Étude comparative d'extraction de règles d'association positives et négatives et optimisations

Sylvie Guillaume et Pierre-Antoine Papon

Clermont Université, Université d'Auvergne, LIMOS
BP 10448, F-63000 Clermont-Ferrand
guillaum@isima.fr,
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 (*recherche de motifs raisonnablement fréquents et utilisation d'une mesure d'intérêt supplémentaire*) et au niveau du parcours de recherche des règles (*étude de la moitié des règles négatives potentiellement valides et proposition de règles d'élagage*). L'étude s'est terminée par une évaluation de cet algorithme sur deux bases de données.

1 Introduction

L'extraction de règles d'association (Agrawal et Srikant, 1994), consistant à découvrir des corrélations entre les attributs (*ou variables*) d'une base de données, est une tâche importante en fouille de données. Une règle d'association est une implication de la forme $X \Rightarrow Y$, où X (*prémisse*) et Y (*conclusion*) sont deux ensembles $X = \{x_1, .., x_i, .., x_p\}$ et $Y = \{y_1, .., y_j, .., y_q\}$ disjoints d'items ($X \cap Y = \varnothing$). Un item x_i ou y_j avec $(i \in \{1, .., p\})$ et $(j \in \{1, .., q\})$ est une variable binaire de la base de données et nous parlons de motif lorsque nous sommes en présence d'un ensemble d'items ; $X = \{x_1, .., x_i, .., x_p\}$ et $Y = \{y_1, .., y_j, .., y_q\}$ sont donc deux motifs. La règle $X \Rightarrow Y$ signifie que les individus qui vérifient tous les items (*ou caractéristiques*) x_i $(i \in \{1, .., p\})$ de la prémisse X vérifient également en général tous les items y_j $(j \in \{1, .., q\})$ de la conclusion Y. Par exemple, $\{crêpes, beurre\} \Rightarrow \{cidre\}$ est une règle d'association révélant que lorsqu'un consommateur achète à la fois des *crêpes* et du *beurre*, alors il achète également en général du *cidre*. Pour simplifier les notations sans nuire à la compréhension du lecteur, nous noterons cette règle $crêpes, beurre \Rightarrow cidre$. Afin de quantifier l'intérêt d'une règle $X \Rightarrow Y$, on utilise en

général deux mesures d'intérêt objectives : le support et la confiance dont nous rappelons la définition.

Définition 1. Le ***support*** de la règle $X \Rightarrow Y$ est le support du motif $X \cup Y = \{x_1, .., x_i, .., x_p,\ y_1, .., y_j, .., y_q\}$ c'est-à-dire le pourcentage d'individus qui vérifient à la fois tous les items de X et tous les items de Y, ou encore la probabilité d'apparition de $X \cup Y$:

$$sup(X \Rightarrow Y) = sup(X \cup Y) = P(X \cup Y)$$

Le support permet d'évaluer la portée de la règle en révélant la proportion d'individus de la base de données concernée par cette règle.

Définition 2. La ***confiance*** de la règle $X \Rightarrow Y$ est le pourcentage d'individus qui vérifient tous les items de Y parmi ceux qui vérifient tous les items de X. C'est également la probabilité conditionnelle de Y sachant X et par conséquent, c'est le rapport du support de $X \cup Y$ sur le support de X :

$$conf(X \Rightarrow Y) = P(Y/X) = \frac{sup(X \cup Y)}{sup(X)}$$

La confiance permet d'évaluer la force de la règle.

Ainsi la règle $crêpes, beurre \Rightarrow cidre$ avec un support de $0,01$ et une confiance de $0,85$ signifie que 1% des individus de la base de données ont acheté des *crêpes*, du *beurre* et du *cidre* ; et que parmi les individus qui ont acheté des *crêpes* et du *beurre*, 85% d'entre eux ont également acheté du *cidre*.

Afin de retenir les règles les plus intéressantes, l'utilisateur fixe deux seuils minimaux min_{sup} et min_{conf} pour respectivement le support et la confiance. Les règles vérifiant ces deux contraintes $(Ct_1) : sup(X \Rightarrow Y) \geqslant min_{sup}$ et $(Ct_2) : conf(X \Rightarrow Y) \geqslant min_{conf}$ seront celles qui seront retenues et restituées par l'algorithme d'extraction. Classiquement, les règles vérifiant ces deux contraintes sont appelées règles valides. Nous donnons ci-dessous une définition plus large d'une règle valide, et également la définition d'une règle valide pour une mesure donnée m.

Définition 3. On entend par règle ***valide***, une règle qui vérifie un ensemble de contraintes (Ct_i). Dans *Apriori* (Agrawal et Srikant, 1994), ces contraintes sont les suivantes :
$(Ct_1) : sup(X \Rightarrow Y) \geqslant min_{sup}$ et $(Ct_2) : conf(X \Rightarrow Y) \geqslant min_{conf}$.

Définition 4. Une règle est dite ***valide pour une mesure*** m si la valeur de cette mesure m pour la règle $X \Rightarrow Y$ vérifie une contrainte (Ct_m) du type "*appartenir à un ensemble V de valeurs*" : $m(X \Rightarrow Y) \in V$

Ainsi si les valeurs des deux seuils retenus par l'utilisateur sont égales à $min_{sup} = 0,05$ et $min_{conf} = 0,80$, la règle $crêpes, beurre \Rightarrow cidre$ n'est pas valide car elle ne vérifie pas la contrainte (Ct_1) pour le support.

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. Ces algorithmes ont essentiellement deux limites : 1) les variables doivent être binaires, 2) seules les règles positives $X \Rightarrow Y$ sont extraites alors que les règles négatives $X \Rightarrow \overline{Y}$, $\overline{X} \Rightarrow Y$ et $\overline{X} \Rightarrow \overline{Y}$ sont ignorées. $\overline{X}$ et $\overline{Y}$ désignent respectivement les négations des motifs X et Y. Ainsi le motif $\overline{X}$ indique l'absence d'au moins un item x_i ($i \in \{1, .., p\}$) composant le motif X (*plus exactement non* x_1 *ou .. ou non* x_i *ou .. ou non* x_p). Par exemple, la règle $crêpes, cidre \Rightarrow \overline{beurre\ doux}, \overline{vin\ rouge}$ révèle que les consommateurs qui achètent à la fois des *crêpes* et du *cidre* ne remplissent pas leur caddie avec du *beurre doux* et / ou du *vin rouge*. 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 par exemple en médecine, la connaissance de caractéristiques chez les patients empêchant une maladie de se déclarer peut être capital ; correspondant à des règles du type $X \Rightarrow \overline{maladie}$ et par conséquent aux règles $X \Rightarrow \overline{Y}$. De la même façon les règles $\overline{X} \Rightarrow Y$ et $\overline{X} \Rightarrow \overline{Y}$ peuvent être très précieuses également en médecine. Par exemple, la règle $\overline{fer} \Rightarrow anémie$ nous indique que la carence en fer est une des causes de l'anémie. Les règles $hepcidine \Rightarrow anémie$ et $\overline{hepcidine} \Rightarrow \overline{anémie}$ sont aussi riches d'enseignement. Les deux règles impliquant l'hormone *hepcidine* dans l'apparition ou la disparition de l'anémie montre son lien étroit avec cette maladie. Les chercheurs ont ainsi découvert que cette hormone régulait l'absorption de fer. Le fer nécessaire à notre organisme provient de l'alimentation et pénètre au niveau intestinal. Cependant il existe un phénomène de régulation : lorsque nous manquons de fer, il est absorbé par la paroi intestinale ; lorsque nous en avons trop, il ne peut plus franchir la barrière intestinale. Or cette hormone libérée par le foie agit directement sur la paroi intestinale en bloquant l'entrée du fer : en son absence, le fer entre librement dans l'intestin d'où la règle $\overline{hepcidine} \Rightarrow \overline{anémie}$; et en sa présence excessive, le fer est arrêté d'où la seconde règle $hepcidine \Rightarrow anémie$.

L'extraction de telles règles est un défi car l'absence de variables binaires pour un individu dans une base de données est en général plus importante que la présence de ces mêmes variables. Ainsi dans les bases de données de la grande distribution, il y a des milliers d'articles et les consommateurs n'achètent qu'une petite partie de ces articles. L'approche naïve consistant à ajouter les négations des variables dans la base de données et à utiliser un algorithme classique d'extraction de règles positives est irréalisable. De plus, beaucoup de règles redondantes et inintéressantes sont extraites. Plusieurs techniques ont été proposées.

(Brin et al., 1997) utilisent le test du χ^2 (Pearson, 1900) 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 positifs avec la connaissance du domaine, connaissance présentée sous forme de taxonomie, afin de détecter les dissociations (*ou associations négatives*).
Nous rappelons la définition d'un motif fréquent.

Définition 5. Un motif X est dit ***fréquent*** si son support est supérieur à un seuil min_{sup} fixé par l'utilisateur : $sup(X) \geqslant min_{sup}$.

Dans un souci de simplification d'écriture et sans nuire à la compréhension du lecteur, nous noterons par la suite $X \cup Y$ par XY. Cette dernière approche est difficile à généraliser puisqu'elle dépend de la connaissance du domaine et nécessite d'avoir au préalable une taxo-

nomie. Une approche similaire se rencontre dans (Yuan et al., 2002) où les auteurs trouvent un sous-ensemble de règles négatives valides. (Boulicaut et al., 2000) recherchent les règles négatives du type $XY \Rightarrow \overline{Z}$ ou $\overline{X}Y \Rightarrow Z$ et pour cela, ils proposent une approche basée sur les contraintes. (Teng et al., 2002) proposent un algorithme détectant les règles négatives du type $X \Rightarrow \overline{Y}$. Cet algorithme recherche tout d'abord les items concrets c'est-à-dire les items dont la valeur du χ^2 est élevée et dont le support est supérieur à la valeur attendue. Une fois ces items découverts, ils calculent le coefficient de corrélation pour chaque paire d'items concrets. Enfin, à partir de ces paires d'items corrélées négativement, ils vont extraire les règles du type $X \Rightarrow \overline{Y}$. L'inconvénient de cette technique est qu'elle est limitée à un seul type de règles négatives. Dans (Missaoui et al., 2008), les auteurs proposent la génération de règles négatives à partir de règles positives mais dans un contexte bien particulier : celui des implications logiques (*i.e. règles ayant une confiance égale à 1*). Quant à (Wu et al., 2004), (Antonie et Zaïane, 2004) et (Cornelis et al., 2006), ils extraient des règles négatives grâce à un algorithme basé sur l'algorithme fondateur *Apriori* (Agrawal et al., 1993). (Wu et al., 2004) utilisent en plus du couple de mesures (*support*, *confiance*), les deux mesures suivantes : une mesure d'intérêt qui est 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 le *facteur de certitude* (Heckerman et Shortliffe, 1992) définie antérieurement. 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 faisons deux propositions :

- **(1)** retenir un sous-ensemble de motifs fréquents, *les motifs raisonnablement fréquents*, en éliminant ceux qui vont conduire à des règles non pertinentes. 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.
- **(2)** utiliser une mesure supplémentaire au couple (*support*, *confiance*) pour sélectionner les règles valides : la mesure M_G (Guillaume, 2010) qui est une amélioration du *facteur de certitude* (Heckerman et Shortliffe, 1992) utilisée par (Wu et al., 2004). Cette mesure plus sélective 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 sont à étudier. De plus, parmi ce sous-ensemble de règles à étudier, nous avons utilisé la propriété d'anti-monotonicité de la confiance, propriété abandonnée par (Antonie et Zaïane, 2004) et (Wu et al., 2004). Pour finir, nous avons ajouté deux nouvelles propriétés d'élagage, propriétés qui ont été dégagées par (Guillaume et Papon, 2012) et qui reposent sur la nouvelle mesure que nous allons utiliser, la mesure M_G.

L'article s'organise donc de la façon suivante. La *section 2* développe les trois propositions majeures existantes dans la littérature et reposant sur l'algorithme *Apriori* en mettant en évidence leurs avantages et leurs inconvénients. La *section 3* présente et motive les choix retenus pour optimiser l'extraction des règles d'association positives et négatives. La *section 4* développe l'algorithme proposé et la *section 5* évalue notre technique sur deux bases de données, et ceci par rapport aux trois techniques existantes développées dans la *section 2*. L'article se termine par une conclusion et des perspectives.

2 Techniques existantes reposant sur l'algorithme *Apriori*

Dans cette section, nous exposons trois techniques majeures existantes et reposant sur l'algorithme *Apriori* : (1) l'algorithme de (Antonie et Zaïane, 2004), (2) l'algorithme de (Wu et al., 2004) et (3) l'algorithme de (Cornelis et al., 2006).

