OKM : une extension des k-moyennes pour la recherche de classes recouvrantes

Guillaume Cleuziou

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

Résumé. Dans cet article nous abordons le problème de la classification (ou clustering) dans le but de découvrir des classes avec recouvrements. Malgré quelques avancées récentes dans ce domaines, motivées par des besoins applicatifs importants (traitements des données multimédia par exemple), nous constatons l'absence de solutions théoriques à ce problème. Notre étude consiste alors à proposer une nouvelle formulation du problème de classification par partitionnement, adaptée à la recherche d'un recouvrement des données en classes d'objets similaires. Cette approche se fonde sur la définition d'un critère objectif de qualité d'un recouvrement et d'une solution algorithmique visant à optimiser ce critère. Nous proposons deux évaluations de ce travail permettant d'une part d'appréhender le fonctionnement global de l'algorithme sur des données simples (vitesse de convergence, visualisation des résultats) et d'autre part d'évaluer quantitativement le bénéfice d'une telle approche sur une application de classification de documents textuels.

1 Introduction

La classification automatique (ou *clustering*) est un domaine d'étude situé à l'intersection de deux thématiques de recherches majeures que sont l'analyse de données et l'apprentissage automatique. Ce domaine est en perpétuelle évolution du fait de l'apparition constante de nouveaux besoins portant à la fois sur la quantité ou la nature des données à traiter (numériques, symboliques, spatiales, histogrammes, etc.) que sur le type de classification attendue (partition, hiérarchie, schéma flou, etc.).

Nombreuses sont les approches proposées afin d'organiser, de résumer ou de simplifier un ensemble de données à l'aide d'une structure de laquelle il est possible de faire émerger des classes d'objets similaires au sens d'un critère de proximité défini ou plus généralement au regard des propriétés que ces objets partagent. Il est de coutume de structurer ces approches en différentes catégories mutuellement non-exclusives (voir Jain et al. (1999)) comme par exemple, pour ne citer que les principales, les approches hiérarchiques, par partitionnement ou encore les modèles de mélanges.

Les approches par partitionnement, dont l'algorithme des k-moyennes (MacQueen, 1967) en est l'un des plus célèbre représentant, consiste le plus souvent à construire une collection de classes disjointes formant une partition des données par optimisation d'un critère objectif.