Étude comparative de deux approches de classification recouvrante : MOC vs. OKM

Guillaume Cleuziou et Jacques-Henri Sublemontier

Laboratoire d'Informatique Fondamentale d'Orléans (LIFO)
Université d'Orléans
Rue Léonard de Vinci - 45067 ORLEANS Cedex 2
prenom.nom@univ-orleans.fr

Résumé. La classification recouvrante désigne les techniques de regroupements de données en classes pouvant s'intersecter. Particulièrement adaptés à des domaines d'application actuels (e.g. Recherche d'Information, Bioinformatique) quelques modèles théoriques de classification recouvrante ont été proposés très récemment parmi lesquels le modèle MOC (Banerjee et al. (2005a)) utilisant les modèles de mélanges et l'approche OKM (Cleuziou (2007)) consistant à généraliser l'algorithme des k-moyennes. La présente étude vise d'une part à étudier les limites théoriques et pratiques de ces deux modèles, et d'autre part à proposer une formulation de l'approche OKM en terme de modèles de mélanges gaussiens, laissant ainsi entrevoir des perspectives intéressantes quant à la variabilité des schémas de recouvrements envisageables.

1 Introduction

La classification recouvrante (en anglais *overlapping clustering*) constitue un domaine de recherche étudié depuis les années 60 et relancé par des besoins applicatifs dans des domaines importants tels que la Recherche d'Information ou encore la Bioinformatique.

Le but recherché est alors d'extraire une collection de classes recouvrantes à partir d'une population d'individus de telle manière que : chaque individu appartienne à une ou plusieurs classes, les individus d'une même classe soient similaires, et deux individus n'appartenant pas au moins à une classe commune soient dissimilaires. Différentes directions ont été prospectées afin d'obtenir ce type de schéma de classification.

Des modèles hiérarchiques ont été proposés ; Jardine et Sibson (1971) ont permis, en introduisant les k-ultramétriques, d'envisager des structures hiérarchiques (ou pseudo-hiérarchiques) moins contraignantes que les arbres, par exemple des pyramides (Diday (1984)) ou encore des hiérarchies dites "faibles" étudiées par Bertrand et Janowitz (2003) notamment. L'un des avantages de ces modèles est de proposer une interprétation visuelle des classes et de leur organisation. En revanche, ces modèles ne permettent pas de prendre en compte la globalité des schémas de recouvrements possibles ; par exemple Bertrand et Janowitz (2003) montrent que dans une k-hiérarchie faible (le modèle hiérarchique le moins contraignant), "l'intersection de $(k+1)$ classes arbitraires peut être réduite à l'intersection de k de ces classes".

Les approches par partitionnement proposées ont consisté dans un premier temps à déterminer des centres, des axes ou des représentants de classes auxquels les individus sont affectés