Robust Clustering in Fuzzy c-Means with Regularization by Cross Entropy [*][†]

Hidetomo ICHIHASHI[†], Katsuhiro HONDA[†]
[†] Graduate School of Engineering, Osaka Prefecture University,
1-1 Gakuen-cho, Sakai, Osaka, 599-8531 Japan

Abstract

Gaussian mixture model(GMM) for density estimation uses maximum likelihood approach, whereas Fuzzy c-Means (FCM) clustering is based on an objective function method. Since the similar algorithm with the one in GMM can be derived from the modified FCM with a regularizer of K-L information (cross entropy), this paper shows two ways to make the FCM with K-L regularizer(KFCM) robust. One is within the scope of noise clustering due to Dave and the other is the addition of a Gustafson and Kessel's constraint. Both the KFCM approachs realize robust clustering methods which produce quite similar results.

1 Introduction

Gaussian mixture density model (GMM) [4] 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. Entropy method that uses an additional term of entropy for fuzzification in the Fuzzy c-Means (FCM) [1] was proposed by Miyamoto *et al.* [9]. A similar entropy term was considered by Dave and Krishnapuram [3] to prevent trivial solution within the scope of Possibilistic c-Means due to Krishnapuram and Keller [7]. A new FCM clustering objective function with an additional term known as Kullback-Leibler information or cross entropy was proposed. Close connection between the GMM Expectation Maximizing(EM) [8] algorithm and the FCM algorithm with K-L regularizer (KFCM) was clarified [6]. In this paper, we present an idea how to make the KFCM clustering which is robust to outliers or noise. This is done by the modification to the KFCM clustering as in the noise clustering by Dave[2], or the addition of a

[*] Proc. of 1st International Conference on Fuzzy Systems and Knowledge Discovery, 1, 471-475 (2002)

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

constraint of Gustafson and Kessel [5]. Simulation experiment shows how well the accumulated points are extracted by the modified KFCM algorithms.

2 GMM density estimation and KFCM clustering

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.

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 [5], 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.* [9], 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 cross entropy or K-L information and including constraint term in a Lagrangian function, we consider the minimization of the following objective function.

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

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 make clusters fuzzier, the second term of Eq.(6) becomes zero if u_{ik}, $k = 1, ..., n$ take exactly the same value as π_i in the ith cluster for all i. The K-L information represents the proximity between the distribution of u_{ik} and that of π_i. If $u_{ik} \simeq \pi_i$ for all k and i, partition becomes very fuzzy but when λ is 0 the optimization problem reduces to a linear one and the solution u_{ik} are obtained at extremal point, i.e., u_{ik} equals 0 or 1. Fuzziness of the clusters can be controlled by λ. We can derive some necessary conditions for optimality of Eq.(6).

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

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

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.

3 Robust Clustering in KFCM

Noise clustering(NC) was proposed by Dave[2] so that the noise data will be involved in the noise cluster. Now we apply the idea of noise cluster due to Dave to KFCM. This approach is referred to as NKFCM. The noise is considered to be a separate class and is represented by a prototype that has a constant distance δ from all data points. When the number c of ordinary clusters is given, the $c+1$th cluster plays the role of the noise cluster. The objective function in Eq.(6) is modified as follows:

$$\begin{aligned} J_{\lambda\delta} &= \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}d_{ik} + \delta\sum_{k=1}^{n} u_{c+1\ k} \\ &+ \lambda\sum_{i=1}^{c+1}\sum_{k=1}^{n} u_{ik}\log\frac{u_{ik}}{\pi_i} \\ &+ \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}\log|A_i| \\ &- \sum_{k=1}^{n}\eta_k(\sum_{i=1}^{c+1} u_{ik} - 1) \\ &- \tau(\sum_{i=1}^{c+1}\pi_i - 1) \end{aligned} \tag{11}$$

from the necessary condition of optimality, for $i \leq c$

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n} u_{ik}\boldsymbol{x}_k}{\sum_{k=1}^{n} u_{ik}} \tag{12}$$

By denoting

$$\begin{aligned} W_k &= \sum_{j=1}^{c}\pi_j\exp\big(-\frac{1}{\lambda}d_{jk}\big)|A_j|^{-1/\lambda} \\ &+ \pi_{c+1}\exp\big(-\frac{\delta}{\lambda}\big) \end{aligned} \tag{13}$$

for $i \le c$

$$u_{ik} = \pi_i \exp\bigl(-\frac{1}{\lambda} d_{ik}\bigr)|A_i|^{-1/\lambda}/W_k \tag{14}$$

and for $i = c+1$

$$u_{ik} = \pi_{c+1} \exp\bigl(-\frac{\delta}{\lambda}\bigr)/W_k \tag{15}$$

Matrix A_i is for $i \le c$

$$A_i = \frac{\sum_{k=1}^{n} u_{ik}(\boldsymbol{x}_k - \boldsymbol{v}_i)(\boldsymbol{x}_k - \boldsymbol{v}_i)^T}{\sum_{k=1}^{n} u_{ik}} \tag{16}$$

and for $i \le c+1$

$$\pi_i = \frac{\sum_{k=1}^{n} u_{ik}}{\sum_{j=1}^{c+1}\sum_{k=1}^{n} u_{jk}} = \frac{1}{n}\sum_{k=1}^{n} u_{ik} \tag{17}$$

The clustering algorithm in this case is also the fixed point iteration as in conventional FCM.

