

Généralisation de la propriété de monotonie de la *all-confidence* pour l'extraction de motifs intéressants non fréquents

Yannick Le Bras^{*,***}, Philippe Lenca^{*,***}, Stéphane Lallich^{**} et Sorin Moga^{*,***}

*Institut TELECOM ; TELECOM Bretagne ;
UMR CNRS 3192 Lab-STICC;
Technopôle Brest Iroise CS 83818
29238 Brest Cedex 3

{yannick.lebras || philippe.lenca || sorin.moga} @telecom-bretagne.eu

**Université de Lyon,

Laboratoire ERIC, Lyon 2, France

stephane.lallich@univ-lyon2.fr

***Université Européenne de Bretagne, France

Résumé. Différentes études ont montré les limites du couple support/confiance dans les algorithmes de type Apriori, tant du point de vue quantitatif (quantité des règles), que qualitatif (redondance, intérêt, pépites de connaissance). Une solution consiste à concentrer au plus tôt la recherche sur les règles intéressantes en utilisant des mesures d'intérêt possédant des propriétés algorithmiques, mais aussi des capacités à mettre en évidence des règles d'un intérêt certain. Celles-ci permettent alors de trouver les règles d'intérêt élevé, sans utiliser un élagage préalable par la condition de support. Elles rendent également possible la recherche efficace de pépites de connaissance. C'est le cas de la *all-confidence* (ou *h-confidence*), transformation antimotone de la confiance. Nous nous intéressons ici à la possibilité d'appliquer une transformation semblable à d'autres mesures, au travers d'une condition nécessaire s'appuyant sur un cadre formel que nous définissons. Nous montrons cependant que parmi les 27 mesures étudiées ici, seules 5 d'entre elles peuvent être transformées en une mesure antimotone. Ainsi bien que très prometteuse, cette propriété d'antimonotonie n'est à l'heure actuelle applicable qu'à peu de mesures.

1 Introduction

La recherche de règles a attiré l'attention de nombreux chercheurs en apprentissage automatique et fouille de données. La principale justification de cet intérêt se trouve dans leur capacité à représenter les connaissances de manière compréhensible. Dans le cadre de l'apprentissage supervisé, les arbres de décision, par exemple CART (Breiman et al. (1984)) et C4.5 (Quinlan (1993)), ou bien les listes de décision, par exemple CN2 (Clark et Niblett (1989)), sont deux approches efficaces pour les tâches de classification. Dans le cadre non supervisé, la recherche de règles d'association est probablement la méthode la plus populaire pour établir des relations entre les différentes variables d'une base de données. Depuis sa première formulation et la proposition de l'algorithme AIS par Agrawal et al. (1993), le problème de la découverte de règles d'association, et par conséquent le problème sous-jacent de la recherche de motifs fréquents, ont fait l'objet de nombreux travaux.

Rappelons dans un premier temps les fondements de la recherche de règles d'association, ainsi que ses principaux défis. Une règle d'association $A \rightarrow B$ est la donnée de deux ensembles d'attributs A et B (aussi appelés motifs) non vides tels que $A \cap B = \emptyset$. Étant donnée un ensemble d'individus (chaque individu étant décrit par un ensemble d'attributs), la règle $A \rightarrow B$ signifie que lorsqu'un individu contient le motif A , alors, probablement, il contient aussi le motif B . Le problème de la découverte de règles d'association, conformément à Agrawal et al. (1993), consiste en la génération de toutes les règles ayant un support et une confiance supérieurs à des seuils fixés au préalable par l'utilisateur. Le support définit la proportion de transactions de la base de données incluant les motifs A et B (noté $supp(A \rightarrow B)$), tandis que la confiance représente cette proportion, mais au sein des transactions contenant A . Elle est notée $conf(A \rightarrow B)$. La recherche de règles d'association se déroule en deux étapes. Dans un premier temps, la contrainte de support est appliquée pour extraire de la base les motifs fréquents, c'est-à-dire des motifs dont le support est supérieur au seuil fixé. Dans un second temps, la contrainte de confiance minimale est appliquée aux motifs fréquents pour former des règles intéressantes. Cette stratégie implique deux problèmes majeurs : l'un concerne la complexité de la méthode, et l'autre la qualité des règles produites. Trouver tous les motifs fréquents dans une base de données comportant k attributs est une tâche coûteuse en temps de calcul, car le nombre de motifs présents dans la base peut atteindre 2^k . Cependant, la propriété de clôture du support,

