Généralisation des k-moyennes pour produire des recouvrements ajustables

Chiheb-Eddine Ben N'Cir*, Guillaume Cleuziou**,***, Nadia Essoussi*

*LARODEC, ISG Tunis, Université de Tunis, Tunisie
chiheb.benncir@isg.rnu.tn
nadia.essoussi@isg.rnu.tn
**LIFO, Université d'Orléans, France
***GREYC, Université de Caen Basse-Normandie, France
guillaume.cleuziou@univ-orleans.fr

Résumé. La recherche de groupes non-disjoints à partir de données non-étiquetées est une problématique importante en classification non-supervisée. La classification recouvrante (Overlapping clustering) contribue à la résolution de plusieurs problèmes réels qui nécessitent la détermination de groupes qui se chevauchent. Cependant, bien que les recouvrements entre groupes soient tolérés voire encouragés dans ces applications, il convient de contrôler leur importance. Nous proposons dans ce papier des généralisations de k-moyennes offrant le contrôle et le paramétrage des recouvrements. Deux principes de régulation sont mis en place, ils visent à contrôler les recouvrements relativement à leur taille et à la dispersion des classes. Les expérimentations réalisées sur des jeux de données réelles, montrent l'intérêt des principes proposés.

1 Introduction

La classification non-supervisée est une tâche importante dans l'exploration de données non-étiquetées, elle vise à les organiser en groupes (ou classes) contenant des données similaires. Cette technique est utilisée avec succès dans de nombreux domaines d'application tels que le marketing et la recherche d'information. Cependant, dans plusieurs de ces applications, les données s'organisent naturellement en groupes non-disjoints nécessitant donc de l'émergence de groupes qui se chevauchent. Le domaine de recherche correspondant à cette problématique est la classification recouvrante (*overlapping clustering*), étudiée à travers différentes approches au cours du dernier demi-siècle (Shepard et Arabie, 1979; Diday, 1987; Banerjee et al., 2005; Cleuziou, 2008; Depril et al., 2008; Fellows et al., 2011).

Le clustering recouvrant trouve ses applications dans de nombreux domaines nécessitant qu'un individu appartienne à plusieurs classes. Par exemple, en analyse des réseaux sociaux, un acteur peut appartenir à plusieurs communautés (Tang et Liu, 2009; Wang et al., 2010; Fellows et al., 2011) ; en classification de vidéos, chaque entrée peut potentiellement avoir plusieurs genres différents (Snoek et al., 2006) ; en détection d'émotions, une pièce de musique peut engendrer plusieurs émotions (Wieczorkowska et al., 2006), dans les systèmes de recherche