Proceedings of the 2007 IEEE Symposium on
Foundations of Computational Intelligence (FOCI 2007)

Aggregation of Standard and Entropy Based Fuzzy c-Means Clustering by a Modified Objective Function

Hidetomo Ichihashi, Katsuhiro Honda, Akira Notsu and Takao Hattori
Graduate School of Engineering, Osaka Prefecture University
1-1 Gakuen-cho, Naka-ku, Sakai, Osaka 599-8531 Japan

***Abstract*— A generalized fuzzy c-means (FCM) clustering is proposed by modifying the standard FCM objective function and introducing some simplifications. FCM clustering results in very fuzzy partitions for data points that are far from all cluster centroids. This property distinguishes FCM from Gaussian mixture models or entropy based clustering. The generalized FCM clustering aims at aggregating standard FCM and entropy based FCM so that the generalized algorithm is furnished with the two distinctive properties for data points that are far from all centroids and for those that are close to any centroid. k-Harmonic means clustering are reviewed from the view point of FCM clustering. Graphical comparisons of the four classification functions are presented.**

I. Introduction

The unsupervised partitioning of data is often called clustering, which forms a significant area of research such as data mining, statistical data analysis, data compression and vector quantization. k-Means or hard c-means clustering method stands out, among the many clustering algorithms, as one of the few most popular algorithms accepted by many application domains. However, k-means does have a widely known problem, i.e., the local minimum it converges to is very sensitive to the initialization. The standard fuzzy c-means (FCM-s) algorithms [1], [2] can alleviate the problem and are robust tools for the problem of clustering objects into groups of similar individuals when the data is available as object data, consisting of a set of feature vectors.

Within the framework of FCM clustering, the methods that uses an additional entropy term [3] or a quadratic term [4] for fuzzification was proposed. The same algorithm as the Gaussian mixture models (GMM) or normal mixture [5], [6] with the expectation maximizing algorithm [7], [8] is derived from the FCM objective function with regularization by entropy term. The difference between the FCM-s and FCM-e comes from the difference of the membership functions as amplified in [3]. FCM-s results in very fuzzy partitions for data points that are far from all cluster centroids. This property distinguishes FCM-s from FCM-e. This paper proposes a generalized FCM clustering (FCM-g) by slightly modifying the objective function of FCM-s and introducing some simplifications. FCM-g clustering aims at aggregating FCM-s and FCM-e so that FCM-g is furnished with the two distinctive properties for data points that are far from all centroids and for those that are close to any centroid.

k-Harmonic means (KHM) clustering [9], [10], [11], [12] is a relatively new iterative unsupervised learning algorithm for clustering. The motivations come principally from an analogy with powerful supervised classification methods known as boosting algorithms [13], [14], [15]. The papers and patents of KHM have emphasized a new trend in clustering. It basically consists of penalizing solutions via weights on the instance points. We review KHM clustering from the view point of FCM clustering.

The paper is organized as follows. Section II gives a brief description of the standard, entropy based and quadratic term based FCM clustering. A modified FCM objective function is proposed and the characteristics of the three algorithms are described in Section III. In Section IV, we give the reinterpretation of k-harmonic means clustering within the framework of FCM clustering. Section V provides graphical comparisons of the classification functions of the four algorithms. Some simplifications of the clustering algorithm will be described in Section VI. Section VII concludes the paper.

II. Unsupervised clustering

We first review the three kinds of objective functions, i.e., the standard [1], entropy-term-based [3], [5], [6], and quadratic-term-based [4] fuzzy c-means. The objective function of FCM-s is:

$$\bar{U} = \arg\min_{U \in \mathcal{U}_f} J_{\rm fcm}(U, \bar{V}). \tag{1}$$

$$J_{\rm fcm}(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m D(x_k, v_i), \quad (m>1), \tag{2}$$

under the constraint:

$$\begin{aligned}\mathcal{U}_f \;=\; & \{\, U=(u_{ik}) \,:\, \sum_{j=1}^{c} u_{kj} = 1,\ 1 \le k \le n; \\ & u_{ik} \in [0,1],\ 1 \le k \le n,\ 1 \le i \le c \,\}. \end{aligned} \tag{3}$$

$D(x_k, v_i)$ denotes the squared distance between feature vector x_k and cluster centroid vector v_i, so the standard objective function is the weighted sum of squared distances.

Following objective function is used for the entropy-based method.

$$J_{\rm efc}(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik} D(x_k, v_i) + \nu \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik} \log u_{ik}. \tag{4}$$

The objective function of the quadratic-term-based method is:

$$J_{\rm qfc}(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}D(x_k,v_i)+\frac{1}{2}\nu\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^2. \tag{5}$$

Combining these three, the objective function can be written as :

$$J(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m D(x_k,v_i)+\nu\sum_{i=1}^{c}\sum_{k=1}^{n}K(u), \tag{6}$$

where both m and ν are the fuzzifiers.

When $m>1$ and $\nu=0$, (6) is the standard objective function. When m=1 and $K(u)=u_{ik}\log u_{ik}$, (6) is the objective function of entropy-based method, whose algorithm is the same as the EM algorithm for GMM or normal mixture with a unit covariance matrix and equal cluster volume. When $m=1$ and $K(u)=(u_{ik})^2$, (6) is the objective function of the quadratic-term-based method.