appelée aussi antimonotonie, a permis l'élaboration d'algorithmes très efficaces tels que APRIORI (Agrawal et Srikant (1994)), PARTITION (Savasere et al. (1995)), ECLAT (Zaki (2000)) ou encore FP-GROWTH (Han et al. (2000)). D'intéressantes synthèses ont été réalisées par Hipp et al. (2000), Goethals et Zaki (2004) et Goethals (2005). Toutes ces approches soulignent le rôle prépondérant du support et du seuil associé. S'affranchir du support et/ou du seuil a été identifié comme un problème majeur, et pour le résoudre, deux grandes lignes se démarquent. L'une d'entre elles consiste à se libérer de l'utilisation du support (Cohen et al. (2001)), quand l'autre tente d'éliminer la tâche de fixation du seuil (Koh (2008)).

Trouver exactement toutes les règles intéressantes est également un problème, car les ensembles de règles générés sont très grands, en particulier avec le couple support/confiance. Une stratégie consiste à augmenter le seuil de support, ce qui entraîne la disparition de règles intéressantes, telles que les règles de faible support et de confiance élevée, appelées pépites de connaissance. Une autre stratégie, qui consiste à augmenter le seuil de confiance, favorise les règles ayant un conséquent fréquent, ce qui donne un grand nombre de règles inintéressantes, au sens où la connaissance de l'antécédent ne caractérise en rien la présence du conséquent.

On peut encore améliorer la recherche en ajoutant des contraintes sur l'apparition des motifs (Pei et Han (2000)), ou bien en utilisant les mesures d'intérêt au moment de la génération de candidats pour réduire à la fois le nombre de règles générées et le temps de parcours des bases de données. Cette technique est exploitée par exemple par Bayardo et al. (1999) dans l'algorithme DENSEMINER. En effet, comme le soulignent Yao et al. (2003), les mesures d'intérêt peuvent jouer des rôles différents : elles peuvent servir à diminuer l'espace de recherche (par exemple le support), sélectionner les règles intéressantes (par exemple la confiance), ou encore quantifier l'utilité et l'efficacité des règles (par exemple le coût). Nous nous limitons aux mesures objectives de l'intérêt des règles, qui ne dépendent que de la contingence des données, et parmi celles-ci, nous nous intéressons plus particulièrement à celles qui permettent à la fois de réduire l'espace de recherche et d'évaluer efficacement la qualité des motifs découverts.

De notre point de vue, cette approche est certainement la plus prometteuse. Pour dépasser le primat de la condition de support, différentes solutions ont été proposées au niveau algorithmique, plus spécialement pour la confiance.

Le reste de l'article est organisé de la façon suivante. Dans la section 2, nous passons en revue de récents travaux se concentrant sur les propriétés algorithmiques. Puis dans la section 3, nous introduisons un nouveau cadre formel permettant de mettre en relation les propriétés algorithmiques et analytiques des mesures. Dans la section 4, nous appliquons ce cadre à la recherche d'une généralisation de la *all-confidence* de Omiecinski (2003), qui permet de rendre antimonotone la confiance, et démontrons une condition suffisante de non-existence de cette généralisation. Forts de ce résultat, nous montrons qu'un grand nombre de mesures sont incompatibles avec cette propriété. Finalement, nous offrons une conclusion de notre travail dans la section 5.

2 Etat de l'art

Pour s'affranchir du support, quelques auteurs ont proposé des travaux se concentrant sur les propriétés algorithmiques de la confiance, le but annoncé étant de parvenir à obtenir toutes les règles de confiance élevée, et plus particulièrement les pépites de connaissance. Nous passons en revue ici certains des travaux proposant une alternative au support. Une partie d'entre eux agissent sur le type de motifs considérés, en minimisant leur nombre. D'autres se focalisent sur les mesures et exploitent leurs propriétés intrinsèques.

