Extraction optimisée de Règles d'Association Positives et Négatives (RAPN)

Sylvie Guillaume* et Pierre-Antoine Papon**

*Clermont Université, Université d'Auvergne, LIMOS, BP 10448, F-63000 Clermont
guillaum@isima.fr,
**Clermont Université, Université Blaise Pascal, LIMOS, BP 10448, F-63000 Clermont
papon@isima.fr

Résumé. La littérature s'est beaucoup intéressée à l'extraction de règles d'association positives et peu à l'extraction de règles négatives en raison essentiellement du coût de calculs et du nombre prohibitif de règles extraites qui sont pour la plupart redondantes et inintéressantes. Dans cet article, nous nous sommes intéressés aux algorithmes d'extraction de RAPN (*Règles d'Association Positives et Négatives*) reposant sur l'algorithme fondateur *Apriori*. Nous avons fait une étude de ceux-ci en mettant en évidence leurs avantages et leurs inconvénients. A l'issue de cette étude, nous avons proposé un nouvel algorithme qui améliore cette extraction au niveau du nombre et de la qualité des règles extraites et au niveau du parcours de recherche des règles. L'étude s'est terminée par une évaluation de cet algorithme sur plusieurs bases de données.

1 Introduction

L'extraction de règles d'association, consistant à découvrir des associations entre les conjonctions de variables binaires (*ou motifs*) d'une base de données, est une tâche importante en fouille de données. La recherche d'algorithmes efficaces de telles règles a été un problème majeur de cette communauté. Depuis le célèbre algorithme Apriori (Agrawal et Srikant, 1994), il y a eu de nombreuses variantes et améliorations. L'importance de l'extraction des règles négatives fut mise en évidence par (Brin et al., 1997) qui indiquent que de la connaissance précieuse peut se cacher dans ces règles. Ainsi (Brin et al., 1997) utilisent le test du χ^2 pour déterminer la dépendance entre deux motifs et ensuite une mesure de corrélation afin de trouver la nature de cette dépendance (*positive ou négative*). (Savasere et al., 1998) combinent les motifs fréquents [1] positifs avec la connaissance du domaine afin de détecter les associations négatives. Cette approche est difficile à généraliser puisqu'elle dépend de la connaissance du domaine. (Boulicaut et al., 2000) recherchent deux types de règles négatives, les règles du type $X \wedge Y \rightarrow \overline{Z}$ et $\overline{X} \wedge Y \rightarrow Z$, et pour cela ils proposent une approche basée sur les contraintes. (Teng et al., 2002) proposent un algorithme détectant uniquement les règles négatives du type $X \rightarrow \overline{Y}$. Quant à (Wu et al., 2004), (Antonie et Zaïane, 2004) et (Cornelis et al.,

1. Un motif X est dit **fréquent** si sa probabilité d'apparition $P(X)$ ou son support $sup(X)$ (*puisque nous avons* $P(X) = sup(X)$) est supérieure à un seuil min_{sup} fixé par l'utilisateur i.e. $sup(X) \geqslant min_{sup}$.

2006), ils extraient des règles négatives grâce à un algorithme basé sur l'algorithme fondateur *Apriori* (Agrawal et Srikant, 1994). (Wu et al., 2004) utilisent en plus du couple de mesures (*support* [2], *confiance* [3]), les deux mesures suivantes : une mesure d'intérêt qui n'est autre que la valeur absolue de la *nouveauté* (Lavrac et al., 1999) et une mesure nommée *ratio incrément de la probabilité conditionnelle* qui n'est autre que la mesure de *Shortliffe* (Shortliffe, 1976). Quant à (Antonie et Zaïane, 2004), ils utilisent comme mesure supplémentaire, le coefficient de corrélation (Pearson, 1896). Nous nous sommes focalisés dans cet article sur les techniques basées sur l'algorithme pionnier *Apriori*, et plus particulièrement sur les travaux de (Wu et al., 2004), (Antonie et Zaïane, 2004) et (Cornelis et al., 2006). A l'issue d'une étude approfondie de chacune des trois techniques, nous avons mis en évidence essentiellement les deux failles suivantes : (1) un nombre encore trop important de règles inintéressantes et (2) un parcours de recherche des règles non optimisé. Pour remédier au premier problème (*nombre important de règles inintéressantes*), nous retenons un sous-ensemble de motifs fréquents, *les motifs raisonnablement fréquents*, en éliminant ceux qui vont conduire à des règles non pertinentes c'est-à-dire les règles éliminées par toute mesure d'intérêt évaluant l'écart à l'indépendance de la règle comme par exemple la mesure de Piatetsky-Shapiro (Piatetsky-Shapiro, 1991). L'avantage de ce choix est que l'élimination intervient dans la première phase de l'algorithme et non plus en deuxième phase (*i.e. l'extraction des règles*) ou dans une phase de post-traitement des règles. De plus, nous utilisons également une mesure supplémentaire au couple de mesures (*support*, *confiance*) pour sélectionner les règles valides [4], la mesure M_G (Guillaume, 2010) qui est une amélioration de la mesure de *Shortliffe* (Shortliffe, 1976) utilisée par (Wu et al., 2004), et qui évalue non seulement l'écart de la règle par rapport à l'indépendance [5] mais également par rapport au point d'équilibre [6] (Blanchard et al., 2005). L'intérêt de prendre en compte le point d'équilibre est développé dans (Blanchard et al., 2005). Cette mesure plus sélective que celle utilisée dans (Wu et al., 2004) va permettre d'éliminer une nouvelle catégorie de règles inintéressantes. Pour remédier au deuxième problème (*parcours de recherche des règles non optimisé*), nous démontrons que seulement la moitié des règles négatives potentiellement valides sont à étudier, et ceci en fonction de la valeur de la confiance de la règle positive par rapport au support de la conclusion. De plus, parmi les 4 règles à étudier, nous avons de nouveau utilisé la propriété d'anti-monotonicité de la confiance [7], propriété abandonnée par (Antonie et Zaïane, 2004) et (Wu et al., 2004), à laquelle nous en avons ajouté une nouvelle dégagée par (Guillaume et Papon, 2012) et qui repose sur la mesure que nous allons utiliser, la mesure M_G.

L'article s'organise donc de la façon suivante. La *section 2* présente et motive les choix retenus pour optimiser l'extraction des règles d'association positives et négatives. La *section 3* développe l'algorithme proposé et la *section 4* évalue notre technique sur plusieurs bases de données. L'article se termine par une conclusion et des perspectives.