2.1 Algorithme de (Antonie et Zaïane, 2004)

Dans (Antonie et Zaïane, 2004), les auteurs génèrent à la fois les règles positives et les règles négatives. Comme cet algorithme repose sur l'algorithme *Apriori* (Agrawal et Srikant, 1994), ils commencent par rechercher les motifs candidats (*i.e. les différentes combinaisons entre les items de la base :* $\{x_1, x_2\}$*,* $\{x_1, x_3\}$*, ..,* $\{x_1, x_2, x_3\}$ *...*), puis à partir des motifs candidats vont générer les différentes règles. Cette recherche des règles s'opère à partir des motifs candidats et non pas à partir des motifs fréquents comme dans l'algorithme fondateur *Apriori*. Pour un niveau k donné, l'ensemble C_k des motifs candidats C est généré grâce au produit cartésien entre l'ensemble des fréquents F_{k-1} de niveau inférieur $(k-1)$ et l'ensemble des *1*-motifs fréquents F_1.
On dit que X est un k-motif, avec k un entier supérieur ou égal à 1, si celui-ci se compose de k items : $X = \{x_1, .., x_i, .., x_k\}$.
Pour extraire les différentes règles (*positives et négatives*) valides à partir des motifs candidats C, ils utilisent tout d'abord le coefficient de corrélation ρ (Pearson, 1896) afin de déterminer les règles à étudier. Ainsi, si le coefficient de corrélation $\rho(X, Y)$ entre les motifs X et Y pour le candidat $C = XY$ est supérieur à un seuil minimum min_ρ défini par l'utilisateur, alors les auteurs vont s'intéresser aux deux règles $X \Rightarrow Y$ et $\overline{X} \Rightarrow \overline{Y}$. Les règles valides sont celles dont le support et la confiance sont supérieures à deux seuils fixés par l'utilisateur : min_{sup} et min_{conf}, contraintes venant s'ajouter à $\rho(X, Y) \geqslant min_\rho$. Si au contraire ce coefficient de corrélation est inférieur à $-\rho_{min}$, les auteurs s'intéressent aux deux règles $\overline{X} \Rightarrow Y$ et $X \Rightarrow \overline{Y}$ et les règles valides sont cette fois-ci celles dont la confiance est supérieure à min_{conf} (*avec la contrainte initiale* $\rho(X, Y) \leqslant -min_\rho$). Dans ce deuxième cas de figure, les auteurs ne vérifient pas la contrainte du support pour ces deux types de règles. De plus, pour tous les candidats C et pour tous les arrangements (X, Y) telles que $C = XY$, ils commencent par calculer le coefficient de corrélation $\rho(X, Y)$ qui est le même que pour la combinaison (Y, X). Les auteurs effectuent donc deux fois le même calcul (*puisque* $\rho(X, Y) = \rho(Y, X)$) et il serait plus judicieux de considérer les règles symétriques lors de l'examen de la combinaison (X, Y). Ainsi, si $\rho(X, Y) \geqslant min_\rho$, les 4 règles $X \Rightarrow Y$, $Y \Rightarrow X$, $\overline{X} \Rightarrow \overline{Y}$ et $\overline{Y} \Rightarrow \overline{X}$ sont à étudier ; et si $\rho(X, Y) \leqslant -min_\rho$, ce sont les 4 autres types de règles $\overline{X} \Rightarrow Y$, $\overline{Y} \Rightarrow X$, $X \Rightarrow \overline{Y}$ et $Y \Rightarrow \overline{X}$ qui sont à examiner. Pour finir, les auteurs n'utilisent pas la propriété d'*anti-monotonicité de*

la confiance dans la phase de recherche des règles comme dans l'algorithme *Apriori* (Agrawal et Srikant, 1994) puisque celle-ci n'est valable que pour les règles positives. Nous rappelons cette propriété.
Propriété. La confiance est ***anti-monotone*** puisque
$\forall X, Y, Z$ tel que $Y \subseteq Z$ si $conf(X \Rightarrow Z) \geqslant min_{conf}$ alors $conf(X \Rightarrow Y) \geqslant min_{conf}$.

Nous étudions maintenant une deuxième proposition : celle de (Wu et al., 2004).

2.2 Algorithme de Wu et al. (2004)

Dans (Wu et al., 2004), les auteurs proposent un algorithme qui génère tout d'abord les règles positives à partir des motifs fréquents et ensuite les règles négatives à partir des motifs restants c'est-à-dire les motifs non fréquents. Pour la recherche des règles positives, celle-ci s'effectue donc à partir des motifs fréquents et grâce à trois mesures : la *confiance*, une mesure d'intérêt qui n'est autre que la valeur absolue de la *nouveauté* (Lavrac et al., 1999) et une troisième mesure qu'ils nomment *ratio incrément de la probabilité conditionnelle* qui n'est autre que le *facteur de certitude* FC (Heckerman et Shortliffe, 1992) défini antérieurement. La mesure d'intérêt évalue si la valeur absolue de la différence entre le support observé $sup(XY)$ et le support attendu $sup(X) \times sup(Y)$ pour le motif XY est supérieure à une valeur fixée par l'utilisateur $min_{interêt}$. Le facteur de certitude détermine la distance de la règle $X \Rightarrow Y$ entre l'indépendance et l'implication logique lorsque la réalisation de X augmente les chances d'apparition de Y ; ou la distance entre l'indépendance et l'incompatibilité lorsque la réalisation de X diminue les chances d'apparition de Y. Deux nouveaux seuils minimaux doivent donc être fixés par l'utilisateur. Afin de bien comprendre la sémantique du facteur de certitude, nous rappelons les définitions de l'implication logique et de l'indépendance.

Définition 6. L'***implication logique*** est le cas où la confiance de la règle est égale à 1.

Définition 7. L'***indépendance*** est le cas où la réalisation de X n'influence pas la réalisation de Y autrement dit, c'est le cas où la confiance de la règle est égale au support de la conclusion i.e. $conf(X \Rightarrow Y) = sup(Y)$.

Ainsi pour un couple (X, Y) de motifs fréquents, les auteurs commencent par vérifier les contraintes du *support*, de la *confiance* et de l'*intérêt*. Si ceux-ci sont vérifiés, ils étudient les deux règles $X \Rightarrow Y$ et $Y \Rightarrow X$ et conservent celles qui vérifient la contrainte pour la troisième mesure qui est le facteur de certitude. Cette procédure pose quelques problèmes. Tout d'abord, ils vérifient une seconde fois la contrainte du support, ce qui est inutile car la recherche des règles positives s'effectue à partir des motifs fréquents. De plus, la confiance n'est pas une mesure symétrique puisque $conf(X \Rightarrow Y) \neq conf(Y \Rightarrow X)$ et par conséquent la contrainte de la confiance de la règle $Y \Rightarrow X$ n'est pas vérifiée. Pour finir, les auteurs n'ont pas utilisé la propriété d'anti-monotonicité de la confiance, propriété pouvant être utilisée ici puisque nous sommes en présence de règles positives uniquement.

Pour la recherche des règles négatives, celle-ci s'effectue à partir des motifs non fréquents. Ainsi, pour tous les motifs non fréquents C et pour toutes les combinaisons (X, Y) telle que $C = X \cup Y$ et $X \cap Y = \varnothing$, ils commencent par vérifier les contraintes du *support*, de la *confiance* et de l'*intérêt* pour le couple de motifs $(X, \overline{Y})$ et les contraintes du *support*

pour les motifs X et Y. Si ces contraintes sont vérifiées, ils étudient les 6 règles suivantes : $\overline{X} \Rightarrow Y$, $Y \Rightarrow \overline{X}$, $X \Rightarrow \overline{Y}$, $\overline{Y} \Rightarrow X$, $\overline{X} \Rightarrow \overline{Y}$ et $\overline{Y} \Rightarrow \overline{X}$ et conservent celles vérifiant la contrainte du facteur de certitude. Cette extraction pose également des problèmes. Tout d'abord les contraintes du *support* et de la *confiance* ne sont vérifiées que pour la règle $X \Rightarrow \overline{Y}$. Ensuite, (Guillaume et Papon, 2012) ont démontré que si les règles $\overline{X} \Rightarrow \overline{Y}$ et $\overline{Y} \Rightarrow \overline{X}$ sont potentiellement intéressantes, alors les règles $X \Rightarrow \overline{Y}$, $\overline{Y} \Rightarrow X$, $\overline{X} \Rightarrow Y$ et $Y \Rightarrow \overline{X}$ ne peuvent pas l'être. (Guillaume et Papon, 2012) entendent par règles potentiellement intéressantes les règles dont la réalisation de la prémisse augmente les chances d'apparition de la conclusion, ce qui est également vérifié par les auteurs grâce à la valeur absolue de la *nouveauté* (Lavrac et al., 1999). Ainsi, les auteurs étudient dans tous les cas au moins deux règles négatives qui ne pourront pas être considérées comme valides, erreur non présente dans (Antonie et Zaïane, 2004). Cependant, et contrairement à (Antonie et Zaïane, 2004), ils évaluent en même temps les règles symétriques puisque la valeur absolue de la *nouveauté* (*comme le coefficient de corrélation utilisé par (Antonie et Zaïane, 2004)*) est un indice symétrique [1] mais ils oublient que la *confiance* n'est pas un indice symétrique et certaines règles non valides vont être extraites puisque la contrainte de la *confiance* n'aura pas été vérifiée. Pour finir, (Wu et al., 2004) imposent que les supports de X et Y soient fréquents pour que les diverses règles négatives soient valides. Ainsi, la règle $\overline{X} \Rightarrow \overline{Y}$ est valide si les conditions suivantes sont réalisées : $sup(XY) \leqslant min_{sup}$, $sup(X\ \overline{Y}) \geqslant min_{sup}$, $conf(X \Rightarrow \overline{Y}) \geqslant min_{conf}$, $|sup(X\ \overline{Y}) - sup(X) \times sup(\overline{Y})| \geqslant min_{interêt}$, $FC(\overline{X} \Rightarrow \overline{Y}) \geqslant min_{FC}$, $sup(X) \geqslant min_{sup}$ et $sup(Y) \geqslant min_{sup}$, ce qui n'est pas logique. Comme nous souhaitons comparer cet algorithme avec celui que nous proposons dans cet article, nous avons corrigé l'erreur des auteurs, à savoir la règle $\overline{X} \Rightarrow \overline{Y}$ est valide si les conditions suivantes sont réalisées : $sup(XY) \leqslant min_{sup}$, $sup(\overline{X}\ \overline{Y}) \geqslant min_{sup}$, $conf(\overline{X} \Rightarrow \overline{Y}) \geqslant min_{conf}$, $|sup(\overline{X}\ \overline{Y}) - sup(\overline{X}) \times sup(\overline{Y})| \geqslant min_{interêt}$, $FC(\overline{X} \Rightarrow \overline{Y}) \geqslant min_{FC}$, $sup(X) \geqslant min_{sup}$ et $sup(Y) \geqslant min_{sup}$.
Nous passons maintenant à l'étude de la troisième technique : celle de (Cornelis et al., 2006).

2.3 Algorithme de Cornelis et al. (2006)

Dans (Cornelis et al., 2006), les auteurs recherchent également les règles positives et négatives grâce au couple de mesures (*support*, *confiance*), sans avoir recours à d'autres mesures comme l'ont fait (Antonie et Zaïane, 2004) avec le *coefficient de corrélation* et (Wu et al., 2004) avec les deux mesures suivantes : la valeur absolue de la *nouveauté* et le *facteur de certitude*.

L'algorithme se déroule en cinq étapes et qui sont les suivantes :
- **Étape (1)** : Recherche de tous les motifs fréquents positifs X.
- **Étape (2)** : Recherche de toutes les négations fréquentes $\overline{X}$.
- **Étape (3)** : Recherche de tous les motifs fréquents du type $\overline{X}\ \overline{Y}$.
- **Étape (4)** : Recherche de tous les motifs fréquents du type $\overline{X}\ Y$.
- **Étape (5)** : Recherche des règles valides (*positives et négatives*) à partir des motifs fréquents trouvés dans les étapes (1), (3) et (4).

1. Un indice m est symétrique s'il évalue de la même manière les règles $X \Rightarrow Y$ et $Y \Rightarrow X$, autrement dit $\forall (X, Y)\ m(X \Rightarrow Y) = m(Y \Rightarrow X)$.

