statistics & Data Analysis, Limassol Chypre, 2005. 47

[Church et Hanks 90] K. W. Church & P. Hanks. *Word association norms, mutual information, and lexicography.* Computational Linguistics, vol. 16, no. 1, pages 22–29, 1990. 27

[Cleverdon et al. 66] C. W. Cleverdon, J. Mills & M. Keen. Factors determining the performance of indexing systems. ASLIB Cranfield project, Cranfield, 1966. 27, 28

[Cohen et al. 01] E. Cohen, M. Datar, S. Fujiwara, A. Gionis, P. Indyk, R. Motwani, J. D. Ullman & C. Yang. *Finding Interesting Associations without Support Pruning.* IEEE Transaction on Knowledge and Data Engineering, vol. 13, no. 1, pages 64–78, 2001. 75

[Cohen 60] J. Cohen. *A Coefficient of Agreement for Nominal Scales.* Educational and Psychological Measurement, vol. 20, no. 1, pages 37–46, April 1960. 27

[Czekanowski 13] J. Czekanowski. *Zarys metod statystycznych (Die Grundzuge der statischen Metoden).* Towarzystwo Naukowe Warszawskie, 1913. 27

[Dougherty et al. 95] J. Dougherty, R. Kohavi & M. Sahami. *Supervised and Unsupervised Discretization of Continuous Features.* In A. Prieditis & S. J. Russell, editeurs, Proceedings of the 12th International Conference on Machine Learning, Tahoe City, California, USA, pages 194–202. Morgan Kaufmann, 1995. 4

[Dudoit et van der Laan 07] S. Dudoit & M. J. van der Laan. Multiple testing procedures with applications to genomics. Springer, New York, NY, USA, 2007. 57

[Efron 79] B. Efron. *Bootstrap Methods : Another Look at the Jackknife.* The Annals of Statistics, vol. 7, no. 1, pages pp. 1–26, 1979. 57

[Fukuda et al. 96] T. Fukuda, Y. Morimoto, S. Morishita & T. Tokuyama. *Data Mining Using Two-Dimensional Optimized Accociation Rules : Scheme, Algorithms, and Visualization.* In H. V. Jagadish & I. S. Mumick, editeurs, ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, pages 13–23. ACM Press, June 1996. 27

[Ganascia 91] J. G. Ganascia. Deriving the learning bias from rule properties, pages 151–167. Clarendon Press, New York, NY, USA, 1991. 27

[Gay et Boullé 11] D. Gay & M. Boullé. *Un critère Bayésien pour évaluer la robustesse des règles de classification.* In A. Khenchaf & P. Poncelet, editeurs, 11th Extraction et Gestion des Connaissances conference, Brest,

France, volume RNTI-E-20 of *Revue des Nouvelles Technologies de l'Information*, pages 539–550. Hermann-Éditions, 2011. 45, 100

[Geng et Hamilton 06] L. Geng & H. J. Hamilton. *Interestingness measures for data mining : A survey.* ACM Computing Surveys, vol. 38, no. 3, Article 9, 2006. 24, 26

[Gini 21] C. Gini. *Measurement of Inequality and Incomes.* The Economic Journal, vol. 31, pages 124–126, 1921. 27

[Goethals 05] B. Goethals. *Frequent Set Mining.* In O. Maimon & L. Rokach, editeurs, The Data Mining and Knowledge Discovery Handbook, pages 377–397. Springer, 2005. 3, 12

[Good 65] I. J. Good. The estimation of probabilities : An essay on modern bayesian methods. M.I.T. Press, Cambridge, Massachusetts, 1965. 28

[Guillaume et al. 10] S. Guillaume, D. Grissa & E. M. Nguifo. *Propriété des mesures d'intérêt pour l'extraction des règles.* In 6th Workshop on Qualité des Données et des Connaissances, in conjunction with the 10th Extraction et Gestion des Connaissances conference, Hammamet, Tunisie, pages 15–28, January 2010. 79

