Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel TIC

Julien Blanchard, Fabrice Guillet, Régis Gras, Henri Briand

IRIN – Ecole polytechnique de l'université de Nantes
La Chantrerie – BP 50609 – 44306 Nantes cedex 3
{julien.blanchard, fabrice.guillet, regis.gras, henri.briand}@polytech.univ-nantes.fr

Résumé. La validation des connaissances est l'une des étapes les plus problématiques d'un processus de découverte de règles d'association. Pour que le décideur (expert des données) puisse trouver des connaissances intéressantes dans les grandes quantités de règles produites par les algorithmes de fouille de données, il est nécessaire de mesurer la qualité des règles. Nous insérant dans le cadre de l'analyse statistique implicative, nous proposons dans cet article d'évaluer les règles en considérant leur contenu informationnel à travers un nouvel indice de qualité fondé sur l'entropie de Shannon : *TIC* (*Taux Informationnel modulé par la Contraposée*). Cet indice a l'avantage d'être bien adapté à la sémantique des règles, puisque d'une part il respecte leur caractère asymétrique et d'autre part il tire profit de leurs contraposées. Par ailleurs, c'est à notre connaissance la seule mesure de qualité de règles qui intègre à la fois indépendance et déséquilibre, c'est-à-dire qui permette de rejeter simultanément les règles entre variables corrélées négativement et les règles qui possèdent plus de contre-exemples que d'exemples. Des comparaisons de *TIC* avec la J-mesure, l'information mutuelle, l'indice de Gini, et la confiance sont réalisées sur des simulations numériques.

1. Introduction

Parmi les modèles de connaissances utilisés en Extraction de Connaissances dans les Données (ECD), les règles d'association [Agrawal *et al.*, 1993] sont devenues un concept majeur qui a donné lieu à de nombreux travaux de recherche. Ces règles sont des tendances implicatives de la forme $a \rightarrow b$ entre les attributs d'une base de données (variables booléennes dénommées items). Une telle règle signifie que la plupart des enregistrements qui vérifient la prémisse a dans la base de données vérifient aussi la conclusion b. L'une des étapes les plus problématiques d'un processus de découverte de règles d'association est la validation des règles après leur extraction. Les algorithmes de *data mining* peuvent en effet produire d'énormes quantités de règles, ce qui empêche le décideur (expert des données étudiées) de pouvoir exploiter les résultats directement à la sortie des algorithmes. Ce problème est lié à la nature non supervisée de la découverte de règles : le décideur n'explicite pas ses buts et ne spécifie aucune variable endogène. Le nombre de conjonctions d'items manipulées par les algorithmes devient alors prohibitif.

Pour assister le décideur dans sa recherche des connaissances intéressantes pour la prise de décision, il est nécessaire de mesurer la qualité des règles extraites. Dans l'importante littérature consacrée à l'évaluation de cette notion complexe de qualité, les mesures sont souvent classées en deux catégories : les subjectives (orientées décideur) et les objectives