Clustering criteria for discrete data and latent class models

Gilles Celeux, Gérard Govaert

INRIA

UNITE DE RECHERCHE
INRIA-LORRAINE

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
BP 105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

N° 1122

Programme 5
Automatique, Productique,
Traitement du Signal et des Données

CLUSTERING CRITERIA FOR DISCRETE DATA AND LATENT CLASS MODELS

Gilles CELEUX
Gérard GOVAERT

Novembre 1989

RR_1122

CLUSTERING CRITERIA FOR DISCRETE DATA AND LATENT CLASS MODELS

CRITERE DE CLASSIFICATION POUR LES DONNEES DISCRETES ET MODELE DES CLASSES LATENTES

G. CELEUX and G. GOVAERT

INRIA Domaine de Voluceau Rocquencourt B.P. 105 78153 Le Chesnay Cedex
Université de Metz et INRIA-Lorraine Ile du Saulcy 57045 Metz Cedex 1

Abstract

We show that some well known clustering criteria for discrete data, the information criterion and the $\chi 2$ criterion, are closely related with the classification maximum likelihood criterion for the latent class model. Emphasis is placed on binary clustering criteria which are analyzed under the maximum likelihood approach for different multivariate Bernoulli mixtures. This alternative form of criteria reveals non-apparent aspects of clustering techniques. All the discussed criteria can be optimized with the alternating optimization algorithm.

Keywords : Binary clustering, L_1 distance, mixture, latent class models.

Résumé

Nous montrons que le critère de l'information et celui du $\chi 2$, critères de classification bien connus dans le cas de données discrètes, sont très liés au critère du maximum de vraisemblance classifiante appliqué au modèle des classes latentes. En particulier, nous traitons la classification de données binaires, analysée ici sous l'approche de l'estimation du maximum de vraisemblance d'un mélange de distributions multivariées de Bernoulli. L'étude de ces liens permet ainsi de mettre en évidence des aspects cachés de certaines techniques de classification.

Mots-clés : Classification binaire, distance L_1, mélange de lois de probabilité, modèle des classes latentes.

1. Introduction

Generally, clustering criteria are expressed in a geometric framework. This way of understanding a clustering criterion does not reveal some of its aspects. Some authors have investigated alternative forms of clustering techniques to improve their understanding, and recently Windham (1987) and Bryant (1988) have proposed a general approach to characterize optimization-based clustering methods. One way to obtain deep insight into clustering criteria based on within-cluster scatter matrix is to consider the classification maximum likelihood (CML) criterion for Gaussian mixtures (Scott and Symons 1971, Marriott 1975). This paper is devoted to studying the relations between some clustering criteria for discrete data and CML criteria for some specific mixtures. Section 2 is concerned with general clustering criteria and section 3 is concerned with clustering and CML methods for binary data. In section 4, we briefly report how the above mentioned criteria can be optimized using the alternating optimization algorithm. But, first, we describe the CML method for mixtures in a general setting.

Let $(\mathbf{x}_1,...,\mathbf{x}_n)$ represent sample values of a random vector whose probability distribution function (p.d.f.) is

$$f(\mathbf{x}) = \sum_{k=1}^{K} p_k \; f(\mathbf{x},\mathbf{a}_k)$$

where $p_k > 0$ for $k=1,...,K$; $p_1+...+p_K = 1$ and $f(.,\mathbf{a}_k)$ is a p.d.f. which has a specified parametric form where $\mathbf{a}_k$ denotes the parameter occuring in $f(.,\mathbf{a}_k)$ for $k=1,...,K$.

In the CML method, $(p_1,...,p_K)$, $(\mathbf{a}_1,...,\mathbf{a}_K)$ are estimated by choosing $\hat{\mathbf{a}}_1,...,\hat{\mathbf{a}}_K$ and $k(i)$ in the set $\{1,...,K\}$, for $i=1,...,n$, to maximize $\prod_{i=1}^{n} f(\mathbf{x}_i,\mathbf{a}_{k(i)})$. The $k(i)$'s define a partition $P=(P_1,...,P_K)$ on $(\mathbf{x}_1,...,\mathbf{x}_n)$ where, for $k=1,...,K$, $P_k = \{\mathbf{x}_i / k(i)=k \}$, and the CML criterion to be maximized can be written

$$C(P,\mathbf{a}) = \sum_{k=1}^{K} \ln (L(P_k,\mathbf{a}_k))$$

$$\text{where} \quad \mathbf{a}=(\mathbf{a}_1,...,\mathbf{a}_K) \quad \text{and} \quad L(P_k,\mathbf{a}_k) = \prod_{\mathbf{x}_i \in P_k} f(\mathbf{x}_i,\mathbf{a}_k)$$

Note that $p_1,...,p_K$ are estimated by the ratio $\text{card } P_1^* / n$,..., $\text{card } P_K^* / n$ where the P_K^*'s are the clusters obtained at the convergence of CML method.

