Contrôle du risque multiple pour la sélection de règles d'association significatives

Stéphane Lallich*, Elie Prudhomme*, Olivier Teytaud**

*Laboratoire E.R.I.C, Université Lumière Lyon 2
5, avenue Pierre Mendès-France, 69676 BRON Cedex – France
stephane.lallich@univ-lyon2.fr, Elie.Prudhomme@etu.univ-lyon2.fr

**Artelys
215 avenue Jean-Jacques Rousseau, 92136 Issy-les-Moulineaux
olivier.teytaud@artelys.com

Résumé. Les algorithmes d'extraction de règles d'association parcourent efficacement le treillis des itemsets pour constituer une base de règles admissibles à des seuils de support et de confiance, mais donnent une multitude de règles peu exploitables. Nous suggérons d'épurer de telles bases en éliminant les règles non statistiquement significatives. La multitude de tests pratiqués conduit mécaniquement à multiplier les règles sélectionnées à tort. Après avoir présenté des procédures issues de la biostatistique qui contrôlent non pas le risque, mais le nombre de fausses découvertes, nous proposons BS_FD, un algorithme original fondé sur le bootstrap qui sélectionne les règles significatives en contrôlant le nombre de fausses découvertes. Des expérimentations montrent l'efficacité de ces procédures.
Mots-clefs : Règle d'association, qualité, contrôle du risque multiple.

1 Admissibilité, intérêt et signification statistique

La recherche des règles d'association intéressantes est un problème classique de l'Extraction des Connaissances à partir des Données à la suite des travaux de [Agrawal et al., 1993] dans le cadre des bases de données transactionnelles. Dans une telle base, un enregistrement est une transaction et les champs correspondent aux articles disponibles. On note n le nombre de transactions et p le nombre d'articles. L'acte d'achat (item) associé à chaque article est une variable booléenne. Sur l'ensemble des transactions, on a une matrice booléenne X, de dimensions n et p. La conjonction des actes d'achat (*itemset*) associés à un ensemble d'articles est vue comme une variable booléenne.

A partir de la matrice booléenne X, on veut extraire des règles du type "si un client achète du pain et du fromage, alors probablement il achète aussi du vin". Une règle d'association est une expression r du type $A \rightarrow B$, où l'antécédent A et le conséquent B sont des itemsets qui n'ont pas d'items communs. On note n_a et n_b les nombres de transactions qui réalisent respectivement les items de A et de B, n_{ab} le nombre de celles qui réalisent à la fois A et B. Les proportions correspondantes sont désignées par p_a, p_b et p_{ab}. Ce formalisme se généralise à toute base de données dont on a extrait une table booléenne cas-attributs.