2. Le support $sup(X \Rightarrow Y)$ d'une règle est la fréquence d'apparition de la règle.

3. La confiance $conf(X \Rightarrow Y)$ d'une règle est la probabilité conditionnelle $P(Y/X)$.

4. On entend par règle **valide**, une règle qui vérifie un ensemble de contraintes. Dans *Apriori*, ces contraintes sont les suivantes : $sup(X \Rightarrow Y) \geqslant min_{sup}$ et $conf(X \Rightarrow Y) \geqslant min_{conf}$.

5. L'indépendance est le cas où $conf(X \Rightarrow Y) = P(Y)$ avec $P(Y) = sup(Y)$.

6. Le point d'équilibre est le cas où lorsque X est réalisé, il y a autant de chances de voir se réaliser Y que $\overline{Y}$, ainsi nous avons les relations suivantes : $conf(X \Rightarrow Y) = \frac{1}{2}$ et $conf(X \Rightarrow \overline{Y}) = \frac{1}{2}$ puisque $conf(X \Rightarrow Y) + conf(X \Rightarrow \overline{Y}) = 1$.

7. $\forall\ (X, Y, Z)\ /\ Y \subsetneq Z \subsetneq X$ et $X \subseteq \mathcal{I}$, si $conf(X \backslash Y \Rightarrow Y) < min_{conf}$ alors $conf(X \backslash Z \Rightarrow Z) < min_{conf}$

2 Optimisations de l'extraction des RAPN

Dans cette section, nous exposons les optimisations apportées par rapport aux techniques existantes. Nous commençons par présenter un moyen de réduire le nombre de règles en éliminant une catégorie de règles non pertinentes grâce à l'extraction de motifs raisonnablement fréquents.

2.1 Extraction de motifs raisonnablement fréquents

Nous recherchons, non plus les motifs fréquents comme dans *Apriori*, mais les motifs raisonnablement fréquents, c'est-à-dire les motifs dont le support est supérieur à un seuil minimal min_{sup} mais également inférieur à un seuil maximal que nous nommerons max_{sup}. Ce nouveau seuil maximal max_{sup}, initialisé par défaut à la valeur $1 - min_{sup}$, sera utilisé pour tous les types de motifs, à savoir les motifs positifs $X \cup Y$, les motifs négatifs du type $\overline{X} \cup \overline{Y}$ et les motifs mixtes $\overline{X} \cup Y$ que nous noterons par simplification respectivement XY, $\overline{X}\ \overline{Y}$ et $\overline{X}\ Y$; et où X et Y sont des conjonctions de variables binaires. Cette proposition se justifie par le fait qu'un motif omniprésent [8] M_1 (*où* M_1 *peut être un motif positif ou négatif c'est-à-dire* $M_1 \in \{X, \overline{X}\}$) est combiné avec presque tous les autres motifs fréquents M_2 ($M_2 \in \{Y, \overline{Y}\}$) car $sup(M_1M_2) \approx sup(M_2)$, et ceci sans pour autant révéler une combinaison M_1M_2 pertinente. Ainsi, beaucoup de règles du type $M_2 \Rightarrow M_1$ vont être extraites puisque $conf(M_2 \Rightarrow M_1) = \frac{sup(M_1M_2)}{sup(M_2)} \approx \frac{sup(M_2)}{sup(M_2)} \approx 1$ sans pour autant être pertinentes comme le montre la valeur de la *nouveauté* (Lavrac et al., 1999) : $nouveauté(M_2 \Rightarrow M_1) = sup(M_1M_2) - sup(M_1) \times sup(M_2) = sup(M_2) - sup(M_1) \times sup(M_2) = sup(M_2)(1 - sup(M_1)) \approx 0$ car $1 - sup(M_1) \approx 0$. Cette valeur proche de zéro pour la *nouveauté* indique que la règle est très proche du cas de l'indépendance entre les motifs M_1 et M_2, donc règle peu pertinente. De plus, $conf(M_1 \Rightarrow M_2) = \frac{sup(M_1M_2)}{sup(M_1)} \approx \frac{sup(M_2)}{sup(M_1)} \ll 1$ puisque le support de M_1 a une valeur élevée. En conclusion, cette recherche des motifs raisonnablement fréquents va nous permettre d'éliminer un certain type de règles non pertinentes, et ceci est d'autant plus intéressant que cela intervient en début de l'algorithme et non plus grâce à une étape de post-traitement des règles.

Nous présentons maintenant une seconde optimisation qui réduit le parcours de recherche des règles valides.

2.2 Parcours optimisé pour la recherche des règles valides

Aucune technique d'élagage pour le parcours des règles n'est utilisée par (Antonie et Zaïane, 2004), (Cornelis et al., 2006) et (Wu et al., 2004). En effet la propriété d'anti-monotonicité de la confiance n'est valable que pour les règles positives. Cependant, il est possible de restreindre et de diviser par 2 le nombre de règles négatives à étudier en fonction (1) soit du signe de la *nouveauté* ; (2) soit de la réponse à la question suivante "*la réalisation de la prémisse augmente-t-elle les chances d'apparition de la conclusion ?*". La réponse à cette question peut être obtenue grâce à la *nouveauté* puisque $sup(XY) - sup(X) \times sup(Y) = P(XY) - P(X)P(Y) = P(X)\left[\frac{P(XY)}{P(X)} - P(Y)\right] = P(X)\left[P(Y/X) - P(Y)\right] =$

8. On entend par motif *omniprésent*, un motif ayant une très forte valeur pour son support.

$P(X)\,[conf(X \Rightarrow Y) - P(Y)]$ ou grâce à la mesure de *Shortliffe*, propriété non exploitée par (Wu et al., 2004) malgré une utilisation de cette mesure. Nous explicitons cette restriction du nombre de règles négatives à évaluer grâce aux liens suivants entre les différentes règles.