Bryant and Williamson (1978) showed that, under general conditions, the CML estimate of $\mathbf{b}=(p_1,...,p_K,\mathbf{a})$ converges a.s. to $\mathbf{b}_0$, with in general $\mathbf{b}_0 \neq \mathbf{b}$, and thus is inconsistent.

2. Discrete data clustering and the latent class model

Let n objects be described by q categorical variables, with respective number of categories $m_1,...,m_q$. We denote $m = \sum_{j=1}^{q} m_j$ the total number of categories. The data can be represented with a n by m matrix $\mathbf{X}$:

$$\mathbf{X} = (x_i^{jh} \; ; \; h=1,...,m_j \; ; \; j=1,...,q \; ; \; i=1,...,n)$$

where $x_i^{jh} = 1$ if the object i belongs to the category h of the variable j and 0 otherwise.

We define

$$x^{jh} = \sum_{i=1}^{n} x_i^{jh} \qquad s = \sum_{i=1}^{n} \sum_{j=1}^{q} \sum_{h=1}^{m_j} x_i^{jh} = n\,q \qquad x_k^{jh} = \sum_{\mathbf{x}_i \in P_k} x_i^{jh} \qquad x_k = \sum_{j=1}^{q} \sum_{h=1}^{m_j} x_k^{jh} = q\,n_k$$

with n_k=card(P_k) and where P=(P_1,...,P_K) is a partition of the n objects.

In this context, two classical clustering criteria (see for instance Benzécri 1973) are

$$H(P) = \sum_{k=1}^{K} \sum_{j=1}^{q} \sum_{h=1}^{m_j} \frac{x_k^{jh}}{s} \ln \left(\frac{s\,x_k^{jh}}{x_k\,x^{jh}}\right) \qquad \text{(information criterion to be maximized)}$$

$$W(P) = \sum_{k=1}^{K} \sum_{j=1}^{q} \sum_{h=1}^{m_j} \frac{(s\,x_k^{jh} - x_k\,x^{jh})^2}{s\,x_k\,x^{jh}} \qquad (\chi^2 \text{ criterion to be minimized})$$

The information and the χ^2 statistics are usually regarded as equivalent within an acceptable limit. Bozdogan (1987) noticed, in another context, that in practice they can give widely different results. But, in the clustering context, many investigations reported in Benzécri (1973) and Govaert (1983) showed that both criteria behaved the same and can be effectively regarded as equivalent.

Now, the χ^2 clustering criterion is more widely used for two reasons : optimizing this criterion appears to be easier, and moreover, it is closely related to correspondence analysis. For algebraic reasons, we focus here on the information criterion H(P). It can be written

$$H(P) = \ln(s) + \sum_{k=1}^{K} \sum_{j=1}^{q} \sum_{h=1}^{m_j} \frac{x_k^{jh}}{s} \ln(x_k^{jh}) - \sum_{k=1}^{K} \frac{n_k}{n} \ln(n_k) - \sum_{j=1}^{q} \sum_{h=1}^{m_j} \frac{x^{jh}}{s} \ln(x^{jh})$$

Hence, maximizing H(P) leads to maximizing

$$H_1(P) = \sum_{k=1}^{K} \sum_{j=1}^{q} \sum_{h=1}^{m_j} x_k^{jh} \ln(x_k^{jh}) - q \sum_{k=1}^{K} n_k \ln(n_k).$$

