Int. J. Open Problems Compt. Math., Vol. 2, No. 4, December 2009
ISSN 1998-6262;
www.i-csrs.org

Cross Clustering of Contingency Table With Multinomial Laws

Khemal Bencheikh yamina

Laboratory of Fundamental and Numerical Mathematics
Department of Mathematics, Faculty of science
Ferhat Abbas University , Setif
e-mail: bencheikh-00@yahoo.fr

Abstract

Clustering is an essential step in data mining. The classical methods are based on metric criteria but, the use of mixture model in clustering is now a classical and powerful approach. Typically, the data that arises in these applications is arranged as a two-way contingency. In this paper, we embed the block clustering problem in the mixture approach. We propose a multinomial block mixture model and adopting the classification maximum likelihood principle we perform a new algorithm.

Keywords: *Contingency table, block clustering, mixture model.*

1 Introduction

Cluster analysis is an important tool in a variety of scientifique areas such as pattern recognition, information retrieval, microarray, data mining and so forth. Although many clustering procedures such as hierarchical clustering, k-means or self-organizing maps, aim to construct an optimal partition of objects or sometimes of variables, there are other methods, called block clustering methods, which consider simultaneously the two sets and organize the data into homogeneous blocks.

Many clustering methods commonly used in practice are based on a distance or a dissimilarity meausure and to cluster the rows and the columns of a contingency table we may use the CROKI2 algorithm (Govaert [5]), that is

to say an adapted version of k-means based on the chi-square distance. Infortunately, unlike the standard k-means algorithm, this algorithm does not correspond to the classification approach associated with a mixture model. To cluster the rows or the columns of a contingency table, we here propose using the mixture model algorithm associated with multinomial laws.

2 Contingency table

2.1 Notations

$X(I, J(= (x_i^j)$ will denote the initial contingency table defined on the two sets I and J of sizes n and p. The usual terminology is used :

- $s = \sum_{i \in I} \sum_{j \in J} x_i^j$
- F is the frequency table ($f_{ij} = \frac{x_i^j}{s}$, $i \in I$ and $j \in J$)

f_{ij} is an estimation of probability that an object has simultaneously the category i and the category j

- $f_{i.}$ and $f_{.j}$ are the marginal frequencies :

$\forall\ i \in I,\ f_{i.} = \sum_{i \in I} f_{ij}$ and $\forall\ j \in J\ ,\ f_{.j} = \sum_{j \in J} f_{ij}$

2.2 The summary table

The summary table associated with the two partitions must also be a contingency table. It is obtained by regrouping the rows and columns according to the partitions P and Q in the following manner : If $P = (P_1, P_2, ..., P_K)$ is a partition of I into K clusters and $Q = (Q^1, Q^2, ..., Q^M)$ a partition of J into M clusters, it becomes possible to define a new contingency table by summing the elements of the initial contingency table corresponding to each pair of clusters (P_k, Q^l), this table denoted $T(P, Q)$ is defined by : $T(k, l) = \sum_{i \in P_k} \sum_{j \in Q^l} x_j^i = x_k^l \qquad \forall\ k = 1, ..., K \qquad \text{and} \quad l = 1, ..., M$

2.3 The objective function

The chosen information measure we would like to preserve is the χ^2 of contingency (or Pearson chi-square statistic) which measures the dependence between I and J by contingency χ^2 :

$$\chi^2(I; J) = s \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}}$$

This measure usually provides statistical evidence of a significant association, or dependence between rows and columns of table.

The χ^2 information associated to a table $T(k, l)$ is :

$$\chi^2(P;Q) = s \sum_{k=1}^{K} \sum_{l=1}^{M} \frac{(f_{kl} - f_{k.} f_{.l})^2}{f_{k.} f_{.l}}$$

where $f_{kl} = \sum_{i \in P_k} \sum_{j \in Q^m} f_{ij}$, $f_{k.} = \sum_{i \in P_k} f_{ij},$, $f_{.l=} \sum_{j \in Q^m} f_{ij}$

3 The mixture model approach

Here we take up Bencheikh [1]and[2] representation. The starting data table X of (n, p) dimension, is considered as a sample $T = I \times J$ of $n \times p$ size (where I set constitutes a sample of n size from Ω population, the same for J set constitutes a sample of p size from Ω' population) of aleatory variable Z with values in R whose probability law admits the distribution function :

$$p(x) = \sum_{k=1}^{K} \sum_{l=1}^{M} p_k^l \;\; p(x, \lambda_k^l)$$

$$\forall\, x \in R \quad \forall\, k = 1, ..., K \quad \forall\, l = 1, ..., M \quad 0 \leq \; p_k^l \leq 1 \text{ et } \sum_{k=1}^{K} \sum_{l=1}^{M} \; p_k^l = 1$$

where $p(., \lambda_k^l)$ is a distribution function on R belonging to a parameterized family of distribution function depending on the λ parameter, p_k^l is the probability that a point of the sample follows the distribution law $p(., \lambda_k^l)$. One will call these p_k^l the proportitions of the mixture.

The problem arising is the estimate of the numbers K and M of components of the mixture and the unknown parameters $q_k^l = \left(p_k^l, \lambda_k^l\right)$; $k = 1, ..., K$ *and* $l = 1, ..., M$ *within sight of the sample* $T = I \times J$.