Nous avons le lien suivant entre les règles antinomiques $X \Rightarrow Y$ et $X \Rightarrow \overline{Y}$: si la règle $X \Rightarrow Y$ est potentiellement intéressante c'est-à-dire si la réalisation de X augmente les chances d'apparition de Y (*question précédente*) ou encore si la confiance de la règle est supérieure à la probabilité d'apparition du motif conclusion Y (*autrement dit si* $conf(X \Rightarrow Y) > P(Y)$), alors la règle antinomique $X \Rightarrow \overline{Y}$ ne pourra pas être intéressante puisque dans ce cas-là, la confiance de la règle antinomique est inférieure à la probabilité d'apparition de la conclusion $\overline{Y}$ c'est-à-dire $conf(X \Rightarrow \overline{Y}) < P(\overline{Y})$ [9]. De la même façon, nous avons le lien suivant entre les règles $X \Rightarrow Y$ et $\overline{Y} \Rightarrow \overline{X}$: si $X \Rightarrow Y$ est potentiellement intéressante alors $\overline{Y} \Rightarrow \overline{X}$ le sera également puisque $conf(\overline{Y} \Rightarrow \overline{X}) > P(\overline{X})$ [10]. Pour finir, nous avons le lien suivant entre les règles symétriques $X \Rightarrow Y$ et $Y \Rightarrow X$: si la règle $X \Rightarrow Y$ est potentiellement intéressante, alors $Y \Rightarrow X$ le sera également puisque $conf(Y \Rightarrow X) > P(X)$ [11]. De ces trois liaisons précédemment établies entre les règles, nous pouvons en déduire que si la règle $X \Rightarrow Y$ est potentiellement intéressante alors les règles $Y \Rightarrow X$, $\overline{X} \Rightarrow \overline{Y}$ et $\overline{Y} \Rightarrow \overline{X}$ le seront également et si la règle $X \Rightarrow \overline{Y}$ est potentiellement intéressante alors les règles $\overline{Y} \Rightarrow X$, $\overline{X} \Rightarrow Y$ et $Y \Rightarrow \overline{X}$ le seront également. Par contre, si la règle $X \Rightarrow Y$ est potentiellement intéressante, les règles $X \Rightarrow \overline{Y}$, $\overline{Y} \Rightarrow X$, $\overline{X} \Rightarrow Y$ et $Y \Rightarrow \overline{X}$ ne seront pas intéressantes. Par conséquent, la connaissance de l'intérêt potentiel (*c'est-à-dire si la réalisation de la prémisse augmente les chances d'apparition de la conclusion*) ou non d'une des 8 règles (i.e. $X \Rightarrow Y$, $Y \Rightarrow X$, $\overline{X} \Rightarrow Y$, $Y \Rightarrow \overline{X}$, $X \Rightarrow \overline{Y}$, $\overline{Y} \Rightarrow X$, $\overline{X} \Rightarrow \overline{Y}$ et $\overline{Y} \Rightarrow \overline{X}$), permet d'éliminer l'examen de 4 autres règles (i.e. soit ($X \Rightarrow Y$, $Y \Rightarrow X$, $\overline{X} \Rightarrow \overline{Y}$ et $\overline{Y} \Rightarrow \overline{X}$), soit ($\overline{X} \Rightarrow Y$, $Y \Rightarrow \overline{X}$, $X \Rightarrow \overline{Y}$, $\overline{Y} \Rightarrow X$)) qui, comme nous venons de le démontrer, ne seront pas intéressantes. C'est ce qui est réalisé en partie par (Antonie et Zaïane, 2004) puisque le calcul du coefficient de corrélation entre X et Y va leur permettre de savoir s'il y a une corrélation positive ou négative entre X et Y. Si la corrélation est positive alors nous avons également une corrélation positive entre $\overline{X}$ et $\overline{Y}$. Si la corrélation est négative entre X et Y, alors nous pouvons en déduire une corrélation positive entre X et $\overline{Y}$ et également entre $\overline{X}$ et Y. Ainsi, si la corrélation est positive entre X et Y et jugée suffisamment élevée, c'est-à-dire si $coefCorr(X,Y) \geqslant min_{coefCorr}$ avec $min_{coefCorr}$ un seuil minimum défini par l'utilisateur, alors les deux règles $X \Rightarrow Y$ et $\overline{X} \Rightarrow \overline{Y}$ sont évaluées pour savoir si elles sont valides. Au contraire, si la corrélation est négative et jugée suffisamment faible, c'est-à-dire si $coefCorr(X,Y) \leqslant -min_{coefCorr}$, ce sont les deux règles $X \Rightarrow \overline{Y}$ et $\overline{X} \Rightarrow Y$ qui sont évaluées pour répondre à la question de leur validité. Concernant les règles manquantes (*c'est-à-dire les règles* $Y \Rightarrow X$, $\overline{Y} \Rightarrow \overline{X}$, $Y \Rightarrow \overline{X}$ et $\overline{Y} \Rightarrow X$), (Antonie et Zaïane, 2004) considèrent ensuite le couple de motifs (Y,X) et refont le calcul du coefficient de corrélation entre X et Y qui est inutile puisque $coefCorr(X,Y) = coefCorr(Y,X)$.

---

9. $conf(X \Rightarrow Y) > P(Y) \iff \frac{P(XY)}{P(X)} > P(Y) \iff \frac{P(X)-P(X\,\overline{Y})}{P(X)} > 1 - P(\overline{Y}) \iff 1 - conf(X \Rightarrow \overline{Y}) > 1 - P(\overline{Y}) \iff conf(X \Rightarrow \overline{Y}) < P(\overline{Y})$.

10. $\frac{P(\overline{X} \wedge \overline{Y})}{P(\overline{Y})} > P(\overline{X}) \iff P(\overline{X} \vee \overline{Y}) > P(\overline{X})P(\overline{Y}) \iff 1 - P(X \vee Y) > (1 - P(X))(1 - P(Y)) \iff 1 - P(X) - P(Y) + P(X \wedge Y) > 1 - P(X) - P(Y) + P(X)P(Y) \iff P(X \wedge Y) > P(X)P(Y) \iff conf(X \Rightarrow Y) > P(Y)$.

11. $\frac{P(XY)}{P(Y)} > P(X) \iff \frac{P(XY)}{P(X)} > P(Y) \iff conf(X \Rightarrow Y) > P(Y)$.