III. A GENERALIZED OBJECTIVE FUNCTION

From the above consideration we can generalize the standard objective function a little further. Let $m>1$ and $K(u)=(u_{ik})^m$, then (6) is the objective function ($J_{\rm gfc}$) from which we can easily derive the necessary condition for the optimality.

We consider minimization of (6) with respect to U under the condition $\sum_{i=1}^{c}u_{ik}=1$ using the method of Lagrange multipliers. Let the Lagrange multiplier be λ_k, $k=1,\ldots,n$, and put

$$\begin{aligned}L &= \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m D(x_k,v_i)+\nu\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m\\ &+ \sum_{k=1}^{n}\lambda_k\Big(\sum_{i=1}^{c}u_{ik}-1\Big)\\ &= J_{\rm gfc}+\sum_{k=1}^{n}\lambda_k\Big(\sum_{i=1}^{c}u_{ik}-1\Big).\end{aligned} \tag{7}$$

For the necessary condition of optimality of (7) we differentiate

$$\frac{\partial L}{\partial u_{ik}}=m(u_{ik})^{m-1}(D(x_k,v_i)+\nu)+\lambda_k=0.$$

$D(x_k,v_i)+\nu>0$ ($i=1,\ldots,c$), if $\nu>0$. To eliminate λ_k, we note

$$u_{kj}=\left[\frac{-\lambda_k}{m(D(x_k,v_j)+\nu)}\right]^{\frac{1}{m-1}}. \tag{8}$$

Summing up for $j=1,\ldots,c$ and taking $\sum_{j=1}^{c}u_{kj}=1$ into account, we have

$$\sum_{j=1}^{c}\left[\frac{-\lambda_k}{m(D(x_k,v_j)+\nu)}\right]^{\frac{1}{m-1}}=1.$$

Using (8) to this equation, we can eliminate λ_k, having

$$u_{ik}=\left[\sum_{j=1}^{c}\left(\frac{D(x_k,v_i)+\nu}{D(x_k,v_j)+\nu}\right)^{\frac{1}{m-1}}\right]^{-1}. \tag{9}$$

This solution satisfies $u_{ik}\geq 0$ and u_{ik} is continuous.

The solution for V is also easily derived by differentiating L with respect to V.

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m}. \tag{10}$$

We now have insight about the property, which distinguishes FCM-g from FCM-s and FCM-e.

Let $\vec{U}^{(i)}_{\rm gfc}(x;V)$ denote the classification function for the generalized method. We use the term "classification function" to signify the function for partitioning the input feature space into clusters.

$$\vec{U}^{(i)}_{\rm gfc}(x;V) = \left[\sum_{j=1}^{c}\left(\frac{D(x,v_i)+\nu}{D(x,v_j)+\nu}\right)^{\frac{1}{m-1}}\right]^{-1}. \tag{11}$$

It should be noted that when $m=2$ and $\nu=1$, $\vec{U}^{(i)}_{\rm gfc}(x;V)$ has a close relationship with Cauchy weight function normalized such that the weights sum to one. The membership function of FCM-s suffers from the singularity which occurs when $D(x_k,v_i)=0$. When $\nu>0$, (9) alleviates the singularity.

Next proposition states that ν is a fuzzifier.

Proposition 1. The function $\vec{U}^{(i)}_{\rm gfc}(x;V)$ is a decreasing function of ν when

$$\|x-v_i\|<\|x-v_j\|,\ \forall j\neq i\quad (x\in R^p),$$

and if $\nu>0$,

$$\max_{x\in R^p}\vec{U}^{(i)}_{\rm gfc}(x;V)=\vec{U}^{(i)}_{\rm gfc}(v_i;V)<1. \tag{12}$$

$\vec{U}^{(i)}_{\rm gfc}(x;V)$ is an increasing function of ν when

$$\|x-v_i\|>\|x-v_j\|,\ \forall j\neq i\quad (x\in R^p).$$

$\vec{U}^{(i)}_{\rm gfc}(x;V)$ tends to $1/c$ as $\nu\to+\infty$.

$$\lim_{\nu\to+\infty}\vec{U}^{(i)}_{\rm gfc}(x;V)=\frac{1}{c}. \tag{13}$$

$\vec{U}^{(i)}_{\rm gfc}(x;V)$ tends to $1/c$ as $\|x\|\to+\infty$.

$$\lim_{\|x\|\to+\infty}\vec{U}^{(i)}_{\rm gfc}(x;V)=\frac{1}{c}. \tag{14}$$

The first part of the proposition immediately follows from

$$1/\vec{U}^{(i)}_{\rm gfc}(x;V)-1 = \sum_{j\neq i}\left(\frac{\|x-v_i\|^2+\nu}{\|x-v_j\|^2+\nu}\right)^{\frac{1}{m-1}}, \tag{15}$$

hence (12) follows.

