Applied Mathematical Sciences, Vol. 2, 2008, no. 50, 2481 - 2488

Block Mixture Models and Discrete Data

Khemal Bencheikh Yamina

Department of Mathematics
Ferhat Abbas University
Setif 19000, Algeria
bencheikh_00@yahoo.fr

Abstract

This work deals with links which exist between cross clustering methods and probabilistic models when data is discrete. For this, we define the notion of metric criterion, then we show that probabilistic criterion can always be considered as a metric criterion, and last, we establish conditions so that the reciprocal is right. These results are then applied to two families of metric criterion : the first ones are defined from quadratic distance and the second one, from the L1 distance.This approach allows to precise, in particular, the differences between adaptative distance method, the identifying method of mixture in gaussian case, and to show that the criteria defined with the city block metric is associated to a bilateral exponential distributions mixture, to propose a new criteria which may improve the qualify of the results.

Mathematics Subject Classification: 6207, 62H30

Keywords: cross automatic clustering, mixture model, city block metric

1 Introduction

One of the principal difficulties with automatic clustering methods is the choice of criterion and the metric used. When it is possible to find a probability laws mixture, such as the estimation of model parameters by clustering approach (Bencheikh [1], Celeux [3], Govaert [6], Schroeder [8], Scott [9]) leads to the optimization of a clustering numeric criterion, we obtain a new light for this criterion and the subjacent metric allowing to justify or eventually reject them. was interested in links which exist between automatic clustering and probabilistic models when data engage one set Govaert [5]. Here, we propose for application when data engage two sets, this is the case of cross automatic clustering. Through the two first paragraphs, we define two types of criterion and we study in which conditions the criteria can be equivalent. In the third paragraph, we presente links which exist between ber laws and adaptive distances. This approch allows to precise in particular differences between adaptive distances method and mixture identification method in bernoulli case.

2 Definition of the two types of criteria

In the following, initial data is supposed to be given as X table of n rows and p columns containing values taken by n individuals for p discrete variables. These values will be noted x_i^j , $i = 1, ..., n$ and $j = 1, ..., p$. Here, we consider two types of criteria: the first one which we call metric criterion uses the notion of dissimilarity measure, and the second one which we call probabilistic criterion, uses the notion of probabilistic mixture. We presente first of all these two types of criteria.

2.1 Metric criteria

Here we take up Govaert [7] representation.In this approch. We present the data in the form of set $T = I \times J$ (this is the cartesian result of I individuals by set J of variables). Each partition class is going to be represented by one element of set L which called set of kernels.

Let us have a function D of $E \times L$ in $\mathbb{R}^+$ which will measure dissimilarity between one element from E and one kernel.

The problem to be solved is to find the partition $P = (P_1, ..., P_K)$ of set I within K classes, the partition $Q = (Q^1, ..., Q^M)$ of J set in M classes and one$K.M$-times (λ_k^m) ; $k = 1, ..., K$ and $m = 1, ..., M$ (one by class) minimizing the following criterion:

$$W(P \times Q, L) = \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i \in P_k} \sum_{j \in Q^m} D(x_i^j, \lambda_k^m)$$

where $L = \{\lambda_k^m, k = 1, ..., K \text{ and } m = 1, ..., M\}$.

This criterion which depends on the D dissimilarity mesure, will be called metric criterion and noted $\mathbf{CM(E, L, D)}$.

Remark: The dynamic clusters algorithm Diday [5] provide a solution to this problem by constructing in a iterative way a successions of kernels-partitions making the criterion $W(P \times Q, L)$ decrease (Govaert [7])

2.2 Equivalent metric criteria

Definition: Let us say that two metric criteria are equivalent if and only if they are defined on the same sets E and L , and if it does exist a bijection ϕ of R strictly increasing verifying :

$$\mathbf{CM}(\mathbf{E},\mathbf{L},\mathbf{D}_1) = \phi \circ \mathbf{CM}(\mathbf{E},\mathbf{L},\mathbf{D}_2)$$

where D_1 and D_2 are the dissimilarity measures associated to the two criteria.

