Méta-règles pour la génération de règles négatives

Sylvie Guillaume* et Pierre-Antoine Papon**

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

Résumé. La littérature s'est beaucoup intéressée à l'extraction de règles classiques (*ou positives*) et peu à l'extraction des règles négatives en raison essentiellement d'une part, du coût de calculs et d'autre part, du nombre prohibitif de règles redondantes et inintéressantes extraites. La démarche que nous avons retenue est de dégager les règles négatives lors de l'extraction des règles positives, et pour cela, nous recherchons les règles négatives que l'on peut inférer ou pas à partir de la pertinence d'une règle positive. Ces différentes inférences vont être formalisées par un ensemble de méta-règles.

1 Introduction

L'extraction de règles d'association est une tâche importante en fouille de données. Ces algorithmes d'extraction ont essentiellement deux limites : les variables doivent être binaires et seules les règles positives sont extraites. L'importance de l'extraction des règles négatives fut mise en évidence par Brin et al. (1997). 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. Pour finir, beaucoup de règles redondantes et inintéressantes sont extraites. Plusieurs techniques ont été proposées.

Brin et al. (1997) utilisent le test du χ^2 pour déterminer la dépendance entre deux motifs et ensuite une mesure de corrélation afin de trouver la nature de cette dépendance (*positive ou négative*). Savasere et al. (1998) combinent les motifs fréquents [1] positifs avec la connaissance du domaine afin de détecter les dissociations (*ou associations négatives*). Boulicaut et al. (2000) recherchent les règles négatives du type $XY \rightarrow \overline{Z}$ ou $\overline{X}Y \rightarrow Z$ en proposant une approche basée sur les contraintes. Wu et al. (2004) et Antonie et Zaïane (2004) utilisent une mesure supplémentaire pour générer les règles positives et 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 [2]. Dans la lignée de cette dernière approche, nous souhaitons inférer les règles négatives à partir des règles positives mais dans un contexte plus large : celui des règles admettant des contre-exemples [3]. Ces différentes inférences vont être formalisées par un ensemble de méta-règles.

1. Un motif X est dit fréquent si sa probabilité d'apparition $P(X)$ est supérieure à un seuil fixé par l'utilisateur.
2. Une implication logique est une règle d'association avec une confiance de 100%.
3. Un contre-exemple est un individu qui vérifie la prémisse X de la règle mais qui ne vérifie pas la conclusion Y.

L'article s'organise donc de la façon suivante. La *section 2* met en évidence les règles négatives inintéressantes et celles qui sont potentiellement intéressantes en fonction de l'intérêt de la règle positive $X \rightarrow Y$. La *section 3* recherche parmi les règles potentiellement intéressantes, celles qui sont réellement intéressantes et celles qui ne le sont pas grâce à l'utilisation d'une mesure d'intérêt, la mesure M_G. L'article se termine par une conclusion et des perspectives.

2 Règles négatives potentiellement intéressantes

Une règle $X \rightarrow Y$ est dite valide Agrawal et Srikant (1994) lorsque le support [4] de la règle est supérieur à un seuil défini par l'utilisateur et la confiance [5] de la règle est supérieure à un autre seuil. Cependant une règle valide n'est pas toujours pertinente ou digne d'intérêt. Si la confiance de la règle a une valeur inférieure à la probabilité $P(Y)$, cette règle n'est pas digne d'intérêt puisque l'apparition de X diminue les chances d'apparition de Y. Il faut que la confiance de la règle soit également supérieure à $P(Y)$. Nous dirons qu'une règle est potentiellement intéressante lorsque $P(Y/X) > P(Y)$. Dans le cas contraire, nous parlerons de règles inintéressantes et nous la noterons $X \not\rightarrow Y$. De plus, une règle sera dite située dans la zone d'attraction si celle-ci est potentiellement intéressante. Dans le cas contraire, la règle sera dite située dans la zone de répulsion.

