Une comparaison de certains indices de pertinence des règles d'association

Marie Plasse* **, Ndeye Niang*
Gilbert Saporta*, Laurent Leblond**

* CNAM Laboratoire CEDRIC 292 Rue St Martin Case 441 Paris Cedex 03
niang@cnam.fr, saporta@cnam.fr
** PSA Peugeot Citroën 45 rue Jean-Pierre Timbaud 78307 Poissy Cedex
marie.plasse@mpsa.com, laurent.leblond1@mpsa.com

Résumé. Cet article propose une comparaison graphique de certains indices de pertinence pour évaluer l'intérêt des règles d'association. Nous nous sommes appuyés sur une étude existante pour sélectionner quelques indices auxquels nous avons ajouté l'indice de Jaccard et l'indice d'accords désaccords (IAD). Ces deux derniers nous semblent plus adaptés pour discriminer les règles intéressantes dans le cas où les items sont des événements peu fréquents. Une application est réalisée sur des données réelles issues du secteur automobile.

1 Introduction

Notre étude a été motivée par le problème suivant : nous disposons de données concernant plusieurs dizaines de milliers d'individus décrits par quelques milliers d'attributs binaires assez rares et nous recherchons les éventuels liens entre certains attributs ou groupes d'attributs. La similitude de nos données avec des données de transactions nous a naturellement amenés à utiliser un algorithme de recherche de règles d'association. Cependant, le nombre élevé d'attributs conjugué à leur rareté conduit à un très grand nombre de règles dont les supports sont très faibles et les confiances très élevées. C'est pourquoi nous avons cherché à compléter l'approche support-confiance pour extraire les règles les plus pertinentes. De nombreux indices ont été proposés dans la littérature pour évaluer l'intérêt des règles d'association. Quelques uns font l'objet d'une analyse graphique à l'aide de courbes de niveaux. Nous exposons ensuite une application sur données industrielles.

2 Contexte

Ce travail est issu d'un projet industriel où l'objectif est d'exploiter une partie de l'informationnel d'un grand constructeur automobile afin d'extraire de nouvelles connaissances. Les données, issues du process de fabrication des véhicules, sont sous la forme d'une matrice où chaque véhicule est décrit par la présence ou l'absence d'attributs binaires. La connaissance d'éventuelles corrélations entre certains attributs ou groupes d'attributs représente un avantage non négligeable pour le constructeur automobile qui met un point d'honneur à améliorer