Gustafson and Kessel[5] introduced matrix A_i which defines Mahalanobis distance and to shape the cluster to better fit a given data distribution. A constraint by Gustafson and Kessel[5] is applicable to KFCM clustering in order to enhance it's robustness. As the fourth term of the objective function of KFCM, G-K constraint is added. By restricting the variation of data in the $c+1th$ cluster or the size of $|A_{c+1}|$, the $c+1th$ cluster works as a nois cluster. $|A_{c+1}|$ is to be ρ and the new objective function becomes

$$\begin{aligned} J_{\lambda\rho} &= \sum_{i=1}^{c+1}\sum_{k=1}^{n} u_{ik}d_{ik} + \lambda\sum_{i=1}^{c+1}\sum_{k=1}^{n} u_{ik}\log\frac{u_{ik}}{\pi_i} \\ &+ \sum_{i=1}^{c+1}\sum_{k=1}^{n} u_{ik}\log|A_i| \\ &+ \gamma(\log|A_{c+1}| - \rho) \\ &- \sum_{k=1}^{n}\eta_k\Bigl(\sum_{i=1}^{c+1} u_{ik} - 1\Bigr) - \tau\Bigl(\sum_{i=1}^{c+1}\pi_i - 1\Bigr) \end{aligned} \tag{18}$$

γ is a Lagrangean function and ρ is a positive constant. Unknown parameters are iteratively updated from the necessary conditions of optimality of the objective function.

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n} u_{ik}\boldsymbol{x}_k}{\sum_{k=1}^{n} u_{ik}} \tag{19}$$

And, by denoting

$$W_k = \sum_{j=1}^{c+1} \pi_j \exp\bigl(-\frac{1}{\lambda} d_{jk}\bigr)|A_j|^{-1/\lambda} \tag{20}$$

we have

$$u_{ik} = \pi_i \exp\bigl(-\frac{1}{\lambda} d_{ik}\bigr)|A_i|^{-1/\lambda}/W_k \tag{21}$$

$$\pi_i = \frac{\sum_{k=1}^{n} u_{ik}}{\sum_{j=1}^{c+1}\sum_{k=1}^{n} u_{jk}} = \frac{1}{n}\sum_{k=1}^{n} u_{ik} \tag{22}$$

Matrix A_i is obtained for $i \leq c$ as:

$$A_i = \frac{\sum_{k=1}^{n} u_{ik}(\boldsymbol{x}_k - \boldsymbol{v}_i)(\boldsymbol{x}_k - \boldsymbol{v}_i)^T}{\sum_{k=1}^{n} u_{ik}} \tag{23}$$

and for $i = c + 1$

$$A_{c+1} \quad = \quad \frac{S_{c+1}}{|S_{c+1}|^{1/p}} e^{\rho/p} \tag{24}$$

where $S_i = \sum_{k=1}^{n} u_{ik}(\boldsymbol{x}_k - \boldsymbol{v}_i)(\boldsymbol{x}_k - \boldsymbol{v}_i)^T$.
λ controls the fuzziness of clusters. The algorithm with additional G-K constraint is referred to as GKFCM. The clustering algorithm is the fixed point iteration as in conventional FCM, i.e., the repetition through Eqs.(20)- (24) . As we will see in the next section, GKFCM approach provides the similar robust clustering results as by the noise clustering.

Another prototype-based robust clustering approach called Possibilistic c-Means (PCM) is attributed to Krishnapuram and Keller [7], in which there are no constraints on the membership except for the requirement that they should be in [0, 1]. This also can be applicable to KFCM, but unlike the NKFCM and GKFCM, this approach need a careful assignment of parameters in the objective function. In our computer simulations, due to the many unknown parameters the possibilistic approach trapped into unfavorable local minimums so frequently that we discarded this approach in our KFCM clustering.

