Classification d'un tableau de contingence et modèle probabiliste

Gérard Govaert*,Mohamed Nadif**

*Heudiasyc, UMR CNRS 6599, Université de Technologie de Compiègne,
BP 20529, 60205 Compiègne Cedex, France
gerard.govaert@utc.fr
**LITA, Université de Metz, Ile du Saulcy, 57045 Metz Cedex, France
mohamed.nadif@univ-metz.fr

Résumé. Ces dernières années, la classification croisée ou classification par blocs, c'est-à-dire la recherche simultanée d'une partition des lignes et d'une partition des colonnes d'un tableau de données, est devenue un outil très utilisé en fouille de données. Dans ce domaine, l'information se présente souvent sous forme de tableaux de contingence ou tableaux de co-occurrence croisant les modalités de deux variables qualitatives. Dans cet article, nous étudions le problème de la classification croisée de ce type de données en nous appuyant sur un modèle de mélange probabiliste. En utilisant l'approche vraisemblance classifiante, nous proposons un algorithme de classification croisée basé sur la maximisation alternée de la vraisemblance associée à deux mélanges multinomiaux classiques et nous montrons alors que sous certaines contraintes restrictives, on retrouve les critères du Chi2 et de l'information mutuelle. Des résultats sur des données simulées et des données réelles illustrent et confirment l'efficacité et l'intérêt de cette approche.

1 Introduction

La classification automatique, comme la plupart des méthodes d'analyse de données peut être considérée comme une méthode de réduction et de simplification des données. Dans le cas où les données mettent en jeu deux ensembles I et J, ce qui est le cas le plus fréquent, la classification automatique en ne faisant porter la structure recherchée que sur un seul des deux ensembles, agit de façon dissymétrique et privilégie un des deux ensembles, contrairement par exemple à l'analyse factorielle des correspondances qui obtient simultanément des résultats sur les deux ensembles ; il est alors intéressant de rechercher *simultanément* une partition des deux ensembles. Ce type d'approche a suscité récemment beaucoup d'intérêt dans divers domaines tels que celui des biopuces où l'objectif est de caractériser des groupes de gènes par des groupes de conditions expérimentales ou encore celui de l'analyse textuelle où l'objectif est de caractériser des classes de documents par des classes de mots. Notons que dans ce domaine, les données se présentent généralement sous forme d'un tableau de contingence où chaque cellule correspond au nombre d'occurrences d'un mot dans un document.