Remark: All optimal solutions corresponding to equivalent criteria are identical. Besides, seaching algorithms of local optima, as dynamic clusters algorithm, will give for equivalent criteria the same results.

We can easily show that if replacing D by an increasing linear function of D, we get an equivalent metric criterion.

Proposition 1: $\forall \alpha \in R^+$ and $\beta \in R$ the criteria $\mathbf{CM}(\mathbf{E},\mathbf{L},\mathbf{D})$ and $\mathbf{CM}(\mathbf{E},\mathbf{L},\alpha\mathbf{D}+\beta)$ are equivalent.

2.3 Probabilistic criterion

Here we take up Bencheikh [2] représentation. 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 and Ω population, the same for J set constitutes a sample of p size and Ω'population) of aleatory variable Z with values in E whose probability law admits the distibution function:

$$p(x) = \sum_{k=1}^{K} \sum_{m=1}^{M} p_k^m p(x/\lambda_k^m)$$

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

where $p(.,\lambda_k^m)$ is a distribution function on E belonging to a parameterized family of distribution function depending on the λ, p_k^m is the probability that a point of the sample follows the distribution law $p(.,\lambda_k^m)$. One will call these p_k^m the proportition of the mixture. One will note L the whole of the parameters.

The problem arising is the estimate of the numbers K and M of components of the mixture and the unknown parameters q_k^m:

$(q_k^m = (p_k^m, \lambda_k^m)\ ;\ k = 1,...,K$ and $m = 1,...,M)$ within sight of the sample $T = I \times J$. In the approach classification (Bencheikh [2], Schroeder [8], Scott et Symons [9]), one replaces the initial problem of estimate by the following problem:

To seek a partition $P \times Q = \{P_k \times Q^m\ ;\ k = 1,...,K$ and $m = 1,...,M\}$, *K and M* being supposed known, such as each class $P_k \times Q^m$ is assimilable to a subsample which follows a law $p(.,\lambda_k^m)$. It is then a question ofmaximizing the criterion of probability classifying according to:

$$VC(P \times Q, L) = \sum_{k=1}^{K} \sum_{m=1}^{M} \log l(P_k \times Q^m, \lambda_k^m)$$

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

$$VC(P \times Q, L) = \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i \in P_k} \sum_{j \in Q^m} \log p(x_i^j, \lambda_k^m)$$

This criterion which depends on the family of function of distributions F defined on E, will be called probabilistic criterion and noted $\mathbf{CP}(\mathbf{E}, \mathbf{F})$.

Remark: To maximize the preceding criterion, one can use the dynamic clusters algorithm Diday [4] provide a solution to this problem by constructing in a iterative way a successions of kernels-partitions making the criterion $VC(P \times Q, L)$ crease (Bencheikh [2]).

3 Links between the two types of criteria

After having defined these two types of criteria, we will generalize the links establish by Govaert [5] for simple automatic classification and will adapt them to the case of cross classification. We study how the two types of criteria defined previously can meet. For this, we define two types of links. The first makes it possible to associate any probabilistic criterion a metric criterion called metric criterion associated the probabilistic criterion; the second allows to extend the concept of equivalent criteria defined in the case of metric and probabilistic criteria. We study then the following problem: Is a given metric criterion associated a probabilistic criterion? this property is not true in general, but we establish a condition necessary and sufficient so that it is checked. Finally we show that it is enough to a weaker condition so that a metric criterion given either simply equivalent, and non associated with a probabilistic criterion.

3.1 Conditions for that a metric criterion has be associated to a probabilistic criterion

Proposition 2: $\mathbf{CP}(\mathbf{E}, \mathbf{F}) = -\mathbf{CM}(\mathbf{E}, \mathbf{L}, \mathbf{D})$

where L is the whole of definition of the parameters of the family F and D is defined by:

$$\forall x \in E, \forall \lambda \in L \qquad D(x, \lambda) = -\log p(x, \lambda)$$

The metric criterion thus defined is called associated metric criterion.

This result thus makes it possible to consider that all the probabilistic criteria are metric criteria.

3.2 Probabilistic criterion equivalent to a metric criterion

Since any probabilistic criterion can be regarded as a metric criterion, one can extend the definition of equivalent metric criteria.

