Fuzzy c-Means Clustering with Regularization by K-L Information

Hidetomo ICHIHASHI
Kiyotaka MIYAGISHI, Katsuhiro HONDA

Industrial Engineering, Graduate School of Engineering, Osaka Prefecture University
1-1 Gakuencho, Sakai, Osaka 599-8531 Japan, ichi@ie.osakafu-u.ac.jp

Proc. of 10th IEEE International Conference on Fuzzy Systems, pp.924-927(2001), partly revised May 2006.

Abstract

Gaussian mixture model or Gaussian mixture density model(GMM) uses the likelihood function as a measure of fit. We show that just the same algorithm as the GMM can be derived from a modified objective function of Fuzzy c-Means (FCM) clustering with the regularizer by K-L information, only when the parameter λ equals 2. Although the fixed-point iteration scheme of FCM is similar to that of the GMM, the FCM has more flexible structure since the algorithm is based on the objective function method. In a slightly different manner such as installing a deterministic annealing or an addition of Gustafson and Kessel's constraint, the proposed algorithm is likely to provide more valid clustering results.

I Introduction

Entropy method that uses an additional term of entropy for fuzzification in the Fuzzy c-Means (FCM) [1] was proposed by Miyamoto *et al.* [14]. A similar entropy term was considered to prevent trivial solution by Dave and Krishnapuram [2] within the scope of Possibilistic c-Means (PCM) due to Krishnapuram and Keller [12]. The thesis of this paper is that there exist a close relationship between the FCM clustering and the Gaussian mixture model (GMM) [5, 7, 17, 19]. Gath and Geba's algorithm [6] is an extension of Gustafson and Kessel' FCM [9] and is apart from density estimation but is similar to GMM algorithm. GMM is to approximate PDFs by a mixture of Gaussian PDFs, i.e., the problem of extracting each Gaussian component in a given data set. We propose a new FCM clustering objective function with an additional term of the Kullback-Leibler information. We show that the same algorithm as the GMM Expectation Maximizing(EM) [4, 13] algorithm is derived from the FCM with K-L information term (KLFCM) only when a parameter λ, which specifies fuzziness of clusters, equals 2, but for other values of λ, no GMM exists which corresponds to the KLFCM. Thus unlike the GMM, the values of λ can be decreased during the iteration through the algorithm, i.e., a simple deterministic annealing (DA) [11, 16, 18], in which λ is regarded as temperature.

Numerical examples show that the annealing tends to avoid trapping into the local extremum for some data set. Further more we show that the proposed KLFCM clustering plays a roll of PCM or noise clustering (NC)[12, 10] by the addition of a constraint of Gustafson and Kessel [9]. Simulation experiment shows how well the accumulated points are extracted by the modified KLFCM.

II GMM algorithm and KLFCM clustering

The Gaussian mixture density model (GMM) [5, 19] is well recognized as a statistical technique for density estimation where the probability density function (PDF) is approximated by a mixture of Gaussian distribution functions rather than a single parametric function. The best fitting PDF for the data set will be defined by a parameter set that maximizes the likelihood. The likelihood function is a function of the model parameters and it gives a measure of how well the PDF defined by the parameters fits the given data set. There is thus a need to find an estimate of these maximum likelihood parameters. The EM algorithm, composed of E-step and M-step, is used to fit a fixed number of Gaussians to a data set. If a parameter set maximizes the likelihood, then these parameters are considered to define the best fitting PDF for the data set. The GMM algorithm represents the data set as a collection of Gaussian distributions whereas the FCM clustering regards it as a collection of clusters. Both of these algorithm aim to find a mathematical function that represents the data distribution most properly. As in the FCM clustering, the GMM algorithm also alternately estimates the group membership of the data points using a previous estimate of the parameters of the model, and then updates this estimate of the parameters using the estimate of the group membership of the data points.