[Han et al. 00] J. Han, J. Pei & Y. Yin. *Mining frequent patterns without candidate generation.* In W. Chen, J. F. Naughton & P. A. Bernstein, editeurs, ACM SIGMOD International Conference on Management of Data, Dallas, Texas, pages 1–12, Dallas, Texas, USA, 2000. ACM New York, NY, USA. 13

[Hébert et Crémilleux 06] C. Hébert & B. Crémilleux. *Optimized Rule Mining Through a Unified Framework for Interestingness Measures.* In A. M. Tjoa & J. Trujillo, editeurs, 8th International Conference on Data Warehousing and Knowledge Discovery, Krakow, Poland, volume 4081 of *Lecture Notes in Computer Science*, pages 238–247. Springer, September 2006. 33

[Hébert et Crémilleux 07] C. Hébert & B. Crémilleux. *A Unified View of Objective Interestingness Measures.* In P. Perner, editeur, 5th International Conference on Machine Learning and Data Mining, Leipzig, Germany, volume 4571 of *Lecture Notes in Computer Science*, pages 533–547. Springer, July 2007. 33

[Jaccard 01] P. Jaccard. *Étude comparative de la distribution florale dans une portion des Alpes et du Jura.* Bulletin de la Société Vaudoise des Sciences Naturelles, vol. 37, pages 547–579, 1901. 27

[Jalali-Heravi et Zaïane 10] M. Jalali-Heravi & O. R. Zaïane. *A study on interestingness measures for associative classifiers.* In Proceedings of the 2010 ACM Symposium on Applied Computing,Sierre, Switzerland, SAC'10, pages 1039–1046, New York, NY, USA, 2010. ACM. 71, 129, 131, 133

[Jaroszewicz et Simovici 04] S. Jaroszewicz & D. A. Simovici. *Interestingness of frequent itemsets using Bayesian networks as background knowledge.* In W. Kim, R. Kohavi, J. Gehrke & W. DuMouchel, editeurs, 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, pages 178–186. ACM, August 2004. 24

BIBLIOGRAPHIE

[Jeffreys 35] H. Jeffreys. *Some Tests of Significance, Treated by the Theory of Probability.* Proceedings of the Cambridge Philosophical Society, vol. 31, pages 203–222, 1935. 27

[Jovanoski et Lavrac 01] V. Jovanoski & N. Lavrac. *Classification Rule Learning with APRIORI-C.* In P. Brazdil & A. Jorge, editeurs, Progress in Artificial Intelligence, volume 2258 of *Lecture Notes in Computer Science*, pages 111–135. Springer Berlin / Heidelberg, 2001. 23

[Klemettinen et al. 94] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen & A. I. Verkamo. *Finding Interesting Rules from Large Sets of Discovered Association Rules.* In 3rd International Conference on Information and Knowledge Management, Gaithersburg, Maryland, pages 401–407. ACM, November - December 1994. 24

[Klösgen 92] W. Klösgen. *Problems for knowledge discovery in databases and their treatment in the statistics interpreter EXPLORA.* International journal of intelligent systems, vol. 7, pages 649–673, 1992. 27

[Kulczynski 27] S. Kulczynski. *Die Pflanzenassoziationen des Pieninen.* Bulletin International de l'Acadḿie Polonaise des Sciences et des Lettres, vol. Ser. B, Suppl. II, pages 57–203, 1927. 27

[Lallich et al. 07a] S. Lallich, O. Teytaud & E. Prudhomme. *Association Rule Interestingness : Measure and Statistical Validation.* In F. Guillet & H. J. Hamilton, editeurs, Quality Measures in Data Mining, volume 43 of *Studies in Computational Intelligence*, pages 251–275. Springer, 2007. 57

[Lallich et al. 07b] S. Lallich, B. Vaillant & P. Lenca. *A probabilistic framework towards the parameterization of association rule interestingness measures.* Methodology and Computing in Applied Probability, vol. 9, pages 447–463, 2007. 57

[Lavrac et al. 99] N. Lavrac, P. A. Flach & B. Zupan. *Rule Evaluation Measures : A Unifying View.* In ILP'99 : Proceedings of the 9th International Workshop on Inductive Logic Programming, pages 174–185, London, UK, 1999. Springer-Verlag. 28

