Méthode de regroupement par graphe de voisinage

Fabrice Muhlenbach

Université de Lyon, Université de Saint-Étienne
UMR CNRS 5516, Laboratoire Hubert Curien
18 rue du Professeur Benoît Lauras, 42000 SAINT-ÉTIENNE, FRANCE
fabrice.muhlenbach@univ-st-etienne.fr, http://labh-curien.univ-st-etienne.fr/ muhlenbach/

Résumé. Ce travail s'inscrit dans la problématique de l'apprentissage non supervisé. Dans ce cadre se retrouvent les méthodes de classification automatique non paramétriques qui reposent sur l'hypothèse que plus des individus sont proches dans l'espace de représentation, plus ils ont de chances de faire partie de la même classe. Cet article propose une nouvelle méthode de ce type qui considère la proximité à travers la structure fournie par un graphe de voisinage.

1 Introduction : classification et graphes de voisinage

Une caractéristique humaine fondamentale est la capacité que nous avons à organiser notre monde, à parvenir à effectuer des catégorisations. Ce phénomène correspond à la faculté de pouvoir regrouper dans de mêmes ensembles (c.-à-d. des classes homogènes) des éléments ayant des traits en commun. Reprise dans le domaine du traitement automatisé de l'information, cette caractéristique englobe les méthodes de classification non supervisée, appelées aussi méthodes de *clustering* (Cleuziou, 2004), une famille de méthodes d'apprentissage automatique qui, à partir des informations connues sur les données, cherchent à retrouver des groupes, à définir des amas, à construire des classes. La classification non supervisée donne lieu à de multiples applications dans le domaine de la fouille de données (en fouille de texte, en bio-informatique, dans le domaine du marketing, en vision par ordinateur, etc.)

Suivant la connaissance existant sur les données, différentes familles de méthodes de classification non supervisée pourront s'appliquer. Dans le cas où il existe a priori une hypothèse sur la distribution des données, il est possible d'employer les méthodes dites « probabilistes » (comme EM). Cependant, en absence de ce genre de connaissance, il faut se limiter aux méthodes dites « non paramétriques » qui reposent sur la seule hypothèse que plus deux individus sont proches, plus ils ont de chances de faire partie du même groupe, de la même classe. Dans ce second cas, nous distinguons principalement trois approches.

Les méthodes de la première approche produisent un partitionnement des données qui sera retenu, parmi les différents regroupements possibles, au moyen d'un critère de qualité donné. Citons parmi ces dernières la méthode des k-Means (Steinhaus, 1956) qui a pour objectif de détecter les différents groupes obtenus à partir d'une partition initiale aléatoire par la recherche des points moyens qui vont minimiser la variance intra-classe de ces différents groupes.

Les méthodes de la deuxième approche produisent une hiérarchie sur les données représentée par un arbre (appelé « dendogramme »). Dans le cas de la classification hiérarchique