OKMED et WOKM : deux variantes de OKM pour la classification recouvrante

Guillaume Cleuziou

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

Résumé. Cet article traite de la problématique de la classification recouvrante (overlapping clustering) et propose deux variantes de l'approche OKM : OKMED et WOKM. OKMED généralise k-médoïdes au cas recouvrant, il permet d'organiser un ensemble d'individus en classes non-disjointes, à partir d'une matrice de distances. La méthode WOKM (Weighted-OKM) étend OKM par une pondération locale des classes ; cette variante autorise chaque individu à appartenir à plusieurs classes sur la base de critères différents. Des expérimentations sont réalisées sur une application cible : la classification de textes. Nous montrons alors que OKMED présente un comportement similaire à OKM pour la métrique euclidienne, et offre la possibilité d'utiliser des métriques plus adaptées et d'obtenir de meilleures performances. Enfin, les résultats obtenus avec WOKM montrent un apport significatif de la pondération locale des classes.

1 Introduction

La classification recouvrante (ou *overlapping clustering*) constitue une problématique particulière dans le domaine de la classification non-supervisée (ou *clustering*). Il s'agit d'organiser un ensemble d'individus en classes d'individus similaires en autorisant chaque donnée à appartenir à plusieurs classes. Ce type de schéma correspond à une organisation naturelle des données pour de nombreuses applications. Par exemple, en Recherche d'Information un même document peut porter sur une ou plusieurs thématiques, en Bioinformatique un même gène peut intervenir dans un ou plusieurs processus métaboliques, en Traitement du Langage un même verbe peut satisfaire une ou plusieurs grammaires de sous-catégorisation, etc.

On parle de "problématique" au même titre que la problématique générale de la classification, puisqu'il n'existe pas d'avantage de solution triviale pour extraire des classes d'individus similaires qui soient indiscutables et universelles. De surcroît, la classification recouvrante offre un espace de solutions plus vaste que dans le cas traditionnel, qu'il est donc encore plus difficile d'explorer.

Durant les quatre dernières décennies, quelques solutions ont été proposées spécifiquement pour la classification recouvrante. Dattola (1968) envisagait une approche de type centres mobiles avec affectation multiple des individus déterminée par un seuil. Jardine et Sibson (1971), en introduisant les k-ultramétriques, ont ouvert la voie des recherches fondamentales sur les