Clustering en haute dimension par accumulation de clusterings locaux

Marc-Ismaël Akodjènou-Jeannin *, Kavé Salamatian*
Patrick Gallinari *

*104, avenue du Président Kennedy
75016 Paris
{Marc-Ismael.Akodjenou, Kave.Salamatian, Patrick.Gallinari}@lip6.fr
http://www.lip6.fr

Résumé. Le clustering est une tâche fondamentale de la fouille de données. Ces dernières années, les méthodes de type *cluster ensembles* ont été l'objet d'une attention soutenue. Il s'agit d'agréger plusieurs clusterings d'un jeu de données afin d'obtenir un clustering "moyen". Les clusterings individuels peuvent être le résultat de différents algorithmes. Ces méthodes sont particulièrement utiles lorsque la dimensionalité des données ne permet pas aux méthodes classiques basées sur la distance et/ou la densité de fonctionner correctement. Dans cet article, nous proposons une méthode pour obtenir des clusterings individuels à faible coût, à partir de projections partielles du jeu de données. Nous évaluons empiriquement notre méthode et la comparons à trois méthodes de différents types. Nous constatons qu'elle donne des résultats sensiblement supérieurs aux autres.

1 Introduction

Le clustering consiste à découvrir automatiquement des groupes ("clusters") présents dans le jeu de données. Une littérature abondante existe sur le sujet (une revue des principales méthodes peut être trouvée dans Rui et Wunsch (2005)). Nous nous plaçons ici dans le cadre des "cluster ensembles" (Strehl et Ghosh (2002)). Les "cluster ensembles" sont une sorte de méta-clustering : à partir de plusieurs clusterings du même jeu de données, on déduit un clustering "moyen" (Strehl et Ghosh (2002)). Plusieurs alternatives ont été proposées pour trouver le clustering moyen (méthodes agglomératives, ou basées sur des graphes). Indépendamment de la méthode de synthèse choisie, il est clair que le clustering moyen dépend fortement de la qualité et de la diversité de chaque clustering individuel (Fern et Brodley (2003)). Par exemple, agréger plusieurs clusterings issus de l'algorithme K-means avec des initialisations différentes atténuera les erreurs particulières dues à chaque clustering individuel ; cependant cela ne permettra pas de contourner les limitations fondamentales de l'algorithme (clusters de forme sphérique, sensibilité à la dimension...). La situation idéale pour les cluster ensembles est celle où les clusterings individuels sont variés, de bonne qualité et obtenus à faible coût.

L'idée explorée par Topchy et al. (2003) est d'obtenir ces clusterings individuels en projetant le jeu de données sur une direction aléatoire, et en faisant un clustering simple sur la