La *figure 1* restitue le résultat de cette étude qui détecte les règles négatives potentiellement intéressantes et les règles négatives inintéressantes en fonction de l'intérêt (*zones attractive ou répulsive*) de la règle positive. Lorsque la règle $X \rightarrow Y$ est dans la zone attractive, les règles $Y \rightarrow X$, $\overline{Y} \rightarrow \overline{X}$ et $\overline{X} \rightarrow \overline{Y}$ sont également dans cette zone et les règles $X \rightarrow \overline{Y}$, $Y \rightarrow \overline{X}$, $\overline{Y} \rightarrow X$ et $\overline{X} \rightarrow Y$ sont dans la zone répulsive. Lorsque la règle $X \rightarrow Y$ n'est pas potentiellement intéressante, les règles $Y \rightarrow X$, $\overline{Y} \rightarrow \overline{X}$ et $\overline{X} \rightarrow \overline{Y}$ ne le sont pas également alors que les règles $X \rightarrow \overline{Y}$, $Y \rightarrow \overline{X}$, $\overline{Y} \rightarrow X$ et $\overline{X} \rightarrow Y$ le sont. La justification de tous ces résultats est présente dans Guillaume et Papon (2011).

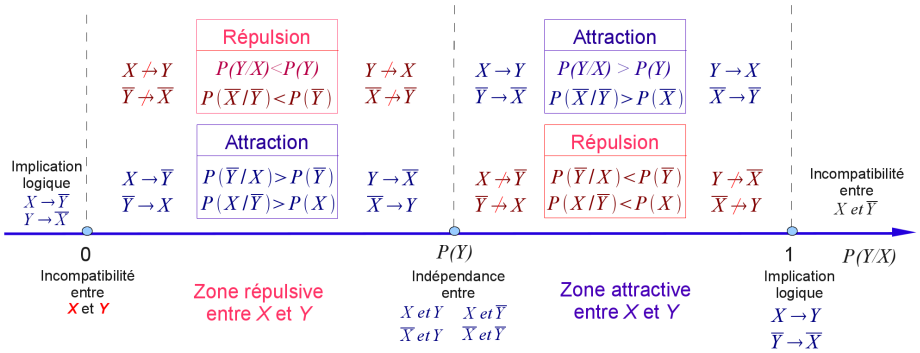

FIG. 1 – *Règles négatives potentiellement intéressantes et inintéressantes.*

Au vu de ces résultats, il paraît intéressant de connaître la zone d'appartenance de la règle $X \rightarrow Y$ afin d'en déduire les *4* règles inintéressantes. La confiance ne nous renseigne pas sur

4. Le support d'un motif X est la probabilité $P(X)$ d'apparition du motif.
5. La confiance d'une règle est la probabilité $P(Y/X)$ d'apparition de Y sachant X.

la zone d'appartenance de la règle $X \rightarrow Y$ et par conséquent une mesure supplémentaire est nécessaire. Dans la section suivante, nous exposons la mesure supplémentaire choisie qui va nous permettre d'inférer les règles négatives à partir de l'intérêt de la règle positive.

3 Recherche de méta-règles

La mesure choisie pour inférer les règles négatives est la mesure M_G Guillaume (2010) :

$$\underline{\text{Zone attractive } max(\tfrac{1}{2}, P(Y)) < P(Y/X)} : M_{G_a}(X \rightarrow Y) = \frac{P(Y/X) - max(P(Y), \frac{1}{2})}{max(P(Y), \frac{1}{2})}$$
$$\underline{\text{Zone répulsive } P(Y/X) < min(\tfrac{1}{2}, P(Y))} : M_{G_r}(X \rightarrow Y) = \frac{P(Y/X) - min(P(Y), \frac{1}{2})}{min(P(Y), \frac{1}{2})}$$
$$\underline{\text{Zone inintéressante } min(\tfrac{1}{2}, P(Y)) \leq P(Y/X) \leq max(\tfrac{1}{2}, P(Y))} : M_{G_i}(X \rightarrow Y) = 0$$

Cette mesure permet de connaître la zone d'appartenance de la règle. Elle présente en plus l'avantage d'affiner ces deux zones. La zone attractive est la zone comprise entre (soit l'indépendance [6], soit l'indétermination [7]) et l'implication logique [8]. L'état (*indépendance ou indétermination*) qui sera retenu sera celui dont la confiance est maximale. Cette nouvelle détermination de la zone attractive permet d'éliminer les règles où $P(Y) < \frac{1}{2}$. La zone répulsive est la zone comprise entre l'incompatibilité [9] et (soit l'indépendance, soit l'indétermination). L'état qui sera retenu sera celui dont la confiance est minimale. Cette nouvelle détermination de la zone répulsive va éliminer les règles où $P(\overline{Y}) < \frac{1}{2}$. Pour la suite, nous retiendrons ces nouvelles définitions pour les deux zones. Une nouvelle zone apparaît lors de la redéfinition des deux zones précédentes : c'est la zone comprise entre l'indépendance et l'indétermination, zone inintéressante. Nous exposons maintenant l'inférence des différentes règles.