For (13) and (14), we can use (15) again:

$$1/\vec{U}^{(i)}_{\rm gfc}(x;V) - 1 \to c - 1 \quad (\nu \to +\infty \text{ or } \|x\| \to +\infty).$$

$\vec{U}^{(i)}_{\rm fcm}(x;V)$ shares the last property (14) but

$$\lim_{x \to v_i} \vec{U}^{(i)}_{\rm fcm}(v_i;V) = 1 \tag{16}$$

holds for $\forall i$. By choosing the values of the fuzzifier m and ν, FCM-g is equipped with the two distinctive properties of FCM-s and FCM-e for data points that are far from all centroids and for those that are close to any centroid.

Table I summarizes the characteristics of the three methods when $\|x\| \to +\infty$ or $x \to v_i$. It is a rational decision to fuzzily partition the data that are far from all centroids. Human intuition may suggests to do so and no one can answer which cluster the distant points should belong. Although FCM-s is favorable in this point, we sometimes want to fuzzily partition even the data points that are very close to any cluster centroid. FCM-g improves this deficiency of FCM-s.

TABLE I
CHARACTERISTICS OF FUZZY CLUSTERINGS

	$\|x\| \to +\infty$	$x \to v_i$
standard (FCM-s)	fuzzy	crisp
entropy-based (FCM-e)	crisp	crisp-fuzzy
generalized (FCM-g)	fuzzy	crisp-fuzzy

Note that the objective function of the possibilistic clustering [16], [17] is written similarly to (7) as:

$$\begin{aligned} J_{\rm pos}(U,V) &= \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m D(x_k,v_i) \\ &+ \nu\sum_{i=1}^{c}\sum_{k=1}^{n}(1-u_{ik})^m \end{aligned} \tag{17}$$

where the condition $\sum_{i=1}^{c} u_{ik} = 1$ is omitted.

As pointed out in [16], [17], the possibilistic clustering is closely related with robust M-estimation [18], [19] and ν in (17) plays the role of robustizer whereas ν in (7) is a fuzzifier as stated in Proposition 1.

IV. CONNECTIONS WITH k-HARMONIC MEANS

k-harmonic means (KHM) [9], [10], [11], [12] is a relatively new iterative unsupervised learning algorithm for clustering. KHM is essentially insensitive to the initialization of the centroids. It basically consists of penalizing solutions via weights on the data points, somehow making the centroids move toward the hardest (difficult) points. The motivations come from an analogy with supervised classifier design methods known as boosting [13], [14], [15].

The harmonic average of c numbers $a_1, ..., a_c$ is defined as $\frac{c}{\sum_{i=1}^{c}\frac{1}{a_i}}$. For clarifying the connection between FCM and KHM, the objective function of KHM is rewritten as:

$$\begin{aligned} J_{\rm KHM}(V) &= \sum_{k=1}^{n}\frac{c}{\sum_{i=1}^{c}\frac{1}{\|x_k - v_i\|^p}} \\ &= \sum_{k=1}^{n}\sum_{i=1}^{c}\left(\frac{D(x_k,v_i)^{\frac{p}{2}}}{\sum_{l=1}^{c}\frac{D(x_k,v_i)^{\frac{p}{2}}}{D(x_k,v_l)^{\frac{p}{2}}}}\right) \\ &= \sum_{k=1}^{n}\sum_{i=1}^{c}\left(\sum_{l=1}^{c}\frac{D(x_k,v_i)^{\frac{p}{2}}}{D(x_k,v_l)^{\frac{p}{2}}}\right)^{-1} D(x_k,v_i)^{\frac{p}{2}}. \end{aligned} \tag{18}$$

When p=2, (18) is the same as $J_{\rm gfc}$ in (7) or $J_{\rm fcm}$ in (2) with m=1 and $\nu = 0$, if substituted with (9) where $m = 2$ and $\nu = 0$. The objective function (18) does not coincide with (2), though the update rule of centroids v is the same as (10) with $m = 2$ and $\nu = 0$ as we will show below.

By taking partial derivative of $J_{\rm KHM}(V)$ with respect to v_i, we have

$$\frac{\partial J_{\rm KHM}(V)}{\partial v_i} = -cp\sum_{k=1}^{n}\frac{x_k - v_i}{D(x_k,v_i)^{\frac{p}{2}+1}\left(\sum_{l=1}^{c}\frac{1}{D(x_k,v_i)^{\frac{p}{2}}}\right)^2} \tag{19}$$

Although $D(x_k,v_i)$ includes v_i, from (19) the iterative update rule can be written as:

$$\begin{aligned} v_i &= \frac{\sum_{k=1}^{n}\frac{1}{D(x_k,v_i)^{\frac{p}{2}+1}\left(\sum_{l=1}^{c}\frac{1}{D(x_k,v_i)^{\frac{p}{2}}}\right)^2}\, x_k}{\sum_{k=1}^{n}\frac{1}{D(x_k,v_i)^{\frac{p}{2}+1}\left(\sum_{l=1}^{c}\frac{1}{D(x_k,v_i)^{\frac{p}{2}}}\right)^2}} \\ &= \frac{\sum_{k=1}^{n}\left(\sum_{l=1}^{c}\frac{D(x_k,v_i)^{\frac{p}{2}}}{D(x_k,v_l)^{\frac{p}{2}}}\right)^{-2} D(x_k,v_i)^{\frac{p}{2}-1}\, x_k}{\sum_{k=1}^{n}\left(\sum_{l=1}^{c}\frac{D(x_k,v_i)^{\frac{p}{2}}}{D(x_k,v_l)^{\frac{p}{2}}}\right)^{-2} D(x_k,v_i)^{\frac{p}{2}-1}} \end{aligned} \tag{20}$$