Wang et al. (2001) introduisent la notion de *Universal Existential Upward Closure* qui exploite une propriété de monotonie de la confiance. Cette propriété s'applique aux règles de classification, et permet une exploration par le haut du treillis des motifs. Cette stratégie est efficace en termes de génération de candidats, et permet de résoudre le problème des pépites de connaissance. En se concentrant sur la mesure de Jaccard, Cohen et al. (2001) proposent un algorithme d'approximation fondé sur les tables de hachage. L'apparition de faux positifs ou négatifs peut être contrôlée par les paramètres de l'algorithme. Sa complexité dépend de la force de ce contrôle, qui ne peut cependant jamais être complet. Zimmermann et Raedt (2004) adaptent quant à eux une technique développée par Morishita et Sese (2000), et proposent l'algorithme CORCLASS pour la classification. Cette technique exploite pleinement une propriété d'antimonotonie liée à la convexité du χ^2 . Li (2006) introduit la notion d'ensemble de règles optimales pour les règles de classification. Un tel ensemble contient toutes les règles, à l'exception de celles qui n'ont pas un intérêt plus grand que l'une de leurs règles plus générales. Ces ensembles de règles ont des propriétés qui sont à mettre en rapport avec les motifs fermés et les règles non redondantes étudiées par Zaki (2004). Leur principal avantage algorithmique se trouve dans la stratégie d'élagage que ces propriétés permettent d'adopter, stratégie qui s'applique à un grand nombre de mesures. L'auteur donne une preuve cas par cas pour 13 mesures. Dans Le Bras et al. (2009) nous généralisons ce résultat en proposant une condition nécessaire et suffisante d'existence de cette stratégie pour les mesures objectives. Cela nous permet alors de vérifier qu'elle peut s'appliquer à 26 des 32 mesures que nous y étudions.

Bonchi et Lucchese (2005) introduisent pour leur part une nouvelle notion d'antimonotonie appelée la *Loose-Antimonotonie*, qui s'applique efficacement à la plupart des contraintes statistiques. L'avantage d'une telle propriété est la

possibilité d’exploiter les travaux menés sur les algorithmes de type Apriori. Omiecinski (2003) introduit les mesures de *any-confidence*, *all-confidence* et *bond*. Les deux dernières ont l’avantage d’être antimonotones, comme le support, et facilement calculables. Il est alors possible de s’appuyer sur les nombreux efforts faits dans le domaine des algorithmes de type Apriori pour utiliser ces deux mesures. La *all-confidence*, définie pour un itemset I sur une base \mathcal{D} par $\frac{|\{d \in \mathcal{D} | I \subseteq d\}|}{\max_{I \subseteq I', I' \neq \emptyset} |\{d \in \mathcal{D} | I' \subseteq d\}|}$ sera en particulier renommée *h-confidence* par Xiong et al. (2003) et étudiée plus en détail. Un motif est intéressant au sens de la *h-confidence* si toutes les règles qu’il peut engendrer ont une confiance supérieure à un seuil donné. La *h-confidence* permet aussi de filtrer les motifs, jugés inintéressants, ayant la propriété de *cross-support*, c’est-à-dire les motifs dont les attributs ont des valeurs de support très différentes. L’algorithme HYPERCLIQUE-MINER (Xiong et al. (2003)) exploite ces deux propriétés.

Notons que les mesures objectives peuvent être considérées comme des fonctions de $\mathbb{R}^3 \rightarrow \mathbb{R}$. Lenca et al. (2008) ont proposé une étude poussée des domaines d’arrivée pour certaines mesures. D’autres proposent de contrôler l’espace d’arrivée en normalisant les mesures (Diatta et al. (2007); Slowiński et al. (2008)).

Nous introduisons ici un nouveau cadre formel proche de celui de Hébert et Crémilleux (2007) et l’utilisons pour faire le lien entre les propriétés analytiques des mesures et la propriété de monotonie, inspirée de la définition de la *all-confidence*, que nous proposons dans cet article.

3 L’espace des règles d’association

Nous proposons ici d’établir un cadre formel qui nous permet d’établir une étude analytique des mesures d’intérêt. Les mesures d’intérêt objectives, sur lesquelles porte notre étude, s’expriment en fonction de la table de contingence en fréquences relatives, et donc de trois variables. L’étude de leurs variations en fonction de ces variables nous permettra par la suite de faire le lien entre les mesures et leurs propriétés algorithmiques, mais elle implique la description d’un domaine de définition nous permettant de n’étudier que des cas réels.

3.1 Règles d’association

Une base de données booléenne \mathcal{DB} est décrite par un triplet $(\mathcal{A}, R, \mathcal{T})$, où \mathcal{A} est un ensemble d’attributs, \mathcal{T} un ensemble de transactions et R une relation binaire sur $\mathcal{A} \times \mathcal{T}$. Une règle d’association est définie par une base de données \mathcal{DB} , un ensemble $A \subset \mathcal{A}$ non vide appelé motif antécédent (ou tout simplement antécédent), et un ensemble $B \subset \mathcal{A}$ non vide appelé motif conséquent (ou tout simplement conséquent), tel que $A \cap B = \emptyset$. On note une règle d’association $A \xrightarrow{\mathcal{DB}} B$ ou simplement $A \rightarrow B$ s’il n’y a aucune ambiguïté possible.