Méta-règles pour déduire les règles symétriques $Y \rightarrow X$

Nous avons la relation suivante entre les règles symétriques : $P(X/Y) = \frac{P(X)}{P(Y)} P(Y/X)$. Si nous faisons l'hypothèse que $P(X) < P(Y)$, nous avons donc $P(X/Y) < P(Y/X)$. Ainsi, si la règle $X \rightarrow Y$ est jugée non pertinente (*i.e. potentiellement intéressante mais pas assez proche de l'implication logique*), nous avons $P(Y/X) < min(\frac{1}{2}, P(Y))$ et par conséquent $P(X/Y) < min(\frac{1}{2}, P(Y))$, donc la règle $Y \rightarrow X$ est également non pertinente, ce qui nous permet d'en déduire la première méta-règle MR_1 suivante :

$$\boxed{(MR_1) : \forall X \rightarrow Y \text{ avec } P(X) < P(Y), X \not\rightarrow Y \Longrightarrow Y \not\rightarrow X.}$$

La contraposée de (MR_1) nous permet d'obtenir la méta-règle (MRC_1) suivante :

$$\boxed{(MRC_1) : \forall X \rightarrow Y \text{ avec } P(X) > P(Y), X \rightarrow Y \Longrightarrow Y \rightarrow X.}$$

Méta-règles pour déduire les règles $\overline{Y} \rightarrow \overline{X}$

Trois cas sont à envisager selon le positionnement de $P(X)$ et $P(Y)$ par rapport à $\frac{1}{2}$. Ce positionnement va permettre de savoir si les règles $X \rightarrow Y$ et $\overline{Y} \rightarrow \overline{X}$ ont leur point

6. Cas où la probabilité $P(Y/X)$ est égale à $P(Y)$.
7. Le point d'équilibre ou d'indétermination (Blanchard et al., 2005) est le cas où le nombre d'exemples est égal au nombre de contre-exemples, ou encore $P(Y/X) = \frac{1}{2}$ et $P(\overline{Y}/X) = \frac{1}{2}$.
8. Cas où la probabilité conditionnelle est égale à 1.
9. Cas où la probabilité conditionnelle $P(Y/X)$ est égale à 0.

d'indépendance avant ou après le point d'équilibre. En effet, si $P(Y) < \frac{1}{2}$ alors nous pouvons en déduire que la règle $X \rightarrow Y$ a son point d'équilibre après l'indépendance. Dans le cas contraire, le point d'équilibre de la règle sera avant l'indépendance.

Cas 1 : $P(X) < P(Y) < \frac{1}{2}$

Dans ce cas, l'équilibre pour la règle $X \rightarrow Y$ est après l'indépendance et l'équilibre pour $\overline{Y} \rightarrow \overline{X}$ est avant l'indépendance. La courbe de gauche de la *figure 2* restitue les courbes d'évolution de M_G pour ces règles dans la zone attractive uniquement. Comme la courbe

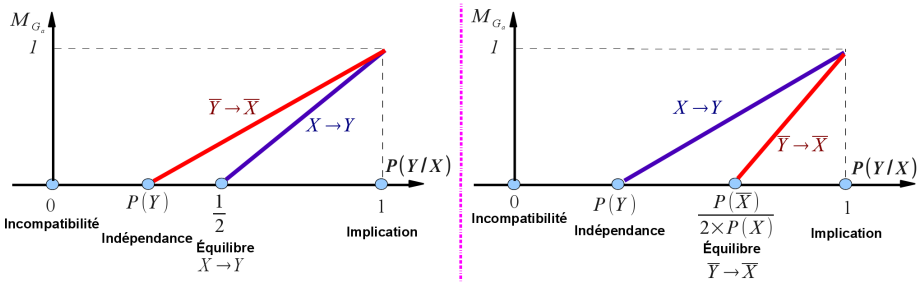

FIG. 2 – *Courbes d'évolution de M_G - cas 1 : courbe gauche, cas 2 : courbe droite.*

d'évolution de la règle $\overline{Y} \rightarrow \overline{X}$ est au dessus de celle de $X \rightarrow Y$, nous pouvons en déduire que : $\forall X \rightarrow Y$ avec $P(X) < P(Y) < \frac{1}{2}, X \rightarrow Y \Longrightarrow \overline{Y} \rightarrow \overline{X}$.

