Formulation Condorcéenne du critère de la modularité

Lazhar Labiod, Nistor Grozavu, Younès Bennani

LIPN UMR 7030, Université Paris 13
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse
Prénom.Nom@lipn.univ-paris13.fr,

Résumé. La mesure de modularité a été utilisée récemment pour la classification de graphes (Newman et Girvan, 2004), (Agarwal et Kempe, 2008). Dans ce papier, nous montrons que la mesure de modularité peut être formellement étendue pour la classification non supervisée des données catégorielles. Nous établissons également des connexions entre le critère de modularité et celui de l'analyse relationnelle qui est basé sur le critère de Condorcet. Nous développons ensuite un algorithme efficace inspiré de l'heuristique de l'analyse relationnelle pour trouver la partition optimale maximisant le critère de modularité. Les résultats expérimentaux montrent l'efficacité de notre approche.

1 Introduction

La classification automatique est une méthode d'apprentissage non supervisé permettant le partitionnement d'un ensemble d'observations en classes. les méthodes de classification automatique conduisent à une partition de la population initiale en groupes disjoints, tels que, selon un critère choisi a priori, deux individus d'un même groupe aient entre eux un maximum d'affinité et deux individus de deux groupes différents aient entre eux un minimum d'affinité. La classification automatique a été largement étudiée en apprentissage automatique, en bases de données et en statistique de divers points de vue.

De nombreuses applications de la classification automatique ont été discutées et de nombreuses techniques ont été développées. Une étape importante dans la conception d'une technique de classification consiste à définir un critère pour mesurer la qualité de partitionnement en termes des deux objectifs cités ci-dessus. Pour la classification des données numériques continues, il est naturel de penser à utiliser une mesure basée sur une distance géométrique. Étant donnée une telle mesure, une partition appropriée peut être calculée par l'optimisation de certaines quantités (par exemple, la somme des distances des observations à leurs centroïdes). Toutefois, si les vecteurs de données contiennent des variables catégorielles, le problème de la classification devient plus difficile et d'autres stratégies doivent être développées. C'est souvent le cas dans de nombreuses applications où les données sont décrites par un ensemble d'attributs descriptifs ou binaires, dont beaucoup ne sont pas numériques. Des exemples de tels attributs sont le pays d'origine et la couleur des yeux dans les données démographiques. De nombreux algorithmes ont été développés pour la classification des données catégorielles, par exemple, (Barbara et al, 2002), (Gibson et al, 1998), (Huang, 1998) et (Ganti et al, 1999).

La mesure de modularité a été utilisée récemment pour la classification de graphes (Agarwal et Kempe, 2008), (Newman et Girvan, 2004) et (White et Smyth, 2005). Dans ce papier,