Détermination du nombre des classes dans l'algorithme CROKI2 de classification croisée

Malika CHARRAD*, Yves LECHEVALLIER**
Gilbert SAPORTA*,** Mohamed BEN AHMED***

*Laboratoire RIADI, Ecole Nationale des Sciences de l'Informatique, Tunis
malika.charrad@riadi.rnu.tn,
mohamed.benahmed@riadi.rnu.tn
**INRIA-Rocquencourt, 78153 Le Chesnay cedex
yves.lechevallier@inria.fr
***CNAM, 292 rue Saint-Martin, 75141 Paris cedex 03
gilbert.saporta@cnam.fr

Résumé. Un des problèmes majeurs de la classification non supervisée est la détermination ou la validation du nombre de classes dans la population. Ce problème s'étend aux méthodes de bipartitionnement ou block clustering. Dans ce papier, nous nous intéressons à l'algorithme CROKI2 de classification croisée des tableaux de contingence proposé par Govaert (1983). Notre objectif est de déterminer le nombre de classes optimal sur les lignes et les colonnes à travers un ensemble de techniques de validation de classes proposés dans la littérature pour les méthodes classiques de classification.

1 Introduction

Comme la qualité d'une partition est très liée au choix du nombre de classes, les auteurs définissent trois types de critères de validation selon que l'on dispose ou pas d'information a priori sur les données : critère interne, critère externe et critère relatif. Dans ce papier, nous proposons d'utiliser ce dernier critère pour déterminer le nombre de classes dans la partition sur les lignes et celles sur les colonnes. Il y a trois familles de critères de validation en Classification : la séparation, l'homogénéité et la dispersion. En se basant sur ces trois familles de critères de validation, plusieurs indices sont construits pour évaluer la qualité des partitions. Nous utilisons quelques uns de ces indices, à savoir l'indice de Davies et Bouldin (1979), l'indice Dunn (1974), l'indice Silhouette, proposé par Rousseeuw (1987), l'indice de séparation S (Separation index) proposé par Xie (1991) et l'indice CS proposé dans chou (2003). Nous appliquons chacun de ces indices sur la partition sur les lignes en fixant la partition sur les colonnes et inversement. Une valeur moyenne des deux valeurs est attribuée à chaque indice. Outre ces indices, nous proposons d'utiliser deux autres indices inspirés des travaux de Govaert (1983). Soit le tableau de contingence $I \times J$. L'algorithme CROKI2 recherche alternativement une partition P de I en K classes et une partition Q de J en L classes. Il applique la méthode des nuées dynamiques en utilisant la métrique de χ^2 et le centre de gravité comme noyau. On considère le nuage $N(I)$ des n vecteurs des profils $f_J^i, i \in I$ munis