Le support d’un motif traduit la fréquence d’apparition du motif dans la base de données. Nous le noterons $supp_{\mathcal{DB}}(A)$, ou, lorsqu’il n’y a pas d’ambiguïté, simplement $supp(A)$. Le support d’une règle $A \rightarrow B$ est le support du motif AB .

3.2 Tables de contingence

Soit A et B deux motifs dans une base de données \mathcal{DB} . La table de contingence en fréquences relatives conjointes de A et B renseigne sur la coprésence de ces deux motifs (figure 1).

| | | | |
|-----------|------------------|------------------------|-----------------|
| | B | \bar{B} | |
| A | $supp(AB)$ | $supp(A\bar{B})$ | $supp(A)$ |
| \bar{A} | $supp(\bar{A}B)$ | $supp(\bar{A}\bar{B})$ | $supp(\bar{A})$ |
| | $supp(B)$ | $supp(\bar{B})$ | 1 |

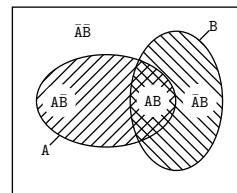


FIG. 1 – Table de contingence de A et B , sous forme de tableau et sous forme graphique.

La table de contingence a 3 degrés de liberté : il faut au moins trois de ses valeurs pour la décrire, et trois valeurs suffisent pour retrouver les autres valeurs. Par exemple, la table de contingence est entièrement décrite par les deux fréquences relatives marginales $supp(A)$ et $supp(B)$, et la fréquence relative conjointe $supp(AB)$.

Une règle d’association sur une base de données étant définie par deux motifs, on peut faire le raccourci de parler de table de contingence d’une règle d’association, ce qui nous amène à proposer la notion de système descripteur.

Définition 1. Nous appelons *système descripteur* de la table de contingence un triplet de fonctions (f, g, h) sur les règles d’association qui permet de décrire entièrement la table de contingence des règles d’association.

Exemple 1. Définissons les fonctions suivantes :

$$ant(A \rightarrow B) = supp(A); \quad cons(A \rightarrow B) = supp(B); \quad conf(A \rightarrow B) = \frac{supp(AB)}{supp(A)}$$

Le triplet $(conf, ant, cons)$ est un système descripteur de la table de contingence. De même, si l'on définit les fonctions $ex(A \rightarrow B) = supp(AB)$ et $c-ex(A \rightarrow B) = supp(A\bar{B})$, les triplets $(ex, ant, cons)$ et $(c-ex, ant, cons)$ sont des systèmes descripteurs.

Une mesure d'intérêt objective est une fonction de l'espace des règles d'association dans l'espace des réels étendus $(\mathbb{R} \cup \{-\infty, +\infty\})$. Elle sert à quantifier l'intérêt des règles. Il existe un grand nombre de mesures objectives (Tan et al. (2004); Yao et al. (2006); Geng et Hamilton (2006); Lenca et al. (2008)). Celles-ci s'expriment en fonction de la table de contingence, et peuvent donc être considérées comme des fonctions de trois variables (un système descripteur de la table de contingence). Dans ce travail, nous nous intéressons uniquement à ce type de mesures en approfondissant notamment le travail de Omiecinski (2003) et Xiong et al. (2003) pour la *all-confidence*.

3.3 Domaine minimal conjoint

Comme il existe des variables aléatoires sur un univers probabiliste, il est possible de définir des variables sur l'espace des règles d'association. Un système descripteur d de la table de contingence est alors un triplet de variables sur cet espace, et une mesure d'intérêt m peut être exprimée à l'aide d'une fonction ϕ_m du triplet. Si l'on veut étudier cette fonction, il suffit de se restreindre à l'étude sur le domaine de variation conjoint de ce triplet. Cette étude n'aura d'ailleurs pas de sens en dehors de ce domaine, où les points ne correspondent à aucune situation réelle.

Définition 2. Nous appelons le couple (ϕ_m, \mathcal{D}_d) , formé par cette fonction et le domaine de variation conjoint (associé à un système descripteur), la *fonction de mesure d-adaptée* à la mesure m .

Il est important de voir que l'écriture de la partie fonctionnelle de cette fonction de mesure dépend du système descripteur choisi. Cependant, une fois ce système fixé, la fonction de mesure adaptée est définie de manière unique. Nous nous permettrons donc, lorsque toute ambiguïté sera levée, d'omettre de préciser le système descripteur choisi.