Nous allons utiliser ce résultat des liaisons d'intérêt entre les règles négatives afin de diminuer le nombre de règles à évaluer en le divisant par 2. Pour cela, nous devons savoir si la confiance de la règle $X \Rightarrow Y$ est supérieure au support de la conclusion (c'est-à-dire si $conf(X \Rightarrow Y) > sup(Y)$), ce qui nous garantit d'obtenir une règle potentiellement intéressante. (Wu et al., 2004) vérifient également le potentiel intérêt des règles grâce à la valeur absolue de la *nouveauté*. Autrement dit, avant de tester la validité des règles au regard du *support* et de la *confiance*, (Wu et al., 2004) vérifient si $\mid sup(XY) - sup(X) \times sup(Y) \mid \geqslant min_{interêt}$ que nous pouvons également écrire par $\mid P(XY) - P(X) \times P(Y) \mid \geqslant min_{interêt}$, ce qui peut se traduire par $\mid \frac{P(XY)}{P(X)} - P(Y) \mid \geqslant min_{\ \%\ interêt}$ et donc $\mid conf(X \Rightarrow Y) - sup(Y) \mid \geqslant min_{\ \%\ interêt}$. Ainsi, la recherche de l'appartenance de la règle à cette zone où $conf(X \Rightarrow Y) > sup(Y)$ avant de tester la contrainte de la confiance nous assure d'éliminer une partie des règles inintéressantes. C'est ce que nous allons retenir dans notre proposition pour répondre non seulement à une exigence de rapidité d'exécution de l'algorithme (*moins de règles à évaluer puisque nous divisons ce nombre par 2*) mais également au problème du nombre important de règles restituées, règles pas toujours pertinentes (*les règles inintéressantes qui sont éliminées sont celles où la prémisse n'augmente pas les chances d'apparition de la conclusion*). Cependant cette zone où les règles sont potentiellement intéressantes est encore trop importante et peut générer encore des règles inintéressantes. C'est le cas où la confiance de la règle est bien supérieure au support de la conclusion mais également inférieure au point d'équilibre (Blanchard et al., 2005), c'est-à-dire lorsque $sup(Y) < conf(X \Rightarrow Y) < \frac{1}{2}$. Dans le cas où la confiance de la règle $X \Rightarrow Y$ est inférieure à $\frac{1}{2}$, nous pouvons en déduire que nous avons plus de contre-exemples [12] que d'exemples [13] puisque $conf(X \Rightarrow \overline{Y}) > conf(X \Rightarrow Y)$, donc nous sommes en présence d'une règle $X \Rightarrow Y$ non pertinente. Nous allons donc retenir cette nouvelle zone d'intérêt potentiel que nous nommerons zone attractive entre X et Y et qui a été prise en compte par la mesure M_G (Guillaume, 2010) dont nous rappelons l'expression.

Zone attractive entre X **et** Y : $max\left(\frac{1}{2}, sup(Y)\right) < conf(X \Rightarrow Y)$

$$M_{G_a}(X \Rightarrow Y) = \frac{conf(X \Rightarrow Y) - max\left(sup(Y), \frac{1}{2}\right)}{1 - max\left(sup(Y), \frac{1}{2}\right)}$$

Zone répulsive entre X **et** Y : $conf(X \Rightarrow Y) < min\left(\frac{1}{2}, sup(Y)\right)$

$$M_{G_r}(X \Rightarrow Y) = \frac{conf(X \Rightarrow Y) - min\left(sup(Y), \frac{1}{2}\right)}{min\left(sup(Y), \frac{1}{2}\right)}$$

Zone inintéressante : $min\left(\frac{1}{2}, sup(Y)\right) \leqslant conf(X \Rightarrow Y) \leqslant max\left(\frac{1}{2}, sup(Y)\right)$

$$M_{G_i}(X \Rightarrow Y) = 0$$

Pour plus de précisions sur la sémantique de cette mesure, nous invitons le lecteur à consulter l'article de référence (Guillaume, 2010). Nous venons de montrer que lorsque les motifs X et Y ont une attraction positive, alors seules les règles $X \Rightarrow Y$, $\overline{X} \Rightarrow \overline{Y}$, $Y \Rightarrow X$ et $\overline{Y} \Rightarrow \overline{X}$ sont à étudier et lorsque les motifs X et Y ont une attraction négative, seules les règles $X \Rightarrow \overline{Y}$, $\overline{X} \Rightarrow Y$, $\overline{Y} \Rightarrow X$ et $Y \Rightarrow \overline{X}$ sont à évaluer. Si nous souhaitons utiliser également la propriété d'anti-monotonicité de la confiance, nous devons étudier séparemment les couples de motifs (X, Y) et (Y, X), ce qui oblige en contrepartie à calculer deux fois le type d'attraction entre X et Y comme le font (Antonie et Zaïane, 2004) lors du calcul du coefficient de corrélation entre les couples de motifs (X, Y) et (Y, X). En conséquence, soit nous n'utilisons pas la propriété

12. Un **contre-exemple** est un individu qui vérifie la prémisse X mais qui ne vérifie pas la conclusion Y, donc qui vérifie à la fois X et $\overline{Y}$.

13. Un **exemple** est un individu qui vérifie à la fois la prémisse X et la conclusion Y.

d'anti-monotonicité de la confiance et pour un couple (X, Y) de motifs, nous recherchons les 4 règles potentiellement valides après avoir déterminé le type d'attraction entre les deux motifs X et Y ; soit nous utilisons la propriété d'anti-monotonicité et nous traitons différemment les couples (X, Y) et (Y, X) et examinons uniquement les deux règles potentiellement valides après avoir déterminé (*certes deux fois*) le type d'attraction entre les motifs X et Y. Ainsi pour le couple (X, Y) et dans le deuxième cas de figure (*utilisation de la propriété d'anti-monotonicité de la confiance*), nous étudions soit les règles $X \Rightarrow Y$ et $\overline{X} \Rightarrow \overline{Y}$ (*attaction positive entre X et Y*), soit les règles $X \Rightarrow \overline{Y}$ et $\overline{X} \Rightarrow Y$ (*attaction négative entre X et Y*). Dans ce dernier cas de figure (*utilisation de la propriété d'anti-monotonicité de la confiance*) et si nous souhaitons poursuivre notre optimisation dans le parcours des règles potentiellement intéressantes, nous devons connaître les conditions qui vont nous permettre d'inférer le potentiel intérêt de la règle $\overline{X} \Rightarrow \overline{Y}$ à partir de celui de la règle $X \Rightarrow Y$. Il en est de même pour le couple de règles $X \Rightarrow \overline{Y}$ et $\overline{X} \Rightarrow Y$. Afin d'y parvenir nous utilisons les méta-règles dégagées par (Guillaume et Papon, 2012) qui permettent de générer les règles négatives à partir des règles positives $X \Rightarrow Y$. Comme notre objectif est de limiter l'espace de recherche des règles, nous utilisons uniquement la méta-règle nous révèlant que si la règle $X \Rightarrow Y$ ne vérifie pas la contrainte du seuil minimal min_{M_G} pour M_G, alors elle ne sera pas vérifiée par la règle $\overline{X} \Rightarrow \overline{Y}$ dans le cas où $\left(\frac{1}{2} < sup(X) < sup(Y)\right)$ et également si $\left(sup(X) < \frac{1}{2} < sup(Y)\right)$.