[Le Bras et al. 09a] Y. Le Bras, P. Lenca & S. Lallich. *On Optimal Rule Mining : A Framework and a Necessary and Sufficient Condition of Antimonotonicity.* In T. Theeramunkong, B. Kijsirikul, N. Cercone & H. T. Bao, editeurs, 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Bangkok, Thailand, volume 5476 of *Lecture Notes in Computer Science*, pages 705–712. Springer, 2009. 1, 73, 97

[Le Bras et al. 09b] Y. Le Bras, P. Lenca, S. Lallich & S. Moga. *Généralisation de la propriété de monotonie de la all-confidence pour l'extraction de motifs intéressants non fréquents.* In 5th Workshop on Qualité des Données et des Connaissances, in conjunction with the 9th Extraction et Gestion des Connaissances conference, Strasbourg, France, pages 17–24, January 2009. 1

[Le Bras et al. 09c] Y. Le Bras, P. Lenca, S. Moga & S. Lallich. *All-Monotony : A Generalization of the All-Confidence Antimonotony.* 4th International Conference on Machine Learning and Applications, pages 759–764, 2009. 73, 83, 84

[Le Bras et al. 10a] Y. Le Bras, P. Lenca & S. Lallich. *Mining interesting rules without support requirement : a general universal existential upward closure property.* Annals of Information Systems, vol. 8, no. Part 2, pages 75–98, 2010. 8232. 1, 73, 78, 89

[Le Bras et al. 10b] Y. Le Bras, P. Meyer, P. Lenca & S. Lallich. *Mesure de la robustesse des règles d'association.* In 6th Workshop on Qualité des Données et des Connaissances, in conjunction with the 10th Extraction et Gestion des Connaissances conference, Hammamet, Tunisie, pages 29–40, January 2010. 43

[Le Bras et al. 10c] Y. Le Bras, P. Meyer, P. Lenca & S. Lallich. *A robustness measure of association rules.* 13rd European Conference on Principles of Data Mining and Knowledge Discovery, Barcelona, Spain, vol. 6322, pages 227–242, 2010. 43, 47, 51

[Le Bras et al. 11] Y. Le Bras, P. Lenca & S. Lallich. *Mining Classification Rules Without Support : an Anti-monotone Property of Jaccard Measure.* vol. 6926, pages 179–193, October 2011. 111

[Le et al. 08] T. T. N. Le, X.-H. Huynh & F. Guillet. *Finding the Most Interesting Association Rules by Aggregating Objective Interestingness Measures.* In D. Richards & B. H. Kang, editeurs, Knowledge Acquisition : Approaches, Algorithms and Applications, Pacific Rim Knowledge Acquisition Workshop, Hanoi, Vietnam, December 15-16, 2008, Revised Selected Papers, volume 5465 of *Lecture Notes in Computer Science*, pages 40–49. Springer, 2008. 30

[Lenca et al. 06] P. Lenca, S. Lallich & B. Vaillant. *On the robustness of association rules.* In 2nd IEEE International Conference on Cybernetics and Intelligent Systems and Robotics, Automation and Mechatronics, Bangkok, Thailand, pages 596 – 601, June 2006. 31, 46

[Lenca et al. 08] P. Lenca, P. Meyer, B. Vaillant & S. Lallich. *On selecting interestingness measures for association rules : user oriented description and Multiple Criteria Decision Aid.* European Journal of Operational Research, vol. 184, no. 2, pages 610–626, 2008. 26, 87, 105, 131, 133

[Lerman et al. 81] I.-C. Lerman, R. Gras. & H. Rostam. *Elaboration d'un indice d'implication pour les données binaires, I et II.* Mathématiques et Sciences Humaines, no. 74, 75, pages 5–35, 5–47, 1981. 27, 57

[Li et al. 01] W. Li, J. Han & J. Pei. *CMAR : Accurate and Efficient Classification Based on Multiple Class-Association Rules.* In N. Cercone, T. Y. Lin & X. Wu, editeurs, 1st IEEE International Conference on Data Mining, San Jose, California, USA, pages 369–376, Washington, DC, USA, 2001. IEEE Computer Society. 23, 46

