Élagage et aide à l'interprétation symbolique et graphique d'une pyramide

Kutluhan Kemal Pak, Mohamed Cherif Rahal, Edwin Diday

CEREMADE – Université Paris Dauphine
Place du Maréchal de Lattre de Tassigny
75775 Paris cedex 16
{Pak, Rahal, Diday}@ceremade.dauphine.fr
www.ceremade.dauphine.fr

Résumé : Le but de ce travail est de faciliter l'interprétation d'une classification pyramidale construite sur un tableau de données symboliques. Alors que dans une hiérarchie binaire le nombre de paliers est égal à ***n-1***, si ***n*** est le nombre d'individus à classer, dans le cas d'une pyramide ce dernier peut atteindre ***n(n-1)/2.*** Afin de réduire ce nombre, on élague la pyramide et on utilise un critère de sélection de paliers basé sur la hauteur. De plus on décrit tous les paliers retenus par des variables que l'on sélectionne également en utilisant "**le degré de généralité**" ainsi que des mesures de dissimilarités de type symbolique-numérique. L'aide à l'interprétation se sert d'outils graphiques et interactifs grâce à la bibliothèque OpenGL. Enfin une simulation montre comment évoluent ces sélections quand le nombre de classes et de variables croit.

Mots clés. Classification pyramidale. Classification hiérarchique. Données symboliques. Élagage d'une pyramide. Sélection de variables. Sélection de classes et description. Interprétation d'une classification.

1. Introduction

La classification automatique a pour but la recherche de groupes homogènes, selon un critère bien déterminé, la proximité entre les objets à classer par exemple. Les méthodes de classification automatique sont généralement applicables sur des ensembles de données ou d'objets décrits par des attributs, les habitants d'une ville, les patients d'un service médical… etc. Chaque méthode de classification a ses propres objectifs et sa propre représentation : Arbre, Graphe, Groupement sous forme d'ensembles (Voir (Jain et Dubes (1988))).

Dans le cas de la classification ascendante pyramidale (CAP) qui a été proposée par (Diday 1984), puis développée par (Bertrand (1986)), (Brito (1991)), (Mfoumoune (1998)), (Rodriguez (2000)), (Pak (2004)), et (Rahal (2004)) généralisant la classification ascendante hiérarchique (CAH) (Benzécri (1973)). Il en résulte qu'une représentation en groupes "non disjoints″ et emboîtés d'une pyramide est plus fidèle et riche en information par rapport aux données initiales qu'une représentation de type hiérarchique . Rappelons qu'une pyramide ***P*** construite sur un ensemble *E = {1,2, …,n}* est un ensemble fini de sous-ensembles non vides *{A, B, ….}, (A,B… Í E)* tel que : 1) *E*∈ *P* (le plus grand palier de la pyramide contient tous les individus), 2)Tous les singletons *{1},{2},…{n}* appartiennent à *P 3) " A,B* deux classes de la pyramide ***P*** on a soit *A ÇB = Æ ou A Ç B* Ì *P.* 4) ∃ un ordre ***q*** compatible avec *P.* Si on définit un index *f(A)* ³ *0* pour chaque classe *A* de *P* tel que *f* est isotonique sur *P : f(A)* £ *f (B)*