Bipartitionnement d'un tableau de contingence

Malika Charrad[*,**], Yves Lechevallier[***]
Mohamed Ben Ahmed[*]

[*]École Nationale des Sciences de l'Informatique
[**]Conservatoire National des Arts et Métiers
malika.charrad@riadi.rnu.tn,
[***]INRIA-Rocquencourt
Yves.lechevallier@inria.fr

Résumé. La recherche simultanée de partitions sur l'ensemble de lignes et l'ensemble de colonnes d'un tableau de données a donné naissance à des méthodes de classification simultanée ou bipartitionnement. On parle aussi de la classification croisée ou la classification par blocs. Plusieurs algorithmes de bipartitionnement ont été proposés dans la littérature selon le type de tableau des données. Nous nous intéressons dans ce papier à l'algorithme Croki2 de classification croisée des tableaux de contingence. Nous proposons dans ce papier une variante plus rapide de cet algorithme que nous comparons à la version originale à travers des expérimentations sur des données présentant une structure de biclasses générées artificiellement selon une méthodologie que nous détaillons.

1 Introduction

Les méthodes de classification automatique appliquées à des tableaux mettant en jeu deux ensembles de données agissent de façon dissymétrique en ne faisant porter la structure recherchée que sur un seul ensemble. L'application d'une classification sur chaque ensemble est possible mais la détermination des liens entre les deux partitions est difficile. La recherche de structures de classes symétriques, plus précisément, la recherche simultanée de partitions sur les deux ensembles a donné naissance à des méthodes de classification simultanée ou de bipartitionnement. On parle aussi de la classification croisée ou la classification par blocs. Ce type de classification est connu dans la littérature anglaise sous différents noms. Souvent on parle de "two-mode clustering", "two-side clustering", "two-way clustering", "direct clustering" (Hartigan, 1972), "biclustering" (Mirkin, 1996) ou encore "co-clustering" (Dhillon et al., 2003). Ces approches diffèrent souvent dans les algorithmes employés, la nature des blocs recherchés qui peuvent être isolés ou imbriqués, le nombre de blocs identifiés dans les données et la nécessité de fixer le nombre de classes sur les lignes et les colonnes. Ce type d'approches a suscité beaucoup d'intérêt dans divers domaines, en particulier celui des biopuces où l'objectif est de caractériser des groupes de gènes par des groupes de conditions expérimentales. Cependant, les travaux de synthèse sur les algorithmes de bipartitionnement sont concentrés sur les algorithmes appliqués en bioinformatique tels que les travaux de Madeira et Oliveira (2004) et Tanay et al. (2004). Par ailleurs, les algorithmes de classification directe (ou block