[Li et al. 05] J. Li, A. W.-C. Fu, H. He, J. Chen, H. Jin, D. McAullay, G. Williams, R. Sparks & C. Kelman. *Mining risk patterns in medical data.* In R. Grossman, R. J. Bayardo & K. P. Bennett, editeurs, 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, pages 770–775, New York, NY, USA, 2005. ACM. 80, 97, 100

[Li 06] J. Li. *On Optimal Rule Discovery.* IEEE Transaction on Knowledge and Data Engineering, vol. 18, no. 4, pages 460–471, 2006. 80, 82, 97, 98, 102, 116, 117

BIBLIOGRAPHIE

[Lin et al. 02] W.-Y. Lin, M.-C. Tseng & J.-H. Su. *A Confidence-Lift Support Specification for Interesting Associations Mining.* In M.-S. Chen, P. Yu & B. Liu, editeurs, Advances in Knowledge Discovery and Data Mining, volume 2336 of *Lecture Notes in Computer Science*, pages 148–158. Springer Berlin / Heidelberg, 2002. 22

[Liu et al. 98] B. Liu, W. Hsu & Y. Ma. *Integrating Classification and Association Rule Mining.* In R. Agrawal, P. E. Stolorz & G. Piatetsky-Shapiro, editeurs, 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York City, USA, pages 80–86. AAAI Press, 1998. 23, 46, 71, 116

[Loevinger 47] J. Loevinger. *A systemic approach to the construction and evaluation of tests of ability.* Psychological monographs, vol. 61, no. 4, 1947. 28

[Morishita et Sese 00] S. Morishita & J. Sese. *Transversing itemset lattices with statistical metric pruning.* In 19th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Dallas, Texas, United States, pages 226–236, New York, NY, USA, 2000. ACM. 76

[Ohsaki et al. 04] M. Ohsaki, S. Kitaguchi, K. Okamoto, H. Yokoi & T. Yamaguchi. *Evaluation of rule interestingness measures with a clinical dataset on hepatitis.* In J.-F. Boulicaut, F. Esposito, F. Giannotti & D. Pedreschi, editeurs, 8th European Conference on Principles of Data Mining and Knowledge Discovery, Pisa, Italy, volume 3202 of *Lecture Notes in Computer Science*, pages 362–373, New York, NY, USA, 2004. Springer. 26

[Omiecinski 03] E. Omiecinski. *Alternative interest measures for mining associations in databases.* IEEE Transaction on Knowledge and Data Engineering, vol. 15, no. 1, pages 57–69, 2003. 76, 77, 83

[Padmanabhan et Tuzhilin 99] B. Padmanabhan & A. Tuzhilin. *Unexpectedness as a measure of interestingness in knowledge discovery.* Decision Support Systems, vol. 27, pages 303–318, December 1999. 24

[Pasquier et al. 99] N. Pasquier, Y. Bastide, R. Taouil & L. Lakhal. *Efficient Mining of Association Rules Using Closed Itemset Lattices.* Information Systems, vol. 24, no. 1, pages 25–46, 1999. 17

[Pasquier 00] N. Pasquier. *Data Mining : algorithmes d'extraction et de réduction des règles d'association dans les bases de données.* PhD thesis, Université Blaise Pascal - Clermont-Ferrand II, January 2000. 3

[Pearson 96] K. Pearson. *Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia.* Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, vol. 187, pages 253–318, 1896. 28

[Pei et al. 01] J. Pei, J. Han & L. V. Lakshmanan. *Mining Frequent Itemsets with Convertible Constraints.* In 17th International Conference on Data Engineering, Heidelberg, Germany, pages 433–442, Washington, DC, USA, 2001. IEEE Computer Society. 18

[Piatetsky-Shapiro 91] G. Piatetsky-Shapiro. *Discovery, Analysis, and Presentation of Strong Rules.* In Knowledge Discovery in Databases, pages 229–248. AAAI/MIT Press, 1991. 26, 28, 32