When p=2, (20) is the same as (10) substituted with (9) where $m = 2$ and $\nu = 0$. Thus, we have the same clustering results as FCM-s with $m = 2$. In (20), $\left(\sum_{l=1}^{c}\frac{D(x_k,v_i)^{\frac{p}{2}}}{D(x_k,v_l)^{\frac{p}{2}}}\right)^{-2}$ is the weight on x_k for computing weighted mean of x_k's.

Let u_{ik} be the membership function as:

$$u_{ik} = \left(\sum_{l=1}^{c}\frac{D(x_k,v_i)^{\frac{p}{2}}}{D(x_k,v_l)^{\frac{p}{2}}}\right)^{-1}, \tag{21}$$

then for $\forall p > 0, u_{ik}$'s sum to one except when $\exists l, D(x_k, v_l) = 0$.

$$\sum_{i=1}^{c} u_{ik} = \sum_{i=1}^{c} \left(\sum_{l=1}^{c} \frac{D(x_k, v_i)^{\frac{p}{2}}}{D(x_k, v_l)^{\frac{p}{2}}} \right)^{-1} = 1 \qquad (22)$$

For $p > 2$, $D(x_k, v_i)^{\frac{p}{2}-1}$ in (20) can be seen as weights on data points, which come from an analogy with supervised classifier design methods known as boosting. This view of the weights is slightly different from [9], [10], [11] but the effect of the weights is the same. As $D(x_k, v_i)$ approaches to zero, the effect of u_{ik} for computing v_i decreases. When $p \leq 2$, similar to FCM-s, KHM clustering also suffers from the singularity which occur when $D(x_k, v_i) = 0$, and the weight $D(x_k, v_i)^{\frac{p}{2}-1}$ mitigates this effect when $p > 2$.

V. Graphical comparisons

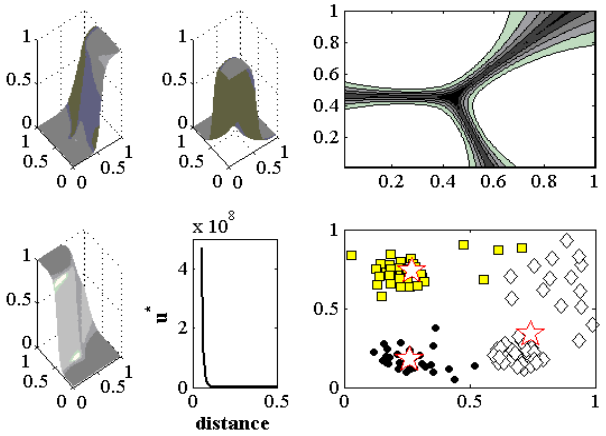

Fig. 1. Rather crisply partitioned result by the standard method (FCM-s) with $m = 1.3$

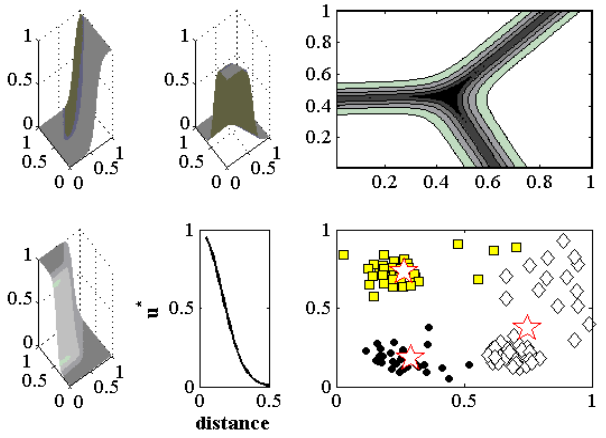

Fig. 2. Rather crisply partitioned result by the entropy-based method (FCM-e) with $\nu = 0.05$

Characteristics of the four clustering methods are compared in Figs.1-7, where $c = 3$ and other parameter values are given in the legend of each figure. In each figure, upper left and middle, and lower left graphs show 3D graphics of the classification functions. Lower middle graph shows the membership function with respect to distance from a cluster centroid. Upper right graph shows the contours of classification functions. The contours of maximum values among the three classification functions are drawn. Lower right graph shows the clustering results where stars mark cluster centroids.

