Étude de l'interaction entre variables pour l'extraction des règles d'influence

L. Nemmiche Alachaher* et S. Guillaume*

*LIMOS, UBP UMR 6158 CNRS
Complexe des Cézeaux
63177 AUBIERE Cedex - France
{nemmiche, sylvie.guillaume}@isima.fr

Résumé. Cet article présente une méthode efficace pour l'extraction de règles d'influence quantitatives positives et négatives. Ces règles d'influence introduisent une nouvelle sémantique qui vise à faciliter l'analyse d'un volume important de données. Cette sémantique fixe la direction de la règle entre deux variables en positionnant, au préalable, l'une comme étant l'*influent* et l'autre comme étant l'*influé*. Elle permet, de ce fait, d'exprimer la nature de l'influence : *positive*, en maximisant le nombre d'éléments en commun ou *négative*, en maximisant le nombre d'éléments qui violent l'influé.
Notre approche s'appuie sur une stratégie qui comporte cinq étapes dont deux exécutées en parallèle. Ces deux étapes constituent les étapes clé de notre approche. La première combine une méthode d'élagage et de regroupement tabulaire basée sur les tableaux de contingence. Cette dernière construit et classe les zones potentiellement intéressantes. La seconde, injecte la sémantique et évalue le degré d'influence que produirait l'introduction d'une nouvelle variable sur un ensemble de variables en utilisant une nouvelle mesure d'intérêt, l'*Influence*. Cette étape vient affiner les résultats de la première étape, et permet de se focaliser sur des zones valides par rapport aux contraintes spécifiées. Enfin, un système de règles d'influence jugées intéressantes est construit basé sur la juxtaposition des résultats des deux étapes clé de notre approche.

1 Introduction

L'extraction de connaissances est un processus qui permet d'analyser une masses de données importante afin d'en extraire des connaissances nouvelles, valides et utiles. Ces connaissances sont ensuite présentées sous différentes formes notamment sous forme de règles d'association. Une règle d'association (RA) (Agrawal et al. (1993)) est une implication de la forme $C_1 \rightarrow C_2$, où C_1 et C_2 sont des conditions C sur les attributs de la base. Soient $minsup$ et $minconf$ des seuils prédéfinis. Une RA est dite forte si elle satisfait deux contraintes :

- son support $supp(C) \geq minsup$, avec $supp(C)$: nombre de transactions dans la base qui satisfont l'ensemble des conditions C tel que $supp(C_1 \rightarrow C_2) = supp(C_1 \wedge C_2)$;
- sa confiance $conf(C_1 \rightarrow C_2) \geq minconf$, avec $conf(C_1 \rightarrow C_2) = \frac{supp(C_1 \rightarrow C_2)}{supp(C_1)}$.