L'**étape (1)** est l'étape de recherche des motifs fréquents, la même étape que celle présente dans (Agrawal et Srikant, 1994).
Quant à l'**étape (2)** de recherche des négations fréquentes à partir des motifs fréquents trouvés à l'**étape (1)**, elle est immédiate grâce à la formule suivante : $sup(\overline{X}) = 1 - sup(X)$, où le support $sup(\overline{X})$ de $\overline{X}$ doit être supérieur au seuil minimal min_{sup} défini par l'utilisateur. Il est à noter que nous avons également $sup(X) \geqslant min_{sup}$ puisque cette recherche des négations fréquentes s'effectue à partir des motifs fréquents. En conséquence, les auteurs recherchent les négations $\overline{X}$ vérifiant $min_{sup} \leqslant sup(X) \leqslant (1 - min_{sup})$.
L'**étape (3)** recherche tous les motifs fréquents négatifs du type $\overline{X}\ \overline{Y}$ où $\overline{X}$ et $\overline{Y}$ sont des motifs fréquents négatifs minimaux. Un motif fréquent négatif minimal $\overline{X}$ est une négation de motif qui est fréquente (*i.e.* $sup(\overline{X}) \geqslant min_{sup}$) et où il n'existe pas un sous-ensemble $X' \subsetneq X$ tel que $\overline{X'}$ soit également fréquent (*autrement dit* $\nexists X' \subsetneq X\ /\ sup(\overline{X'}) \geqslant min_{sup}$). L'**étape (3)** se déroule de la manière suivante : les auteurs génèrent les candidats puis recherchent les fréquents parmi ces candidats, procédure classique pour les algorithmes issus de *Apriori* (Agrawal et al., 1993). Ici ce sont les motifs $\overline{X}\ \overline{Y}$ de taille (*ou niveau*) k ***non*** fréquents qui servent à générer les candidats (*c'est-à-dire les conjonctions de 2 motifs négatifs minimaux*) de niveau supérieur $(k+1)$ et non pas les motifs $\overline{X}\ \overline{Y}$ fréquents comme dans *Apriori*. L'**étape (3)** s'arrête lorsqu'aucun candidat ne peut plus être généré. La génération des candidats s'effectue de la façon suivante. Tout d'abord, les *2*-motifs candidats $\overline{i}\ \overline{j}$ sont générés à partir des *1*-motifs fréquents i et j trouvés à l'**étape (1)** de l'algorithme. Pour les niveaux supérieurs $(k > 2)$, la génération des candidats $\overline{X}\ \overline{Y}$ s'effectue en ajoutant à l'un des deux motifs négatifs (*soit à* $\overline{X}$, *soit à* $\overline{Y}$) un item i non présent dans les motifs X et Y (*c'est-à-dire* $i \notin \{X \cup Y\}$). Pour que ce candidat $\overline{X\{i\}}\ \overline{Y}$ ou $\overline{X}\ \overline{Y\{i\}}$ soit conservé, il faut tout d'abord que le motif $X\{i\}$ ou $Y\{i\}$) ainsi généré soit un motif fréquent (*i.e.* $sup(X\{i\}) \geqslant min_{sup}$ *ou* $sup(Y\{i\}) \geqslant min_{sup}$) et également que le candidat soit minimal (*i.e.* $\nexists X' \subsetneq X$, $\nexists Y' \subsetneq Y\ /\ \overline{X'}\ \overline{Y'}$ *soit déjà un motif fréquent*).
L'**étape (4)** de l'algorithme recherche tous les motifs fréquents du type $\overline{X}\ Y$. Pour cela, ils recherchent les fréquents parmi les candidats. La génération des candidats s'effectue tout d'abord en augmentant la taille de X puis ensuite en augmentant celle de Y. La génération des candidats par augmentation de la taille de Y s'effectue de la même façon que pour l'algorithme *Apriori* (Agrawal et Srikant, 1994) puisque le motif Y est un motif positif. Ainsi, ils commencent par générer les candidats potentiels $\overline{X}\ Y$ par jointure entre les motifs $\overline{X}\ Y_1$ et $\overline{X}\ Y_2$ si $taille(Y_1) = taille(Y_2)$ et si les $(taille(Y_1)-1)$ premiers items de Y_1 et Y_2 sont identiques. La jointure donne lieu aux candidats $\overline{X}\ Y_1\ \{i\}$ où i est le dernier item de Y_2. Ensuite ils vérifient que ce sont des candidats minimaux en opérant les vérifications suivantes : (1) s'il n'existe pas un sous-ensemble X' de X tel que $\overline{X'}\ Y$ soit un motif fréquent ; (2) s'il n'existe pas un sous-ensemble Y' de Y tel que $\overline{X}\ Y'$ soit non fréquent. Quant à la génération des motifs par augmentation de la taille de $\overline{X}$, ils vérifient que le motif candidat généré $\overline{X\{i\}}\ Y$ soit minimal.
L'**étape (5)** génère toutes les règles valides à partir des motifs fréquents c'est-à-dire à partir des motifs fréquents positifs obtenus à l'**étape (1)**, des motifs fréquents $\overline{X}\ \overline{Y}$ obtenus à l'**étape (3)** et des motifs fréquents $\overline{X}\ Y$ obtenus à l'**étape (4)**. La recherche des règles positives valides à partir des motifs fréquents positifs obtenus à l'**étape (1)** est la même que celle d'*Apriori* (Agrawal et Srikant, 1994). La recherche des règles négatives du type $\overline{X} \Rightarrow \overline{Y}$ à partir des motifs fréquents $\overline{X}\ \overline{Y}$ s'effectue simplement en vérifiant que la confiance de la règle est supérieure au seuil minimum min_{conf} fixé par l'utilisateur. Pour finir, la recherche des règles négatives

à partir du motif $\overline{X}\ Y$ s'effectue de la manière suivante : en génèrant les règles $\overline{X} \Rightarrow Y$ et $X \Rightarrow \overline{Y}$ et en retenant également celles qui ont des valeurs pour la confiance supérieures au seuil minimum. Or, les auteurs ne se sont pas assurés que les motifs $X\ \overline{Y}$ soient fréquents et minimaux, ce qui va produire obligatoirement des règles non valides.
L'**étape (2)** de l'algorithme pose la question de son intérêt car les auteurs recherchent toutes les négations fréquentes de motifs fréquents et pas seulement les motifs négatifs fréquents minimaux, les seuls utilisés ensuite pour les **étapes (3)** et **(4)** de l'algorithme. Cette étape pourrait ainsi être modifiée en recherchant ces motifs négatifs fréquents minimaux afin de simplifier l'**étape (3)** par une combinaison de tous les minimaux négatifs fréquents et cela éviterait également de vérifier que le motif associé positif est fréquent. De plus, en procédant de la manière indiquée précédemment (*i.e. combinaison des minimaux négatifs fréquents*), nous éliminerions les redondances et tests inutiles car à partir d'un candidat $\overline{X}\ \overline{Y}$ non fréquent, les auteurs génèrent tous les candidats possibles (*production de redondances*) et vérifient ensuite que le nouveau motif ainsi constitué est minimal (*test inutile avec la procédure préconisée*). Ainsi, des candidats identiques vont être générés comme par exemple, à partir du candidat non fréquent $\overline{a}\ \overline{b}$, on peut générer le candidat $\overline{a}\ \overline{bc}$ en ajoutant l'item c, et à partir du candidat non fréquent $\overline{a}\ \overline{c}$, on peut également générer le candidat $\overline{a}\ \overline{bc}$ en ajoutant l'item b. Deux tests vont ensuite avoir lieu pour vérifier que le motif $\overline{bc}$ est minimal.

Après avoir exposé trois techniques majeures existantes dans la littérature en mettant en évidence leurs avantages et leurs inconvénients, nous présentons et justifions les choix retenus pour l'algorithme que nous proposons.

3 Optimisations de l'extraction des RAPN

Dans cette section, nous exposons les optimisations apportées aux techniques existantes reposant sur *Apriori* et qui sont : (1) la réduction du nombre de règles par élimination de celles qui sont inintéressantes et (2) l'optimisation du parcours de recherche des règles. Nous commençons par présenter un moyen de réduire le nombre de règles grâce à l'extraction de motifs raisonnablement fréquents.

3.1 Réduction du nombre de règles

3.1.1 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} (*ce qui correspond à la définition d'un motif fréquent*) mais également inférieur à un seuil maximal que nous nommerons max_{sup}. Ce nouveau seuil max_{sup} sera utilisé pour tous les types de motifs M, à savoir les motifs positifs $M = XY$, les motifs négatifs du type $M = \overline{X}\ \overline{Y}$ et les motifs mixtes $M = \overline{X}\ Y$.

Définition 8. Soient deux seuils min_{sup} et max_{sup} définis par l'utilisateur. Un motif M positif (XY), négatif ($\overline{X}\ \overline{Y}$) ou mixte ($\overline{X}\ Y$) sera dit ***raisonnablement fréquent*** si :

$$min_{sup} \leqslant sup(M) \leqslant max_{sup}$$

La valeur par défaut de max_{sup} est égale à $1 - min_{sup}$ mais l'utilisateur peut définir un autre seuil.

Cette proposition se justifie par le fait qu'un motif omniprésent [2] M_1 positif ou négatif (*c'est-à-dire* $M_1 \in \{X, \overline{X}\}$) est combiné avec presque tous les autres motifs fréquents M_2 positifs ou négatifs ($M_2 \in \{Y, \overline{Y}\}$) car $sup(M_1M_2) \approx sup(M_2)$, et ceci sans pour autant révéler une combinaison $M = M_1M_2$ pertinente. L'exemple bancaire suivant illustre cette situation : en général le produit financier "*compte bancaire*" est associé avec tous les autres produits financiers proposés par la banque sans pour autant révéler une combinaison pertinente puisque la plupart des gens dans une agence bancaire commence par ouvrir un compte courant.

Nous démontrons dans ce qui suit que les règles $M_1 \Rightarrow M_2$ et $M_2 \Rightarrow M_1$ ne sont pas intéressantes lorsque le motif M_1 est omniprésent. Dans le cas de la règle $M_1 \Rightarrow M_2$, celle-ci n'est pas intéressante car elle vérifie rarement le seuil minimum pour la confiance et pour la règle $M_2 \Rightarrow M_1$, elle n'est pas intéressante car elle est trop proche du cas de l'indépendance.

Nous prouvons dans ce qui suit la nécessité de supprimer les motifs M_1 omniprésents.

Démonstration 1. $M_1 \Rightarrow M_2$ **non valide pour la confiance**
$conf(M_1 \Rightarrow M_2) = \frac{sup(M_1M_2)}{sup(M_1)} \approx \frac{sup(M_2)}{sup(M_1)} \ll 1$ puisque le support de M_1 a une valeur élevée.

Démonstration 2. $M_2 \Rightarrow M_1$ **non valide pour toute mesure d'écart à l'indépendance**
Beaucoup de règles du type $M_2 \Rightarrow M_1$ vont être extraites par toute approche reposant sur le couple (*support*, *confiance*) puisque la confiance de cette règle est proche de 1 : $conf(M_2 \Rightarrow M_1) = \frac{sup(M_1M_2)}{sup(M_2)} \approx \frac{sup(M_2)}{sup(M_2)} \approx 1$. Cependant ces règles ne sont pas intéressantes car trop proche du cas de l'indépendance. Pour le prouver, nous utilisons la *nouveauté* (Lavrac et al., 1999) qui est une mesure qui évalue l'écart de la règle par rapport à l'indépendance. Plus la valeur de la *nouveauté* est éloignée de la valeur 0, plus la règle sera jugée intéressante. Or la valeur de la *nouveauté* pour cette règle est proche de 0 comme le prouve les égalités suivantes :

$$\begin{aligned} nouveauté(M_2 \Rightarrow M_1) &= sup(M_1M_2) - sup(M_1) \times sup(M_2) \\ &= sup(M_2) - sup(M_1) \times sup(M_2) \\ &= sup(M_2)(1 - sup(M_1)) \approx 0 \text{ car } 1 - sup(M_1) \approx 0. \end{aligned}$$

En conclusion, cette recherche des motifs raisonnablement fréquents va nous permettre d'éliminer un 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.

Remarque. (Cornelis et al., 2006) appliquent ce principe puisque d'une part, les motifs négatifs $\overline{X}$ sont recherchés à partir des motifs fréquents X, et par conséquent le support de $\overline{X}$ ne doit pas dépasser le seuil $1 - min_{sup}$ (*voir l'étape 2 de leur algorithme*) et d'autre part, le motif $\overline{X}\ \overline{Y}$ est un motif candidat si les motifs X et Y sont fréquents (*voir l'étape 3 de leur algorithme*).

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

Après avoir exposé une première optimisation éliminant un certain type de règles inintéressantes, nous en présentons une seconde qui va écarter un autre type de règles non pertinentes grâce à l'utilisation de la mesure d'intérêt M_G.