Let s dimensional vector $\boldsymbol{x}_k$ represents the kth object or sample from a given set of n unlabelled objects. Each feature vector consists of s real-valued measurements describing the features of the object represented by $\boldsymbol{x}$. The means of c Gaussian distributions are denoted by $\boldsymbol{v}_i$. ϕ^* is a set of parameters with estimated values. ϕ is a set of updated parameters. In the Gaussian mixture model, the PDF $g(\boldsymbol{x})$, is approximated by a mixture of PDF denoted by $g(\boldsymbol{x}|\phi) = \sum_{i=1}^{c} \pi_i p_i(\boldsymbol{x}|\phi_i)$, The covariance matrix A_i, mean $\boldsymbol{v}_i$ of Gaussian PDF $p_l(\boldsymbol{x}|\phi_i)$ and ratio π_i are estimated by the maximum likelihood approach. When $\boldsymbol{x}_k$ is given, the posteriori probability is

$$u_{lk} = \frac{\pi_l^* p_l(\boldsymbol{x}_k|\phi_l^*)}{\sum_{j=1}^{c} \pi_j^* p_j(\boldsymbol{x}_k|\phi_j^*)} \tag{1}$$

The proportion π_l represents the contribution of the lth Gaussian PDF. Then, the EM algorithm maximizes log-likelihood,

$$Q(\phi|\phi^*) = \sum_{i=1}^{c}\sum_{k=1}^{n} \log[\pi_i p_i(\boldsymbol{x}_k|\phi_i)]u_{ik} \tag{2}$$

The algorithm is the repetition through E-step and M-step. In the GMM, the covariance matrix A_i is decision variable. To be confident that the resulting parameters are at a global maximum of the likelihood function it is desired to run the GMM algorithm a number of times using different initializations.

The FCM clustering partition the data set by introducing the membership to fuzzy clusters. p dimensional vector $\boldsymbol{v}_i$ denotes prototype parameter (i.e., cluster center), which is used instead of the mean of the Gaussian distribution. The u_{ik} denotes the membership of the kth data to the ith cluster. The clustering criterion used to define good clusters for fuzzy c-means partitions is the FCM objective function:

$$J_m = \sum_{i=1}^{c}\sum_{k=1}^{n} (u_{ik})^m d_{ik} \tag{3}$$

where m is the weighting exponent on each fuzzy membership. The larger m is, the fuzzier the partition becomes. The nonnegative membership u_{ik} sum to one with respect to c clusters for each object.

$$d_{ik} = (\boldsymbol{x}_k - \boldsymbol{v}_i)^T A_i^{-1} (\boldsymbol{x}_k - \boldsymbol{v}_i) \tag{4}$$

is a measure of the distance from $\boldsymbol{x}$ to the ith cluster prototype. The Euclidean distance metric is often used where A_i is a diagonal matrix. In the modified FCM by Gustafson and Kessel [9], the matrices A_i are also decision variables and the size of $|A_i|$ is constrained to a certain value.

The optimal u_{ik} and $\boldsymbol{v}_i$ for all i and k are sought using a fixed-point iteration scheme, which is similar to the GMM algorithm. There is one technical trick in the basic FCM. When $\boldsymbol{x}_k$ and $\boldsymbol{v}_i$ assume the same value and the distance d_{ik} between them equals 0, then the membership u_{ik} goes to infinite. In Miyamoto *et al.* [14], an entropy term K and a positive parameter λ are introduced and $J_\lambda = J_1 + \lambda K$ is minimized instead of J_m.

$$J_\lambda = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik} d_{ik} + \lambda \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik} \log u_{ik} \tag{5}$$

This approach is referred to as entropy regularization. The trick in the basic FCM is not needed. By replacing the entropy term in Eq.(5) with K-L information and including constraint term in a Lagrangian function, we consider the minimization of the following objective function.

$$\begin{aligned} J_{\lambda\tau} &= \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik} d_{ik} + \lambda \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik} \log \frac{u_{ik}}{\pi_i} \\ &+ \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik} \log |A_i| - \sum_{k=1}^{n} \eta_k (\sum_{i=1}^{c} u_{ik} - 1) \\ &- \tau(\sum_{i=1}^{c} \pi_i - 1) \end{aligned} \tag{6}$$

d_{ik} is as in Eq.(4), η_k and τ are Lagrangian multipliers whose corresponding terms represent the constraints that both the sum of u_{ik} and the sum of π_i with respect to i equal one respectively. As the entropy term in Eq.(5) forces memberships u_{ik} to take similar values, i.e., to obtain fuzzy clusters, the second term of Eq.(6) becomes zero if u_{ik}, $k = 1, ..., n$ take the same value π_i within the ith cluster for all i. Thus the term represents the proximity between the two distributions of u_{ik} and π_i, or K-L information. If $u_{ik} \simeq \pi_i$ for all k and i, partition becomes very fuzzy but when λ is 0 the optimization problem with respect to u_{ik} reduces to a linear one and the solution u_{ik} are obtained at extremal point (i.e., 0 or 1). Fuzziness of the clusters can be controlled by λ. We can derive some necessary conditions for optimality of Eq.(6).