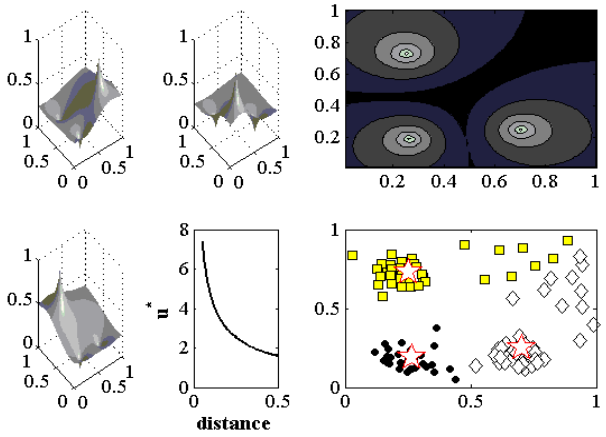

Fig. 3. Fuzzily partitioned result by the standard method (FCM-s) with $m = 4$

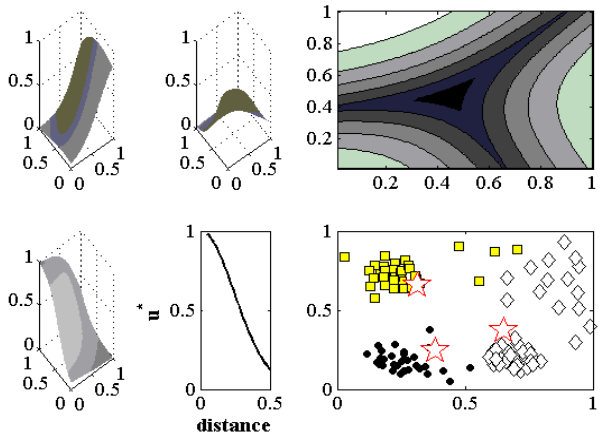

Fig. 4. Fuzzily partitioned result by the entropy-based method (FCM-e) with $\nu = 0.12$

Figs.1 and 2 show rather crisply partitioned results by the standard method (FCM-s) with $m = 1.3$ and by the entropy-based method (FCM-e) with $\nu = 0.05$ respectively. The contours are different from each other at the upper right coner of the graphs since the points near (1.0, 1.0) are far from all the cluster centroids. Figs.3 and 4 show fuzzily partitioned results by FCM-s with $m = 4$ and by FCM-e with $\nu = 0.12$ respectively. These two methods produce quite different contours of classification functions when the fuzzifier is relatively large. Fig.3 shows the robustness of FCM-s where all centroids are located at densely accumulated areas. FCM-s suffers from the problem called singularity

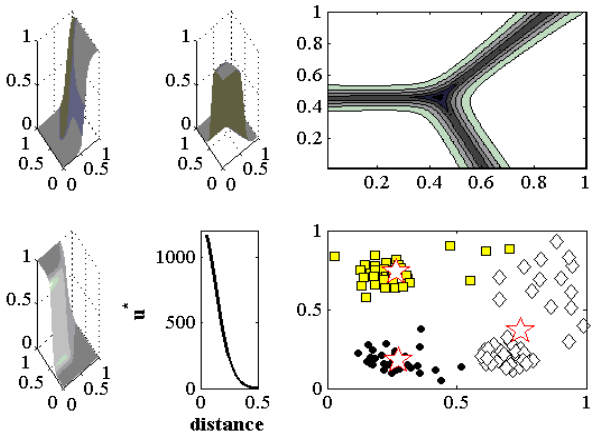

Fig. 5. Rather crisply partitioned result by the generalized method (FCM-g) with $m = 1.05$ and $\nu = 0.7$

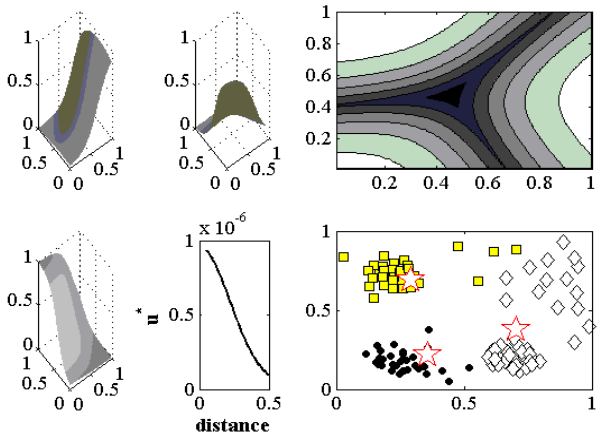

Fig. 6. Fuzzily partitioned result by the generalized method (FCM-g) with $m = 1.05$ and $\nu = 2$

when $D(x_k, v_i) = 0$, which thus results in the singularity in the shape of classification function $\vec{U}^{(i)}_{\rm fcm}(x; V)$. When the fuzzifier m is large, the classification function appears to be spiky at the centroids as shown by 3D graphics in Fig.3 and this is the singularity in shape.

Fig.5 shows rather crisply partitioned result by FCM-g with $m = 1.05$ and $\nu = 0.7$. The result is similar to one by FCM-e in Fig.2.

Fig.6 shows fuzzily partitioned result by FCM-g with $m = 1.05$ and $\nu = 2$. The result also is similar to one by FCM-e in Fig.4.

Since FCM-g is reduced to FCM-s when the fuzzifier $\nu = 0$, FCM-g can produce the same results with those by FCM-s. Therefore, FCM-g is an aggregated algorithm of FCM-s and FCM-e.