In the approach classification (Bencheikh [1], Celeux [3] , Govaert [4], Schroeder [6], Scott and Symons [7]), one replaces the initial problem of estimate by the following problem :

To seek a partition $P \times Q = \left\{P_k \times Q^l \;;\; k = 1, ..., K \;\; and \;\; l = 1, ..., M\right\}$, *K and M being supposed known, such as each class* $P_k \times Q^l$ *is assimilable to a subsample which follows a law* $p(., \lambda_k^l)$.

It is then a question of maximizing the classification likelihood criterion $VC(P \times Q, H) = \sum_{k=1}^{K} \sum_{l=1}^{M} \log L(P_k \times Q^l, \lambda_k^l)$

where H is the $K.M - times$ (λ_k^l, $k = 1, ..., K$ *and* $l = 1, ..., M$) and $L(P_k \times Q^l, \lambda_k^l)$ is the likelihood of subsample $P_k \times Q^l$ who follows the law $p(., \lambda_k^l)$.

4 Contingency table and multinomial laws

It is supposed that the data of the table $T(k,l)$ forms a sample of size $K.M$ and comes from only one law multinomial of parameter H.
$H = \Big\{ \lambda_k^l, k = 1, ..., K \text{ and } l = 1, ..., M \Big\}$.

The likelihood associated with this sample is

$$L(x_1^1, ..., x_k^l, ..., x_K^M; \lambda_1^1, ..., \lambda_k^l, ..., \lambda_K^M) = s! \prod_{k=1}^{K} \prod_{l=1}^{M} \frac{(\lambda_k^l)^{x_k^l}}{(x_k^l)!}$$

Let us pose $\lambda_k^l = \lambda_k.\lambda^l$ with $\sum\_k = 1^K \lambda_k = 1$ and $\sum_{l=1}^{M} \lambda^l = 1$

$$\begin{aligned} VC(P \times Q, H) &= \log\ L(x_1^1, ..., x_k^l, ..., x_K^M; \lambda_1^1, ..., \lambda_k^l, ..., \lambda_K^M) \\ &= \log s! \prod_{k=1}^{K} \prod_{l=1}^{M} \frac{(\lambda_k^l)^{x_k^l}}{(x_k^l)!} \\ &= \log s! + \sum_{k=1}^{K} \sum_{l=1}^{M} \Big[x_k^l \log \lambda_k + x_k^l \log \lambda^l - \log(x_k^l)! \Big] \end{aligned}$$

by optimizing the criterion $VC(P \times Q, H)$ and taking into account the constraints $(\sum_{k=1}^{K} \lambda_k = 1$ and $\sum_{l=1}^{M} \lambda^l = 1)$ we obtain : $\lambda_k = \frac{x_k}{s}$ and $\lambda^l = \frac{x^l}{s}$
where $x_k = \sum_{l=1}^{M} x_k^l$ and $x^l = \sum_{k=1}^{K} x_k^l$

If all the x_k^l are too large, calculations of approximations allow to write :
$p(x, \lambda) = s! \prod_{k=1}^{K} \prod_{l=1}^{M} \frac{(\lambda_k^l)^{x_k^l}}{(x_k^l)!} \approx cte \exp - \sum_{k=1}^{K} \sum_{l=1}^{M} \frac{(x_k^l - s\lambda_k^l)^2}{s\lambda_k^l}$.

The maximization of $VC(P \times Q, H)$ thus returns to the minimization of criterion $C(P \times Q) = \sum_{k=1}^{K} \sum_{l=1}^{M} \frac{(sx_k^l - x_k x^l)^2}{s x_k x^l}$

5 Conclusion

This last, corresponds to the criterion of the χ^2 listed in the paragraph 2.3 which has been posed without referring to any concept of model. Through this approach, we established an approximate bond between the multinomial laws and the χ^2 metric. Thus, the χ^2 criterion which was suggested within a purely geometrical framework without referring to any concept of model, has been interpreted and cleared up by the model approach proposed in this paper.

6 Open Problem

In this work, devoted to a mixture modeling of block clustering algorithms, we have only considered the classification likelihood approach of mixture model.

The next step will be to study this problem under the likelihood approach and to propose an EM algorithm to estimate the parameters of the model.

References

[1] Y. Bencheikh, Classification croisée et modèles. Rairo operations research, vol.33, n°4 p 525-541,1999.

[2] Y. Khemal Bencheikh, Block mixture modeles and discrete data. Applied Mathematical Sciences, n°50,vol 2, p 2481-2488, 2008.

[3] G.Celeux, Classification et modèles. Rev statist. Appl, n°4, p 43-58, 1988.

[4] G. Govaert, Classification binaire et modèles, Rev.Statistique Appliquées , vol 38, p 67-81, 1990.

[5] G. Govaert, Simultaneous clustering of rows and columns, Control and Cybernetics, 24(4), p 437-458, 1995.

[6] A. Schroeder, Reconnaissance des composants d'un mélange, Thèse de Doctorat de 3ème cycle Université de Paris 6, 1974.

[7] A. Scott & Symons, Clustering methods based on likelihood ratio criteria, Biometrics 27. 387-397, 1971.