$$u_{ik} = \frac{\pi_i \exp\left(-\frac{1}{\lambda} d_{ik}\right) |A_i|^{-\frac{1}{\lambda}}}{\sum_{j=1}^{c} \pi_j \exp\left(-\frac{1}{\lambda} d_{jk}\right) |A_j|^{-\frac{1}{\lambda}}} \tag{7}$$

$$\pi_i = \frac{\sum_{k=1}^{n} u_{ik}}{\sum_{j=1}^{c}\sum_{k=1}^{n} u_{jk}} = \frac{1}{n}\sum_{k=1}^{n} u_{ik} \tag{8}$$

The above equation means that π_i signifies the volume or ratio of the data involved in the ith fuzzy cluster.

$$A_i = \frac{\sum_{k=1}^{n} u_{ik} (\boldsymbol{x}_k - \boldsymbol{v}_i)(\boldsymbol{x}_k - \boldsymbol{v}_i)^T}{\sum_{k=1}^{n} u_{ik}} \tag{9}$$

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n} u_{ik} \boldsymbol{x}_k}{\sum_{k=1}^{n} u_{ik}} \tag{10}$$

The algorithm is the repetition through Eqs.(7)-(10).

As shown above, when the parameter λ equals 2, we have the same algorithm as one in the GMM. As in the GMM algorithm, the solution of the KLFCM clustering algorithm often traps into a local extremum and A_i may become singular. Our attempt to improve this deficiency by a simple DA is discussed in section 5.

III Comparison between GMM and KLFCM

As we stated in the previous section, when $\lambda = 2$, the KLFCM algorithm is the same as one in the GMM. We discuss here whether there exist any mixture model other than Gaussian one, when λ is not equal to two. Eq.(7) is a well-known Bayes rule to obtain a posteriori probability from a priori probability. We now consider the following function by which Eq. (7) representing a posteriori probability, can be derived in E-step. Let

$$p_i(\boldsymbol{x}) \quad = \quad \frac{|A_i^{-1}|^{1/\lambda}}{K} \exp\left(-\frac{1}{\lambda} d_{ik}\right) \tag{11}$$

where K is a constant. We discuss here whether the function (11) is a probability density function or not. By properly specifying the value K, if the integral of Eq. (11) attains one, the above function can be some probability density function and Eq.(7) is obtained by the Bayes theorem. For simplicity's shake let us confine the discussion only to two-dimensional case. The integral of the function (11) over the entire domain is

$$\frac{|A_i^{-1}|^{1/\lambda}}{K} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{\lambda}(\boldsymbol{x}-\boldsymbol{v}_i)^T A_i^{-1}(\boldsymbol{x}-\boldsymbol{v}_i)\right) d\boldsymbol{x} = \frac{\lambda\pi}{K}|A_i|^{\frac{\lambda-2}{2\lambda}} \tag{12}$$

If $K = \lambda\pi$ and $\lambda = 2$, then Eq.(11) is Gaussian, but if $\lambda \neq 2$, in order that the integral attains one, $K = \lambda\pi|A_i|^{\frac{\lambda-2}{2\lambda}}$ must be satisfied. Since K depends $|A_i|$ and $|A_i|$ might be different for each i, K cannot assume a single constant value for all i. Eq.(11) is a sole formula, which can derive Eq.(7) by using Bayes rule in E-step. Thus when $\lambda \neq 2$ there is no corresponding mixture density and the proposed KLFCM clustering is unique and novel one so long as we confine ourselves to a rigid definition of the probability. It is easy to generalize the discussion to high dimensional case.

In KLFCM clustering, parameter λ can be changed freely during the iteration of the algorithm, we can gradually reduce the value as the deterministic annealing procedure [16] to obtain better solution. $\sum_{k=1}^{n} u_{ik}$ indicates the number of data included in the ith cluster. From $J_{\lambda\tau}$, A_i becomes a fuzzy variance-covariance matrix as in Eq.(9). Let A_i has m linearly independent eigenvectors, i.e., principal component vectors $(\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_m)$. Then $\log|A_i| = \log\prod_{j=1}^{m}\delta_j^2 = \sum_{j=1}^{m}\log\delta_j^2$ where δ_j^2 is an eigen value of A_i. Therefore $\log|A_i|$ is equivalent to the sum of the log transformed variances of principal components. The third term of $J_{\lambda\tau}$, $\sum_{k=1}^{n} u_{ik}\log|A_i|$, represents the variation weighted by the number of data. Mahalanobis distance of Eq.(4) is defined by A_i^{-1}.