Fig.7 shows the clustering result of KHM with $p = 2.2$. The 3D graphics shows the weight $\left(\sum_{l=1}^{c} \frac{D(x_k,v_i)^{\frac{p}{2}}}{D(x_k,v_l)^{\frac{p}{2}}}\right)^{-2} \times D(x_k, v_i)^{\frac{p}{2}-1}$. A dent is seen on each centroid of the cluster, though the clustering result is similar to one by FCM-s.

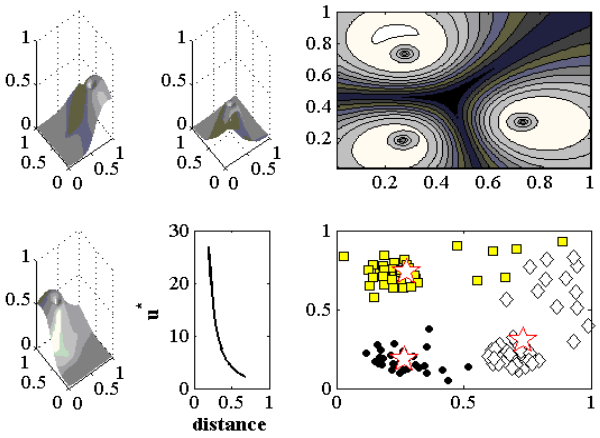

Fig. 7. Result by KHM with $p = 2.2$

VI. Clustering with Iteratively Reweighted Least Square Technique

By replacing the entropy term of the entropy-based method in (4) with K-L information term, we can consider the minimization of the following objective function under the constraints that both the sum of u_{ik} and the sum of π_i with respect to i equal one respectively.

$$\begin{aligned} J_{\rm efc}(U, V, S, \Pi) &= \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik} D(x_k, v_i) \\ &+ \nu \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik} \log \frac{u_{ik}}{\pi_i} \\ &+ \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik} \log |S_i|, \end{aligned} \tag{23}$$

where

$$D(x_k, v_i) = (x_k - v_i)^\top S_i^{-1} (x_k - v_i) \tag{24}$$

is Mahalanobis distance from x_k to i-th cluster prototype, and S_i is a covariance matrix of data samples of the i-th cluster. From this objective function, we can derive an iterative algorithm of the normal mixture or Gaussian mixture model when $\nu = 2$. From the necessary condition for the optimality of the objective function, we can derive:

$$S_i = \frac{\sum_{k=1}^{n} u_{ik} (x_k - v_i)(x_k - v_i)^\top}{\sum_{k=1}^{n} u_{ik}}. \tag{25}$$

$$v_i = \frac{\sum_{k=1}^{n} u_{ik} x_k}{\sum_{k=1}^{n} u_{ik}}. \tag{26}$$

$$\pi_i = \frac{\sum_{k=1}^{n} u_{ik}}{\sum_{j=1}^{c}\sum_{k=1}^{n} u_{jk}} = \frac{1}{n}\sum_{k=1}^{n} u_{ik}. \tag{27}$$

This is the only case known to date, where covariance matrices (S_i) are taken into account in the objective function $J(U, V)$ in (6). Although Gustafson and Kessel's modified

FCM [20] is derived from an objective function with fuzzifier m, we need to specify the values of determinant $|S_i|$ for all i. In order to deal with covariance structure more freely within the scope of fuzzy c-means clustering, we need some simplifications based on the iteratively reweighted least square (IRLS) technique [18]. Runkler and Bezdek's [21] fuzzy clustering scheme called alternating cluster estimation (ACE) is this kind of simplification.

Now we consider to deploy a technique from the robust M-estimation [18], [19]. The M-estimators try to reduce the effect of outliers by replacing the squared residuals with ρ-function, which is chosen to be less increasing than square. Instead of solving directly this problem, we can implement it as IRLS. While the IRLS approach does not guarantee the convergence to a global minimum, experimental results have shown reasonable convergence points. If one is concerned about local minima, the algorithm can be run multiple times with different initial conditions.

We implicitly define ρ-function through the weight function. Let us consider a clustering problem whose objective function is written as:

$$J_\rho = \sum_{i=1}^{c}\sum_{k=1}^{n}\rho(d_{ik}) \tag{28}$$

where $d_{ik} = \sqrt{D(x_k, v_i)}$ is a square root of the distance given by (24). Let v_i be the parameter vector to be estimated. The M-estimator of v_i based on the function $\rho(d_{ik})$ is the vector which is the solution of the following $m \times c$ equations:

$$\sum_{k=1}^{n}\psi(d_{ik})\frac{\partial d_{ik}}{\partial v_{ij}} = 0, i = 1, ..., c, j = 1, ..., m \tag{29}$$

where the derivative $\psi(z) = d\rho/dz$ is called the influence function. We can define the weight function as:

$$w(z) = \psi(z)/z. \tag{30}$$

Since

$$\frac{\partial d_{ik}}{\partial v_i} = -\left((x_k - v_i)^\top S_i^{-1}(x_k - v_i)\right)^{-\frac{1}{2}} S_i^{-1}(x_k - v_i),$$