We turn now to the CML criterion for the latent class model. The basic idea of the latent class model is that the observed associations between the q categorical variables are generated by the presence of several different "latent" classes within which the variables are independent. This may be formulated in mixture terms (Everitt 1984) by supposing that the $\{0,1\}^m$ valued vectors which characterize the n objects is a sample of a mixture of mutivariate multinomial distributions. A random vector arising from such a structure has a p.d.f. given by

$$f(\mathbf{x}) = \sum_{k=1}^{K} p_k \; f(\mathbf{x},\mathbf{a_k})$$

with

$$f(\mathbf{x},\mathbf{a_k}) = \prod_{j=1}^{q} \prod_{h=1}^{m_j} (a_k^{jh})^{x^{jh}} \quad \text{for } k=1,\ldots,K$$

where $\mathbf{a_k} = (a_k^{jh};\ h=1,\ldots,m_j\ ;\ j=1,\ldots,q)$ and a_k^{jh} gives the probability of category h of the variable j in the class k ; we have $\sum_{h=1}^{m_j} a_k^{jh} = 1$ for $j=1,\ldots,d$ and $k=1,\ldots,K$.

The CML criterion takes the form

$$C(P,\mathbf{a}) = \sum_{k=1}^{K} \sum_{\mathbf{x_i}\in P_k} \sum_{j=1}^{q} \sum_{h=1}^{m_j} x_i^{jh} \ln (a_k^{jh}) = \sum_{k=1}^{K} \sum_{j=1}^{q} \sum_{h=1}^{m_j} x_k^{jh} \ln (a_k^{jh})$$

where $\mathbf{a} = (\mathbf{a_1},\ldots,\mathbf{a_K})$.

It is easily seen (cf. Celeux 1988 for details) that the maximization of this criterion is equivalent to the maximization of the criterion

$$H_1(P) = \sum_{k=1}^{K} \sum_{j=1}^{q} \sum_{h=1}^{m_j} x_k^{jh} \ln (x_k^{jh}) \; - \; q \sum_{k=1}^{K} n_k \ln (n_k).$$

Notice that H_1 does not depend explicitly on the parameter $\mathbf{a}$.

This presentation of the information criterion and, also, of the χ^2 criterion, since both criteria are quasi-equivalent, deserves some remarks :

- For any clustering method based on the information criterion, it appears that each cluster is characterized by its center. This is not apparent in the initial form of this criterion.
- When using the χ^2 criterion, it is implicitly assumed that, within each cluster, the q categorical variables are mutually independent.
- The parallel between the χ^2 clustering criterion and the latent class model suggests a way to assess a relevant number of clusters when using the χ^2 criterion. Following Goodman (1974), it is possible to calculate the chi-squared statistic based upon the likelihood ratio associated to the latent class model and to choose the smallest number of clusters, in care of parsimony, which provides a reasonable goodness-of-fit chi-squared of the clustering structure.

$$c(P,\mathbf{a},\varepsilon) = \ln \frac{\varepsilon}{1-\varepsilon} \sum_{k=1}^{K} \sum_{\mathbf{x_i} \in P_k} d(\mathbf{x_i},\mathbf{a_k}) \;+\; n\,q\,\ln(1-\varepsilon).$$

This last form of the CML criterion shows that maximizing in ε and $\mathbf{a}$ can be performed separately. For fixed ε in $]0,1/2[$, $\ln(\varepsilon/(1-\varepsilon))$ is negative and maximizing $C(P,\mathbf{a},\varepsilon)$ leads to minimizing $W_1(P,\mathbf{a})$. After the optimal P^* and $\mathbf{a}^*$ which minimize $W_1(P,\mathbf{a})$ have been found, it is easy to show that $\varepsilon^* = W_1(P^*,\mathbf{a}^*)/(n\,q)$ maximizes $C(P^*,\mathbf{a}^*,\varepsilon)$.

The maximum likelihood approach for the criterion W_1 reveals that this clustering criterion can be thought of as inadequate in many situations since the parameters (ε and $1-\varepsilon$) of the Bernoulli distributions involved in the mixture model do not depend upon the variables or upon the clusters.

We consider, now, two Bernoulli mixture models which overcome this limitation. In the first one, the ε's depend on variables, and, in the second one the ε's depend on variables and on clusters. Hence the associated CML criteria give rise to two useful binary clustering criteria.

In the first model, the Bernoulli mixture components p.d.f.'s take the form

$$f(\mathbf{x},\mathbf{a_k},\varepsilon) = \prod_{j=1}^{q} \{ (\varepsilon^j)^{|x^j - a_k^j|} (1-\varepsilon^j)^{1-|x^j - a_k^j|} \} \quad \text{for } k=1,\ldots,K$$

where $\varepsilon = (\varepsilon^1,\ldots,\varepsilon^q)$ and $\varepsilon^j \in]0,1/2[$ for j=1,...,q.

