Evaluation d'une approche de classification conceptuelle

Marie Chavent *, Yves Lechevallier **

* Mathématiques Appliquées de Bordeaux, UMR 5466 CNRS
Université Bordeaux1, 351, Cours de la libération, 33405 Talence Cedex
chavent@math.u-bordeaux1.fr,
http://www.math.u-bordeaux1.fr/ chavent
**Institut National de Recherche en Informatique et en Automatique,
Domaine de Voluceau-Rocquencourt B.P.105, 78153 Le Chesnay Cedex
Yves.Lechevallier@inria.fr
http://www-rocq.inria.fr/axis/personnel/Yves.Lechevallier/yves.html

Résumé. L'objectif de ce travail est d'évaluer la perte d'information au sens de l'inertie entre des méthodes de partitionnement ou de classification hiérarchiques et une approche de classification conceptuelle. Nous voulons répondre à la question suivante : l'aspect simpliste du processus monothétique d'une méthode conceptuelle implique-t-il des partitions de moins bonne qualité au sens du critère de l'inertie ? Nous proposons de réaliser cette expérience sur 6 bases de l'UCI, trois de ces bases sont des tableaux de données quantitatives, les trois autres sont des tableaux de données qualitatives.

1 Introduction

Les méthodes conceptuelles de classification imposent une description monothétique des classes, cette contrainte forte devant a priori engendrer une perte de la qualité des partitions au sens des critères optimisés par les méthodes de partitionnement. Pour étudier ce phénomène nous avons comparé une méthode de classification conceptuelle appelée DIVCLUS-T qui est une méthode de classification descendante hiérarchique monothétique avec la méthode ascendante hiérarchique WARD et la méthode de partitionnement des k-means. Cette méthode conceptuelle s'applique à des données quantitatives et des données qualitatives. Cette comparaison a pu être effectuée car cette méthode conceptuelle optimise le même critère que les méthodes de WARD et des k-means. Ainsi comme WARD et les k-means, elle est basée sur la minimisation de l'inertie des classes mais à la différence de WARD et des k-means elle fournit par construction une interprétation simple et naturelle des classes. La question à laquelle nous allons chercher à répondre est la suivante : quel est le prix payé, en terme d'inertie, pour cette interprétation sous forme de règles des classes ?

Avant de présenter un peu plus en détail la méthode DIVCLUS-T et de comparer à partir de 6 bases de l'UCI les performances en terme d'inertie de DIVCLUS-T avec WARD et les k-means, nous donnons un petit exemple introductif. Il s'agit des données "protéines" de Hand et al. (1994) qui est un exemple classique de données quantitatives où 25 pays européens sont décrits en fonction de leur consommation en protéines dans neuf types d'aliments (viande