Comparaison de bornes théoriques pour l'accélération du clustering incrémental en une passe

Nicolas Labroche, Marcin Detyniecki, Thomas Baerecke

UPMC Paris 6, LIP6 UMR CNRS 7606
BC 169, 4 place Jussieu 75252 Paris Cedex 05, France
{prenom.nom@lip6.fr}

Résumé. Le clustering incrémental en une passe repose sur l'affectation efficace de chaque nouveau point aux clusters existants. Dans le cas général, où les clusters ne peuvent être représentés par une moyenne, la détermination exhaustive du cluster le plus proche possède une complexité quadratique avec le nombre de données. Nous proposons dans ce papier une nouvelle méthode d'affectation stochastique à chaque cluster qui minimise le nombre de comparaisons à effectuer entre la donnée et chaque cluster pour garantir, étant donné un taux d'erreur acceptable, l'affectation au cluster le plus proche. Plusieurs bornes théoriques (Bernstein, Hoeffding et Student) sont comparées dans ce papier. Les résultats sur des données artificielles et réelles montrent que la borne de Bernstein donne globalement les meilleurs résultats (notamment lorsqu'elle est réduite) car elle permet une accélération forte du processus de clustering, tout en conservant un nombre très faible d'erreurs.

1 Introduction

Le clustering permet l'exploration d'ensembles de données en les résumant sous la forme de groupes homogènes plus facilement caractérisables et interprétables. Récemment, de nouveaux algorithmes ont été proposés pour répondre aux problèmes de traitement des grands volumes ou des flux de données (Aggarwal et al., 2003; Cao et al., 2006; Philipp Kranen et Seidl, 2011). Ces méthodes reposent généralement sur des algorithmes qui ne réalisent qu'une seule passe sur les données initiales. Ceux-ci ne sont malheureusement applicables qu'à des données vectorielles, pour lesquelles des structures incrémentales de description des clusters existent (Zhang et al., 1996).

Dans ce travail, nous nous intéresssons au cas général où les données ne sont pas nécessairement vectorielles, et où il n'est donc pas possible d'utiliser de telles structures. Une solution consiste à résumer chaque cluster par un sous-ensemble des données qui le compose, possiblement un seul point, appelé médoide, qui est le plus similaire aux autres données du cluster. Le problème est que la détermination des médoides, et donc l'affectation d'un nouveau point aux clusters existants dans un contexte incrémental, possède une complexité quadratique avec le nombre de données. Cela n'étant pas envisageable dans des cas d'usage réels, les algorithmes implémentent généralement des mécanismes d'échantillonnage pour réduire le coût des calculs