Optimisation de Fuzzy C-Means (FCM) clustering par la méthode des directions alternées (ADMM)

Benoit Albert*, Violaine Antoine*, Jonas Koko*

*LIMOS, Université Clermont Auvergne, France
{benoit.albert,violaine.antoine,jonas.koko}@uca.fr

Résumé. Parmi les méthodes de classification non supervisée, K-Means et ses variantes sont très populaires. Ces méthodes résolvent à chaque itération les conditions d'optimalité du premier ordre. Cependant dans certains cas, la fonction à minimiser n'est pas convexe, comme pour la version Fuzzy C-Mean avec la distance de Mahalanobis (FCM-GK). Dans cette étude, nous appliquons la méthode des directions alternées (ADMM) afin d'assurer une bonne convergence. ADMM est une méthode souvent appliquée à la résolution d'un problème de minimisation convexe séparable avec des contraintes linéaires. ADMM est une méthode de décomposition/coordination avec une étape de coordination assurée par des multiplicateurs de Lagrange. En introduisant avec justesse des variables auxiliaires, cette méthode permet de décomposer le problème en sous problèmes convexes faciles à résoudre tout en gardant la même structure itérative. Les résultats numériques ont démontré la performance significative de la méthode proposée par rapport à la méthode standard surtout pour des données de grandes dimensions.

1 Introduction

Le partitionnement de données est un processus d'analyse des données qui consiste à partager les n objets d'un jeu de données en c sous-ensembles, dans le but que chaque groupe (sous ensemble) possède des objets similaires et que les groupes soient bien distincts entre eux (Jain et Dubes, 1988). Il permet de détecter des structures cachées dans les jeux de données sans connaissance préalable. Plusieurs approches différentes existent, les méthodes se distinguent par la nature des partitions créées. Elles peuvent être certaines : chaque objet appartient à un unique groupe. *k-means* est la plus célèbre des méthodes formant ce genre de partition. Chaque groupe est représenté par un centroïde (objet moyen), la notion de similarité est définie par la distance entre les objets et les centroïdes. Grâce à sa faible complexité et sa simplicité, cette méthode est très utilisée (Jain, 2010). Les partitions peuvent être floues permettant de modéliser l'incertitude. La variante floue des k-means est Fuzzy C-Means (FCM) (Bezdek, 1973; Bezdek et Dunn, 1975), chaque object a un degré d'appartenance à chaque groupe. FCM est encore utilisée dans divers domaines (Anter et al., 2019; Yin et Li, 2020; Cai et al., 2021). La similarité entre les objets et les centroïdes dans l'algorithme FCM