Equation (29) becomes

$$\sum_{k=1}^{n} w(d_{ik})S_i^{-1}(x_k - v_i) = 0, i = 1, ..., c, \tag{31}$$

where we set as $w(d_{ik}) = u_{ik}$. (31) is equivalent to (26), which is exactly the solution to the following IRLS problem. We minimize

$$J_{ifc} = \sum_{i=1}^{c}\sum_{k=1}^{n} w(d_{ik})\left(D(x_k, v_i) + \log|S_i|\right). \tag{32}$$

Covariance matrix S_i in (25) can also be derived from (32). The weight w should be recomputed after each iteration in order to be used in the next iteration. In robust M-estimation, the function $w(d_{ik})$ provides adaptive weighting. The influence from x_k is decreased when $|x_k - v_i|$ is very large and suppressed when it is infinitely large. While IRLS approaches in general do not guarantee the convergence to a global minimum, experimental results have shown reasonable convergence points.

To facilitate competitive movements of cluster centroids, we need to define the weight function to be normalized as:

$$u_{ik} = \frac{u^*_{ik}}{\sum_{l=1}^{c} u^*_{lk}}. \tag{33}$$

We confine our numerical comparisons to the following two membership functions $u^{*(1)}$ and $u^{*(2)}$.

$$u^{*(1)}_{ik} = \frac{\pi_i|S_i|^{-1/\gamma}}{(D(x_k, v_i)/0.1 + \nu)^{1/m}}, \tag{34}$$

where $D(x_k, v_i)$ is divided by a scaling factor (0.1) so that the proper value of ν is around 1 when the variance of each element of x is 1.

$$u^{*(2)}_{ik} = \pi_i \exp(-D(x_k, v_i)/\nu)|S_i|^{-1/\gamma}. \tag{35}$$

Especially for (34), u_{ik} of (33) can be rewritten as:

$$u_{ik} = \pi_i|S_i|^{-1/\gamma} \times \left[\sum_{j=1}^{c}\left(\frac{D(x_k, v_i)/0.1 + \nu}{D(x_k, v_j)/0.1 + \nu}\right)^{\frac{1}{m}} \pi_j|S_j|^{-1/\gamma}\right]^{-1} \tag{36}$$

$u^{*(1)}$ is a modified and parameterized multivariational version of Cauchy's weight function in M-estimator or of the probability density function (PDF) of Cauchy distribution. It should be noted that in this case, (33) corresponds to (9), but (26) is slightly simplified from (10). $u^{*(2)}$ is a modified Welsch's weight function in M-estimator. Both the functions take into account covariance matrices in an analogous manner with Gaussian PDF. If we choose $u^{*(2)}$ in (35) with $\lambda = 2, \gamma = 2$, then the IRLS-FCM is the same as GMM.

Algorithm IRLS-FCM: *Procedure of IRLS Fuzzy c-Means.*

IFC1. [Generate initial value:] Generate $c \times n$ initial values for u_{ik} $(i = 1, 2, \ldots, c, k = 1, 2, \ldots, n)$.
IFC2. Calculate $v_i, i = 1, ..., c$ by using (26).
IFC3. Calculate S_i and $\pi_i, i = 1, ..., c$ by using (25) and (27).
IFC4. Calculate u_{ik}, $i = 1, ..., c, k = 1, ..., n$ by using (33) and (34).
IFC5. [Termination:] If the objective function (32) is convergent then terminate, else go to **IFC2**.

End IFC.

Since d^2_{ik} is Mahalanobis distance, $\sum_{i=1}^{c}\sum_{k=1}^{n} w(d_{ik})d^2_{ik}$ converges to a constant value, (i.e., the number of instances (n) $\times$ the number of variates).

Fig.8 shows the clustering rezult by IRLS-FCM with $u^{*(2)}$, $\nu = 2.9$ and $\gamma = 2$. The result is quite the same as the result by GMM, though the fuzzifier ν is larger than GMM (i.e., $\nu = 2$ in GMM). The data set is partitioned with ellipsoidal clusters. The points located at the left bottom of the graph (near the origin) crisply belong to the upper ellipsoidal cluster. This is a significant difference from the result shown in Fig.9 where $u^{*(1)}$ is used and $m = 1, \gamma = 2$

and $\nu = 1$. Those points are fuzzily partitioned and the membership values are around 0.5 in Fig.9. For the points that are distant from the two cluster centroids, no one knows which cluster the points should belong. IRLS-FCM with $u^{*(1)}$ is based on the FCM-g and reflects the rational intuition of human beings.

