KYBERNETIKA — VOLUME 34 (1998), NUMBER 4, PAGES 393–398

FUZZY CLUSTERING OF SPATIAL BINARY DATA

Mô Dang and Gérard Govaert

An iterative fuzzy clustering method is proposed to partition a set of multivariate binary observation vectors located at neighboring geographic sites. The method described here applies in a binary setup a recently proposed algorithm, called Neighborhood EM, which seeks a a partition that is both well clustered in the feature space and spatially regular [2]. This approach is derived from the EM algorithm applied to mixture models [9], viewed as an alternate optimization method [12]. The criterion optimized by EM is penalized by a spatial smoothing term that favors classes having many neighbors. The resulting algorithm has a structure similar to EM, with an unchanged M-step and an iterative E-step. The criterion optimized by Neighborhood EM is closely related to a posterior distribution with a multilevel logistic Markov random field as prior [5, 10]. The application of this approach to binary data relies on a mixture of multivariate Bernoulli distributions [11]. Experiments on simulated spatial binary data yield encouraging results.

1. INTRODUCTION

A fuzzy clustering method is proposed to partition a set of n binary observations vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ ($\boldsymbol{x}_i \in \{0,1\}^d$, $1 \leq i \leq n$) located at neighboring geographic sites. For instance, it may be applied in biogeography to cluster n contiguous quadrats over which the occurrences of d animal species have been recorded. The aim is twofold: produce clusters that are homogeneous in the feature space, and account for some *a priori* hypothesis of spatial smoothness.

Numerous clustering methods have been proposed to take into account the spatial information of the data. Using the geographic coordinates as an additional pair of variates [4] or hierarchical clustering with contiguity constraints [13] tend to enforce the clusters to be spatially connected. Thus, for applications where the same class may appear in separate geographic regions, it seems more suitable to use methods like those of unsupervised image segmentation based on Markov Random Fields modeling [10, 8]; most of these techniques require however computationally intensive Monte-Carlo simulations.

In this work, the approach introduced in Ambroise et al [2] is adapted to the case of binary data. This approach is derived from the EM algorithm applied to mixture models [9]. It mainly consists in adding a spatial regularizing term to the criterion optimized by EM, and optimizing the new criterion by an iterative algorithm similar to EM.

2. CLUSTERING OF BINARY DATA USING MIXTURE MODELS

In clustering based on mixture models, the observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are assumed to be independently drawn from a mixture of k subpopulations in proportions $(p_1, \ldots, p_k)$, each subpopulation having probability distribution function $f_h(\cdot, \boldsymbol{\theta}_h)$ with unknown parameters $\boldsymbol{\theta}_h$ $(1 \leq h \leq k)$.

In the case of binary data, a mixture of k multivariate Bernoulli laws comes as a natural assumption. Following [11, 7], distribution f_h is characterized by its *center* $\boldsymbol{a}_h \in \{0,1\}^d$ and its *dispersion* $\boldsymbol{\varepsilon}_h \in]0; \frac{1}{2}[^d$ — i.e. $\boldsymbol{\theta}_h = (\boldsymbol{a}_h, \boldsymbol{\varepsilon}_h)$ — so that any observation $\boldsymbol{y} \in \{0,1\}^d$ belonging to group h occurs with probability

$$f_h(\boldsymbol{y} \; ; \; \boldsymbol{a}_h, \boldsymbol{\varepsilon}_h) \;=\; \prod_{j=1}^{d} \varepsilon_{hj}^{|y_j - a_{hj}|}(1-\varepsilon_{hj})^{1-|y_j - a_{hj}|}. \tag{2.1}$$

Expression (2.1) means that given class h, observation $\boldsymbol{y}$ arises from independent drawings of d univariate Bernoulli laws with parameters $1 - \varepsilon_{hj}$ if $a_{hj} = 1$, or ε_{hj} if $a_{hj} = 0$ $(1 \leq j \leq d)$. Thus, given class h, for each variable j $(1 \leq j \leq d)$, a_{hj} represents the value that occurs with highest probability, while ε_{hj} represents the probability that observation y_j differs from a_{hj}, thence the terminology of center for $\boldsymbol{a}_h$ and dispersion for $\boldsymbol{\varepsilon}_h$.

