Évaluation de la Résistance au Bruit de quelques Mesures Quantitatives

Jérôme Azé*, Sylvie Guillaume**, Philippe Castagliola***

*CNRS, LRI - Université Paris Sud - 91405 Orsay Cedex
Jerome.Aze@lri.fr
http://www.lri.fr/~aze
**Laboratoire LIMOS, UMR 6158 CNRS
Université Blaise Pascal - Complexe scientifique des Cézeaux
63177 Aubiere Cedex
Sylvie.Guillaume@isima.fr
***École des Mines de Nantes / IRCCyN
4, rue Alfred Kastler - 44307 Nantes Cedex 3
Philippe.Castagliola@emn.fr

Résumé. L'extraction de connaissances dans des données réelles est difficile car les données sont rarement parfaites. L'étude de l'impact du bruit contenu dans les données sur la qualité des résultats obtenus permet de mieux comprendre le comportement des mesures de qualité. Dans cet article, nous présentons différentes mesures quantitatives permettant d'extraire des connaissances dans les données. Pour chacune de ces mesures, une étude empirique de l'impact d'un bruit relativement réaliste sur une base de données bancaires est réalisée. Le comportement des différentes mesures en présence des données bruitées permet d'établir un critère de qualité supplémentaire. Ce nouveau critère lié à la sensibilité des mesures aux données bruitées permet de mieux contrôler le choix des mesures lors du processus d'extraction des connaissances.

1 Introduction

De nombreuses études (Agrawal et al. 1996 ; Mannila et al. 1994 ; Srikant et Agrawal 1996 ; Park et al. 1995 ; Pasquier 2000) ont été réalisées sur la recherche d'algorithmes efficaces d'extraction de règles d'association pour des données discrètes. Pour les autres types de données, un codage disjonctif complet, précédé pour les variables quantitatives d'une discrétisation, est nécessaire afin d'utiliser ces algorithmes (Srikant et Agrawal 1996). Cette transformation des données a deux inconvénients, tout d'abord une augmentation du nombre de règles extraites et ensuite une perte de la structure d'ordre des variables ordinales (variables qualitatives ordinales et variables quantitatives). Afin de remédier à ce problème et de pouvoir extraire des associations directement sur les variables quantitatives sans avoir à les transformer, une étude de différentes mesures quantitatives (coefficient de corrélation linéaire, mesure de vraisemblance du lien, intensité d'implication de A. Larher, intensité de propension et intensité d'inclination) a été effectuée (Guillaume et Castagliola 2003). Cette étude s'est intéressée à deux problématiques, tout d'abord au comportement de ces mesures dans plusieurs situa-