3.1.2 Utilisation de la mesure M_G

Nous commençons par présenter la mesure M_G (Guillaume, 2010) en expliquant sa sémantique et nous terminons cette section en mettant en évidence les règles non pertinentes éliminées par celle-ci.

Mesure M_G
Cette mesure est une extension du facteur de certitude (Heckerman et Shortliffe, 1992) qui d'une part, permet d'éliminer certaines règles non pertinentes et d'autre part, prend en compte l'existence des règles négatives du type $X \Rightarrow \overline{Y}$. Nous rappelons la définition de cette mesure.

Définition 9. Mesure M_G (Guillaume, 2010)

Zone attractive : $max\left(\frac{1}{2}, sup(Y)\right) < conf(X \Rightarrow Y)$
$$M_{G_a}(X \Rightarrow Y) = \frac{conf(X \Rightarrow Y) - max\left(sup(Y), \frac{1}{2}\right)}{1 - max\left(sup(Y), \frac{1}{2}\right)}$$
Zone répulsive : $conf(X \Rightarrow Y) < min\left(\frac{1}{2}, sup(Y)\right)$
$$M_{G_r}(X \Rightarrow Y) = \frac{conf(X \Rightarrow Y) - min\left(sup(Y), \frac{1}{2}\right)}{min\left(sup(Y), \frac{1}{2}\right)}$$
Zone inintéressante : $min\left(\frac{1}{2}, sup(Y)\right) \leqslant conf(X \Rightarrow Y) \leqslant max\left(\frac{1}{2}, sup(Y)\right)$
$$M_{G_i}(X \Rightarrow Y) = 0$$

Avant d'expliquer cette mesure et d'en donner sa sémantique, nous devons définir l'équilibre d'une règle.

Définition 10. 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 la relation suivante :
$conf(X \Rightarrow Y) = conf(X \Rightarrow \overline{Y}) = \frac{1}{2}$ puisque $conf(X \Rightarrow Y) + conf(X \Rightarrow \overline{Y}) = 1$.

L'intérêt de prendre en compte le point d'équilibre a été développé dans (Blanchard et al., 2005).

Cette mesure commence par déterminer la zone d'appartenance (*c'est-à-dire la zone attractive, répulsive ou inintéressante*) de la règle $X \Rightarrow Y$ en calculant la confiance $conf(X \Rightarrow Y)$ de celle-ci et en comparant cette dernière à $min\left(\frac{1}{2}, sup(Y)\right)$ et à $max\left(\frac{1}{2}, sup(Y)\right)$.

• Si la règle $X \Rightarrow Y$ se situe dans la **zone attractive** $\left(i.e.\ max\left(\frac{1}{2}, sup(Y)\right) < conf(X \Rightarrow Y)\right)$ c'est-à-dire si elle est à la fois au delà de l'indépendance $(sup(Y) < conf(X \Rightarrow Y))$ et de l'équilibre $\left(\frac{1}{2} < conf(X \Rightarrow Y)\right)$, cela nous révèle que :
(1) la réalisation de X augmente les chances d'apparition de Y puisque
$$sup(Y) < conf(X \Rightarrow Y) \Longleftrightarrow P(Y) < conf(X \Rightarrow Y),$$

(2) nous avons plus d'exemples[3] que de contre-exemples[4] puisque l'inégalité $\frac{1}{2} < conf(X \Rightarrow Y)$ est vérifiée et par conséquent nous pouvons en déduire l'inégalité $conf(X \Rightarrow \overline{Y}) < \frac{1}{2}$ car nous avons la relation $conf(X \Rightarrow Y) + conf(X \Rightarrow \overline{Y}) = 1$.

La mesure M_G évalue donc la distance de la règle $X \Rightarrow Y$ entre soit l'équilibre soit l'indépendance (*l'état qui sera retenu sera celui dont la valeur du support sera la plus élevée*) et l'implication logique. Ainsi, plus la valeur de M_G est proche de 1, plus la règle est proche de l'implication logique ; et plus la valeur de M_G est proche de 0, plus la règle est proche soit de l'indépendance, soit de l'équilibre.

• Dans le cas où la règle $X \Rightarrow Y$ est dans la **zone répulsive** $\big(\, i.e.\, conf(X \Rightarrow Y) < min\left(\frac{1}{2}, sup(Y)\right)\big)$, la réalisation de X diminue les chances d'apparition de Y (*puisque nous avons* $conf(X \Rightarrow Y) < sup(Y) \Leftrightarrow conf(X \Rightarrow Y) < P(Y)$) ou encore la réalisation de X augmente les chances d'apparition de $\overline{Y}$ puisque nous avons les inégalités suivantes :
$conf(X \Rightarrow Y) < sup(Y) \Longleftrightarrow conf(X \Rightarrow Y) < P(Y) \iff \frac{P(XY)}{P(X)} < P(Y)$
$\iff \frac{P(X)-P(X\,\overline{Y})}{P(X)} < 1 - P(\overline{Y}) \iff 1 - conf(X \Rightarrow \overline{Y}) < 1 - P(\overline{Y})$
$\iff conf(X \Rightarrow \overline{Y}) > P(\overline{Y}) \Longleftrightarrow conf(X \Rightarrow \overline{Y}) > sup(\overline{Y})$.
Dans ce cas-là, nous avons également plus de contre-exemples que d'exemples (*puisque nous avons* $conf(X \Rightarrow Y) < \frac{1}{2}$) ce qui est cohérent puisque c'est la règle $X \Rightarrow \overline{Y}$ qui est la plus pertinente.

La mesure M_G évalue donc la distance de la règle $X \Rightarrow Y$ entre soit le point d'équilibre, soit le point d'indépendance (*celui dont la valeur du support est la plus faible*) et l'incompatibilité. Cette dernière phrase peut également se traduire par la mesure M_G évalue la distance de la règle $X \Rightarrow \overline{Y}$ entre soit le point d'équilibre, soit le point d'indépendance (*celui dont la valeur du support est la plus forte*) et l'implication logique. En effet, nous avons la relation suivante entre les règles antinomiques pour la mesure M_G : $M_{G_a}(X \Rightarrow \overline{Y}) = -M_{G_r}(X \Rightarrow Y)$, $M_{G_r}(X \Rightarrow \overline{Y}) = -M_{G_a}(X \Rightarrow Y)$ et $M_{G_i}(X \Rightarrow \overline{Y}) = -M_{G_i}(X \Rightarrow Y)$.

• La zone restante c'est-à-dire celle où la confiance de la règle $X \Rightarrow Y$ est comprise entre $min\left(\frac{1}{2}, sup(Y)\right)$ et $max\left(\frac{1}{2}, sup(Y)\right)$ est une zone où les règles sont jugées inintéressantes et la valeur de la mesure M_G est égale à 0.

Après avoir présenté la mesure M_G et donné sa sémantique, nous allons mettre en évidence son intérêt en montrant notamment le type de règles qu'elle élimine.

Intérêt de la mesure M_G et règles éliminées par celle-ci

Le couple de mesures (*support*, *confiance*) présente deux atouts majeurs. Tout d'abord, ce sont des mesures facilement interprétables ; et ensuite, ce sont des mesures qui ont une propriété intéressante, la propriété d'anti-monotonicité, qui permet de ne pas parcourir tout l'espace de recherche. Cependant la confiance peut extraire des règles inintéressantes malgré un seuil minimum assez élevé. Soit une base de données décrivant des caractéristiques d'individus tous originaires de Suède. Nous avons les deux items suivants : "*être blond*" symbolisé

3. Un ***exemple*** est un individu qui vérifie à la fois la prémisse X et la conclusion Y.
4. 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 $\overline{Y}$.

par B et "*parler chinois*" symbolisé par C. Les différentes valeurs de support pour ces items sont $sup(B) = 0,90$ et $sup(C) = 0,05$. La valeur du support pour le motif BC est de $0,04$. La règle $parler\ chinois \Rightarrow \hat{e}tre\ blond$ est une règle ayant un support de 4% et une confiance de 80%, donc règle valide si l'utilisateur arrête les seuils suivants : $min_{sup} = 0,03$ et $min_{conf} = 0,80$. Or la connaissance de la prémisse C n'augmente pas les chances d'apparition de la conclusion B puisque sans aucune connaissance, nous avons une probabilité plus élevée d'avoir B et qui est égale à 90%. Pour que la règle soit pertinente, il faut que $conf(C \Rightarrow B) > sup(B)$. C'est la règle $parler\ chinois \Rightarrow \hat{e}tre\ \overline{blond}$ qui est la plus pertinente malgré une faible valeur pour la confiance puisque la connaissance de C augmente les chances d'apparition de $\overline{B}$: $conf(C \Rightarrow \overline{B}) = 0,20$ et $sup(\overline{B}) = 0,10$. Cependant cette condition n'est pas suffisante puisqu'il existe des cas où la règle $X \Rightarrow Y$ vérifie bien la condition $conf(X \Rightarrow Y) > sup(Y)$ et pourtant c'est la règle $X \Rightarrow \overline{Y}$ qui a une confiance plus forte. Prenons cette fois-ci un ensemble d'individus originaires de l'Irlande. On s'intéresse aux caractéristiques suivantes : "*avoir les cheveux roux*" symbolisée par R et "*parler gaélique* (*ou irlandais*)" symbolisée par G. Nous avons les supports suivants : $sup(R) = 0,30$, $sup(G) = 0,10$ et $sup(GR) = 0,04$. La règle $parler\ gaélique \Rightarrow \hat{e}tre\ roux$ est pertinente au regard des critères précédents puisque la confiance de la règle est de 40% et est supérieure au support de la conclusion égal à 30%. La prémisse G augmente les chances d'apparition de la conclusion R, mais cependant la règle $parler\ gaélique \Rightarrow \overline{\hat{e}tre\ roux}$ a une confiance supérieure à la règle antinomique $parler\ gaélique \Rightarrow \hat{e}tre\ roux$ puisque sa confiance est de 60% contre 40%. En conclusion, pour que la règle $X \Rightarrow Y$ soit pertinente, il faut non seulement que la réalisation de X augmente les chances d'apparition de Y mais également que la confiance de la règle $X \Rightarrow Y$ soit supérieure à la confiance de la règle antinomique $X \Rightarrow \overline{Y}$. C'est pour ces raisons que la zone attractive de la mesure M_G vérifie la condition suivante : $max\left(\frac{1}{2}, sup(Y)\right) < conf(X \Rightarrow Y)$.

Cette mesure M_G présente un autre avantage. C'est la relation simple existante entre les règles antinomiques qui est la suivante : $M_{G_a}(X \Rightarrow \overline{Y}) = -M_{G_r}(X \Rightarrow Y)$, $M_{G_r}(X \Rightarrow \overline{Y}) = -M_{G_a}(X \Rightarrow Y)$ et $M_{G_i}(X \Rightarrow \overline{Y}) = -M_{G_i}(X \Rightarrow Y)$. Ainsi, si la valeur de la mesure M_G pour la règle $X \Rightarrow Y$ est strictement négative, nous savons que c'est la règle $X \Rightarrow \overline{Y}$ qui est pertinente et nous avons la valeur de cette règle sans devoir la recalculer grâce à cette relation simple. De plus, la zone répulsive présente les mêmes caractéristiques que la zone attractive : la règle $X \Rightarrow \overline{Y}$ sera pertinente si d'une part, la confiance de la règle est supérieure au support de la conclusion $\overline{Y}$ et d'autre part, si la confiance de la règle $X \Rightarrow \overline{Y}$ est supérieure à la confiance de la règle $X \Rightarrow Y$. C'est pour ces raisons que la zone répulsive est la zone où : $conf(X \Rightarrow Y) < min\left(\frac{1}{2}, sup(Y)\right)$

$$\begin{aligned} &\iff 1 - conf(X \Rightarrow \overline{Y}) < min\left(1 - \tfrac{1}{2}, 1 - sup(\overline{Y})\right) \\ &\iff 1 - conf(X \Rightarrow \overline{Y}) < 1 - max\left(\tfrac{1}{2}, sup(\overline{Y})\right) \\ &\iff conf(X \Rightarrow \overline{Y}) > max\left(\tfrac{1}{2}, sup(\overline{Y})\right). \end{aligned}$$

Après avoir montré l'intérêt de la mesure M_G dans l'élagage des règles inintéressantes, nous passons à la deuxième optimisation, celle du parcours de recherche des règles valides.

3.2 Parcours optimisé de recherche des règles

Dans un premier temps, nous montrons qu'il est inutile d'étudier l'ensemble des règles positives et négatives, l'étude de la moitié des règles est suffisante puisque comme nous allons le montrer dans la section suivante, l'autre moitié ne peut pas conduire à des règles intéressantes.