Maintenant, nous devons optimiser le parcours des règles dans le cas répulsif et par conséquent, utiliser une méta-règle permettant de passer de la règle $X \Rightarrow \overline{Y}$ à la règle $\overline{X} \Rightarrow Y$. Comme (Guillaume et Papon, 2012) ont dégagé des méta-règles uniquement à partir des règles positives $X \Rightarrow Y$, pour pouvoir faire cette transition des règles $X \Rightarrow \overline{Y}$ aux règles $\overline{X} \Rightarrow Y$, nous allons utiliser la méta-règle permettant de passer de la règle $X \Rightarrow Y$ à la règle $\overline{X} \Rightarrow \overline{Y}$, et par conséquent, celle que nous venons de décrire précédemment.
Nous résumons les deux méta-règles qui vont être utilisées pour optimiser la recherche des règles :

$(MR_1) : \forall\, (X, Y, Z)\ /\ Y \subsetneq Z \subsetneq X$ et $X \subseteq \mathcal{I}$,
si $conf(X \backslash Y \Rightarrow Y) < min_{conf}$ alors $conf(X \backslash Z \Rightarrow Z) < min_{conf}$.
$(MR_2) : \forall X \Rightarrow Y$ avec $\left(\frac{1}{2} < sup(X) < sup(Y)\right)$ ou $\left(sup(X) < \frac{1}{2} < sup(Y)\right)$,
si $M_G(X \Rightarrow Y) < min_{M_G}$ alors $M_G(\overline{X} \Rightarrow \overline{Y}) < min_{M_G}$.

Pour finir, les algorithmes existants reposant sur le couple (*support*, *confiance*) extraient des règles du type $X \Rightarrow Y$, $X \Rightarrow \overline{Y}$, $\overline{X} \Rightarrow Y$ et $\overline{X} \Rightarrow \overline{Y}$ et aucun des algorithmes n'extraient des règles du type $\overline{X_1}..\overline{X_p} \Rightarrow \overline{Y_1}..\overline{Y_q}$ et de façon plus générale, des règles du type $\overline{X_1}X_2..\overline{X_p} \Rightarrow Y_1\overline{Y_2}..Y_q$ où la prémisse et la conclusion de la règle sont des conjonctions de motifs à la fois positifs et négatifs. Dans un premier temps, nous allons nous intéresser à ce premier type de règles (*à savoir les règles* $\overline{X_1}..\overline{X_p} \Rightarrow \overline{Y_1}..\overline{Y_q}$) car cette recherche supplémentaire de règles de ce type va renforcer les liens entre la partie gauche et la partie droite des règles lors de la recherche simultanée des motifs raisonnablement fréquents avec ce nouveau type de motifs ($\overline{X_1}..\overline{X_p}$) comme nous l'expliquons dans la *section 2.3*.

2.3 Extension de l'extraction aux règles du type $\overline{X_1}..\overline{X_p} \Rightarrow \overline{Y_1}..\overline{Y_q}$

Lors de la recherche des motifs raisonnablement fréquents, nous allons rechercher en même temps ces conjonctions de motifs négatifs, motifs que nous noterons $\ddot{X}$. Cette recherche simultanée va renforcer notre souhait d'extraire des règles les plus pertinentes possibles. En effet, la

contrainte supplémentaire suivante $sup(\ddot{X}) \geqslant min_{s\ddot{u}p}$ sur les motifs X impose, comme pour la deuxième contrainte des motifs raisonnablement fréquents (*à savoir* $sup(X) \leqslant max_{sup}$), d'être en présence de motifs X non omniprésents. Pour un seuil d'exigence identique (*c'est-à-dire* $min_{sup} = min_{s\ddot{u}p}$ *et* $max_{sup} = 1 - min_{sup}$), cette nouvelle contrainte est plus restrictive que la deuxième contrainte (*à savoir* $sup(X) \leqslant max_{sup}$) puisque si nous avons $sup(X) \leqslant max_{sup}$, alors nous avons les équivalences suivantes : $1 - sup(\overline{X}) \leqslant max_{sup} \Leftrightarrow sup(\overline{X}) \geqslant 1 - max_{sup} \Leftrightarrow sup(\overline{X}) \geqslant min_{sup}$. Comme $sup(\ddot{X}) \leqslant sup(\overline{X})$ et dans ce cas particulier où $min_{sup} = min_{s\ddot{u}p}$, la contrainte $sup(\ddot{X}) \geqslant min_{s\ddot{u}p}$ prouve que le niveau d'exigence en matière de recherche de motifs non omniprésents est plus important. De plus, cette nouvelle contrainte va permettre d'éliminer un autre type de règles pas nécessairement intéressantes et ne conserver que les règles $X \Rightarrow Y$ où les motifs X et Y sont relativement bien corrélés puisque à la fois les motifs XY et $\overline{X}\ \overline{Y}$ doivent être fréquents. Afin de justifier nos propos, prenons l'exemple illustré sur la *Figure 1* et qui représente la contingence d'une règle $X \Rightarrow Y$ matérialisée par la surface des différents ensembles.

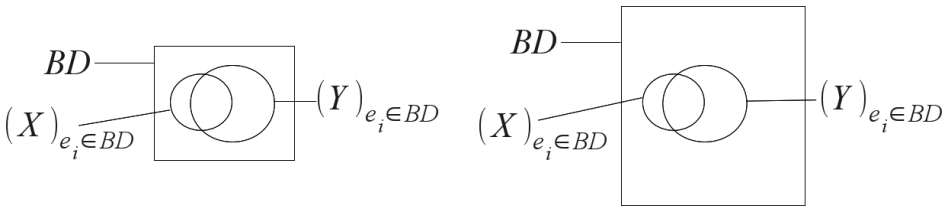

FIG. 1 – *Exemple de règles où les motifs ont des supports relativement élevés (courbe de gauche) et où les motifs ont des supports proches du seuil minimal* min_{sup} *(courbe de droite).*

La contingence des ensembles $X_{e_i \in BD}$ [14], $Y_{e_i \in BD}$ et $(X_{e_i \in BD} \cap Y_{e_i \in BD})$ est la même pour les deux courbes sauf pour l'ensemble $(\overline{X}_{e_i \in BD} \cap \overline{Y}_{e_i \in BD})$ qui est plus faible pour la courbe de gauche. Comme la contingence des ensembles $X_{e_i \in BD}$ et $(X_{e_i \in BD} \cap Y_{e_i \in BD})$ est la même dans les deux cas de figure, les deux règles $X \Rightarrow Y$ associées à ces deux contingences ont la même valeur pour la confiance. Cependant la règle associée à la courbe de droite de la *Figure 1* est plus pertinente que celle de la courbe de gauche puisque la probabilité d'avoir une intersection aussi importante entre $X_{e_i \in BD}$ et $Y_{e_i \in BD}$ est plus faible que pour le cas de la courbe de gauche. Nous savons que la confiance ne peut pas discerner ces deux types de règles et l'ajout de cette nouvelle contrainte sur les motifs $\ddot{X}$ nous assure d'éliminer un certain type de règles non pertinentes. Nous n'ajouterons pas, comme pour les motifs positifs, une valeur maximale à ne pas dépasser sur les supports des motifs $\ddot{X}$ car elle est en partie présente avec la contrainte du support minimum sur les motifs positifs. Nous présentons maintenant notre algorithme.

