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

des masses f avec la métrique quadratique définie par la matrice diagonale $(1/f_{.J})$. Govaert démontre la relation suivante $\chi^2(I,J) = SW(P) + \chi^2(P,J)$ avec S est la somme des éléments du tableau de contingence. SW(P) représente l'information perdue en effectuant les regroupements de la partition P alors que $\chi^2(P,J)$ c'est l'information conservée. $\chi^2(I,J)$ ne dépendant pas de la partition P, la recherche de la partition minimisant le critère W(P) est équivalente à la recherche de la partition maximisant $\chi^2(P,J)$. Ce résultat nous permet de proposer I(P) comme indice de validation, avec $I(P) = W(P)/\chi^2(P,J)$. Par suite de symétrie et en considérant le nuage N(J), le même indice est proposé pour la partition en colonnes : $I(Q) = W(Q)/\chi^2(I,Q)$. Parallèlement, en considérant le nuage N($I \times J$), Govaert démontre la relation suivante :$\chi^2(I,J) = SW(P \times Q) + \chi^2(P,Q)$. Comme $\chi^2(I,J)$ étant constante, la partition $P \times Q$ minimisant $W(P \times Q)$ maximise $\chi^2(P,Q)$. De même, nous proposons d'utiliser l'indice $I(P \times Q) = W(P \times Q)/\chi^2(P,Q)$ pour déterminer la meilleure partition. Une bonne partition sur les lignes et les colonnes donne une valeur minimale de $I(P \times Q)$.

En utilisant un jeux de données issu d'un tableau présentant le nombre d'occurrences des mots figurant dans les pages aux pages Web du site de Metz Charrad et al. (2008), nous faisons varier le nombre de classes de 2 à n sur les lignes et les colonnes et nous enregistrons les valeurs des indices pour opter pour un couple de classes en lignes et en colonnes. Nous identifions les cinq meilleurs couples pour chaque indice. En remarquant que le couple (9,8) est le meilleur pour la majorité des indices, nous optons pour une partition de 9 classes en lignes et 8 classes en colonnes.

Références

Charrad, M., Y. Lechevallier, G. Saporta, et M. B. ahmed (2008). La classification croisée pour l'analyse textuelle d'un site web. *Revue des Nouvelles Technologies Informatiques 1*, 53–54.

Davies, D. et D. Bouldin (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Machine Intell. 4*, 224–227.

Dunn, J. (1974). Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics 4*, 95–104.

Govaert, G. (1983). *Ćlassification croisée*. Thèse de doctorat, Université Pierre et Marie Curie Paris VI.

Rousseeuw, P. (1987). Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*, 53–65.

Xie, X. G. (1991). A validity measure for fuzzy clustering. *IEEE Trans.Pattern Analysis and Machine Intelligence 13*, 841–847.

Summary

One of the major problems of clustering is how to identify the number of clusters. This problem is also present in biclustering or block clustering methods. In this paper, we are interested by CROKI2 algorithm proposed by Govaert (1983) for contingency tables. Our goal is to find the optimal number of clusters on rows and columns by using some techniques of cluster validity.