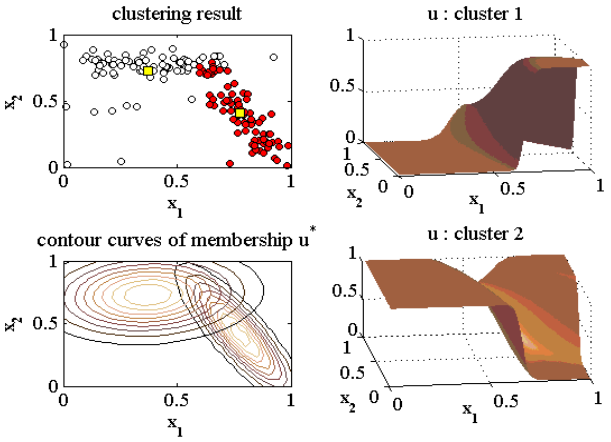

Fig. 8. Result by IRLS-FCM with $u^{*(2)}$, $\nu = 2.9$ and $\gamma = 2$

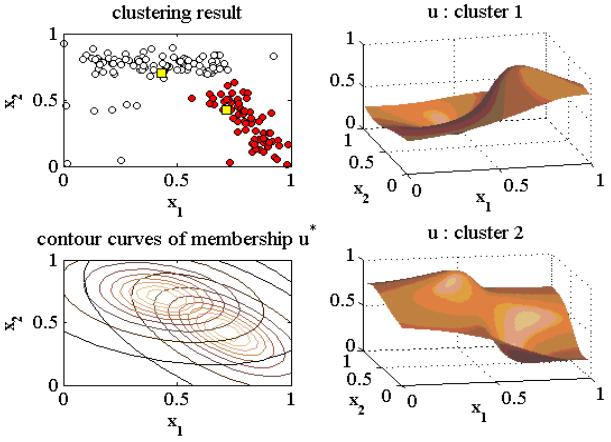

Fig. 9. Result by IRLS-FCM with $u^{*(1)}$, $m = 1, \nu = 1$ and $\gamma = 2$

VII. Conclusion

In this paper we presented a modified FCM clustering method for enhancing the property which is typical in the clustering based on entropy such as GMM or normal mixture. A slightly modified FCM objective function resolves the problem of singularity in FCM-s while maintaining the robustness of FCM-s. KHM clustering is reinterpreted as a kind of FCM clustering. The weighting approach to IRLS-FCM clustering by penalizing solutions via weights on the data points as KHM clustering may be a prospective modification of fuzzy clustering.

IRLS-FCM clustering based on FCM-g is applied to a post-supervised classifier design in [22], and achieves high classification performance.

References

[1] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, 1981.
[2] F. Höppner, F. Klawonn, R. Kruse, T. Runkler : *Fuzzy Cluster Analysis, (Methods for Classification, Data Analysis and Image Recognition)*, John Wiley & Sons, 1999.
[3] Z.-Q. Liu and S. Miyamoto (Eds.), *Softcomputing and Human-Centered Machines*, Springer-Verlag, 2000.
[4] S. Miyamoto, D. Suizu, O. Takata, "Methods of fuzzy c-means and possibilistic clustering using a quadratic term, *Scientiae Mathematicae Japonicae* vol.60, no.2, pp.217-233, 2004.
[5] R. O. Duda and P. E. Hart: *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.
[6] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proc. of the IEEE*, vol.86, no.11, pp.2210-2239, 1998.
[7] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society*, vol.B-39, pp.1-38, 1977.
[8] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions*, John Wiley and Sons, 1997.
[9] B. Zhang, "Generalized k-harmonic means," *Technical Report TRHPL-2000-137*, Hewlett Packard Labs, 2000.
[10] B. Zhang, M. Hsu, and U. Dayal, "K-Harmonic means - spatial clustering algorithm with boosting," *Temporal, Spatial, and Spatio-Temporal Data Mining*, pp.31-45, 2000.
[11] B. Zhang, M. Hsu, and U. Dayal, "Harmonic average based clustering method and system," *US Patent 6, 584, 433*, 2000.
[12] R. Nock and F. Nielsen, "On weighting clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.28, no.8, pp.1223-1235, 2006.
[13] J. Kivinen and M. Warmuth, "Boosting as entropy projection," *Proc. 12th Ann. Conf. Computational Learning Theory*, pp. 134-144,1999.
[14] R.E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Proc. 11th Int'l Conf. Computational Learning Theory*, pp.80-91, 1998.
[15] M.J. Kearns, "Thoughts on hypothesis boosting," *ML class project*, 1988.
[16] R. Krishnapuram and J. Keller, "A possibilistic approach to clustering, "*IEEE Transactions on Fuzzy Systems*, vol.1, pp.98-110, 1993
[17] R. N. Davé and R. Krishnapuram, "Robust clustering methods, A unified approach, "*IEEE Trans. Fuzzy Syst.*, vol.5, no.2, pp.270-293, 1997.
[18] P. W. Holland and R. E. Welsch, "Robust regression using iteratively reweighted least-squares, "*Communications in Statistics*, vol. A6, no. 9, pp. 813-827, 1977.
[19] P. J. Huber. *Robust Statistics*. New York:Wiley, first edition, 1981.
[20] D. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix, "*Proc. IEEE CDC*, vol.2, pp.761-766, 1979.
[21] T. A. Runkler and J. C. Bezdek, "Alternating cluster estimation: a new tool for clustering and function approximation," *IEEE Trans. Fuzzy Syst.*, vol. 7, no. 4, pp. 377-393, 1999.
[22] H. Ichihashi, K. Honda, A. Notsu and T. Yagi, "Fuzzy c-Means classifier with deterministic initialization and missing value imputation," *Proc. of the 2007 IEEE Symposium on Foundations of Computational Intelligence*, Hawaii, April, 2007 (to appear).