Two probabilistic criteria are equivalent if the metric criteria associated are equivalent.

A probabilistic criterion CP_1 and a metric criterion CM_2 are equivalent if the metric criterion CM_1 associated CP_1 is equivalent to the metric criterion CM_2.

3.3 Conditions for that a metric criterion has be associated to a probabilistic criterion

Proposition 3: A metric criterion $\mathbf{CM}(\mathbf{E}, \mathbf{L}, \mathbf{D})$ is associated to a probabilistic criterion if and only if $\forall\ \lambda \in L$ the function $x \longmapsto e^{-D(x,\lambda)}$ is continuous and verify: $\sum_{x \in E} e^{-D(x,\lambda)} dx = 1$.

3.4 Probabilistic criterion equivalent to a metric criterion

By using proposition 2, one can obtain a weaker condition allowing to show than a metric criterion is equivalent (and non associated) with a probabilistic criterion.

Proposition 4: Being given the metric criterion $\mathbf{CM}(\mathbf{E}, \mathbf{L}, \mathbf{D})$, if there exist $r \succ 1$ such as the quantity $s = \sum_{x \in E} r^{-D(x,\lambda)}$ that is to say independent of λ, then the probabilistic criterion $\mathbf{CP}(\mathbf{E}, \mathbf{F})$ where F is defined by the following probabilities distributions: $p(x, \lambda) = \frac{1}{s} r^{-D(x,\lambda)}$ is an equivalent criterion.

The metric criterion associated to the family suggested is defined by the function of following :

$D'(x, \lambda) = -Log\left\{\frac{1}{s} r^{-D(x,\lambda)}\right\} = Logs + D(x, \lambda)Logr.$

the proposition 2 makes it possible to affirm that the metric criteria associated with D and with D' are equivalent. from where the announced result

After having studied the two types of criteria and the conditions under which these criteria can meet, we are interested now in the links existing between the laws of Bernoulli and the distance city-block or distance L_1?

4 Metric L_1 and distribution of Bernoulli

We will study a certain number of criteria resulting from the L_1 distance and definite on a whole of binary data and we show how these distances are directly related to the distributions of Bernoulli.

4.1 Distance L_1 fixed and identical for all the classes

Let us consider $T \subset E = \{0,1\}$ and the whole of the kernels coincide with E.

Let be the following distance:

$\forall x$ and $\lambda_k^m \in E \qquad D(x,\lambda_k^m) = \alpha\,|x-\lambda_k^m| + \beta$

where α is a positive real constant and β an unspecified reality. Since all the metric criteria are equivalent (proposition 1), one limits oneself then to the study of the following metric criterion:

$\forall x$ and $\lambda_k^m \in E \qquad D(x,\lambda_k^m) = \alpha\,|x-\lambda_k^m|$

by applying proposition 3 and by posing $r = e$, one notices that following quantity:

$s = \sum_{x\in E} e^{-\alpha|x-\lambda_k^m|} = 1 + e^{-\alpha}$ is independent of λ_k^m , one can affirm whereas there is a probabilistic criterion are equivalent to the criterion defined by . This one is defined by the following probability distribution:

$p(x,\lambda_k^m) = \begin{cases} \frac{1}{1+e^{-\alpha}} & si \quad x-\lambda_k^m = 0 \\ \frac{e^{-\alpha}}{1+e^{-\alpha}} & si \quad x-\lambda_k^m = 1 \end{cases}$

or while posing $\varepsilon = \frac{1}{1+e^{-\alpha}}$,the distribution takes the form :

$p(x,\lambda_k^m) = \varepsilon^{1-|x-\lambda_k^m|}(1-\varepsilon)^{|x-\lambda_k^m|}$

this expression of the distribution corresponds to one of the two laws of Bernoulli following:

1 with the probability ε and 0 with the probability $1-\varepsilon$

1 with the probability $1-\varepsilon$ and 0 with the probability ε

where $\varepsilon \in \left]\frac{1}{2},1\right[$, i.e. the law of Bernoulli of parameter $1-\varepsilon$ and the law of Bernoulli of parameter ε.