Si d est un système de descripteurs de la table de contingence, le domaine de variation conjoint associé à ce système est défini par les contraintes imposées par les valeurs entre elles.

Exemple 2. Soit la fonction $ex : (\mathbb{A} \xrightarrow{\mathcal{DB}} \mathbb{B}) \mapsto supp_{\mathcal{DB}}(AB)$. On définit, sur la base des fonctions introduites précédemment, le système descripteur $d_{ex} = (ex, ant, cons)$. Nous avons sur ce système l'ensemble de contraintes suivant :

$$\begin{aligned} 0 &< ant < 1 \\ 0 &< cons < 1 \\ \max\{0, ant + cons - 1\} &\leq ex \leq \min\{ant, cons\} \end{aligned}$$

Détaillons ici la moins triviale de ces contraintes, $ant + cons - 1 \leq ex$:

$$supp(AB) = supp(A) - supp(A\bar{B}) \geq supp(A) - supp(\bar{B}) \geq supp(A) - 1 + supp(B)$$

Définissons alors le domaine :

$$D = \left\{ (x, y, z) \in \mathbb{Q}^3 \mid \left\{ \begin{array}{l} 0 < y < 1 \\ 0 < z < 1 \\ \max\{0, y + z - 1\} \leq x \leq \min\{y, z\} \end{array} \right. \right\}$$

Nous avons pour le moment l'affirmation : $\mathcal{D}_{d_{ex}} \subseteq D$ (D est complet). Pour montrer l'inclusion réciproque (i.e. D est minimal), il faut prouver que pour tout élément de D , il existe une règle d'association (et donc une base de données) qui lui correspond.

Soit à cet effet $(x, y, z) \in D$, posons $n \in \mathbb{N}$ tel que $(x \cdot n, y \cdot n, z \cdot n) \in \mathbb{N}^3$ (un tel n existe car x, y et z sont rationnels). Construisons la base de données \mathcal{B} du tableau 1. La définition du domaine donne un sens à cette base de données en assurant les inégalités : $0 < x < y < y + z - x < 1$. Alors dans \mathcal{B} , on a les égalités $supp(AB) = x$; $supp(A) = y$; $supp(B) = z$. Ainsi (x, y, z) est bien une valeur possible du triplet $(ex, ant, cons)$. Finalement, nous avons montré que le domaine de variation conjoint du triplet d_{ex} est exactement D . L'étude des variations des fonctions ϕ_m peut donc se limiter à ce domaine.

| | | | | | | | | | | | | |
|---|----|-----|-------------|-----|-----|-------------|-----|-----|-----------------------|---|-----|-----|
| | 1 | | $x \cdot n$ | | | $y \cdot n$ | | | $(y + z - x) \cdot n$ | | | n |
| A | 1 | ... | 1 | 1 | ... | 1 | 0 | ... | 0 | 0 | ... | 0 |
| B | 1 | ... | 1 | 0 | ... | 0 | 1 | ... | 1 | 0 | ... | 0 |
| | AB | | | A→B | | | B→A | | | | | |

TAB. 1 – Base de données représentant le triplet (x, y, z) . Pour des raisons de place, les individus sont donnés en colonne, les attributs en ligne. On a $\text{supp}(AB) = x$; $\text{supp}(A) = y$; $\text{supp}(B) = z$.

4 Généralisation de la all-confiance

Dans la suite, nous nous concentrons sur le système de descripteurs d_{ex} . Il ne sera donc plus précisé. Lorsque nous étudions les variations d'une fonction f définie sur le domaine D "par rapport à sa 1ère variable", nous étudions en fait les variations de la fonction définie pour tout couple (y, z) sur le domaine $\{x | (x, y, z) \in D\}$ par $x \mapsto f(x, y, z)$.

Définition 3. (all-mesure) Soit m une mesure d'intérêt des règles d'association. Nous définissons la mesure *all-m* sur un motif I à partir de m de la façon suivante :

$$\text{all-}m(I) = \min_{AB=I, A \neq \emptyset, B \neq \emptyset} \{m(A \rightarrow B)\}.$$

La all-mesure d'un motif I est donc la plus petite valeur que puisse prendre la mesure considérée sur une règle extraite du motif I .

Définition 4. (all-monotonie) m est dite all-monotone si la mesure *all-m* est antimonotone.

Cette propriété très intéressante du point de vue algorithmique a été démontrée par Omiecinski (2003) pour la confiance. Nous indiquons que ce résultat est vérifié pour les transformées monotones de la confiance :

Propriété 1. La confiance, et toutes les mesures fonctions croissantes de la confiance seule sont all-monotones.