In clustering applications, the parameters of the mixture

$$\Phi = (p_1, \ldots, p_{k-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$$

are usually unknown. The EM algorithm has become a standard method to estimate these parameters from unlabeled data [9]. This iterative algorithm produces parameters estimate $\hat{\Phi}$ that locally optimize the log-likelihood function

$$L(\Phi) = \sum_{i=1}^{n} \log f(\boldsymbol{x}_i \; ; \; \Phi) = \sum_{i=1}^{n} \log \left(\sum_{h=1}^{k} p_h f_h(\boldsymbol{x}_i \; ; \; \boldsymbol{\theta}_h) \right).$$

In order to take into account the assumption of spatial smoothness on the classification, we take advantage of a relationship exhibited by Hathaway [12], where the EM algorithm applied to mixture models is proved to be equivalent to a grouped coordinate ascent on a function $D(\boldsymbol{c}, \Phi)$ of the parameters Φ and a fuzzy classification matrix $\boldsymbol{c} = (c_{ih})_{1 \leq h \leq k}^{1 \leq i \leq n}$ [12] :

$$D(\boldsymbol{c}, \Phi) \triangleq \sum_{h=1}^{k} \sum_{i=1}^{n} c_{ih} \log(p_h f_h(\boldsymbol{x}_i; \boldsymbol{\theta}_h)) - \sum_{h=1}^{k} \sum_{i=1}^{n} c_{ih} \log(c_{ih}).$$

Notice that in the case of a hard classification matrix and for Bernoulli mixtures having equal proportions and dispersions, the criterion $-D(\boldsymbol{c}, \Phi)$ is akin to a sum of intraclass inertia with a L_1 norm, so that its alternate optimization yields a *k-means* like algorithm using L_1 distances and binary kernels.

3. SPATIAL REGULARIZATION

The criterion $D(\boldsymbol{c}, \Phi)$ optimized by EM favors the homogeneity of the clusters in the variables space, but does not take into account the spatial information of the data. This second point can be addressed by adding the following spatial regularizing term to $D(\boldsymbol{c}, \Phi)$ [1, 2]:

$$G(\boldsymbol{c}) = \frac{1}{2} \sum_{h=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ih} c_{jh} v_{ij},$$

where v_{ij} are the weights of the geographic neighborhood system ($v_{ij} > 0$ if observation i is neighbor of observation j, $v_{ij} = 0$ otherwise). $G(\boldsymbol{c})$ is an increasing function of the number of neighbor pairs having same class. The degree of spatial smoothing is controlled via a weighting coefficient β, so that the new criterion to be optimized is defined as

$$U(\boldsymbol{c}, \Phi) = D(\boldsymbol{c}, \Phi) + \beta \, G(\boldsymbol{c}).$$

Optimizing alternatively criterion $U(\boldsymbol{c}, \Phi)$ over $\boldsymbol{c}$ and Φ yields an iterative algorithm having the same structure as EM, called Neighborhood EM. A neighborhood matrix $\boldsymbol{V} = (v_{ij})_{1 \leq i \leq n}^{1 \leq j \leq n}$ must first be computed according to the spatial relationships. The calculation is then initialized by choosing arbitrary initial values for the parameters of the mixtures, Φ^0, and the classification matrix, $\boldsymbol{c}^0$. The two following steps are then iteratively repeated until convergence is reached ($m + 1$ denotes the current iteration):

1. **E-step:**

$$\boldsymbol{c}^{m+1} = \arg \max_{\boldsymbol{c}} U(\boldsymbol{c}, \Phi^m).$$

The following equations are obtained, for $1 \leq i \leq n$ and $1 \leq h \leq k$:

$$c_{ih}^{m+1} = g_{ih}(\boldsymbol{c}^{m+1}) = \frac{p_h^m f_h(\boldsymbol{x}_i | \theta_h^m) \cdot \exp\{\beta \sum_{j=1}^{n} c_{jh}^{m+1} v_{ij}\}}{\sum_{\ell=1}^{k} p_\ell^m f_\ell(\boldsymbol{x}_i | \theta_\ell^m) \cdot \exp\left\{\beta \sum_{j=1}^{n} c_{j\ell}^{m+1} v_{ij}\right\}} \tag{3.1}$$

suggesting an iterative computing algorithm of the form $\boldsymbol{c} = g(\tilde{\boldsymbol{c}})$, where $\tilde{\boldsymbol{c}}$ is the old classification matrix. The convergence of this fixed point procedure can be proved under a bounding condition on β; the convergence conditions and its proof will be published in a forthcoming paper [3]. From a practical point of view, satisfying results are obtained using only one iteration of this procedure to compute the new classification matrix $\boldsymbol{c}^{m+1}$.

