SSC : Statistical Subspace Clustering

Laurent Candillier[1,2], Isabelle Tellier[1], Fabien Torre[1], Olivier Bousquet[2]

[1] GRAppA - Université Charles de Gaulle - Lille 3
candillier@grappa.univ-lille3.fr
http ://www.grappa.univ-lille3.fr
[2] Pertinence - 32 rue des Jeûneurs -75002 Paris
olivier.bousquet@pertinence.com
http ://www.pertinence.com

Résumé. Cet article se place dans le cadre du *subspace clustering*, dont la problématique est double : identifier simultanément les clusters et le *sous-espace spécifique* dans lequel *chacun* est défini, et caractériser chaque cluster par un nombre minimal de dimensions, permettant ainsi une présentation des résultats compréhensible par un expert du domaine d'application.

Les méthodes proposées jusqu'à présent pour cette tâche ont le défaut de se restreindre à un cadre numérique. L'objectif de cet article est de proposer un algorithme de *subspace clustering* capable de traiter des données décrites à la fois par des attributs continus et des attributs catégoriels.

Nous présentons une méthode basée sur l'algorithme classique EM mais opérant sur un modèle simplifié des données et suivi d'une technique originale de sélection d'attributs pour ne garder que les dimensions pertinentes de chaque cluster. Les expérimentations présentées ensuite, menées sur des bases de données aussi bien artificielles que réelles, montrent que notre algorithme présente des résultats robustes en termes de qualité de la classification et de compréhensibilité des clusters obtenus.

Introduction

Face aux quantités d'informations qui ne cessent d'augmenter dans les bases de données du monde entier, l'extraction automatique de connaissances à partir de ces bases et les techniques de visualisation des résultats sont devenues indispensables. C'est la raison d'être de la *fouille de données*. Dans ce cadre, l'apprentissage non supervisé (ou *clustering*) est depuis longtemps utilisé pour identifier les groupes (ou *clusters*) d'éléments similaires (cf. survey de Berkhin 2002). Une problématique supplémentaire apparaît face à des bases de données de grande dimensionnalité : dans ce cas, les groupes peuvent être caractérisés uniquement par certains sous-ensembles de dimensions et ces dimensions pertinentes peuvent être différentes d'un groupe à l'autre. Sur de tels problèmes, les techniques classiques de *clustering* fonctionnent mal car, fondées sur une distance entre objets définie globalement dans l'espace de description, elles ne peuvent pas appréhender le fait que la notion de similarité varie d'un groupe à l'autre.

Une nouvelle problématique a donc émergé récemment, celle du *subspace clustering*, dont l'enjeu est de cibler les groupes d'objets et, pour chacun, le *sous-espace spécifique*