3 Algorithme

Tout d'abord, nous définissons ce que nous entendons par règle valide et donc les 6 contraintes Ct_1 à Ct_6 que doivent vérifier les règles.

14. $X_{e_i \in BD}$ est l'ensemble des individus e_i de la base de données BD vérifiant le motif X.

Algorithm 1 : Extraction des RAPN

Input : BD (*Base de Données*), min_{sup}, max_{sup}, $min_{s\ddot{u}p}$, min_{conf} et min_{M_G}
Output : R (*ensemble des règles valides*)

1: {Recherche des motifs Raisonnablement Fréquents (RF)}
$RF = funct\_RF(BD, min_{sup}, max_{sup}, min_{s\ddot{u}p})$
2: {Recherche des motifs Négatifs Raisonnablement Fréquents Minimaux (NRFM)}
$NRFM = funct\_NRFM(BD, RF)$
{Extraction des RAPN valides}
3: **for all** motif raisonnablement fréquent $X \in RF$ où $size(X) > 1$ **do**
4: **for all** conclusion $Y \subsetneq X$ / $size(Y) \nearrow$ **do**
5: Détermination du type d'attraction entre X et Y
6: **if** attraction positive **then**
7: $[\ \lceil](MR_1)\ ]$ Etude de la règle $X \backslash Y \Rightarrow Y$
8: $[(X \backslash Y \in NRFM) \wedge (\overline{Y} \in NRFM) \wedge (\lceil](MR_2))\ ]$ Etude de la règle $\overline{X \backslash Y} \Rightarrow \overline{Y}$
9: **else if** attraction négative **then**
10: $[(\overline{Y} \in NRFM)]$ Etude de la règle $X \backslash Y \Rightarrow \overline{Y}$
11: $[(\overline{X \backslash Y} \in NRFM)\ \wedge\ (\lceil](MR_2))\ )\ ]$ Etude de la règle $\overline{X \backslash Y} \Rightarrow Y$
12: **end if**
13: Etude de la règle $X \ddot{\backslash} Y \Rightarrow \ddot{Y}$
14: **end for**{conclusion Y}
15: **end for**{motif raisonnablement fréquent X}

Une RAPN **valide** est une expression du type $C_1 \Rightarrow C_2$ où $C_1 \in \{X, \overline{X}, \ddot{X}\}$, $C_2 \in \{Y, \overline{Y}, \ddot{Y}\}$, $X \subseteq \mathcal{I}$, $Y \subseteq \mathcal{I}$, $X \cap Y = \varnothing$, $\left(C_1 = \ddot{X} \Longleftrightarrow C_2 = \ddot{Y}\right)$, et telle que

$Ct_1 : min_{sup} \leqslant sup(XY) \leqslant max_{sup}$, $Ct_2 : min_{s\ddot{u}p} \leqslant sup(\ddot{X}\ddot{Y})$,
$Ct_3 : sup(C_1 \Rightarrow C_2) \geqslant min_{sup}$ si $(C_1, C_2) \neq (X, Y)$ et $(C_1, C_2) \neq (\ddot{X}, \ddot{Y})$,
$Ct_4 : conf(C_1 \Rightarrow C_2) \geqslant min_{conf}$, $Ct_5 : M_G(C_1 \Rightarrow C_2) \geqslant min_{M_G}$,
$Ct_6 : C_1 \Rightarrow C_2$ est minimal au regard des motifs négatifs raisonnablement fréquents $\overline{X}$ ou $\overline{Y}$.

La contrainte Ct_6 est celle présente dans (Cornelis et al., 2006) où les motifs C_1 et C_2 lorsqu'ils sont des motifs raisonnablement fréquents négatifs $\overline{X}$ et $\overline{Y}$, doivent également être minimaux c'est-à-dire qu'il n'existe pas par exemple pour le motif X un sous-ensemble $X' \subsetneq X$ tel que $\overline{X'}$ soit également raisonnablement fréquent.

L'algorithme d'extraction des RAPN (*voir l'algorithme 1*) commence par rechercher les motifs raisonnablement fréquents grâce à la fonction *funct_RF* (*ligne 1*). Cette recherche est similaire à celle utilisée par (Agrawal et Srikant, 1994) pour générer les motifs fréquents en rajoutant deux contraintes supplémentaires : un seuil maximal max_{sup} qui ne doit pas être dépassé par le support de X et un seuil minimum $min_{s\ddot{u}p}$ pour le support des motifs $\ddot{X}$, ce qui permet de vérifier les *contraintes* Ct_1 et Ct_2 des règles valides définies dans cette même section. A partir des motifs raisonnablement fréquents, on va rechercher les motifs négatifs raisonnablement fréquents minimaux grâce à la fonction *func_NRFM* (*ligne 2*). Cette recherche sert ensuite à s'assurer que la règle vérifie la *contrainte* Ct_6. Cette fonction est similaire à celle exposée dans (Cornelis et al., 2006) en rajoutant la contrainte du support maximum (*i.e.*

