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
**IUT de Metz, LITA, Université de Metz,
Ile du Saulcy, 57045 Metz Cedex, France
nadif@iut.univ-metz.fr

Résumé. Les modèles de mélange, qui supposent que l'échantillon est formé de sous-populations caractérisées par une distribution de probabilité, constitue un support théorique intéressant pour étudier la classification automatique. On peut ainsi montrer que l'algorithme des k-*means* peut être vu comme une version classifiante de l'algorithme d'estimation EM dans un cas particulièrement simple de mélange de lois normales.

Lorsque l'on cherche à classifier les lignes (ou les colonnes) d'un tableau de contingence, il est possible d'utiliser une variante de l'algorithme des k-*means*, appelé Mndki2, en s'appuyant sur la notion de profil et sur la distance du khi-2. On obtient ainsi une méthode simple et efficace pouvant s'utiliser conjointement à l'analyse factorielle des correspondances qui s'appuie sur la même représentation des données.

Malheureusement et contrairement à l'algorithme des k-*means* classique, les liens qui existent entre les modèles de mélange et la classification ne s'appliquent pas directement à cette situation. Dans ce travail, nous montrons que l'algorithme Mndki2 peut être associé, à une approximation près, à un modèle de mélange de lois multinomiales.

1 Introduction

Les modèles de mélange, qui supposent que l'échantillon est formé de sous-populations caractérisées par une distribution de probabilité, sont des modèles très souples permettant de prendre en compte des situations variées comme la présence de populations hétérogènes ou d'éléments atypiques. Grâce à l'algorithme d'estimation EM, particulièrement adapté à cette situation, les modèles de mélange ont fait l'objet de nombreux développements en statistique et en particulier en classification automatique. On peut ainsi montrer que l'algorithme des k-*means* peut être vu comme une version classifiante de l'algorithme EM, appelé CEM, dans un cas particulièrement simple de mélange de lois normales. Dans ce travail, on étudie comment ces propriétés peuvent être étendues aux tableaux de contingence.

Rappelons qu'un tableau de contingence est obtenu à partir du croisement de 2 variables qualitatives ; par exemple, si on note I et J les ensembles de r et s modalités de chaque variable, chaque élément x_{ij} de la matrice de données contiendra le nombre