Cas 2 : $\frac{1}{2} < P(X) < P(Y)$

L'équilibre pour la règle $X \rightarrow Y$ est avant l'indépendance et l'équilibre pour $\overline{Y} \rightarrow \overline{X}$ est après l'indépendance. La courbe de droite de la *figure 2* restitue les courbes d'évolution de M_G. Comme la courbe d'évolution de la règle $X \rightarrow Y$ est au dessus de celle de $\overline{Y} \rightarrow \overline{X}$, nous pouvons en déduire que : $\forall X \rightarrow Y$ avec $\frac{1}{2} < P(X) < P(Y), X \not\rightarrow Y \Longrightarrow \overline{Y} \not\rightarrow \overline{X}$.

Cas 3 : $P(X) < \frac{1}{2} < P(Y)$

Dans ce cas, l'équilibre pour les règles $X \rightarrow Y$ et $\overline{Y} \rightarrow \overline{X}$ est avant l'indépendance. Comme le point d'indépendance est le même pour les deux types de règles, les courbes d'évolution de la mesure M_G sont donc confondues dans la zone attractive.

Nous pouvons donc modifier les deux méta-règles précédentes afin de prendre en compte ce *cas 3*, méta-règles que nous nommerons respectivement (MR_2) et (MR_3).

$(MR_2) : \forall X \rightarrow Y$ / $P(X) < P(Y) < \frac{1}{2}$ ou $P(X) < \frac{1}{2} < P(Y), X \rightarrow Y \Longrightarrow \overline{Y} \rightarrow \overline{X}$.

$(MR_3) : \forall X \rightarrow Y$ / $\frac{1}{2} < P(X) < P(Y)$ ou $P(X) < \frac{1}{2} < P(Y), X \not\rightarrow Y \Longrightarrow \overline{Y} \not\rightarrow \overline{X}$.

Méta-règles pour déduire les règles $\overline{X} \rightarrow \overline{Y}$

Cas 1 : $P(X) < P(Y) < \frac{1}{2}$

Dans ce cas, l'équilibre pour la règle $X \rightarrow Y$ est après l'indépendance et l'équilibre pour $\overline{X} \rightarrow \overline{Y}$ est avant l'indépendance. La courbe de gauche de la *figure 3* restitue les courbes d'évolution de M_G pour ces règles. Nous ne pouvons donc faire aucune déduction.

Cas 2 : $\frac{1}{2} < P(X) < P(Y)$

Nous pouvons appliquer (MR_3) puis (MR_1) pour en déduire la méta-règle : $\forall X \rightarrow Y$ avec $\frac{1}{2} < P(X) < P(Y), X \not\rightarrow Y \Longrightarrow \overline{X} \not\rightarrow \overline{Y}$. La contraposée de cette méta-règle est :

$\forall X \rightarrow Y$ avec $\frac{1}{2} < P(X) < P(Y)$, $\overline{X} \rightarrow \overline{Y} \Longrightarrow X \rightarrow Y$ que nous pouvons transformer en : $\forall X \rightarrow Y$ avec $P(Y) < P(X) < \frac{1}{2}$, $X \rightarrow Y \Longrightarrow \overline{X} \rightarrow \overline{Y}$.

Cas 3 : $P(X) < \frac{1}{2} < P(Y)$

L'équilibre pour la règle $X \rightarrow Y$ est avant l'indépendance et l'équilibre pour la règle $\overline{X} \rightarrow \overline{Y}$ est après l'indépendance. La courbe de droite de la *figure 3* restitue les courbes d'évolution de M_G pour ces règles.