Hence the associated CML criterion can be written

$$C(P,\mathbf{a},\varepsilon) = \sum_{k=1}^{K} \sum_{\mathbf{x_i} \in P_k} \sum_{j=1}^{q} \{ \ln \left(\frac{\varepsilon^j}{1-\varepsilon^j}\right) |x_i^j - a_k^j| + \ln(1-\varepsilon^j) \}$$

Maximizing $C(P,\mathbf{a},\varepsilon)$ is equivalent to minimizing

$$W_2(P,\mathbf{a},\varepsilon) = \sum_{k=1}^{K} \sum_{\mathbf{x_i} \in P_k} d_\varepsilon(\mathbf{x_i},\mathbf{a_k}) \;-\; A$$

where $d_\varepsilon(\mathbf{x_i},\mathbf{a_k}) = \sum_{j=1}^{q} \ln \left(\frac{1-\varepsilon^j}{\varepsilon^j}\right) |x_i^j - a_k^j|$ and $A = n \sum_{j=1}^{q} \ln(1-\varepsilon^j)$.

Notice that d_ε appears to be a weighted L_1-distance.

In the second model, the Bernoulli mixture components p.d.f.'s take the form

$$f(\mathbf{x},\mathbf{a_k},\varepsilon_k) = \prod_{j=1}^{q} \{ (\varepsilon_k^j)^{|x^j - a_k^j|} (1-\varepsilon_k^j)^{1-|x^j - a_k^j|} \}$$

where $\varepsilon = (\varepsilon_k^j;\ j=1,...,q\ ;\ k=1,...,K)$ and $\varepsilon_k^j \in\ ]0,1/2[$ for $j=1,...,q\ ;\ k=1,...,K$.

The associated CML criterion can be written

$$C(P,\mathbf{a},\varepsilon) = \sum_{k=1}^{K} \sum_{\mathbf{x_i}\in P_k} \{ \sum_{j=1}^{q} \ln \left(\frac{\varepsilon_k^j}{1-\varepsilon_k^j}\right) |x_i^j - a_k^j| + \ln(1-\varepsilon_k^j) \}$$

Maximizing $C(P,\mathbf{a},\varepsilon)$ is equivalent to minimizing

$$W_3(P,\mathbf{a},\varepsilon) = \sum_{k=1}^{K} \sum_{\mathbf{x_i}\in P_k} \{d_{\varepsilon_k}(\mathbf{x_i},\mathbf{a_k}) - A_k\}$$

where $d_{\varepsilon_k}(\mathbf{x_i},\mathbf{a_k}) = \sum_{j=1}^{q} \ln \left(\frac{1-\varepsilon_k^j}{\varepsilon_k^j}\right) |x_i^j - a_k^j|$ and $A_k = \sum_{j=1}^{q} \ln(1-\varepsilon_k^j)$.

Here again d_{ε_k} is a weighted L_1-distance.

Some comments are noteworthy :

- It is easy to show that the last Bernoulli mixture model is exactly the latent class model (see Govaert 1989 for details). Thus, to classifying binary data, we recommend employing the criterion W_3 rather than the χ^2 criterion on account of the attractive form of the clusters representation : each cluster is characterized by a binary vector.

- We can parallel the degree of complexity of the clustering criteria defined in the Bernoulli mixture context with clustering criteria defined in the Gaussian mixture context. The criteria W_1, W_2 and W_3 are, respectively, analogous to the within cluster inertia, the trace criterion and the within cluster determinant criterion (cf. Windham 1987 or Celeux 1988).

- Here again, the two remarks which ended section 2 are noteworthy for the W_1, W_2 and W_3 criteria.

4. The alternating optimization algorithm

All the clustering criteria discussed in this paper can be optimized by the so-called alternating optimization algorithm defined by Windham (1987) in a general setting. In the partitioning context, this algorithm is simply the dynamical clustering algorithms (see, for instance, Diday and Simon 1976). We sketch briefly this algorithm and we give the details of its implementation to optimize the criterion W_3, which appears to be the most complicated one.

