Classification de variables et classification croisée utilisées préalablement à la recherche de règles d'association

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

* CNAM Laboratoire CEDRIC 292 Rue St Martin Case 441 Paris Cedex 03
niang@cnam.fr
saporta@cnam.fr
** PSA Peugeot Citroën, Zone d'Activité Louis Bréguet, 78943 Vélizy Villacoublay
marie.plasse@mpsa.com
alexandre.villeminot@mpsa.com
laurent.leblond1@mpsa.com

Résumé. La recherche de règles d'association conduit souvent à l'obtention d'un très grand nombre de règles, alors inexploitables. De plus, il est parfois difficile de faire varier les paramètres d'extraction des règles : le support et la confiance. En effet, dans le cas où les variables sont des événements rares, il est nécessaire de choisir des seuils de support très faibles. Nous avons proposé d'utiliser de manière conjointe la classification de variables et la recherche de règles d'association. La classification préalable des variables permet de construire des groupes homogènes où les variables sont liées. La recherche de règles à l'intérieur de chaque groupe conduit à réduire le nombre de règles à analyser. Les techniques de classification croisée permettent un double partitionnement sur les variables et sur les individus. Nous souhaitons étudier les apports d'une telle classification à notre approche. Cet article présente une comparaison des deux types de classification utilisés préalablement à la recherche de règles d'association. Nous présentons les résultats obtenus sur plusieurs échantillons de données issues de l'industrie automobile.

1 Introduction

Ce travail a pour objectif la découverte d'éventuels liens entre variables ou groupes de variables binaires représentant des événements rares au sein d'une base de données industrielle de taille importante. Plusieurs dizaines de milliers d'individus sont décrits par la présence ou l'absence de plusieurs milliers d'attributs. Les données pouvant se présenter sous la forme d'un tableau de données de transactions, une idée naturelle consiste à utiliser la méthode de recherche de règles d'association. Cependant, le nombre élevé de variables conjugué à la rareté des occurrences conduit à un très grand nombre de règles dont les supports sont très faibles et les confiances très élevées.

Nous avons proposé de réaliser une classification préalable des variables afin de construire des groupes homogènes d'attributs à l'intérieur desquels la recherche de règles d'association est plus pertinente. Cette approche appliquée à nos données nous a permis de diminuer de manière très significative le nombre et la complexité des règles obtenues.

Les techniques de classification croisée suscitent de plus en plus d'intérêt, notamment en bioinformatique où le nombre de variables est souvent très important. Ces méthodes conduisent, par permutation des lignes et des colonnes de la matrice initiale, à des blocs homogènes de données où les individus ont des profils semblables au regard des variables qui les décri-