[Quinlan 93] J. R. Quinlan. C4.5 : Programs for machine learning. Morgan Kaufmann, San Mateo, CA, 1993. 23

[Ragel et Crémilleux 98] A. Ragel & B. Crémilleux. *Treatment of Missing Values for Association Rules.* In X. Wu, K. Ramamohanarao & K. B. Korb, editeurs, 2nd Pacific-Asia Conference on Knowledge Discovery and Data Mining, Melbourne,Australia, volume 1394 of *Lecture Notes in Computer Science*, pages 258–270. Springer, April 1998. 46

[Reiter 87] R. Reiter. A logic for default reasoning, pages 68–93. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1987. 24

[Salton et McGill 83] G. Salton & M. J. McGill. Introduction to modern retrieval. McGraw-Hill Book Company, 1983. 27

[Savasere et al. 95] A. Savasere, E. Omiecinski & S. B. Navathe. *An Efficient Algorithm for Mining Association Rules in Large Databases.* In U. Dayal, P. M. D. Gray & S. Nishio, editeurs, 21th International Conference on Very Large Data Bases, Zurich, Switzerland, pages 432–444. Morgan Kaufmann, September 1995. 16

[Schneider et Eberly 02] P. J. Schneider & D. Eberly. Geometric tools for computer graphics. Elsevier Science Inc., New York, NY, USA, 2002. 61

[Sebag et Schoenauer 88] M. Sebag & M. Schoenauer. *Generation of Rules with Certainty and Confidence Factors from Incomplete and Incoherent Learning Bases.* In J. Boose, B. Gaines & M. Linster, editeurs, European Knowledge Acquisition Workshop, pages 28.1–28.20. Gesellschaft für Mathematik und Datenverarbeitung mbH, Sankt Augustin, Germany, 1988. 28

[Smyth et Goodman 91] P. Smyth & R. M. Goodman. *Rule Induction Using Information Theory.* In Knowledge Discovery in Databases, pages 159–176. AAAI/MIT Press, 1991. 27

[Surana et al. 10] A. Surana, U. Kiran & P. K. Reddy. *Selecting a Right Interestingness Measure for Rare Association Rules.* In S. Chawla, K. Karlapalem & V. Pudi, editeurs, Proceedings of the 16th International Conference on Management of Data, December 8-10, 2010, Nagpur, India, pages 115–124. Computer Society of India, 2010. 113, 132

[Szathmary et al. 06] L. Szathmary, S. Maumus, P. Petronin, Y. Toussaint & A. Napoli. *Vers l'extraction de motifs rares.* In G. Ritschard & C. Djeraba, editeurs, 6th Extraction et Gestion des Connaissances conference, Lille, France, volume RNTI-E-6 of *Revue des Nouvelles Technologies de l'Information*, pages 499–510. Cépaduès-Éditions, 2006. 18

[Szathmary et al. 07] L. Szathmary, A. Napoli & P. Valtchev. *Towards Rare Itemset Mining.* In 19th IEEE International Conference on Tools with Artificial Intelligence, October 29-31, 2007, Patras, Greece, Volume 1, pages 305–312. IEEE Computer Society, 2007. 18

[Szathmary et al. 10] L. Szathmary, P. Valtchev & A. Napoli. *Finding Minimal Rare Itemsets and Rare Association Rules.* In Y. Bi & M.-A. Williams, editeurs, Knowledge Science, Engineering and Management, 4th International Conference, KSEM 2010, Belfast, Northern Ireland, UK, September 1-3, 2010. Proceedings, volume 6291 of *Lecture Notes in Computer Science*, pages 16–27. Springer, 2010. 113

[Tan et al. 04] P.-N. Tan, V. Kumar & J. Srivastava. *Selecting the Right Objective Measure for Association Analysis.* Information Systems, vol. 4, no. 29, pages 293–313, 2004. 26, 28, 131