$sup(\overline{X}) \leqslant max_{sup}$). Vient ensuite la phase d'extraction des règles valides (*lignes 3 à 15*) grâce aux motifs extraits précédemment par les fonctions *funct_RF* et *func_NRFM*. Ainsi, pour chaque motif raisonnablement fréquent $X \in RF$ de taille strictement supérieure à 1 (*ligne 3*) et pour chaque conclusion possible Y (*ligne 4*) ordonnée par taille croissante (*comme pour l'algorithme Apriori*) et telle que $Y \subsetneq X$, on commence par déterminer le type d'attraction entre X et Y (*ligne 5*) grâce à la mesure M_G. Si c'est une attraction positive (*ligne 6*) i.e. $max\left(\frac{1}{2}, sup(Y)\right) < conf(X \Rightarrow Y)$, alors on s'assure que les règles $X \backslash Y \Rightarrow Y$ (*ligne 7*) et $\overline{X \backslash Y} \Rightarrow \overline{Y}$ (*ligne 8*) sont valides et pour cela, on vérifie les *contraintes* Ct_4 et Ct_5 définies précédemment. Avant cette vérification des *contraintes* Ct_4 et Ct_5, on s'assure que la règle $X \backslash Y \Rightarrow Y$ est candidate grâce à la propriété d'anti-monotonicité de la confiance $((MR_1)\ pas\ v\acute{e}rifi\acute{e}e)$ et que la règle $\overline{X \backslash Y} \Rightarrow \overline{Y}$ est également une règle candidate si d'une part, les motifs prémisse et conclusion sont des motifs minimaux i.e. $(\overline{X \backslash Y} \in NRFM)$ et $(\overline{Y} \in NRFM)$ et d'autre part, si la méta-règle 2 n'est pas vérifiée ($\neg(MR_2)$). Pour vérifier que la règle $X \backslash Y \Rightarrow Y$ est une règle candidate, on procède de la même manière que pour l'algorithme *Apriori* c'est-à-dire en vérifiant que tous les sous-ensembles T de Y ont conduit à une règle $X \backslash T \Rightarrow T$ ayant une confiance supérieure au seuil minimum. Lors de l'étude de la règle $\overline{X \backslash Y} \Rightarrow \overline{Y}$, on vérifiera également la *contrainte* Ct_3. Si c'est une attraction négative (*ligne 9*) i.e. $conf(X \Rightarrow Y) < min\left(\frac{1}{2}, sup(Y)\right)$, alors on étudie la règle $X \backslash Y \Rightarrow \overline{Y}$ si le motif $\overline{Y}$ est minimal i.e. $(\overline{Y} \in NRFM)$ (*ligne 10*) ainsi que la règle $\overline{X \backslash Y} \Rightarrow Y$ si le motif $\overline{X \backslash Y}$ est minimal i.e. $(\overline{X \backslash Y} \in NRFM)$ et si la méta-règle 2 n'est pas vérifiée ($\neg(MR_2)$) (*ligne 11*). L'étude des règles $X \backslash Y \Rightarrow \overline{Y}$ et $\overline{X \backslash Y} \Rightarrow Y$ consiste à vérifier les *contraintes* Ct_3, Ct_4 et Ct_5. Une fois l'étude des deux types de règles réalisée (*soit lignes 7 et 8, soit lignes 10 et 11*), nous étudions la règle $X \ddot{\backslash} Y \Rightarrow \ddot{Y}$ (*ligne 13*) en vérifiant les *contraintes* Ct_4 et Ct_5.
Après avoir exposé l'algorithme d'extraction des RAPN, nous présentons les expérimentations qui ont été réalisées sur 5 bases de données.

4 Expérimentations

Les quatre algorithmes ont été développés en Java et incorporés au logiciel libre WEKA (*Waikato Environment for Knowledge Analysis*) (Witten et Frank, 2005). Les expérimentations ont été effectuées sur les 4 bases de données UCI KDD (Hettich et Bay, 1999) suivantes : *Abalone* (*4177 individus et 24 variables binaires*), *Ecoli* (*336 individus et 29 variables binaires*), *Iris* (*150 individus et 15 variables binaires*) et *Wages* (*534 individus et 32 variables binaires*). Tout d'abord, nous avons effectué une étude comparative des 4 algorithmes sur la base de données *Abalone* dont les résultats sont résumés dans la *figure 2* et où nous avons fait varier les valeurs du seuil minimum pour le support et la confiance comme indiqué dans les deux premières colonnes du tableau. Pour chacun des algorithmes, nous avons restitué le temps d'exécution total en secondes (*colonne Temps*) et le nombre total de règles négatives extraites (*colonne # Négatives*). Par manque de place, nous n'avons pas fait resortir le nombre de règles positives. De plus pour notre algorithme, nous avons restitué dans la dernière colonne (*colonne # Nouvelles R.*) le nombre de règles extraites du type $\ddot{X} \Rightarrow \ddot{Y}$, règles non présentes dans les 3 algorithmes existants. Pour finir, nous avons retenu comme seuil minimum pour le coefficient de corrélation nécessaire pour l'algorithme de (Antonie et Zaïane, 2004), la valeur 0,60 et la valeur 0,10 pour la mesure d'intérêt utilisée dans (Wu et al., 2004). La mesure de *Shortliffe*

utilisée dans (Wu et al., 2004) a le même seuil que celui de la confiance. Quant à notre algorithme, nous avons retenu les valeurs suivantes pour les différents seuils : $max_{sup} = 0,80$, $min_{s\ddot{u}p} = min_{sup}$ et $min_{M_G} = 0,60$.

		(Antonie et Zaïane, 2004)		(Cornelis et al., 2006)		(Wu et al.,2004)		Notre algorithme		
minsup	minconf	Temps	# Négatives	Temps	# Négatives	Temps	# Négatives	Temps	# Négatives	# Nouvelles R.
0,01	0,80	28,02	2 435	4,01	10 986	42,61	2 198	2,48	541	6 791
	0,90	28,10	1 780	3,46	10 750	42,84	2 290	2,33	174	4 356
	0,95	28,67	1 069	3,60	9 193	41,87	1 866	2,31	80	3 144
0,10	0,80	4,54	2 327	0,66	1 757	7,24	2 960	0,48	16	1 853
	0,90	4,54	1 672	0,69	1 779	7,21	2 683	0,46	11	1 135
	0,95	4,56	961	0,69	1 510	7,23	1 775	0,47	9	865

FIG. 2 – *Etude comparative des 4 algorithmes sur la base Abalone.*

Nous remarquons que c'est notre algorithme qui restitue, pour tous les cas de figure de cette base *Abalone*, le nombre le plus faible de règles négatives. Nous observons également que le nombre de règles du type $\ddot{X} \Rightarrow \ddot{Y}$ restituées par uniquement notre algorithme est conséquent mais reste globalement inférieur au nombre de règles restituées par (Cornelis et al., 2006). Quant au temps d'extraction, notre algorithme arrive en première place suivi de (Cornelis et al., 2006), (Antonie et Zaïane, 2004) et (Wu et al., 2004).

		Abalone			Ecoli			Iris			Wages		
minsup	minconf	Temps	# Négatif	# Nouv. R	Temps	# Négatif	# Nouv. R	Temps	# Négatif	# Nouv. R	Temps	# Négatif	# Nouv. R
0,01	0,80	2,48	541	6 791	0,50	375	708	0,11	78	65	38,90	1 217	12 260
	0,90	2,33	174	4 356	0,50	149	519	0,13	28	31	38 ,42	30	6 881
	0,95	2,31	80	3 144	0,49	32	449	0,11	6	8	38,85	3	5 411
0,10	0,80	0,48	16	1 853	0,09	0	27	0,07	0	4	0,10	1	24
	0,90	0,46	11	1 135	0,08	0	24	0,07	0	3	0,10	1	15
	0,95	0,47	9	865	0,08	0	18	0,07	0	3	0,10	1	15

FIG. 3 – *Résultats de notre algorithme sur les 4 bases de données.*

La deuxième étude réalisée s'est concentrée sur notre algorithme et nous avons souhaité connaître les temps d'extraction (*colonne Temps*) et le nombre de règles négatives (*colonne # Négatif*) extraites sur différentes bases de données UCI, et cela pour différents seuils pour le support et la confiance comme indiqué dans la *figure 3*. Les autres seuils nécessaires pour notre algorithme sont les mêmes que pour l'étude précédente. Nous constatons des temps d'extraction raisonnables et le nombre de règles négatives extraites est raisonnable en général sauf pour ce nouveau type de règles $\ddot{X} \Rightarrow \ddot{Y}$ où une étude complémentaire est nécessaire afin de ne retenir que les plus pertinentes.

