Classification des données catégorielles via la maximisation spectrale de la modularité

Lazhar Labiod*, Younès Bennani**

* LIPADE, University of Paris Descartes,
45 rue des Saints Pères 75006 Paris, France.
email: Prénom.Nom@parisdescartes.fr,
** LIPN UMR 7030, Université Paris 13
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse
Prénom.Nom@lipn.univ-paris13.fr,

Résumé. Ce papier présente un algorithme spectrale pour maximiser le critère de la modularité étendu à la classification des données catégorielles. Il met en evidence la connexion formelle entre la maximisation de la modularité et la classification spectrale, il présente en particulier le problème de maximisation de la modularité sous forme d'un problème algèbrique de maximisation de la trace. Nous développons ensuite un algorithme efficace 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.

La mesure de modularité a été utilisée récemment pour la classification de graphes [Agarwal and Kempe, 2008], [Newman and Girvan, 2004] et [White and Smyth, 2005]. Dans ce papier, nous montrons que le critère de modularité peut être formellement etendu pour la classification des données catégorielles. Nous développons ensuite une procédure spectrale efficace pour trouver la partition optimale maximisant le critère de modularité. Les résultats expérimentaux montrent l'efficacité de notre approche. La première contribution de ce papier est l'introduction d'une mesure de modularité étendue pour la classification des données catégorielles. La deuxième contribution est la présentation du problème de maximisation de la mesure de modularité étendue sous la forme d'un problème de maximisation de trace.
Le reste du papier est organisé comme suit: la section 2 introduit quelques notations et définitions. La section 3 présente la mesure de modularité étendue. Des discussions sur la connexion