BIBLIOGRAPHIE

[Toivonen 96] H. Toivonen. *Sampling Large Databases for Association Rules.* In T. Vijayaraman, A. P. Buchmann, C. Mohan & N. Sarda, editeurs, 22nd International Conference on Very Large Data Bases, Bombay, India, pages 134–145. Morgan Kaufman, 1996. 17

[Vaillant et al. 06] B. Vaillant, S. Lallich & P. Lenca. *Modeling of the counter-examples and association rules interestingness measures behavior.* In S. Crone, S. Lessmann & R. Stahlbock, editeurs, The 2006 International Conference on Data Mining, Las Vegas, Nevada, USA, pages 132–137. CSREA Press, June 26-29 2006. 31, 46, 47

[Vaillant 06] B. Vaillant. *Mesurer la qualité des règles d'association : Études formelles et expérimentales.* PhD thesis, ENST Bretagne, Brest, France, Décembre 2006. 26

[Wang et al. 98] K. Wang, S. H. W. Tay & B. Liu. *Interestingness-Based Interval Merger for Numeric Association Rules.* In 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, pages 121–128, New York, NY, USA, August 1998. ACM. 27

[Wang et al. 01] K. Wang, Y. He & D. W. Cheung. *Mining confident rules without support requirement.* In 10th International Conference on Information and Knowledge Management, Atlanta, Georgia, USA, pages 89–96, New York, NY, USA, 2001. ACM. 77, 78, 79, 89

[Webb et Zhang 05] G. I. Webb & S. Zhang. *K-Optimal Rule Discovery.* Data Mining and Knowledge Discovery, vol. 10, no. 1, pages 39–79, 2005. 46

[Xiong et al. 03] H. Xiong, P.-N. Tan & V. Kumar. *Mining Strong Affinity Association Patterns in Data Sets with Skewed Support Distribution.* In 3rd IEEE International Conference on Data Mining, Melbourne, Florida, USA, pages 387–394, Washington, DC, USA, 2003. IEEE Computer Society. 76, 77

[Yang et Wu 06] Q. Yang & X. Wu. *10 Challenging Problems in Data Mining Research.* International Journal of Information Technology and Decision Making, vol. 5, no. 4, pages 597–604, 2006. 75

[Yao et Liu 97] J. Yao & H. Liu. *Searching Multiple Databases for Interesting Complexes.* In H. Lu, H. Motoda & H. Liu, editeurs, 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining, Singapore, KDD : Techniques and Applications, pages 198–210. World Scientific Publishing Company, February 1997. 28

[Yin et Han 03] X. Yin & J. Han. *CPAR : Classification based on Predictive Association Rules.* In D. Barbará & C. Kamath, editeurs, 3dr SIAM International Conference on Data Mining, San Francisco, CA, USA, pages 331–335. SIAM, May 2003. 46

[Yule 00] G. U. Yule. *On the Association of Attributes in Statistics : With Illustrations from the Material of the Childhood Society.* Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, vol. 194, pages 257–319, 1900. 28

[Zaki et al. 97] M. J. Zaki, S. Parthasarathy & W. Li. *A Localized Algorithm for Parallel Association Mining.* In Proceedings of the 9th Annual ACM Symposium on Parallel Algorithms and Architectures, Newport, RI, USA, pages 321–330. ACM Press, June 1997. 9, 10

[Zaki 00] M. J. Zaki. *Scalable Algorithms for Association Mining.* IEEE Transaction on Knowledge and Data Engineering, vol. 12, no. 3, pages 372–390, 2000. 7, 12

[Zhang 00] T. Zhang. *Association Rules.* In T. Terano, H. Liu & A. L. P. Chen, editeurs, 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Kyoto, Japan, volume 1805 of *Lecture Notes in Computer Science*, pages 245–256. Springer, April 2000. 28

[Zimmermann et De Raedt 04] A. Zimmermann & L. De Raedt. *CorClass : Correlated Association Rule Mining for Classification.* In E. Suzuki & S. Arikawa, editeurs, 7th International Conference on Discovery Science, Padova, Italy, volume 3245 of *Lecture Notes in Computer Science*, pages 60–72. Springer, October 2004. 76