3.2.1 Étude de la moitié des règles

La restriction de cette étude est possible grâce à la réponse à la question suivante "*la réalisation de la prémisse X augmente-t-elle les chances d'apparition de la conclusion Y ?*", question qui peut se traduire par l'inégalité suivante : $conf(X \Longrightarrow Y) > P(Y) \iff conf(X \Longrightarrow Y) > sup(Y)$?
Nous allons donc étudier l'incidence de la réponse positive à cette question sur trois types de règles, les autres pouvant en être déduits.

Lien entre les règles antinomiques $X \Rightarrow Y$ et $X \Rightarrow \overline{Y}$.
Si la réalisation de X augmente les chances d'apparition de Y (*réponse positive à la question précédente*) alors la réalisation de X diminue les chances d'apparition de $\overline{Y}$.
Démonstration.
$conf(X \Rightarrow Y) > P(Y) \iff \frac{P(XY)}{P(X)} > P(Y) \iff \frac{P(X)-P(X\,\overline{Y})}{P(X)} > 1 - P(\overline{Y})$
$\iff 1 - conf(X \Rightarrow \overline{Y}) > 1 - P(\overline{Y}) \iff conf(X \Rightarrow \overline{Y}) < P(\overline{Y})$.
L'inégalité $conf(X \Rightarrow \overline{Y}) < P(\overline{Y})$ nous indique donc que la réalisation de X diminue les chances d'apparition de $\overline{Y}$.

Lien entre les règles $X \Rightarrow Y$ et $\overline{Y} \Rightarrow \overline{X}$.
Si la réalisation de X augmente les chances d'apparition de Y alors la réalisation de $\overline{Y}$ augmente les chances d'apparition de $\overline{X}$.
Démonstration.
$conf(X \Rightarrow Y) > P(Y) \iff P(XY) > P(X)P(Y)$
$\iff 1 - P(X) - P(Y) + P(XY) > 1 - P(X) - P(Y) + P(X)P(Y)$
$\iff 1 - P(X \vee Y) > (1 - P(X))(1 - P(Y))$
$\iff P(\overline{X \vee Y}) > P(\overline{X})P(\overline{Y}) \iff \frac{P(\overline{X}\,\overline{Y})}{P(\overline{Y})} > P(\overline{X}) \iff conf(\overline{Y} \Rightarrow \overline{X}) > P(\overline{X})$.

Lien entre les règles symétriques $X \Rightarrow Y$ et $Y \Rightarrow X$.
Si la réalisation de X augmente les chances d'apparition de Y alors la réalisation de Y augmente les chances d'apparition de X.
Démonstration.
$conf(X \Rightarrow Y) > P(Y) \iff \frac{P(XY)}{P(X)} > P(Y)$
$\iff \frac{P(XY)}{P(Y)} > P(X) \iff conf(Y \Rightarrow X) > P(X)$.

Si la réponse à la question précédente est positive alors nous pouvons en déduire que la règle $X \Rightarrow Y$ est potentiellement intéressante comme l'a souligné (Piatetsky-Shapiro, 1991). 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 $\overline{Y} \Rightarrow X$,

$conf(X \Rightarrow Y) < sup(Y)$	$conf(X \Rightarrow Y) > sup(Y)$
$X \Rightarrow \overline{Y}$	$X \Rightarrow Y$
$\overline{Y} \Rightarrow X$	$Y \Rightarrow X$
$\overline{X} \Rightarrow Y$	$\overline{X} \Rightarrow \overline{Y}$
$Y \Rightarrow \overline{X}$	$\overline{Y} \Rightarrow \overline{X}$

TAB. 1 – *Ensemble des règles à étudier en fonction de la valeur de la confiance de la règle positive par rapport au support de la conclusion.*

$\overline{X} \Rightarrow Y$ 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 c'est-à-dire soit ($X \Rightarrow Y$, $Y \Rightarrow X$, $\overline{X} \Rightarrow \overline{Y}$, $\overline{Y} \Rightarrow \overline{X}$), soit ($\overline{X} \Rightarrow Y$, $Y \Rightarrow \overline{X}$, $X \Rightarrow \overline{Y}$, $\overline{Y} \Rightarrow X$). Le *tableau 1* résume l'ensemble des règles à étudier en fonction de la réponse à la question précédente c'est-à-dire si $conf(X \Rightarrow Y) > sup(Y)$ ou $conf(X \Rightarrow Y) < sup(Y)$.
Le cas où $conf(X \Rightarrow Y) = sup(Y)$ correspond au cas de l'indépendance et aucune des règles ne sera intéressante puisque nous avons l'indépendance entre X et Y, et par conséquent avec également $\overline{X}$ et $\overline{Y}$.

Après avoir démontré que l'étude de la moitié des règles est suffisante, nous discutons des choix retenus pour les trois algorithmes existants.

Travaux de (Antonie et Zaïane, 2004)
Cette restriction de l'étude des règles est en partie réalisée 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 $coefCorr(X,Y) = coefCorr(Y,X)$.

Travaux de (Wu et al., 2004)
(Wu et al., 2004) recherchent les règles positives à partir des motifs fréquents XY et les règles négatives à partir des motifs non fréquents XY. Ainsi, à partir d'un motif fréquent, ils évaluent la validité des règles $X \Rightarrow Y$ et $Y \Rightarrow X$ en omettant les règles $\overline{X} \Rightarrow \overline{Y}$ et $\overline{Y} \Rightarrow \overline{X}$ qui peuvent être potentiellement intéressantes. Pour finir, à partir d'un motif non fréquent

XY, ils étudient les 6 règles négatives $X \Rightarrow \overline{Y}$, $\overline{Y} \Rightarrow X$, $\overline{X} \Rightarrow Y$, $Y \Rightarrow \overline{X}$, $\overline{X} \Rightarrow \overline{Y}$ et $\overline{Y} \Rightarrow \overline{X}$ dont 2 ou 4 ne pourront pas être potentiellement intéressantes.
(Wu et al., 2004) utilisent une mesure d'intérêt (*la valeur absolue de la nouveauté*) pour sélectionner les règles valides qui comme nous l'avons dit est la valeur absolue de la *nouveauté*. Cette mesure peut donner une réponse à la question précédente puisque la *nouveauté* permet d'évaluer un écart à l'indépendance comme le montre les égalités suivantes : $sup(XY) - sup(X) \times sup(Y) = P(XY) - P(X)P(Y) = P(X)\left[\frac{P(XY)}{P(X)} - P(Y)\right] = P(X)\left[P(Y/X) - P(Y)\right] = P(X)\left[conf(X \Rightarrow Y) - sup(Y)\right]$. Malheureusement (Wu et al., 2004) n'utilisent pas cette propriété.

Travaux de (Cornelis et al., 2006)
(Cornelis et al., 2006) recherchent les règles valides à partir des motifs fréquents XY, $\overline{X}\,\overline{Y}$ et $\overline{X}\,Y$ précédemment définis. Cette recherche des règles valides s'effectue indépendamment de ce qui a été trouvé avec les autres motifs issus du même motif positif XY. En conséquence, l'ensemble des règles négatives est étudié.

Nous venons de démontrer qu'il est inutile d'étudier l'ensemble des règles à partir d'un motif XY. Nous pouvons encore restreindre ce nombre de règles à étudier grâce à l'utilisation des méta-règles dégagées par (Guillaume et Papon, 2012) et reposant sur la mesure M_G.

3.2.2 Règles d'élagage

Nous souhaitons poursuivre l'optimisation de la recherche des règles valides en restreignant encore l'ensemble des règles à étudier. Nous avons vu dans la *section 3.1.2* que les règles vérifiant $conf(X \Rightarrow Y) > sup(Y)$ ne sont pas toutes pertinentes et qu'il est préférable de retenir les règles vérifiant la condition suivante $conf(X \Rightarrow Y) > max\left(\frac{1}{2}, sup(Y)\right)$, condition plus restrictive que celle de la section précédente. Pour vérifier cette condition plus restrictive, nous allons utiliser la mesure M_G précédemment définie dans la *section 3.1.2*.

• Ainsi dans la **zone attractive**, nous aimerions connaître les conditions qui vont nous permettre d'écarter l'étude des règles $\overline{X} \Rightarrow \overline{Y}$ et $Y \Rightarrow X$ à partir de la validité ou non de la règle $X \Rightarrow Y$ pour la mesure M_G. Nous pourrons en déduire l'intérêt de la règle $\overline{Y} \Rightarrow \overline{X}$ à partir des résultats obtenus sur la règle $\overline{X} \Rightarrow \overline{Y}$ puisque nous connaitrons les conditions de rejet d'étude d'une règle à partir de sa règle symétrique.
• Pour la zone **répulsive**, nous allons procéder de même et étudier les conditions pour lesquelles les règles $\overline{X} \Rightarrow Y$ et $\overline{Y} \Rightarrow X$ n'ont pas besoin d'être étudiées à partir de la validité ou non de la règle $X \Rightarrow \overline{Y}$ pour la mesure M_G. De la même façon, nous pourrons en déduire l'intérêt de la règle $Y \Rightarrow \overline{X}$ à partir des résultats obtenus sur la règle $\overline{X} \Rightarrow Y$ puisque nous connaitrons les conditions de rejet d'étude d'une règle à partir de sa règle symétrique.

Afin d'y parvenir nous utilisons les méta-règles dégagées par (Guillaume et Papon, 2012) et qui permettent d'en déduire la validité ou non des règles négatives pour la mesure M_G à partir de la validité ou non validité des règles positives $X \Rightarrow Y$ pour cette même mesure. Comme notre objectif est de limiter l'espace de recherche des règles, nous utilisons uniquement les méta-règles permettant de conclure sur la non validité de la règle au regard de la mesure M_G.

Nous présentons les deux méta-règles qui vont être utilisées pour optimiser la recherche des règles :
$(MR_1) : \forall X \Rightarrow Y$ 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 $M_G(\overline{X} \Rightarrow \overline{Y}) < min_{M_G}$.
$(MR_2) : \forall X \Rightarrow Y$ avec $sup(X) < sup(Y)$,
si $M_G(X \Rightarrow Y) < min_{M_G}$ alors $M_G(Y \Rightarrow X) < min_{M_G}$.

La première méta-règle (MR_1) nous révèle que si la règle $X \Rightarrow Y$ n'est pas valide pour la mesure M_G, alors la règle $\overline{X} \Rightarrow \overline{Y}$ ne sera également pas valide pour cette même mesure dans le cas où $\left(\frac{1}{2} < sup(X) < sup(Y)\right)$ et également si $\left(sup(X) < \frac{1}{2} < sup(Y)\right)$. Quant à la deuxième méta-règle (MR_2), elle nous révèle que si la règle $X \Rightarrow Y$ n'est pas valide pour la mesure M_G, alors la règle $Y \Rightarrow X$ ne sera également pas valide pour cette même mesure dans le cas où $sup(X) < sup(Y)$. Ces deux méta-règles seront utilisées dans le cas attractif et répulsif. Dans le cas répulsif, la déduction de la non validité de la règle $\overline{X} \Rightarrow Y$ à partir de la règle $X \Rightarrow \overline{Y}$ sera effectué grâce à la méta-règle (MR_1).

Nous formalisons la propriété d'anti-monotonicité de la confiance à l'aide d'une méta-règle que nous nommerons (MR_3).

$(MR_3) : \forall\, (X, Y, Z)\ /\ Y \subsetneq Z \subsetneq X$ et $X \subseteq \mathcal{I}$,
si $conf(X \backslash Y \Rightarrow Y) < min_{conf}$ alors $conf(X \backslash Z \Rightarrow Z) < min_{conf}$.

Remarque. Seuls (Cornelis et al., 2006) utilisent une technique d'élagage pour la recherche des règles : ils utilisent la propriété d'anti-monotonicité de la confiance lors de la recherche des règles positives c'est-à-dire (MR_3).

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 3.3*.

3.3 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 importante. 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.

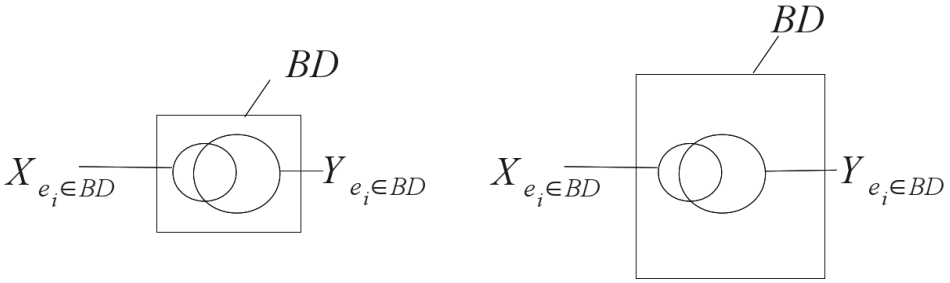

FIG. 1 – *Exemple de règles où les motifs ont des supports relativement élevés (courbe de gauche) et où les motifs ont des supports proches du seuil minimal* min_{sup} *(courbe de droite).*

La contingence des ensembles $X_{e_i \in BD}$ [5], $Y_{e_i \in BD}$ et $(X_{e_i \in BD} \cap Y_{e_i \in BD})$ est la même pour les deux courbes. Seul l'ensemble $(\overline{X}_{e_i \in BD} \cap \overline{Y}_{e_i \in BD})$ a une contingence différente, 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.