2. **M-step:**

$$\Phi^{m+1} = \arg \max_{\Phi} U(\boldsymbol{c}^{m+1}, \Phi) = \arg \max_{\Phi} D(\boldsymbol{c}^{m+1}, \Phi).$$

Thus, to compute the parameters of the mixture, one can use the same formulae as in the M-step of the EM algorithm. More specifically, for a Bernoulli mixture model, the following re-estimation formulae are obtained, for $1 \leq h \leq k$ and $1 \leq j \leq d$:

$$n_h = \sum_{i=1}^{n} c_{ih}^{m+1} \tag{3.2}$$

$$p_h^{m+1} = \frac{n_h}{n} \tag{3.3}$$

$$a_{hj}^{m+1} = \text{rounded value of } \frac{1}{n_h} \sum_{i=1}^{n} c_{ih}^{m+1} x_{ij} \tag{3.4}$$

$$\varepsilon_{hj}^{m+1} = \frac{1}{n_h} \sum_{i=1}^{n} c_{ih}^{m+1} |x_{ij} - a_{hj}^{m+1}|. \tag{3.5}$$

a_{hj}^{m+1} can be interpreted as the most frequently occuring value within class h for variable j, and and ε_{hj}^{m+1} as the proportion of observations that differ from this value.

A hard classification can be obtained at the convergence of NEM by assigning to observation i the class in which it has the highest grade of membership ($\ell = \arg\max_{1 \leq h \leq k} c_{ih}$).

4. NUMERICAL EXPERIMENTS

The behavior of the Neighborhood EM algorithm will be illustrated on a simple artificial data set. The $n = 400$ observations are spatially located on a regular grid of 20 lines by 20 columns. Their class was randomly generated using a Gibbs sampler algorithm [10], according to a Markov random field with $k = 4$ levels, 4 nearest-neighbors contexts and $\beta = 1.2$ (see Figure 4.1.a). Each observation, consisting of a vector of $d = 5$ binary values, has been drawn according to the Bernoulli distribution of its class. The centers of the 4 classes are respectively $a_1 = (01111)$, $a_2 = (11100)$, $a_3 = (00011)$, and $a_4 = (10000)$. and the dispersion is the same for all classes and variables, $\varepsilon_{hj} = \varepsilon = 0.15$ ($1 \leq h \leq k, 1 \leq j \leq d$).

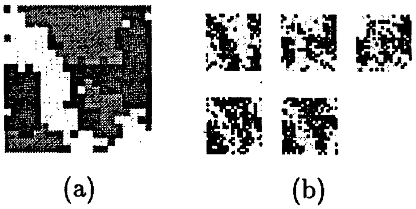

Fig. 4.1. Simulated spatial binary data. (a) Simulated partition with 4 classes using a Gibbs sampler (20 × 20 pixels image). (b) Simulated data with Bernoulli distribution in dimension $d = 5$ (ones are represented by a dark pixel, zeros by a clear pixel).

The NEM algorithm was initialized by drawing randomly the centers out of the observations. The initial partition was then computed from the initial parameters by a “blind” classification using no spatial context. The final result was obtained by retaining the solution that provided the highest criterion $U(c, \Phi)$ out of 30 randomly initialized trials.

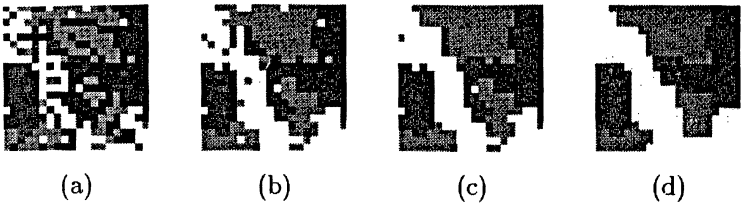

Fig. 4.2. Partitions obtained by NEM with different values of β. (a) $\beta = 0$, error = 23.2 %; (b) $\beta = 0.5$, error = 10.2 %; (c) $\beta = 1.4$, error = 5.2 %; (d) $\beta = 4$, error = 11.5 %.

