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