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

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

□

En revanche, nos résultats (théorèmes 1 et 2) montrent qu'un grand nombre de mesures classiquement étudiées dans la littérature ne vérifient pas cette propriété.

Théorème 1. Soit m une mesure d'intérêt des règles d'association, et $(\phi_m, \mathcal{D}_{d_{ex}})$ 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 all-monotone.

Démonstration. **Supposons que cette fonction de mesure soit décroissante strictement en la deuxième et la troisième variable.** Nous allons montrer que pour de telles mesures, il est possible de construire une base de données simple qui sert de contre-exemple à la all-monotonie. Choisissons de travailler sur le nombre minimal d'attributs dont nous avons besoin pour étudier une situation de all-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 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 :

$$\begin{aligned} \text{supp}(AB) &= x, \text{supp}(A) = y, \text{supp}(B) = z \\ \text{supp}(ABC) &= x, \text{supp}(C) = t, \text{supp}(AC) = y', \text{supp}(BC) = z' \end{aligned}$$

Les propriétés des fréquences impliquent :

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

Nous posons des inégalités strictes pour pouvoir exploiter pleinement la stricte décroissance de la mesure. Supposons de plus par convention 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 :

$$\begin{aligned}
 y \in]0, 1[& \quad (1) & \quad 0 < y' < y & \quad (3) & \quad y' + z' - x < t & \quad (6) \\
 & & \quad 0 < z' < z & \quad (4) & \quad t \leq z & \quad (7) \\
 z \in]0, 1[& \quad (2) & \quad 0 < x \leq \min\{z', y'\} & \quad (5) & \quad t + (y - y') + (z - z') \leq 1 & \quad (8)
 \end{aligned}$$

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 2.

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| B | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| C | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |

TAB. 2 – Base de données à 16 transactions

Remarque 2. D'une manière plus générale, on s'aperçoit que pour un entier n donné, toute base de données à trois attributs $\{A, B, C\}$ et $n + 3$ transactions T_1, \dots, T_{n+3} réparties comme suit : $T_1 = \{A\}$, $T_2 = \{B\}$, $T_3 = \{C\}$, $T_4 = T_5 = \dots = T_{n+3} = \{A, B, C\}$, vérifie les contraintes.

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 :

$$\begin{aligned}
 m(A \rightarrow B) &= \phi_m(x, y, z) & \begin{matrix} < \\ < \\ < \end{matrix} & \begin{matrix} 3 \\ 4 \\ 5+3 \end{matrix} & \phi_m(x, y', z) &= m(AC \rightarrow B) \\
 & & & & \phi_m(x, y, z') &= m(A \rightarrow BC) \\
 m(B \rightarrow A) &= \phi_m(x, z, y) & \begin{matrix} < \\ < \\ < \end{matrix} & \begin{matrix} 4 \\ 3 \\ 5+4 \end{matrix} & \phi_m(x, z, y') &= m(BC \rightarrow A) \\
 & & & & \phi_m(x, z, y) &= m(B \rightarrow AC) \\
 & & & & \phi_m(x, z, x) &= m(C \rightarrow AB)
 \end{aligned}$$

Ces inégalités nous donnent la conclusion : $all-m(\{A, B\}) < all-m(\{A, B, C\})$, **ce qui contredit l'antimonotonie.** \square

Exemple 3. La fonction de mesure adaptée à la mesure du facteur de Bayes à 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 all-monotone. Calculons BF sur la base de la table 2 :

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

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

Remarque 3. L'hypothèse de stricte monotonie est trop forte. Voyons la mesure de Loevinger : $L(A \rightarrow B) = 1 - \frac{supp(A\bar{B})}{supp(A)supp(\bar{B})}$. Sa fonction de mesure adaptée sur \mathcal{D} 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 :

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

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

Théorème 2. Soit m une mesure d'intérêt des règles d'association, et $(\phi_m, \mathcal{D}_{d_{ex}})$ 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 all-monotone.

Démonstration. La démonstration est la même que pour le théorème 1, les inégalités \leq devenant des égalités. \square

Remarque 4. La réciproque du théorème 2 n'est pas vérifiée et nous n'avons donc pas de condition nécessaire et suffisante. Le cas de la mesure *contramin* (Azé et Kodratoff (2004)) justifie cette affirmation. On l'applique à une base de données de taille n décrite remarque 2. Puisque $\text{contr}(A \rightarrow B) = \frac{\text{supp}(AB) - \text{supp}(A\bar{B})}{\text{supp}(B)}$, on mesure alors : $\text{all-contr}(\{A, B\}) = \frac{n-1}{n+1}$ et $\text{all-contr}(\{A, B, C\}) = \frac{n-1}{n}$. C'est-à-dire $\text{all-contr}(\{A, B\}) < \text{all-contr}(\{A, B, C\})$ dès que $n > 1$. Ainsi *contr* n'est pas all-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.