4 Simulation Experiment

The numerical example shown in Fig. 1 consists of complex four linear clusters surrounded by noise data denoted by ×. Fig. 1 shows the clustering results by Noise approach(NKFCM) and G-K constraint approach(GKFCM). Single noise cluster is specified in each case. In the GKFCM, G-K constraint is given only for the noise cluster. Noisy data depicted by × are widely scattered throughout the entire space. In the figure, data points are crisply partitioned by selecting the maximum membership. ◯, △, $\cdots$, × represent linearly accumulated clusters. In the noise cluster the points depicted by × are regarded as noise by the algorithm. The boldface arrow is depicted to show the direction of principal component vector (eigenvector) of fuzzy covariance matrix in each cluster. The two approaches produced fairly similar partitionings that agree with our intuitive judgment of clustering.

Fig. 2 shows the result applied to Anderson's iris data which is a well known classification benchmark. By the regular KFCM approach, in which no noise cluster is taking into account, hyper-elliptic clusters that approximately separate the three Iris species were produced. The result by the two robust clustering approach in this paper are shown in Fig. 2. Although there seems to be no great differences between the two results, since in the G-K approach distance between the noise data and the center of noise cluster is measured in Mahalanobis distance, data in the vicinity of the noise cluster center belong to the cluster. It should be noted that the clusters are depicted crisply, but the eigenvectors are calculated taking fuzzy memberships into account. When δ is decreased in noise approach(NKFCM) or ρ is increased in G-K approach(GKFCM), data points that are relatively far from cluster center tend to be in the noise cluster(Fig. 3). The change in the direction and length of the principal eigenvector clearly show the effect of these parameters. It also should be noted that the vector are depicted only by the two component, i.e., x_3 and x_4. The eigenvalues and their associated eigenvectors are shown in Table 1- 2.

cluster	PC	eigenvalue	eigenvector			
1	1	177.533	-0.675	-0.729	-0.096	-0.066
	2	38.386	-0.036	-0.024	-0.231	0.972
circle	3	27.888	-0.445	0.287	0.827	0.188
	4	4.382	0.587	-0.621	0.503	0.126
2	1	233.487	-0.368	0.475	0.407	-0.688
	2	141.106	0.470	-0.561	-0.067	-0.678
triangle	3	8.439	-0.530	-0.662	0.514	0.130
	4	0.622	-0.602	-0.146	-0.752	-0.224
3	1	45.105	0.322	-0.456	-0.534	0.635
	2	22.646	-0.252	0.801	-0.284	0.464
box	3	14.760	0.756	0.357	-0.351	-0.422
	4	1.571	-0.512	-0.153	-0.714	-0.452

Table 1: Noise approach(Iris,δ=2.0)

cluster	PC	eigenvalue	eigenvector			
1	1	114.853	0.652	0.754	0.054	0.049
	2	59.265	-0.020	-0.011	-0.422	0.906
circle	3	33.105	-0.163	0.050	0.894	0.414
	4	4.296	0.740	-0.655	0.137	0.073
2	1	83.318	0.501	-0.216	-0.556	0.627
	2	23.144	0.287	-0.872	0.282	-0.279
triangle	3	13.535	0.700	0.362	-0.176	-0.590
	4	1.243	-0.421	-0.248	-0.762	-0.425
3	1	37.688	0.680	-0.533	-0.418	0.281
	2	21.269	-0.345	0.158	-0.149	0.913
box	3	7.761	-0.410	-0.815	0.407	0.052
	4	0.916	-0.500	-0.166	-0.799	-0.291

Table 2: G-K approach(Iris,ρ=2.0)

References

[1] J. C. Bezdek: *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press(1981)

[2] R. N. Dave: Characterization and detection of noise in clustering, *Pattern Recognition Letters*, Vol.12, pp.657-664(1991).

[3] R. N. Dave and R. Krishnapuram: Robust clustering methods: A unified approach, *IEEE Trans. Fuzzy Syst.*, Vol.5, No.2, pp.270-293(1997).

[4] R. O. Duda and P. E. Hart: *Pattern Classification and Scene Analysis*, Wiley, New York(1973).

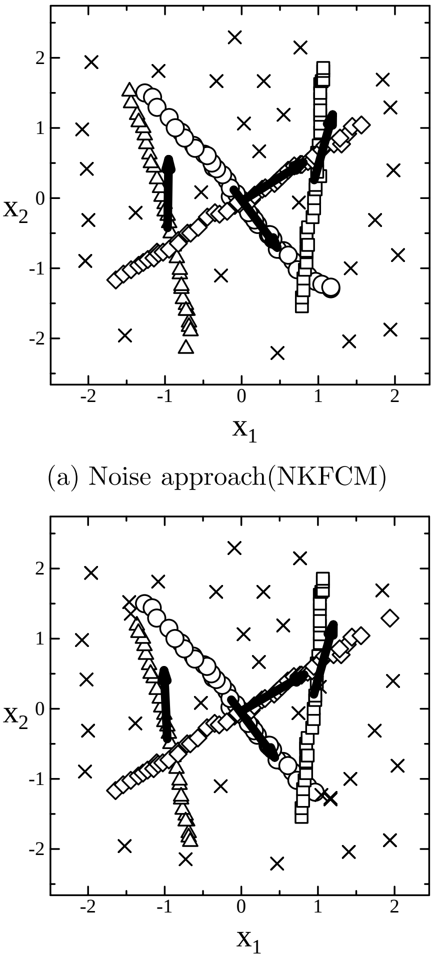
[5] D. E. Gustafson and W. C. Kessel: Fuzzy clustering with a fuzzy covariance matrix, *Proc. IEEE CDC*, Vol.2, pp.761-766(1979)

