Une Méthode de Catégorisation Multi-Echelle

A Multi-Scale Clustering Algorithm

Arnaud Ribert, Abdel Ennaji, Yves Lecourtier

P.S.I. Faculté des Sciences, Université de Rouen
76 821 Mont Saint Aignan Cédex, France
Arnaud.Ribert@univ-rouen.fr

Résumé

Nous proposons dans cet article une méthode originale de détermination du nombre et de la composition des agrégats (i.e. classes au sens non supervisé) présents dans une base de données à partir de l'analyse d'une hiérarchie indicée. Notre méthode est basée sur le principe d'une coupure multi-niveaux dans la hiérarchie permettant d'adapter l'échelle à laquelle considérer les données pour déterminer les agrégats. Pour cela, nous proposons un critère permettant de détecter la présence d'un agrégat unique dans un sous-arbre de la hiérarchie. Ceci permet d'aborder les configurations (très fréquentes dans la pratique) présentant des agrégats de densités variables. De plus, notre algorithme ne nécessite pas de retour sur les données, mais exploite uniquement l'information disponible dans la hiérarchie.

Abstract

This article deals with an original hierarchical clustering algorithm which automatically determines the number and composition of clusters in a database. The method is based upon a multi-level cutting in the hierarchy which allows one to deal with high density variations in the data (frequently encountered in real databases). The algorithm proceeds by an in depth exploration of the hierarchy, deciding a cutting when a single cluster is detected in the current sub-tree. This detection is performed by an original statistical criterion. Moreover, the clustering algorithm is particularly fast, since it does not require any computation involving original data, but exclusively exploits the information provided by the hierarchy.

1. Introduction

Nous proposons dans cet article une méthode originale de détermination du nombre et de la composition des agrégats présents dans un espace euclidien. Cette opération, appelée catégorisation, est souvent utilisée pour la segmentation d'image ainsi qu'en reconnaissance des formes. Les algorithmes de catégorisation les plus fréquemment rencontrés sont généralement des méthodes de partitionnement inspirées de la méthode des k-moyennes [11][7]. Le principal inconvénient de ces approches est la nécessité de connaître à l'avance le nombre d'agrégats ainsi que leur position approximative dans l'espace de représentation.

La méthode de catégorisation que nous proposons est basée sur l'analyse d'une hiérarchie indicée construite par Classification Ascendante Hiérarchique [3][2]. L'arbre ainsi obtenu présente la propriété de posséder des noeuds dont la hauteur (appelée indice) est proportionnelle à la dissemblance entre les groupes qu'ils réunissent. L'intérêt de cette approche réside dans le fait que l'analyse de la forme d'une hiérarchie peut permettre une catégorisation automatique des données considérées, sans a priori sur le nombre et la position des agrégats. La hiérarchie de la Figure 1 montre en effet sans ambiguïté la présence de six agrégats dans les données. Une étude comparative très complète, menée par Milligan et Cooper [12] a montré que de nombreuses méthodes ont été proposées pour exploiter ce potentiel [9][6]. La plupart d'entre elles sont, comme les méthodes de partitionnement, basées sur des critères de minimisation de la variance intra-agrégat et de maximisation de la variance inter-agrégat [5][4]. Elles nécessitent donc un retour sur les données. Ce point est un inconvénient non négligeable, car il implique un surcroît de calculs déjà très nombreux pour construire une hiérarchie indicée. De plus, il est connu que l'utilisation de la variance sur ce type de problème favorise la formation d'agrégats hyper-sphériques. Une analyse de données ne vérifiant pas cette hypothèse implicite est donc biaisée.