Ces travaux nous permettent d'apporter trois types de conclusion sur les mesures issues de la littérature (Tan et al. (2004); Geng et Hamilton (2006); Lenca et al. (2008)), la première s'appuyant sur le théorème 2, la deuxième sur la propriété 1 concernant les mesures fonctions de la confiance, et la dernière mettant en avant le lien entre *recall* et *confiance* :

- En combinant le résultat du théorème 2 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 all-monotonie : *spécificité, précision, lift, levier, confiance centrée, Jaccard, confiance positive, odds ratio, Klogen, contramin, valeur ajoutée, conviction, one way support, J₁-mesure, Piatetsky-Shapiro, cosine, Loevinger, gain informationnel, facteur bayésien, Zhang, index d'implication et kappa* ;
- A l'opposé, les mesures de *confiance, Sebag-Shoenauer, Ganascia* et le *taux d'exemples-contre exemples* 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. (2008)), il est donc possible de les utiliser dans un algorithme de type Apriori pour réaliser un élagage du treillis des motifs ;
- Enfin, concernant la mesure de *recall*, on remarque que pour deux itemsets A et B , $\text{rec}(A \rightarrow B) = \frac{\text{supp}(AB)}{\text{supp}(B)} = \text{conf}(B \rightarrow A)$. Ainsi, pour un itemset I , $\{\text{rec}(A \rightarrow B)|_{AB=I}\} = \{\text{conf}(A \rightarrow B)|_{AB=I}\}$. Donc $\text{all-rec}(I) = \text{all-conf}(I)$, et la mesure de *recall* est all-montone.

5 Conclusion et perspectives

Les mesures objectives d'intérêt des règles possèdent quelques propriétés algorithmiques qui n'ont pour l'instant été que peu exploitées, mis à part pour la confiance. Nous proposons dans cet article un cadre formel pour l'étude analytique des mesures. À l'aide de ce cadre, nous pouvons mettre en évidence les propriétés intrinsèques des mesures, et faire le lien avec des propriétés algorithmiques existantes. La all-confiance est l'une de ces propriétés, puisqu'elle permet de rendre antimonotone la confiance. Nous appliquons ce cadre pour répondre à la question suivante : existe-t-il d'autres mesures pouvant être rendues antimonotones de la même manière ? Nos résultats montrent qu'un grand nombre de mesures ne vérifient pas la propriété de all-monotonie que nous avons définie par extension de la all-confiance. Cependant, quelques mesures peuvent être rendues antimonotones par ce procédé.

Nous montrons également que certaines propriétés algorithmiques dépendent en fait de propriétés analytiques des mesures. Notre cadre formel fournit un moyen systématique d'étude des mesures, le but principal étant, pour une mesure donnée, de pouvoir déterminer les propriétés algorithmiques qu'elle vérifie, et donc les algorithmes qui seront utilisables avec cette mesure.

Références

- Agrawal, R., T. Imieliski, et A. Swami (1993). Mining association rules between sets of items in large databases. In *ICMD*, pp. 207–216.
- Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules in large databases. In *VLDB*, pp. 478–499.
- Azé, J. et Y. Kodratoff (2004). Extraction de "pépites" de connaissance dans les données : une nouvelle approche et une étude de la sensibilité au bruit. *RNTI 1*, 247–270.
- Bayardo, Jr, R. J., R. Agrawal, et D. Gunopulos (1999). Constraint-based rule mining in large, dense databases. In *ICDE*, pp. 188–197.
- Bonchi, F. et C. Lucchese (2005). Pushing tougher constraints in frequent pattern mining. In *PAKDD*, pp. 114–124.
- Breiman, L., J. Friedman, R. Olshen, et C. Stone (1984). *Classification and Regression Trees*. Wadsworth and Brooks.
- Clark, P. et T. Niblett (1989). The CN2 induction algorithm. *ML 3*, 261–283.
- Cohen, E., M. Datar, S. Fujiwara, A. Gionis, P. Indyk, R. Motwani, J. D. Ullman, et C. Yang (2001). Finding interesting associations without support pruning. *TKDE 13*(1), 64–78.