[6] H.Ichihashi, K.Miyagishi, K.Honda: Fuzzy c-means clustering with regularization by K-L Information, Proc. of FUZZ-IEEE2001, Melbourne, November(2001)
http://www.ie.osakafu-u.ac.jp/ ichi/KFCM.html

[7] R. Krishnapuram and J. Keller: A possibilistic approach to clustering, *IEEE Transactions on Fuzzy Systems*, Vol.1, pp.98-110(1993)

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

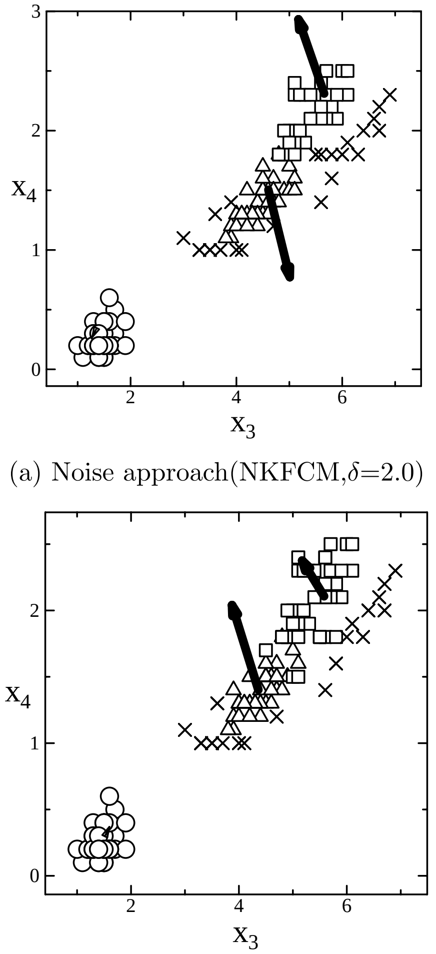
[9] S. Miyamoto and M. Mukaidono: Fuzzy *c*-means as a regularization and maximum entropy approach, *Proc. of the 7th International Fuzzy Systems Association World Congress(IFSA '97)*, Vol.II, pp.86-92(1997)



(a) Noise approach(NKFCM)

(b) G-K constraint approach(GKFCM)

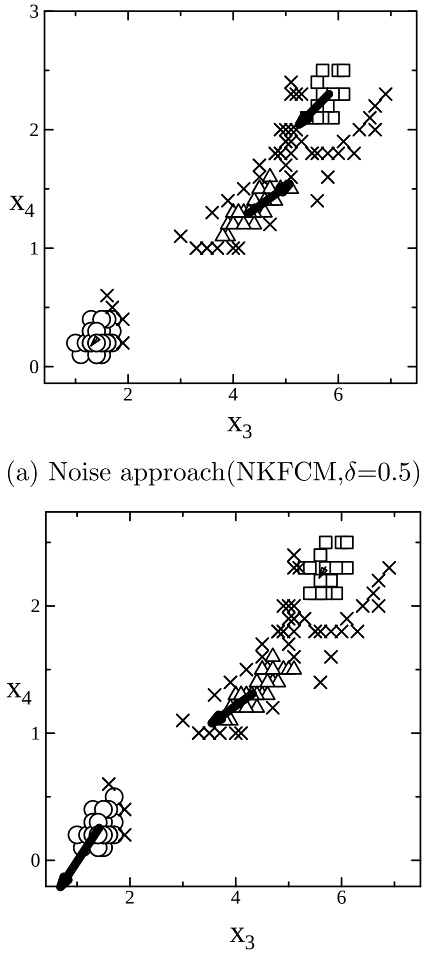
Figure 1: Example 1



(a) Noise approach(NKFCM,δ=2.0)

(b) G-K approach(GKFCM,ρ=2.0)

Figure 2: Clustering Andersen's Iris data : slight noise



(a) Noise approach(NKFCM,δ=0.5)

(b) G-K approach(GKFCM,ρ=2.5)

Figure 3: Clustering Andersen's Iris data : heavy noise