Outil de classification et de visualisation de grands volumes de données mixtes

Christophe CANDILLIER*, Noureddine MOUADDIB**

*Entreprise GÉOBS SA, 8 avenue des Thébaudières, 44800 SAINT-HERBLAIN
christophe.candillier@lina.univ-nantes.fr
http://c.candillier.free.fr/
**LINA (Laboratoire d'Informatique de Nantes Atlantique)
2 rue de la Houssinière, 44322 Nantes cedex 3
mouaddib@lina.univ-nantes.fr
http://www.sciences.univ-nantes.fr/lina/

Résumé. Nous avons conçu un outil de classification de données original que nous détaillons dans le présent article. Cet outil comporte un module de création de résumés et un module d'affichage. Le module de création de résumés prend en charge les données mixtes (qualitatives et quantitatives) ainsi que les grands volumes de données en utilisant une méthode de classification incrémentale et agglomérative originale. Le module de visualisation permet une lecture aisée des résumés grâce à une interface graphique évoluée permettant la présentation et l'exploration des résumés sous forme d'une hiérarchie de profils ou d'un tableau de profils. Chaque profil donne de manière claire les informations importantes relatives au résumé de données correspondant. La lecture de la hiérarchie et du tableau est aussi grandement facilitée par le choix d'un ordre optimal pour la présentation des variables et des résumés.

1 Introduction

Nous discuterons d'abord de l'algorithme de classification utilisé, de ses avantages et de ses inconvénients. Nous nous intéresserons ensuite à la visualisation des résumés produits, cela comprendra le calcul de l'ordre optimal des résumés et des variables ainsi que la visualisation sous la forme d'une hiérarchie des profils des résumés et sous la forme d'un tableau de profils. Finalement, nous illustrerons le fonctionnement de l'outil par l'analyse des données socioprofessionnelles de Paris et sa petite couronne.

2 Outil de Classification

2.1 Préliminaires

Les outils de classification sont divers et variés, ils ont pour but de regrouper les individus les plus semblables dans une même classe (Jain et al. 1999, Berkin 2002). Les deux principales familles sont les méthodes par partitionnement et les méthodes hiérarchiques. Les premières construisent directement les partitions et cherchent ensuite à les améliorer. Les dernières peuvent être scindées entre les méthodes par agglomération qui créent une