IV Noise Clustering with G-K constraint

In this section we propose to adopt a constraint by Gustafson and Kessel[1, 9] to restrict the variation of data in the $c + 1th$ cluster or the size of $|A_{c+1}|$. We add it as the fourth term of the objective function of KLFCM and call it GKFCM. The condition such that sum of the memberships of each datum to the cluster is 1, corresponds to a normalization of the memberships per datum. Due to the constraint, the basic FCM is often referred to as probabilistic clustering. We can drop this normalization condition so that the more distant data from each cluster such as the outlier can receive lower degree of membership and thus avoid the influence of noise data. The FCM clustering without the normalization is called PCM. NC was proposed by Ohashi[15] and independently by Dave[2] so that the noise data will be involved in the noise cluster. GKFCM clustering can produce similar results as the PCM or NC due to the additional constraint of the type by Gastafson and Kessel. Next section provides a simulation result, which is similar to one by PCM or NC. Gustafson and Kessel[9] introduced matrix A_i which defines Mahalanobis distance and to shape the cluster to better fit a given data distribution. $|A_c + 1|$ is to be ρ and the new objective function becomes

$$\begin{aligned} J_{\lambda\tau} \quad &= \quad \sum_{i=1}^{c+1}\sum_{k=1}^{n} u_{ik}d_{ik} + \lambda\sum_{i=1}^{c+1}\sum_{k=1}^{n} u_{ik}\log\frac{u_{ik}}{\pi_i} \\ &+ \quad \sum_{i=1}^{c+1}\sum_{k=1}^{n} u_{ik}\log|A_i| \\ &+ \quad \gamma(\log|A_{c+1}| - \rho) \\ &- \quad \sum_{k=1}^{n}\eta_k(\sum_{i=1}^{c+1} u_{ik} - 1) - \tau(\sum_{i=1}^{c+1}\pi_i - 1) \end{aligned} \tag{13}$$

γ is a Lagrangian function and ρ is a positive constant.

V Simulation Experiment

We have developed KLFCM Simulator, which is available on the web site
`http://www.cs.osakafu-u.ac.jp/hi/ichi/ichi_j.htm`

KLFCM clustering tends to form clusters including sparsely scattered wide area data when λ is relatively large. The GMM is equivalent to KLFCM with $\lambda = 2$ and when λ is greater than 2, it is efficient to extract locally dense clusters. Now we regard λ as temperature. The annealing schedule is in mimicry of Geman and Geman[8]. We set λ^*=8, and $\lambda(t) = \lambda^*/\ln(2+t)$ where t denotes iteration number. λ was fixed after it reached to 2, i.e., when the iteration number was 53 and $\lambda(53) = 1.99$, λ was fixed to 2. We have conducted a series of tests and have chosen the best results in which the objective function assumed the smallest value among 100 trials. In table 1 "G-K" means to apply Gustafson-Kessel's constraints ($\log|A_i| = \rho_i$) for all c clusters. Table 1 shows that KLFCM with the deterministic annealing [16] tends to converge to the best solution in most cases and thus surpasses other approaches. In the table, GMM without annealing is the worst. We have also conducted hard clustering and λ was decreased gradually to zero to obtain clusters with memberships of zero or one. The KLFCM clustering clearly surpassed the GMM without annealing in the comparison. GKFCM clustering, which includes an additional constraint given by Gustafson and Kessel, provides clustering results similar to those by PCM. We have tested the case in which noise data are widely recognized and clear point accumulations present as shown in Fig. 1. Fig. 1 shows the results by GMM ($c = 4$). Fig. 2 shows that by GKFCM ($c + 1 = 5$). The contour curves of noise cluster are not depicted in Fig. 2 to clarify the result. Fig. 2 indicates that the GKFCM can extract clear point accumulations from noise data. This kind of characteristics is endemic to PCM[12, 10] or NC, which abolish the effect of widely spread noise data. By the computation of a covariance matrix for each cluster the FCM is enhanced with more flexible structure than GMM since the algorithm is based on the objective function method.