This mixture of law of Bernoulli depends on the parameter ε, this one measures the variation of a class in its center and depends neither on the variables nor of the classes ; what in certain situations can prove to be unrealistic. for that we propose metric the more general allowing to vary the parameter ε according to the partitions in lines and the partitions in columns.

4.2 Distance L_1 variable and dependent on each classes

the kernels are form $\lambda_k^m = (\alpha_k^m, \beta_k^m, \gamma_k^m)$ and the metric L_1 depends on each class $P_k \times Q^m$ and is defined by: $D(x,(\alpha_k^m,\beta_k^m,\gamma_k^m)) = \alpha_k^m\,|x-\lambda_k^m| + \gamma_k^m$

if $\gamma_k^m = \log(1+e^{-\alpha_k^m})$ i.e $s = 1$, the métric criterion (4.2) is associated to a probabilistic criterion of which the probability distribution is:

$$p(x,\lambda_k^m) = \frac{1}{1+e^{-\alpha_k^m}} e^{-\alpha_k^m|x-\beta_k^m|}$$

where $x \in \{0,1\}$ and $\beta_k^m \in \{0,1\}$. From where $p(x,\lambda_k^m)$ can be written in the form

$$p(x,\lambda_k^m)=\begin{cases}\frac{1}{1+e^{-\alpha_k^m}} & si \qquad x-\beta_k^m=0\\ \frac{e^{-\alpha_k^m}}{1+e^{-\alpha_k^m}} & si \qquad x-\beta_k^m=1\end{cases}$$

or while posing $\varepsilon_k^m=\frac{1}{1+e-\alpha_k^m}$, the distribution takes the form

$$p(x,\lambda_k^m)=(\varepsilon_k^m)^{1-\left|x-\beta_k^m\right|}(1-\varepsilon_k^m)^{\left|x-\beta_k^m\right|}$$

This time the parameter ε_k^m depends on each class $P_k\times Q^m$. It is then a mixture of Bernoulli 's laws with the parameter $\{\varepsilon_k^m,1-\varepsilon_k^m\}$ or $\{1-\varepsilon_k^m,\varepsilon_k^m\}$; where $\varepsilon_k^m\in\left]\frac{1}{2},1\right[$.

One then finds the most general method which one had developped in the approach of the models of mixture (Bencheikh [2]), where one could compare the traditional algorithms using metric fixed and identical for all the classes, and the algorithms using of the adaptive distances which adapt to each iteration with the various classes. These tests showed the interest to use the metric ones which depend on the variables and the classes in the case of simple classification, and of the metric ones which depends on the classes in lines and columns in the case of crossed classification, as it is the case in this study.

5 Conclusion

We thus saw that the comparison of the metric and probabilistic criteria makes it possible to bring a new lighting of many methods of classification to justify a posteriori certain constraints often imposed for technical reasons of optimization, to propose new criteria, but can be even more, this comparison makes it possible to explain the interest and the flexibility of the dynamic cluster algorithm ; the essential idea is the use of the concept of kernel associated with each class.this kernel corresponds quite naturally with the probabilistic criterion to the parameters of law probability associated for itch class.

References

[1] Y. Bencheikh, Classification croisée et modèles, *Rairo operations research,vol.* **33** (1999), 525 - 541.

[2] Y. Bencheikh, Classification croisée et distance L_1adaptative, *Revue de statistique appliquée,vol.* **03** (2002), 53 - 72.

[3] G. Celeux, Classification et modèles, *Revue de statistique appliquée,* **04** (1988), 43 - 58.

[4] E. Diday, Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes, *Thèse d'Etat Université Paris 6*, (1972).

[5] G. Govaert, Modèle de classification et distance dans le cas continu, *rapport de recherche INRIA*, (1989) n°988.

[6] G. Govaert, Classification binaire et modèles, *Revue de statistique appliquée, vol.* **38** (1990), 67 - 81.

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

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

[9] A. Scott, M.J. Symons, Clustering methods based on likelihood ratio criteria, *Biometrics* 27 (1971), 387-397.

Received: February 21, 2008