Après avoir exposé les différentes optimisations retenues pour extraire les RAPN, nous présentons notre algorithme.

4 Algorithme

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

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

Une règle d'association positive ou négative (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 lorsque les motifs C_1 et C_2 sont des motifs raisonnablement fréquents négatifs $\overline{X}$ et $\overline{Y}$, ceux-ci 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.

Après avoir défini ce qu'est une RAPN valide, nous présentons l'algorithme avec au préalable la définition de toutes les notations utilisées.

- BD : base de données ;
- $\mathcal{I}$: ensemble des items ;
- $\overline{\mathcal{I}}$: ensemble des négations d'items ;
- X, Y : motifs positifs ;
- $\overline{X}, \overline{Y}$: motifs négatifs ;
- $\ddot{X}, \ddot{Y}$: conjonctions de motifs négatifs ;
- i : item ou *1*-motif ;
- min_{sup}, max_{sup} : seuils respectivement minimum et maximum pour le support des motifs positifs ;
- $min_{s\ddot{u}p}$: seuil minimum pour le support des motifs $\ddot{X}$;
- min_{conf}, min_{M_G} : seuils minimum pour respectivement la confiance et la mesure M_G ;
- $taille(X)$: nombre d'items composant un motif ;
- R : ensemble des *R*ègles valides ;
- RF : ensemble des motifs *R*aisonnablement *F*réquents ;
- $NRFM$: ensemble des motifs *N*égatifs *R*aisonnablement *F*réquents *M*inimaux ;
- $NNRF_k$: ensemble des k-motifs *N*égatifs *N*on *R*aisonnablement *F*réquents ;
- CP_k : ensemble des k-motifs *C*andidats *P*otentiels ;
- C_k : ensemble des k-motifs *C*andidats ;
- F_k : ensemble des k-motifs *F*réquents X et $\ddot{X}$;
- F : ensemble de tous les motifs *F*réquents ;
- s : support de la règle étudiée
- c : confiance de la règle étudiée
- m : mesure M_G de la règle étudiée

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 exposée dans (Agrawal et Srikant, 1994) pour générer les motifs fréquents en

rajoutant deux contraintes supplémentaires : (1) un seuil maximal max_{sup} qui ne doit pas être dépassé par le support de X et (2) un seuil minimal $min_{s\ddot{u}p}$ pour le support des motifs $\ddot{X}$; ce qui permet de vérifier les contraintes (Ct_1) et (Ct_2). Cette fonction sera développée dans la *section 4.1* grâce à l'*algorithme 2*.

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.* $sup(\overline{X}) \leqslant max_{sup}$). Cette fonction sera développée dans la *section 4.2* grâce à l'*algorithme 3*.

Vient ensuite la phase d'extraction des règles valides (*lignes 3 à 41*) 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*) où la taille de Y est inférieure ou égale à la taille de $X \backslash Y$ [6] et tel que Y soit un sous-ensemble de X (*i.e.* $Y \subsetneq X$), on commence par déterminer le type d'attraction entre X et Y (*ligne 5*) en comparant la confiance $conf(X \backslash Y \Rightarrow Y)$ de la règle $X \backslash Y \Rightarrow Y$ au support $sup(Y)$ de sa conclusion Y.

Si c'est une **attraction positive** (*ligne 6*) c'est-à-dire si $conf(X \backslash Y \Rightarrow Y) > sup(Y)$, alors on s'assure que les règles $X \backslash Y \Rightarrow Y$, $Y \Rightarrow X \backslash Y$, $\overline{X \backslash Y} \Rightarrow \overline{Y}$ et $\overline{Y} \Rightarrow \overline{X \backslash Y}$ sont valides (*lignes 7 à 17*) (*étude d'un sous-ensemble de règles ; section 3.2.1*). Afin de déterminer l'ordre d'étude des différentes règles, nous testons si $sup(X \backslash Y) \leqslant sup(Y)$ (*ligne 7*) dans le but d'appliquer les méta-règles (MR_1) et (MR_2) exposées précédemment. Si la réponse est positive, alors l'ordre d'étude des différentes règles est le suivant : $X \backslash Y \Rightarrow Y$ (*ligne 8*), $Y \Rightarrow X \backslash Y$ (*ligne 8*), $\overline{X \backslash Y} \Rightarrow \overline{Y}$ (*ligne 10*) et $\overline{Y} \Rightarrow \overline{X \backslash Y}$ (*ligne 10*). Si la réponse est négative (*ligne 12*), l'ordre d'étude est le suivant : $Y \Rightarrow X \backslash Y$ (*ligne 13*), $X \backslash Y \Rightarrow Y$ (*ligne 13*), $\overline{Y} \Rightarrow \overline{X \backslash Y}$ (*ligne 15*) et $\overline{X \backslash Y} \Rightarrow \overline{Y}$ (*ligne 15*). La différence est la règle de référence étudiée en premier lieu qui est la règle $X \backslash Y \Rightarrow Y$ dans le cas où $sup(X \backslash Y) \leqslant sup(Y)$, et la règle $Y \Rightarrow X \backslash Y$ dans le cas où $sup(X \backslash Y) > sup(Y)$. Afin de ne pas effectuer deux fois la vérification des motifs négatifs minimaux $\overline{X \backslash Y}$ et $\overline{Y}$, nous regroupons l'étude des règles négatives $\overline{X \backslash Y} \Rightarrow \overline{Y}$ et $\overline{Y} \Rightarrow \overline{X \backslash Y}$ (*soit lignes 9 à 11 ; soit lignes 14 à 16*).

Si c'est une **attraction négative** (*ligne 18*) c'est-à-dire si $conf(X \backslash Y \Rightarrow Y) < sup(Y)$, alors on s'assure que les règles $X \backslash Y \Rightarrow \overline{Y}$, $\overline{Y} \Rightarrow X \backslash Y$, $\overline{X \backslash Y} \Rightarrow Y$ et $Y \Rightarrow \overline{X \backslash Y}$ sont valides (*lignes 19 à 38*) (*étude d'un sous-ensemble de règles ; section 3.2.1*). Afin de déterminer l'ordre d'étude des différentes règles, nous testons si $sup(X \backslash Y) \leqslant sup(\overline{Y})$ (*ligne 20*) dans le but toujours d'appliquer les méta-règles (MR_1) et (MR_2). Si la réponse est positive, alors l'ordre d'étude des différentes règles est le suivant : $X \backslash Y \Rightarrow \overline{Y}$ (*ligne 21*), $\overline{Y} \Rightarrow X \backslash Y$ (*ligne 21*), $\overline{X \backslash Y} \Rightarrow Y$ (*ligne 23*) et $Y \Rightarrow \overline{X \backslash Y}$ (*ligne 23*). Si la réponse est négative (*ligne 25*), l'ordre d'étude est le suivant : $\overline{Y} \Rightarrow X \backslash Y$ (*ligne 26*), $X \backslash Y \Rightarrow \overline{Y}$ (*ligne 26*), $Y \Rightarrow \overline{X \backslash Y}$ (*ligne 28*) et $\overline{X \backslash Y} \Rightarrow Y$ (*ligne 28*). La différence est la règle de référence qui est la règle $X \backslash Y \Rightarrow \overline{Y}$ dans le cas où $sup(X \backslash Y) \leqslant sup(\overline{Y})$, et la règle $\overline{Y} \Rightarrow X \backslash Y$ dans le cas où $sup(X \backslash Y) > sup(\overline{Y})$. Avant d'étudier ces différentes règles négatives, on doit s'assurer que le motif $\overline{Y}$ est un motif négatif minimal (*ligne 19*) ainsi que le motif $\overline{X \backslash Y}$ (*lignes 22, 27, 31*). Dans le cas où $\overline{Y}$ n'est pas un motif négatif minimal, on étudie les règles $\overline{X \backslash Y} \Rightarrow Y$ et $Y \Rightarrow \overline{X \backslash Y}$ (*ligne 33 ou*

6. Cette contrainte évite d'étudier deux fois les mêmes règles puisque pour le couple de motifs $(X \backslash Y, Y)$, nous étudions les règles symétriques, et par conséquent, il est inutile d'étudier le couple de motifs $(Y, X \backslash Y)$.

ligne 35). De la même façon, avant cette étude on s'assure que $\overline{X \backslash Y}$ est un motif négatif minimal (*ligne 31*) et on détermine le sens d'étude des deux règles en fonction du résultat de la comparaison entre $sup(\overline{X \backslash Y})$ et $sup(Y)$ (*lignes 32, 34*).
Après l'étude des règles positives et des règles totalement négatives et mixtes, on étudie la règle $X \ddot{\backslash} Y \Rightarrow \ddot{Y}$ (*ligne 39*) en vérifiant les contraintes (Ct_4) et (Ct_5).

L'étude de la règle positive référence suit le processus suivant. On commence par vérifier que la méta-règle (MR_3) ne peut pas s'appliquer (*propriété d'anti-monotonicité de la confiance*). Si c'est le cas, on vérifie ensuite qu'elle est valide pour la confiance (*contrainte* (Ct_4)) et la mesure M_G (*contrainte* (Ct_5)). Une fois l'étude de la règle positive référence effectuée, on s'intéresse à la règle positive symétrique. Pour cela on commence par vérifier que la méta-règle (MR_2) ne peut pas s'appliquer. Si c'est le cas, on vérifie sa validité pour la confiance (*contrainte* (Ct_4)) et la mesure M_G (*contrainte* (Ct_5)).

Le passage à la première règle négative du type $\overline{X \backslash Y} \Rightarrow \overline{Y}$ s'effectue en vérifiant que la méta-règle (MR_1) ne peut pas s'appliquer. Ensuite, on vérifie la validité de la règle au regard de la confiance et de la mesure M_G. Pour finir, l'étude de la dernière règle négative du type $\overline{Y} \Rightarrow \overline{X \backslash Y}$ vérifie les contraintes (Ct_4) et (Ct_5) sans pouvoir appliquer de méta-règle.

Le principe d'étude des règles situées dans la zone répulsive est le même que celui qui vient d'être décrit dans la zone attractive.

Après avoir présenté l'algorithme général d'extraction des RAPN, nous allons détailler la première phase qui est l'extraction des motifs raisonnablement fréquents.