Let $W(P,\mathbf{a})$ a clustering criterion to be minimized where $P=(P_1,...,P_k) \in \mathbf{P}_k$ is a partition on n objects and $\mathbf{a}=(\mathbf{a_1},...,\mathbf{a_k}) \in A$ is a representation of the partition. We adopt the formal description of the procedure given by Windham (1987) and adapt it to the partitioning context.

Choose $P^0 \in \mathbf{P}_k$, let $N=1$

1. Representation step : find $\mathbf{a}^N = \underset{\mathbf{a} \in A}{\text{argmin}}\ W(P^{N-1},\mathbf{a})$
2. Assignation step : find $P^N = \underset{P \in \mathbf{P}k}{\text{argmin}}\ W(P,\mathbf{a}^N)$
3. If $P^N=P^{N-1}$ then STOP else $N=N+1$: GOTO1.

It can be shown that a sequence of iterated $(P^N,\mathbf{a}^N)$ converges to a critical point which can be expected to be a local minimum (see Diday and Simon 1976, Bezdek et al. 1987 for conditions that ensure that this algorithm produces a local optimum).

To minimize the criterion $W_3(P,\mathbf{a},\varepsilon)$, the alternating optimization algorithm can be described as follows (see Govaert 1989 for details). For simplicity, we omit the iteration index N.

Representation step

For $k=1,...,K$; $j=1,...,q$

a_k^j is the value of highest frequency taken by the objects in P_k for the variable j

and

$$\varepsilon_k^j = e_k^j / n_k \quad \text{where} \quad n_k = \text{card } P_k \quad \text{and} \quad e_k^j = \sum_{\mathbf{x}_i \in P_k} |x_i^j - a_k^j|.$$

Assignation step

For $i=1,...,n$ $\mathbf{x}_i$ is assigned to the cluster P_k such that $d_{\varepsilon k}(\mathbf{x}_i,\mathbf{a}_k) - \sum_{j=1}^{q} \ln (1- \varepsilon_k^j)$ is minimum.

Remark :

In the general latent class model, which gave rise to the criterion W_3, the clusters representation is somewhat complicated and hence the solution obtained by the alternating optimization algorithm depends greatly on its initial position P^0. For this reason, we recommend choosing P^0 as the solution obtained when minimizing the simpler criterion W_1.

References

Aitchison, J. and Aitken, C.G.G. (1976), Multivariate binary discrimination by the kernel method. *Biometrika* **63**, 413-420.

Benzécri, J.P. (1973), Théorie de l'information et classification d'après un tableau de contingence. *L'Analyse des données*, tome 1, Dunod.

Bezdek, J.C., Hathaway, R.J., Howard, R.E., Wilson, C.A. and Windham, M.P. (1987), Local convergence analysis of a grouped variable version of coordinate descent. *JOTA* **54** n°3, 471-477.

Bozdogan, H. (1987), Selecting loglinear models and subset selection of variables in multiway contingency tables using Akaike's information criterion. *Classification and related methods of Data Analysis*, North Holland, 609-616.

Bryant, P. (1988), On characterizing optimisation-based clustering methods. *Journal of Classification* **5**, 81-84.

Bryant, P. and Williamson, J.A. (1978), Asymptotic behaviour of classification maximum likelihood estimates. *Biometrika* **65**, 273-281.
Celeux, G. (1988), Classification et modèles. *R.S.A.* **36** n°4, 43-58.
Diday, E. and Simon, J.C. (1976), Clustering analysis. *Digital Pattern Recognition*. Springer-Verlag, 47-94.
Everitt, B. (1984), *An introduction to latent variable models*. Chapman and Hall.
Goodman, L.A. (1974), Exploratory latent structure models using both identifiable and unidentifiable models. *Biometrika* **61**, 215-231.
Govaert, G. (1983), *Classification croisée*. Thesis Université Paris 6.
Govaert, G. (1989), Classification binaire et modèles, *R.S.A.* **37** (to appear).
Marriott, F.M.C. (1982), Separating mixtures of normal distributions. *Biometrics* **31**, 767-769.
Scott, A.J. and Symons, M.J. (1971), Clustering methods based on likelihood ratio criteria. *Biometrics* **27**, 387-397.
Windham, M.P. (1987), Parameter modification for clustering. *Journal of Classification* **4**, 191-214.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique