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

Sylvie Guillaume* et Pierre-Antoine Papon**

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

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

1 Introduction

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

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