5 Conclusion

Dans cet article, nous avons proposé un algorithme d'extraction de RAPN optimisé par rapport à ceux présents dans la littérature et reposant sur l'algorithme fondateur *Apriori*. Les deux optimisations ont porté sur une diminution du nombre de règles et sur un parcours optimisé de recherche des règles valides. La diminution du nombre de règles a été rendue possible en éliminant certains motifs fréquents qui ne pouvaient pas conduire à des règles intéressantes car ayant soit une valeur pour la confiance trop faible, soit un écart à l'indépendance trop faible.

C'est la recherche des motifs raisonnablement fréquents qui a permis cette diminution du nombre de règles et qui présente l'avantage d'intervenir tout au début du processus d'extraction. L'utilisation de la mesure M_G, plus sélective que les mesures utilisées par (Wu et al., 2004), a également permis d'éliminer un autre type de règles non pertinentes : les règles ayant un écart trop faible par rapport au point d'équilibre. Quant à la recherche optimisée des règles potentiellement valides, nous avons montré que seulement la moitié sont à prendre en considération et que parmi ces règles restantes, nous pouvions également les restreindre grâce non seulement à la propriété d'anti-monotonicité de la confiance, plus utilisée dans les algorithmes existants d'extraction de RAPN, mais également grâce à une méta-règle permettant d'inférer la non validité des règles $\overline{X} \Rightarrow \overline{Y}$ à partir de la non validité des règles $X \Rightarrow Y$ au regard de la mesure M_G. Les expérimentations ont mis en valeur l'intérêt de notre algorithme en terme de temps de calculs et de nombre de règles extraites malgré l'incorporation d'un nouveau type de règles intégré à notre algorithme. Ce dernier type de règles devra cependant faire l'objet d'une étude complémentaire afin d'éliminer celles qui ne sont pas pertinentes. Nous souhaitons poursuivre l'optimisation de notre algorithme en nous penchant sur le problème des règles redondantes, problème non abordé à notre connaissance par les techniques d'extraction de RAPN. Pour finir, nous aimerions étendre notre algorithme à la recherche des règles du type $X_1 \wedge X_2 \vee X_3 \Rightarrow Y_1 \wedge Y_2 \vee Y_3$, c'est-à-dire des règles ayant en prémisse et / ou en conclusion des conjonctions ou disjonctions d'items pouvant être positifs ou négatifs.

Remerciements. Nous remerçions Marion Carrier et Jérémy Blanc pour leur participation à la programmation de l'algorithme de (Wu et al., 2004).

Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th Very Large Data Bases Conference*, pp. 487–499.

Antonie, M.-L. et O. Zaïane (2004). Mining positive and negative association rules: an approach for confined rules. In *Proceedings on Principles and Practice of Knowledge Discovery in Databases*, pp. 27–38.

Blanchard, J., F. Guillet, et R. Briand, H. nd Gras (2005). Ipee : Indice probabiliste d'ecart à l'equilibre pour l'évaluation de la qualité des règles. In *Atelier Qualité des Données et des Connaissances*, pp. 26–34.

Boulicaut, J.-F., A. Bykowski, et B. Jeudy (2000). Towards the tractable discovery of association rules with negations. In *Proceedings of the Fourth International Conference on Flexible Query Answering Systems FQAS'00*, pp. 425–434.

Brin, S., R. Motwani, et C. Silverstein (1997). Beyond market baskets : Generalizing association rules to correlation. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, ACM, pp. 265–276.

Cornelis, C., P. Yan, X. Zhang, et G. Chen (2006). Mining positive and negative association rules from large databases. In *Proceedings of International Conference on Cybernetics and Intelligent Systems (CIS'06)*, IEEE, pp. 613–618.

Guillaume, S. (2010). Améliorations de la mesure d'intérêt m_{GK}. In *Actes des XVIIèmes rencontres de la Société Francophone de Classification*, pp. 41–45.

Guillaume, S. et P. Papon (2012). Méta-règles pour la génération de règles négatives. In RNTI (Ed.), *Actes de la 12ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC 2012)*, Volume E-23 of *Revue des Nouvelles Technologies de l'Information*, pp. 231–236. Hermann.

Hettich, S. et S. D. Bay (1999). The uci kdd archive [http://kdd.ics.uci.edu]. Irvine, CA: University of California, Department of Information and Computer Science.

Lavrac, N., P. Flach, et B. Zupan (1999). Rule evaluation measures: a unifying view. In *Ninth International Worshop on Inductive Logic Programming*, Volume 1634 of *RNTI*, pp. 174–185. Mineau, G. and Ganter, B.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. iii. regression, heredity and panmixia. In *Philosophical Transactions of the Royal Society*.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases 1991*, pp. 229–248. MIT Press.

Savasere, A., E. Omiecinski, et S. Navathe (1998). Mining for strong negative associations in a large database of customer transactions. In *Proceedings of the 14th International Conference on Data Engineering (ICDE'98)*, pp. 494–502. IEEE Computer Society.

Shortliffe, E. (1976). Computer-based medical consultations: Mycin. Elsevier computer science library/North-Holland, New York.

Teng, W.-G., M.-J. Hsieh, et M.-S. Chen (2002). On the mining of substitution rules for statistically dependent items. In *Second IEEE International Conference on Data Mining (ICDM'02)*, pp. 442–449. IEEE Computer Society.

Witten, I. et E. Frank (2005). In *Data Mining, practical machine learning tools and techniques with Java implementations*. Morgan Kauffman.

Wu, X., C. Zhang, et S. Zhang (2004). Efficient mining of both positive and negative association rules. *ACM Transactions on Information Systems (TOIS) 22*, 381–405.

Summary

The literature has been heavily involved in the extraction of classic rules and few in negative rules extraction owing essentially to the calculations cost and to the prohibitive number of extracted rules that are for the most part redundant and uninteresting. In this paper, we take an interest in algorithms that mine PNAR (*Positive and Negative Association Rules*) based on the famous Apriori algorithm. We conducted a study of these algorithms and highlight the strengths and weaknesses of each. At the end of this study, we propose a new algorithm that improve the mining relative to the number and the quality of the extracted rules and also relative to search path of rules. The study concludes by evaluating this algorithm on several databases.