The partitions obtained using four representative values of β arc displayed in Figure 4.2. When $\beta = 0$, the NEM algorithm is identical to EM, i.e. the spatial information is not used at all; due to the dispersion of the classes in the feature space, the misclassification error is quite high (23.2 %) (Figure 4.2.a). When $\beta = 0.5$, the classification is more accurate because the spatial context reduces ambiguities during the clustering process (Figure 4.2.b). One of the best classifications is obtained with $\beta = 1.4$, yielding only 5.8 % of misclassified pixels (compare Figure 4.2.c and the simulated partition Figure 4.1.a). When $\beta = 4$, the error is higher, because of a slight over-smoothing effect (Figure 4.2.d).

5. CONCLUSION

This study has shown that the clustering of a set of multivariate binary observations, taking into account their spatial relationships, may be achieved by combining the Neighborhood EM algorithm with Bernoulli mixture models. The formulation of the algorithm for Bernoulli mixtures has been given, and its practical relevance has been illustrated on a simulated data set. Compared to image segmentation techniques based on Markov random fields, to which it is closely related, Ambroise [1] points out that the NEM algorithm produces segmentations roughly equivalent to those of the Gibbsian EM algorithm [8], its main advantage being its relative efficiency due to its deterministic, iterative scheme. In comparison with other known fuzzy clustering methods, such as Bezdek's fuzzy C-means [6], this “binary” version of the NEM algorithm provides two features of interest: a) it relies on a statistical model of Bernoulli mixtures suited to binary data; b) the spatial information of the data is taken into account without enforcing the clusters to be made of one unique geographic region.

The work presented here suggests that the mixture-based approach of the NEM algorithm should be able to cluster spatial observations containing both categorical *and* continuous data. This could be simply done by considering class distributions mixing, for instance, multinomial probabilities and normal densities.

As is apparent from the simulated example above, the clustering result of the Neighborhood EM algorithm depends largely on the choice of the spatial coefficient

β. The problem of determining automatically the spatial coefficient is currently being investigated. A heuristic method based on the likelihood of the class parameters is being tested, and displays encouraging results on simulated data sets. A real case study is underway on an ecological dataset.

(Received December 18, 1997.)

REFERENCES

[1] C. Ambroise: Approche probabiliste en classification automatique et contraintes de voisinage. PhD Thesis, Université de Technologie de Compiègne 1996.

[2] C. Ambroise, M. V. Dang and G. Govaert: Clustering of spatial data by the EM algorithm. In: Amílcar Soares (J. Gómez-Hernandez and R. Froidevaux, eds.), geoENV I – Geostatistics for Environmental Applications, Kluwer Academic Publisher 1997, pp. 493–504.

[3] C. Ambroise and G. Govaert: Convergence of an EM type algorithm for spatial clustering. Pattern Recognition Lett., to appear.

[4] B. J. L. Berry: Essay on Commodity Flows and the Spatial Structure of the Indian Economy. Research paper 111, Departement of Geography, University of Chicago 1966.

[5] J. E. Besag: Spatial analysis of dirty pictures. J. Roy. Statist. Soc. *48* (1986), 259–302.

[6] J.C. Bezdek and P. F. Castelaz: Prototype classification and feature selection with fuzzy sets. IEEE Trans. Systems Man Cybernet. *SMC-7* (1977), 2, 87–92.

[7] G. Celeux and G. Govaert: Clustering criteria for discrete data and latent class models. J. Classification *8* (1991), 157–176.

[8] B. Chalmond: An iterative gibbsian technique for reconstruction of m-ary images. Pattern Recognition *22* (1989), 6, 747–761.

[9] A. P. Dempster, N. M. Laird and D. B. Rubin: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. *39* (1977), 1–38.

[10] S. Geman and D. Geman: Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Trans. Pattern Analysis Machine Intelligence *PAMI-6* (1984), 721–741.

[11] G. Govaert: Classification binaire et modéles. Rev. Statist. Appl. *38* (1990), 1, 67–81.

[12] R. J. Hathaway: Another interpretation of the EM algorithm for mixture distributions. Statist. Probab. Lett. *4* (1986), 53–56.

[13] P. Legendre: Constrained clustering. Develop. Numerical Ecology. NATO ASI Series *G 14* (1987), 289–307.

Dr. Mô Dang and Pr. Gérard Govaert, UMR CNRS 6599 Heudiasyc, Université de Technologie de Compiègne, B.P. 20529, 60205 Compiègne cedex. France.
e-mails: Van.Mo.Dang@utc.fr, Gerard.Govaert@utc.fr