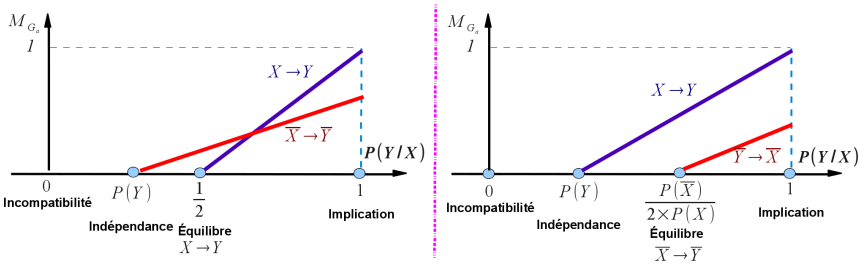

FIG. 3 – *Courbes d'évolution de M_G - cas 1 : courbe gauche, cas 3 : courbe droite.*

Comme la courbe d'évolution de la règle $X \rightarrow Y$ est au-dessus de celle de $\overline{X} \rightarrow \overline{Y}$, nous avons : $\forall X \rightarrow Y$ avec $P(X) < \frac{1}{2} < P(Y)$, $X \not\rightarrow Y \Longrightarrow \overline{X} \not\rightarrow \overline{Y}$. La contraposée de cette méta-règle conduit à : $\forall X \rightarrow Y$ avec $P(X) < \frac{1}{2} < P(Y)$, $\overline{X} \rightarrow \overline{Y} \Longrightarrow X \rightarrow Y$. En la modifiant, nous avons : $\forall X \rightarrow Y$ avec $P(Y) < \frac{1}{2} < P(X)$, $X \rightarrow Y \Longrightarrow \overline{X} \rightarrow \overline{Y}$.

Synthèse : Nous pouvons en déduire les méta-règles (MR_4) et (MRC_4) suivantes :

$(MR_4) : \forall X \rightarrow Y$ / $\frac{1}{2} < P(X) < P(Y)$ ou $P(X) < \frac{1}{2} < P(Y)$, $X \not\rightarrow Y \Longrightarrow \overline{X} \not\rightarrow \overline{Y}$.

$(MRC_4) : \forall X \rightarrow Y$ / $P(Y) < P(X) < \frac{1}{2}$ ou $P(Y) < \frac{1}{2} < P(X)$, $X \rightarrow Y \Longrightarrow \overline{X} \rightarrow \overline{Y}$.

Méta-règles pour déduire les règles antinomiques $X \rightarrow \overline{Y}$

Nous avons la relation suivante entre les règles antinomiques : $M_{G_a}(X \rightarrow \overline{Y}) = -M_{G_r}(X \rightarrow Y)$, $M_{G_i}(X \rightarrow \overline{Y}) = -M_{G_i}(X \rightarrow Y)$ et $M_{G_r}(X \rightarrow \overline{Y}) = -M_{G_a}(X \rightarrow Y)$.

Nous pouvons en déduire les quatre méta-règles suivantes :

$(MR_5) : X \rightarrow Y \Longrightarrow X \not\rightarrow \overline{Y}$.

$(MR_6) : X \rightarrow \overline{Y} \Longrightarrow X \not\rightarrow Y$.

$(MR_7) : X \not\rightarrow Y \Longrightarrow X \not\rightarrow \overline{Y}$.

$(MR_8) : X \not\rightarrow \overline{Y} \Longrightarrow X \not\rightarrow Y$.

4 Conclusion

Notre objectif est de générer les règles négatives à partir de l'extraction des règles d'association positives. Pour cela, nous avons recherché les règles négatives potentiellement intéressantes et celles qui ne le sont pas en fonction de l'intérêt de la règle positive. Ensuite, par l'utilisation d'une mesure d'intérêt appropriée, nous avons recherché parmi les règles négatives potentiellement intéressantes, celles qui l'étaient réellement et celles qui ne pouvaient

pas l'être. Ces 8 inférences ont été dégagées dans le but de trouver un algorithme efficace d'extraction de règles négatives, suite de notre travail.

Références

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

Antonie, M. et O. Zaïane (2004). Mining positive and negative association rules: An approach for confined rules. In *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 27–38.

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.

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

Guillaume, S. et P. Papon (2011). Méta-règles pour la génération de règles négatives. Technical report, LIMOS, RR-11-04, 12 pages.

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.

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.

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

Summary

The literature has been heavily involved in the extraction of classic rules (*or positive*) and few in negative rules extraction owing to essentially on the one hand, calculations cost and on the other hand, the prohibitive number of extracted redundant and uninteresting rules. The approach which we have adopted is to identify negative rules during positive rules extraction, and for that, we look for the negative rules that we can be inferred or not from the interest of a positive rule.