- Diatta, J., H. Ralambondrainy, et A. Totohasina (2007). Towards a unifying probabilistic implicative normalized quality measure for association rules. In *QMDM*, pp. 237–250. Springer.
- Geng, L. et H. J. Hamilton (2006). Interestingness measures for data mining : A survey. *ACM* 38(3, Article 9).
- Goethals, B. (2005). Frequent set mining. In *The Data Mining and Knowledge Discovery Handbook*, pp. 377–397. Springer.
- Goethals, B. et M. J. Zaki (2004). Advances in frequent itemset mining implementations : report on FIMI'03. *SIGKDD* 6(1), 109–117.
- Han, J., J. Pei, et Y. Yin (2000). Mining frequent patterns without candidate generation. In *ICMD*, pp. 1–12.
- Hébert, C. et B. Crémilleux (2007). A unified view of objective interestingness measures. In *MLDM*, pp. 533–547.
- Hipp, J., U. Güntzer, et G. Nakhaeizadeh (2000). Algorithms for association rule mining — a general survey and comparison. *SIGKDD* 2(1), 58–64.
- Koh, Y. S. (2008). Mining non-coincidental rules without a user defined support threshold. In *PAKDD*, pp. 910–915.
- Le Bras, Y., P. Lenca, et S. Lallich (2009). Optimal rules discovery: a framework and a necessary and sufficient condition of antimonotonicity. In *PAKDD (accepted)*.
- Lenca, P., P. Meyer, B. Vaillant, et S. Lallich (2008). On selecting interestingness measures for association rules : user oriented description and multiple criteria decision aid. *EJOR* 184(2), 610–626.
- Li, J. (2006). On optimal rule discovery. *TKDE* 18(4), 460–471.
- Morishita, S. et J. Sese (2000). Transversing itemset lattices with statistical metric pruning. In *PODS*, pp. 226–236.
- Omicinski, E. (2003). Alternative interest measures for mining associations in databases. *TKDE* 15(1), 57–69.
- Pei, J. et J. Han (2000). Can we push more constraints into frequent pattern mining ? In *KDD*, pp. 350–354.
- Quinlan, J. R. (1993). *C4.5 : Programs for Machine Learning*. Morgan Kaufmann.
- Savasere, A., E. Omicinski, et S. B. Navathe (1995). An efficient algorithm for mining association rules in large databases. In *VLDB*, pp. 432–444.
- Slowiński, R., S. Greco, et I. Szczęch (2008). Analysis of monotonicity properties of new normalized rule interestingness measures. In *HCP*, Volume 1, pp. 231–242.
- Tan, P.-N., V. Kumar, et J. Srivastava (2004). Selecting the right objective measure for association analysis. *IS* 4(29), 293–313.
- Wang, K., Y. He, et D. W. Cheung (2001). Mining confident rules without support requirement. In *IKM*, pp. 89–96.
- Xiong, H., P.-N. Tan, et V. Kumar (2003). Mining strong affinity association patterns in data sets with skewed support distribution. In *ICDM*, pp. 387–394.
- Yao, Y., Y. Chen, et X. Yang (2003). A measurement-theoretic foundation for rule interestingness evaluation. In *ICDM*, pp. 221–227.
- Yao, Y., Y. Chen, et X. D. Yang (2006). A measurement-theoretic foundation of rule interestingness evaluation. In *FNADM*, pp. 41–59. Springer.
- Zaki, M. J. (2000). Scalable algorithms for association mining. *TKDE* 12(3), 372–390.
- Zaki, M. J. (2004). Mining non-redundant association rules. *DMKD* 9(3), 223–248.
- Zimmermann, A. et L. D. Raedt (2004). Corclass : Correlated association rule mining for classification. In *DS*, pp. 60–72.

Summary

Many studies have shown the limits of the support/confidence framework used in Apriori-like algorithms, both from the quantitative point of view (quantity of rules) and from the qualitative (redundance, interest, nuggets of knowledge). One solution is to focus as soon as possible on interesting rules, by using interestingness measures possessing algorithmic properties, but also a high power of highlighting very interesting rules. They allow rules of high interest to be found without a preliminary support pruning. They thus make the mining of nuggets of knowledge possible. This is the case for all-confidence, or h-confidence, an antimonotonic transformation of confidence. We here study the possibility of applying such a transformation to a large panel of measures thanks to a necessary condition based on a formal framework that we proposed. However, we show that among the 27 measures studied, only five of them are antimonotone. Then even though very interesting, this antimonotone property is not applicable to a large set of measures.