Table I Simulation Results

Method	Freq.	J
GMM ($\lambda = 2$)	11	3.89×10^{10}
KLFCM (Annealing)	75	3.89×10^{10}
G-K (Annealing)	7	3.85×10^{10}

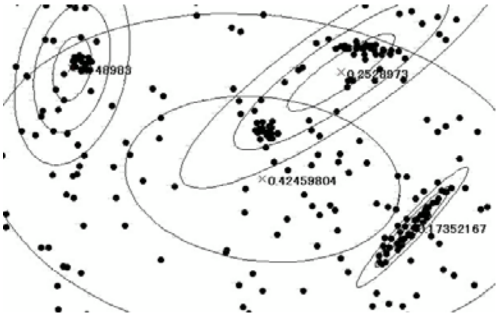

Fig. 1 Result by GMM

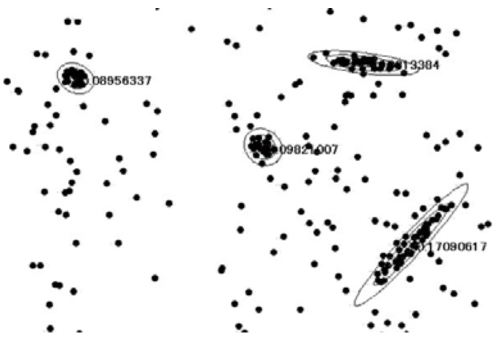

Fig. 2 Result by GKFCM

VI Conclusion

Since KLFCM is an objective function method rather than a maximum likelihood approach, it is easy to introduce additional objectives or constraints e.g. a G-K constraint to abolish the effect of noise. In KLFCM, the parameter λ controls the fuzziness of clusters whereas in GMM, λ must be 2. The initial values of the decision variables do not strongly affect clustering results since a simple DA can be applicable by regarding parameter λ as temperature.

Accknowledgement: This work was supported through the scientific research fund from the Ministry of Education, Culture, Sports, Science and Technology of Japan (#13680375)

References

1. J. C. Bezdek: *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press(1981)
2. R. N. Dave: *Pattern Recognition Letters*, Vol.12, pp.657-664(1991).
3. R. N. Dave and R. Krishnapuram: *IEEE Trans. Fuzzy Syst.*, Vol.5, No.2, pp.270-293(1997).
4. A. P. Dempster, N. M. Laird and D. B. Rubin: *J. Royal Statist. Soc.*, Vol.B-39, pp.1-38(1977).
5. R. O. Duda and P. E. Hart: *Pattern Classification and Scene Analysis*, Wiley, New York(1973).
6. I. Gath and A. B. Geba: *IEEE Trans. Pattern Anal. Machine Intell.*, Vol.11, pp.773-781(1989)
7. R. A. Redner and H. F. Walker: *SIAM Review*, Vol.26, No.2, pp195-239(1984)
8. S. Geman and D. Geman: *IEEE Trans. Pattern Anal. Machine Intell.*, Vol.PAMI-6, pp.721-741(1984).
9. D. E. Gustafson and W. C. Kessel: in *Proc. IEEE CDC*, Vol.2, pp.761-766(1979)
10. F. Höppner, F. Klawonn, R. Kruse, T. Runkler : *Fuzzy Cluster Analysis, (Methods for Classification, Data Analysis and Image Recognition)*, John Wiley & Sons, Ltd.(1999)

11. S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi: *Science*, Vol.220, pp.671-680(1983)

12. R. Krishnapuram and J. Keller: *IEEE Trans. Fuzzy Syst.*, Vol.1, pp.98-110(1993)

13. G. J. McLachlan and T. Krishnan: *The EM algorithm and extensions*, John Wiley and Sons(1997).

14. S. Miyamoto and M. Mukaidono: in *Proc. 7th Int. Fuzzy Syst. Assoc. World Congress(IFSA '97)*, Vol.II, pp.86-92(1997)

15. Y. Ohashi:*9th Meeting SAS User Grp. Int.*, Hollywood Beach, Florida(1984).

16. K. Rose, E. Gurewitz, and G. C. Fox: *Pattern Recognition Letters*, Vol. 11, pp.589-594(1990).

17. R. L. Streit and T. E. Luginbuhl:*IEEE Trans. Neural Networks*, Vol.5, No.5, pp.764-783(1994)

18. N. Ueda and R. Nakano: *Neural Networks*, Vol.11, No.2, pp.271-282(1998).

19. X. Zhuang, Y. Huang, K. Palaniappan, and Y. Zhao:*IEEE Tran. Image Processing*, Vol.5, pp.1293-1302(1996)