Accélération de k-means par pré-calcul dynamique d'agrégats

Nabil El Malki[*,**], Franck Ravat [*], Olivier Teste [*]

[*] IRIT (CNRS/UMR5505)
[**]Capgemini (www.capgemini.com), Toulouse, France
prenom.nom@irit.fr, prenom.nom@capgemini.com

Résumé. L'algorithme de classification non supervisé 'k-means' nécessite un accès itératif et répétitif aux données allant jusqu'à effectuer plusieurs fois le même calcul sur les mêmes données. Ces calculs répétés peuvent s'avérer coûteux lorsqu'il s'agit de classifier des données massives. Nous proposons d'étendre l'algorithme de k-means en introduisant une approche d'optimisation basée sur le pré-calcul dynamique d'agrégats pouvant ensuite être réutilisés afin d'éviter des calculs redondants.

1 Introduction

Dans le cadre des approches non supervisées, un algorithme de classification tente de diviser les données en plusieurs classes de sorte à ce que les données (appelés également individus, observations ou points) qui se trouvent dans la même classe soient les plus similaires possibles, et inversement, les points appartenant à des classes différentes soient les plus dissemblables possibles.

Parmi les algorithmes de classification, l'algorithme de k-means (centres fixes) est probablement un des plus connus (Forgy, 1965) et constitue l'objet de notre étude. Sa première version est apparue dès les années 50 (Jain, 2010). Ce dernier, repose sur un traitement itératif (i.e. les instructions de l'algorithme doivent être réalisées plusieurs fois avant de converger vers une classification stable) et répétitif (i.e. un même calcul est potentiellement effectué plusieurs fois sur les mêmes données). Dans 'k-means', les résultats d'une itération ne sont pas conservés pour alimenter l'itération suivante. Cette caractéristique engendre des dégradations de performances notamment lorsque la dimension (nombre d'attributs) est important et lorsque ces dimensions possèdent de nombreuses valeurs de densité variable.

Contribution : Afin de permettre à k-means d'offrir de meilleures performances notamment sur de gros volumes de données, nous proposons une nouvelle version de k-means basée sur un principe de pré-agrégats. Cette extension repose sur les principes sont les suivants : (i) pré-calculer et stocker les différents calculs effectués lors des itérations successives, (ii) réutiliser les pré-calculs stockés pour accélérer les itérations futures. En section 2, nous discutons l'état de l'art. Ensuite, nous exposons notre modèle (section 3) et nos expérimentations (section 4).