4.1 Recherche des motifs raisonnablement fréquents

Cette recherche des motifs raisonnablement fréquents (*voir l'algorithme 2*) et qui est effectuée par la fonction *funct_RF* (*ligne 1 de l'algorithme 1*) est similaire à celle utilisée par (Agrawal et Srikant, 1994) pour générer les motifs fréquents. Elle rajoute deux contraintes supplémentaires sur les motifs X : 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 $\ddot{X}$. Cette fonction commence par attribuer des valeurs aux deux seuils max_{sup} et $min_{s\ddot{u}p}$ si l'utilisateur ne l'a pas fait (*lignes 1 et 2*). Il est à noter que la valeur du seuil min_{sup} doit obligatoirement être renseignée. Ensuite, on initialise l'ensemble RF des motifs raisonnablement fréquents à l'ensemble vide (*ligne 3*) et l'ensemble C_1 des *1*-candidats est initialisé à l'ensemble de tous les items i de la base BD (*ligne 4*). Le processus suivant (*lignes 5 à 20*) va être réitéré jusqu'à ce qu'on n'obtienne plus de candidat $(C_k \neq \varnothing)$ à partir de l'ensemble F_{k-1} des motifs fréquents X et $\ddot{X}$ de niveau inférieur (*ligne 17*). En effet, la génération des candidats va s'effectuer à partir uniquement des motifs fréquents X et $\ddot{X}$ en raison de la propriété d'anti-monotonicité du support que nous rappelons : $\forall\ Y \supsetneq X \ \ sup(Y) \leqslant sup(X)$. Nous savons que pour tout sur-ensemble Y de X $(Y \supsetneq X)$ et pour tout sur-ensemble $\ddot{Y}$ de $\ddot{X}$ $(\ddot{Y} \supsetneq \ddot{X})$, si X et $\ddot{X}$ ne vérifient le support minimum (*c'est-à-dire si* $sup(X) < min_{sup}$ *et* $sup(\ddot{X}) < min_{s\ddot{u}p}$), alors Y et $\ddot{Y}$ ne pourront pas le vérifier car $sup(Y) \leqslant sup(X)$ et $sup(\ddot{Y}) \leqslant sup(\ddot{X})$. Il n'en est pas de même pour le support maximal de Y car comme le support de Y est inférieur au support de X, le motif Y peut vérifier la contrainte du support maximal même si le motif X ne le vérifie pas. C'est pour cette raison que nous travaillons avec deux ensembles : l'ensemble F_k des motifs fréquents X et $\ddot{X}$ qui va servir uniquement à générer l'ensemble CP_{k+1} des candidats potentiels de niveau supérieur (*ligne 17*) et l'ensemble recherché RF des motifs raisonnable-

ment fréquents (*ligne 13*). Pour un niveau donné k, on commence par initialiser l'ensemble F_k à l'ensemble vide (*ligne 6*). Ensuite pour tous les candidats X de C_k (*ligne 7*), on calcule le support s de X (*ligne 8*) et le support $\ddot{s}$ de $\ddot{X}$ (*ligne 9*). Le calcul du support de $\ddot{X}$ ne nécessite pas de parcourir la base BD mais peut être calculé à partir des supports des motifs positifs. Ce calcul du support de $\ddot{X}$ est réalisé grâce à la fonction $computeSupport(X, s, F)$. Si les motifs X et $\ddot{X}$ sont fréquents (*ligne 10*), alors on stocke le motif X dans l'ensemble F_k des fréquents (*ligne 11*) qui servira ensuite à générer les candidats potentiels CP_{k+1} de niveau supérieur (*ligne 17*). Ensuite dans le cas où les motifs X et $\ddot{X}$ sont fréquents, on vérifie que le support de X n'est pas trop élevé (*ligne 12*), c'est-à-dire inférieur au seuil max_{sup}. Si c'est le cas, le motif X est un motif raisonnablement fréquent et sera mémorisé dans l'ensemble RF (*ligne 13*) des motifs raisonnablement fréquents accompagné de ses deux valeurs de support : le support s de X et le support $\ddot{s}$ de $\ddot{X}$. Une fois que tous les candidats X de C_k ont été parcourus (*lignes 7 à 16*), on génère l'ensemble CP_{k+1} des candidats potentiels (*ligne 17*) à partir de l'ensemble des fréquents F_k, comme effectué avec l'algorithme *Apriori*. Ensuite, on vérifie que tous les sous-ensembles de X sont des fréquents (*ligne 18*), ce qui conduit à l'ensemble des candidats C_{k+1}, de nouveau comme pour l'algorithme *Apriori*. On termine par la mémorisation des fréquents de niveau k dans l'ensemble F de tous les fréquents (*ligne 19*), l'ensemble F servant à la recherche des motifs candidats C_{k+1} (*ligne 18*).
Après avoir exposé la recherche des motifs raisonnablement fréquents, nous expliquons la recherche des motifs négatifs raisonnablement fréquents minimaux.

4.2 Recherche des motifs négatifs raisonnablement fréquents minimaux

Comme nous l'avons dit précédemment, la recherche des motifs négatifs raisonnablement fréquents minimaux va servir pour vérifier la contrainte (Ct_6) de la validité des règles.

Cette fonction *funct_NRFM* (*ligne 2 de l'algorithme 1*) et exposé dans l'*algorithme 3* commence par initialiser l'ensemble recherché $NRFM$ des motifs Négatifs Raisonnablement Fréquents Minimaux à l'ensemble de toutes les négations d'items de taille 1 raisonnablement fréquentes (*ligne 1*). Cette initialisation est rendue possible grâce à la connaissance des supports des items i calculés lors de l'exécution de la fonction *funct_RF* (*ligne 1 de l'algorithme 1*) puisque nous avons la relation suivante : $sup(\bar{i}) = 1 - sup(i)$. Puis, nous stockons dans l'ensemble $NNRF_1$ les items négatifs non raisonnablement fréquents $(i.e\ \overline{\mathcal{I}} \setminus NRFM)$ auquel on enlève les négations d'items $\bar{i}$ dont le support est supérieur au seuil maximal (*ligne 2*) car tout sur-ensemble $\{\overline{i,j}\}$ aura une valeur de support supérieure à celui de $\bar{i}$ ou de $\bar{j}$. Ensuite, on génère l'ensemble C_2 des motifs candidats de taille 2 à partir de l'ensemble $NNRF_1$ des négations d'items non raisonnablement fréquentes (*ligne 3*). Le processus suivant va être réitéré jusqu'à ce que l'on n'arrive plus à générer de candidats $(C_k \neq \varnothing)$ (*lignes 4 à 16*). On commence par initialiser l'ensemble $NNRF_k$ des motifs $\overline{X}$ ayant un support inférieur à min_{sup} (*car si le support de $\overline{X}$ est supérieur à max_{sup} alors tout sur-ensemble $\overline{XY}$ aura un support encore plus élevé*) à l'ensemble vide (*ligne 5*). On parcourt tous les motifs candidats $\overline{X}$ (*ligne 6*) afin de détecter ceux qui sont raisonnablement fréquents (*lignes 7* et 8). Si ce n'est pas le cas, on s'assure que $\overline{X}$ n'a pas un support supérieur à max_{sup} (*par conséquent on teste si son support est inférieur à min_{sup}*) (*ligne 10*) et on le stocke dans $NNRF_k$ (*ligne 11*) comme motif pouvant générer au niveau supérieur $(k+1)$ un motif candidat (*ligne 15*).

Après avoir exposé l'algorithme d'extraction des RAPN, nous présentons les expérimentations qui ont été réalisées sur deux bases de données.

5 Expérimentations

Les quatre algorithmes présentés dans cet article 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 3 bases de données UCI KDD *Ecoli*, *Iris* et *Abalone*.

Nous avons fait une première extraction pour comparer les temps d'exécution (*en secondes*) des différents algorithmes sur les bases de données UCI. Nous connaissons l'impact du support minimum min_{sup} sur les temps d'extraction des différents algorithmes dérivant de *Apriori*, et c'est pour cette raison que nous faisons varier ce paramètre. Nous retenons pour tous les algorithmes, la valeur de 0,80 pour la confiance minimum. Pour les autres seuils, nous avons retenu les valeurs suivantes : 0,60 pour le coefficient de corrélation nécessaire à l'algorithme de (Antonie et Zaïane, 2004) ; 0,10 pour la mesure d'intérêt et 0,60 pour le facteur de certitude utilisées dans (Wu et al., 2004) ; 0,60 pour la mesure M_G, 0,80 pour le seuil maximal du support et $min_{s\ddot{u}p} = min_{sup}$ pour l'algorithme que nous avons proposé. Nous avons volontairement choisi des seuils suffisamment bas pour le coefficient de corrélation, la mesure d'intérêt, le facteur de certitude et la mesure M_G afin d'obtenir des résultats au niveau des temps d'extraction relativement comparables. La *figure 2* nous restitue les résultats pour les deux bases de données *Ecoli* et *Iris*.

	Ecoli				Iris			
minsup	Antonie	Cornelis	Wu	RAPN	Antonie	Cornelis	Wu	RAPN
0,01	14,74	18,16	0,84	0,47	0,12	0,18	0,07	0,11
0,05	3,44	8,45	0,35	0,12	0,13	0,12	0,05	0,10
0,10	2,99	6,43	0,19	0,08	0,14	0,09	0,06	0,07
0,20	0,73	3,33	0,08	0,03	0,10	0,07	0,03	0,05

FIG. 2 – *Etude comparative des temps d'exécution en secondes des 4 algorithmes sur les bases de données Ecoli et Iris.*

Nous vérifions que les optimisations apportées ont eu un impact significatif sur les temps d'extraction puisque notre algorithme RAPN a des temps d'extraction faibles pour ces deux bases de données, et ceci malgré l'extraction d'un nouveau type de règles négatives. Sur la base de données *Iris*, les temps d'extraction sont relativement similaires pour les 4 algorithmes, ce qui n'est pas le cas pour la base de données *Ecoli* où notre algorithme ainsi que l'algorithme de (Wu et al., 2004) corrigé à des temps d'extraction inférieurs à la seconde.

Après avoir vérifié la rapidité de notre algorithme sur deux bases de données, nous étudions maintenant le nombre de règles extraites, et ceci par type de règles. Cette comparaison a été effectuée avec la base de données *Abalone* et les résultats sont synthétisés dans les *figures 3* et *4*. La colonne étiquetée "+" correspond aux règles positives $X \Rightarrow Y$, la colonne étiquetée "- +" correspond aux règles négatives mixtes du type $\overline{X} \Rightarrow Y$, la colonne "+ -" correspond aux règles mixtes du type $X \Rightarrow \overline{Y}$ et pour finir la colonne "- -" correspond aux règles entièrement négatives du type $\overline{X} \Rightarrow \overline{Y}$. L'avant-dernière colonne restitue le nombre total de règles néga-

tives extraites et la dernière colonne, le nombre de règles total. Pour notre algorithme RAPN, nous avons une colonne supplémentaire nommée "*Nouvel*" restituant le nombre de règles extraites du type $\ddot{X} \Rightarrow \ddot{Y}$. Cette extraction a été effectuée pour différentes valeurs de seuil pour le support (*0,01 ; 0,05 ; 0,10 et 0,20*) et pour la confiance (*0,80 ; 0,90 et 0,95*).

		Antonie						Cornelis					
minsup	minconf	+	- +	+ -	- -	TotalN	Total	+	- +	+ -	- -	TotalN	Total
0,01	0,80	7 535	947	4 784	1 032	6 763	14 298	46 645	223	27 815	150	28 188	74 833
	0,90	4 034	397	4 136	1 032	5 565	9 599	31 274	128	24 453	115	24 696	55 970
	0,95	2 962	145	3 190	1 032	4 367	7 329	24 087	93	21 695	89	21 877	45 964
0,05	0,80	7 508	947	4 784	6	5 737	13 245	28 571	220	10 233	87	10 540	39 111
	0,90	4 007	397	4 136	6	4 539	8 546	18 596	125	8 525	54	8 704	27 300
	0,95	2 935	145	3 190	6	3 341	6 276	14 412	90	7 561	38	7 689	22 101
0,10	0,80	3 794	947	4 784	690	6 421	10 215	12 784	218	3 826	55	4 099	16 883
	0,90	2 549	397	4 136	495	5 028	7 577	9 174	123	3 162	34	3 319	12 493
	0,95	1 477	145	3 190	258	3 593	5 070	6 313	88	2 624	24	2 736	9 049
0,20	0,80	3 013	947	4 784	2 545	8 276	11 289	6 038	212	1 862	66	2 140	8 178
	0,90	2 338	397	4 136	1 921	6 454	8 792	4 712	123	1 552	36	1 711	6 423
	0,95	1 363	145	3 190	751	4 086	5 449	3 506	88	1 350	24	1 462	4 968

FIG. 3 – *Etude comparative du nombre de règles extraites pour l'algorithme de (Antonie et Zaïane, 2004) et (Cornelis et al., 2006) pour la base de données Abalone.*

		Wu						RAPN						
minsup	minconf	+	- +	+ -	- -	TotalN	Total	+	- +	+ -	- -	TotalN	Total	Nouvel
0,01	0,80	12 668	462	3 452	1 865	5 779	18 447	12 185	32	495	35	562	12 747	9 911
	0,90	7 648	462	1 878	1 519	3 859	11 507	7 892	21	142	31	194	8 086	6 775
	0,95	5 549	462	969	1 423	2 854	8 403	5 956	12	59	21	92	6 048	5 115
0,05	0,80	8 139	328	336	2 062	2 726	10 865	8 767	7	70	35	112	8 879	6 115
	0,90	5 002	318	204	1 170	1 692	6 694	5 605	4	4	31	39	5 644	4 091
	0,95	3 308	318	139	711	1 168	4 476	4 262	0	0	21	21	4 283	3 150
0,10	0,80	5 596	119	143	1 015	1 277	6 873	2 942	0	2	35	37	2 979	2 415
	0,90	3 733	67	112	628	807	4 540	2 098	0	0	31	31	2 129	1 614
	0,95	2 622	49	67	374	490	3 112	1 443	0	0	21	21	1 464	1 251
0,20	0,80	3 537	41	46	175	262	3 799	264	0	0	30	30	294	208
	0,90	2 241	23	38	110	171	2 412	228	0	0	26	26	254	179
	0,95	1 442	15	21	75	111	1 553	150	0	0	21	21	171	140

FIG. 4 – *Etude comparative du nombre de règles extraites pour l'algorithme de (Wu et al., 2004) et l'algorithme que nous proposons pour la base de données Abalone.*

Nous vérifions que notre algorithme est le plus sélectif au niveau du nombre total de règles extraites comme le synthétise la *figure 5*. La *figure 5* restitue pour chacun des algorithmes, le nombre de règles positives et négatives (*axe des ordonnées*) et ceci pour les différentes valeurs de seuil pour le support (*axe des abscisses*). La courbe de gauche montre ces différentes courbes pour le seuil de la confiance à 0,80 et la courbe de droite pour la valeur de 0,90. Nous n'avons pas tracé ces courbes pour le seuil de confiance égal à 0,95 car nous obtenons les mêmes courbes que celles de la *figure 5*. Nous constatons que les optimisations que nous avons apportées aux algorithmes reposant sur *Apriori* pour réduire le nombre de règles extraites ont donné des résultats satisfaisants.

Pour finir, nous avons souhaité connaître la répartition du nombre de règles négatives et du nombre de règles positives extraites pour les 4 algorithmes, et ceci dans un cas de figure : pour le seuil égal à 0,80 pour la confiance (*nous avons les mêmes courbes pour les autres valeurs de seuil*). La *figure 6* restitue cette répartition, les courbes de gauche concernent les règles négatives et les courbes de droite, les règles positives. Notre algorithme reste bien placé en terme de nombre de règles extraites, surtout pour les règles négatives mais une étude plus

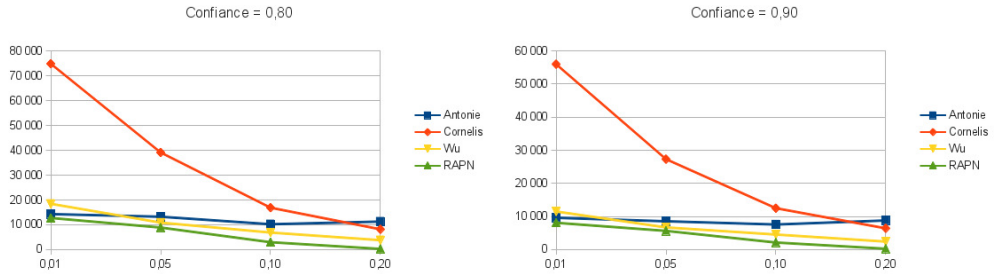

FIG. 5 – *Etude comparative du nombre total de règles extraites pour chacun des algorithmes.*

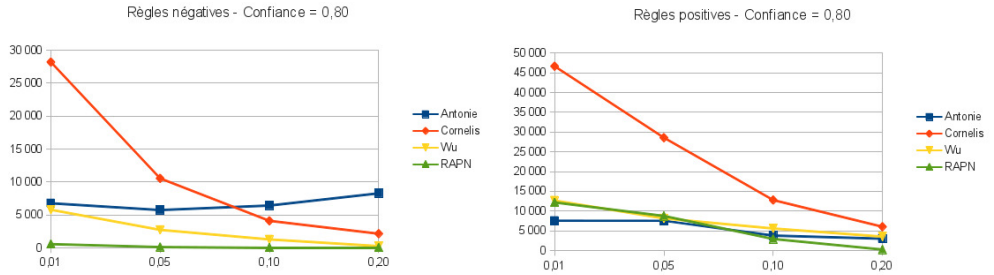

FIG. 6 – *Etude comparative du nombre de règles négatives extraites pour chacun des algorithmes (courbe de gauche) et du nombre de règles positives (courbe de droite).*

approfondie est nécessaire et plus particulièrement sur la qualité des règles extraites ainsi que sur la complétude de celles-ci. Il est à noter que le nombre de nouvelles règles extraites du type $\ddot{X} \Rightarrow \ddot{Y}$ est du même ordre de grandeur (*en nombre*) que les règles positives extraites par notre algorithme RAPN. Une étude complémentaire est également nécessaire et notamment pour vérifier qu'elles sont toutes pertinentes et qu'il n'y a pas de redondance.

6 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 rendu 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 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 (Antonie et Zaïane, 2004) et (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, abandonnée dans les algorithmes existants d'extraction de RAPN, mais également grâce à deux méta-règles permettant d'inférer la non validité des règles $\overline{X} \Rightarrow \overline{Y}$ et des règles $Y \Rightarrow X$ à partir des règles $X \Rightarrow Y$. 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. 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.

Références

Agrawal, R., T. Imielsky, et A. Swami (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 8th ACM SIGKDD Conference on Knowledge Discovery and Data Mining 1993*, ACM, pp. 207–216.

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, H. Briand, et R. 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.

Heckerman, D. et E. Shortliffe (1992). From certainty factors to belief networks. In *Artificial Intelligence in Medicine*, 4, pp. 35–52.

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.

Missaoui, R., L. Nourine, et Y. Renaud (2008). Generating positive and negative exact rules using formal concept analysis : problems and solutions. In *Proceedings of the Sixth International Conference on Formal Concept Analysis*, pp. 169–181.

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

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In *Philosophical Magazine*, Series 5 50 (302), pp. 157–175.

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.

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.

Yuan, X., B. Buckles, Z. Yuan, et J. Zhang (2002). Mining negative association rules. In *Proceedings of the Seventh International Symposium on Computers and Communications (ISCC'02)*, pp. 623–628.

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 two databases.

Algorithm 1 : Extraction des RAPN

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

1: $RF = funct\_RF(BD, min_{sup}, max_{sup}, min_{süp})$
2: $NRFM = funct\_NRFM(BD, RF)$
3: **for all** motif raisonnablement fréquent $X \in RF$ où $taille(X) > 1$ **do**
4: **for all** conclusion $Y \subsetneq X$ / $taille(Y) \leqslant taille(X \backslash Y)$ **do**
5: Détermination du type d'attraction entre X et Y
6: **if** $conf(X \backslash Y \Rightarrow Y) > sup(Y)$ **then**
7: **if** $sup(X \backslash Y) \leqslant sup(Y)$ **then**
8: [ ] (MR_3)] Etude de $X \backslash Y \Rightarrow Y$; [ ] (MR_2)] Etude de $Y \Rightarrow X \backslash Y$
9: **if** $\left(\overline{X \backslash Y} \in NRFM\right) \wedge \left(\overline{Y} \in NRFM\right)$ **then**
10: [ ] (MR_1)] Etude de $\overline{X \backslash Y} \Rightarrow \overline{Y}$; Etude de $\overline{Y} \Rightarrow \overline{X \backslash Y}$
11: **end if**
12: **else if** $sup(X \backslash Y) > sup(Y)$ **then**
13: [ ] (MR_3)] Etude de $Y \Rightarrow X \backslash Y$; [ ] (MR_2)] Etude de $X \backslash Y \Rightarrow Y$
14: **if** $\left(\overline{X \backslash Y} \in NRFM\right) \wedge \left(\overline{Y} \in NRFM\right)$ **then**
15: [ ] (MR_1)] Etude de $\overline{Y} \Rightarrow \overline{X \backslash Y}$; Etude de $\overline{X \backslash Y} \Rightarrow \overline{Y}$
16: **end if**
17: **end if**
18: **else if** $conf(X \backslash Y \Rightarrow Y) < sup(Y)$ **then**
19: **if** $\left(\overline{Y} \in NRFM\right)$ **then**
20: **if** $sup(X \backslash Y) \leqslant sup(\overline{Y})$ **then**
21: Etude de $X \backslash Y \Rightarrow \overline{Y}$; [ ] (MR_2)] Etude de $\overline{Y} \Rightarrow X \backslash Y$
22: **if** $\left(\overline{X \backslash Y} \in NRFM\right)$ **then**
23: [ ] (MR_1)] Etude de $\overline{X \backslash Y} \Rightarrow Y$; Etude de $Y \Rightarrow \overline{X \backslash Y}$
24: **end if**
25: **else if** $sup(X \backslash Y) > sup(\overline{Y})$ **then**
26: Etude de $\overline{Y} \Rightarrow X \backslash Y$; [ ] (MR_2)] Etude de $X \backslash Y \Rightarrow \overline{Y}$
27: **if** $\left(\overline{X \backslash Y} \in NRFM\right)$ **then**
28: [ ] (MR_1)] Etude de $Y \Rightarrow \overline{X \backslash Y}$; Etude de $\overline{X \backslash Y} \Rightarrow Y$
29: **end if**
30: **end if**
31: **else if** $\left(\overline{X \backslash Y} \in NRFM\right)$ **then**
32: **if** $sup(\overline{X \backslash Y}) \leqslant sup(Y)$ **then**
33: Etude de $\overline{X \backslash Y} \Rightarrow Y$; [ ] (MR_2)] Etude de $Y \Rightarrow \overline{X \backslash Y}$
34: **else if** $sup(\overline{X \backslash Y}) > sup(Y)$ **then**
35: Etude de $Y \Rightarrow \overline{X \backslash Y}$; [ ] (MR_2)] Etude de $\overline{X \backslash Y} \Rightarrow Y$
36: **end if**
37: **end if**
38: **end if**
39: Etude de $X \ddot{\backslash} Y \Rightarrow \ddot{Y}$
40: **end for**{conclusion Y}
41: **end for**{motif raisonnablement fréquent X}

Algorithm 2 : function $funct\_RF$ (*Recherche des motifs Raisonnablement Fréquents*)

Input : BD (*not null*), min_{sup} (*not null*), max_{sup} (*null*) et $min_{s\ddot{u}p}$ (*null*)
Output : RF (*ensemble des motifs Raisonnablement Fréquents*)

1: **if** max_{sup} est non défini **then** $max_{sup} = 1 - min_{sup}$
2: **if** $min_{s\ddot{u}p}$ est non défini **then** $min_{s\ddot{u}p} = min_{sup}$
3: $RF = \varnothing$ {initialisation}
4: $C_1 = \{i \in \mathcal{I}\}$ {ensemble des *1*-candidats}
5: **for** $(k = 1 \,;\, C_k \neq \varnothing \,;\, k++)$ **do**
6: $F_k = \varnothing$;
7: **for all** $X \in C_k$ **do**
8: $s = support(BD, X)$ {calcul du support de X}
9: $\ddot{s} = computeSupport(X, s, F)$ {calcul du support de $\ddot{X}$}
10: **if** $(min_{sup} \leqslant s) \wedge (min_{s\ddot{u}p} \leqslant \ddot{s})$ **then**
11: $F_k \leftarrow F_k \cup \{X\}$
12: **if** $s \leqslant max_{sup}$ **then**
13: $RF \leftarrow RF \cup \{(X, s, \ddot{s})\}$
14: **end if**
15: **end if**
16: **end for**
17: $CP_{k+1} = F_k \bowtie F_k$ {Génération des $(k+1)$-motifs candidats potentiels}
18: $C_{k+1} = candidat(CP_{k+1}, F)$ {Recherche des $(k+1)$-motifs candidats}
19: $F \leftarrow F \cup \{F_k\}$
20: **end for**
21: **return** RF

Algorithm 3 : function $funct\_NRFM$ (*Recherche des motifs Négatifs Raisonnablement Fréquents Minimaux*)

Input : BD et RF

Output : $NRFM$ (*ensemble des motifs Négatifs Raisonnablement Fréquents Minimaux*)

1: $NRFM = \{\overline{i} \in \overline{\mathcal{I}} \,/\, min_{sup} \leqslant sup(\,\overline{i}\,) \leqslant max_{sup}\}$
{items Négatifs Raisonnablement Fréquents Minimaux de taille 1}
2: $NNRF_1 = \overline{\mathcal{I}} \setminus NRFM - \{\overline{i} \in \overline{\mathcal{I}} \,/\, sup(\,\overline{i}\,) \geqslant max_{sup}\}$
{items Négatifs Non Raisonnablement Fréquents de taille 1}
3: $C_2 = NNRF_1 \bowtie NNRF_1$ {Candidats de taille 2}
4: **for** $(k = 2\,; C_k \neq \varnothing\,; k++)$ **do**
5: $NNRF_k = \varnothing$
6: **for all** $\overline{X} \in C_k$ **do**
7: **if** $min_{sup} \leqslant sup(\overline{X}) \leqslant max_{sup}$ **then**
8: $NRFM \leftarrow NRFM \cup \{\overline{X}\}$
9: **else**
10: **if** $sup(\overline{X}) \leqslant min_{sup}$ **then**
11: $NNRF_k \leftarrow NNRF_k \cup \{\overline{X}\}$
12: **end if**
13: **end if**
14: **end for**{$all\ \overline{X} \in C_k$}
15: $C_{k+1} = NNRF_k \bowtie NNRF_k$ {Génération des candidats}
16: **end for**{$(k = 2\,; C_k \neq \varnothing\,; k++)$}
17: **return** $NRFM$