Fuzzy C-means algorithm based on standard mahalanobis distances

Hsiang-Chuan Liu
Asia University

ISBN 978-952-5726-02-2 (Print), 978-952-5726-03-9 (CD-ROM)
Proceedings of the 2009 International Symposium on Information Processing (ISIP'09)
Huangshan, P. R. China, August 21-23, 2009, pp. 422-427

Fuzzy C-Means Algorithm Based on Standard Mahalanobis Distances

Hsiang-Chuan Liu[1], Bai-Cheng Jeng[2], Jeng-Ming Yih[3], and Yen-Kuei Yu[2]
[1] Department of Bioinformatics, Asia University, Taichung, Taiwan, ROC
Email: lhc@asia.edu.tw
[2] Graduate School of Educational Measurement and Statistics, National Taichung University, Taichung, Taiwan, ROC
Email: jason5907@yahoo.com.tw, haikuei@gmail.com
[3] Department of Mathematics Education, National Taichung University, Taichung, Taiwan, ROC
Email: yih@mail.ntcu.edu.tw

***Abstract*—Some of the well-known fuzzy clustering algorithms are based on Euclidean distance function, which can only be used to detect spherical structural clusters. Gustafson-Kessel clustering algorithm and Gath-Geva clustering algorithm were developed to detect non-spherical structural clusters. However, the former needs added constraint of fuzzy covariance matrix, the later can only be used for the data with multivariate Gaussian distribution. Two improved Fuzzy C-Means algorithm based on different Mahalanobis distance, called FCM-M and FCM-CM were proposed by our previous works, In this paper, A improved Fuzzy C-Means algorithm based on a standard Mahalanobis distance (FCM-SM) is proposed The experimental results of three real data sets show that our proposed new algorithm has the better performance.**

***Index Terms*—GK-algorithm; GG-algorithm; FCM-M algorithm; FCM-CM algorithm; FCM-SM algorithm**

I. Introduction

Clustering technique plays an important role in data analysis and interpretation. It groups data into clusters so that the data objects within a cluster have high similarity in comparison to one another, but are very dissimilar to those data objects in other clusters. Fuzzy clustering is a branch in clustering analysis and it is widely used in the pattern recognition field. The well-known ones, such as Bezdek's Fuzzy C-Means (FCM) and Li et al's Fuzzy Weighted C-Means (FWCM) [1,2], are based on Euclidean distance. These fuzzy clustering algorithms can only be used to detect the data classes with the same super spherical shapes.

To overcome the drawback due to Euclidean distance, we could try to extend the distance measure to Mahalanobis distance (MD). However, Krishnapuram and Kim (1999) [3] pointed out that the Mahalanobis distance can not be used directly in clustering algorithm. Gustafson-Kessel (GK) clustering algorithm [4] and Gath-Geva (GG) clustering algorithm [5] were developed to detect non-spherical structural clusters. In GK-algorithm, a modified Mahalanobis distance with preserved volume was used. However, the added fuzzy covariance matrices in their distance measure were not directly derived from the objective function. In GG algorithm, the Gaussian distance can only be used for the data with multivariate normal distribution.

In our two previous works, to add a regulating factor of Each covariance matrix to each class in the objective function, and deleted the constraint of the determinants of covariance matrices in the GK algorithm, the Fuzzy C-Means algorithm based on adaptive Mahalanobis distances and common Mahalanobis distance, respectively (FCM-M and FCM-CM), [8-12] were proposed, and then, the fuzzy covariance matrices in the Mahalanobis distance can be directly derived by minimizing the objective function.

In this paper, replacing the common covariance matrix with the correlation matrix in the objective function in the FCM-CM algorithm, and then, a new fuzzy clustering method, called the Fuzzy C-Means algorithm based on standard Mahalanobis distance (FCM-SM), is proposed.

II. Fuzzy C-Means Algorithm

Fuzzy C-Means Algorithm (FCM) is first developed by Dunn [6] and improved by Bezdek [1].The objective function of FCM is given in Equation (1)

$$J_{FCM}^{m}\left(U,A,X\right)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}d_{ij}^{2} \tag{1}$$

Where

$$U=\left[\mu_{ij}\right]_{c\times n},\ \mu_{ij}\in\left[0,1\right],\ i=1,2,...,c,\ j=1,2,...,n$$

$$A=\left[\underline{a}_1,\underline{a}_2,...,\underline{a}_c\right],\ \underline{a}_i\in R^p$$

$$X=\left[\underline{x}_1,\underline{x}_2,...,\underline{x}_n\right],\ \underline{x}_j\in R^p,\ j=1,2,...,n$$

μ_{ij} is the membership degree of data object $\underline{x}_j$ in cluster i and it satisfies the following constraint given by Equation (2).

$$\sum_{i=1}^{c}\mu_{ij}=1,\ \forall j=1,2,...,n \tag{2}$$

c is the number of clusters, m is the fuzzifier, m>1, which controls the fuzziness of the method. They are both parameters and need to be specified before running the algorithm. $d_{ij}^2=\left\|\underline{x}_j-\underline{a}_i\right\|^2$ is the square Euclidean distance between data object $\underline{x}_j$ to center $\underline{a}_i$. Minimizing the objective function, Equation (1), with the constraint, Equation (2), is a non-trivial constraint nonlinear optimization problem with continuous parameters $\underline{a}_i$ and μ_{ij}. Therefore, an analytical solution is absent. As a result, an alternating optimization scheme, alternatively optimizing one set of parameters while the other set of parameters are considered as fixed, is used here. Then the updating functions for $\underline{a}_i$ and μ_{ij} are obtained as Equation (4) and (5).

The steps of the FCM are listed as follows:

Step 1: Determining the number of cluster; c and m-value (let m=2), given converging error, $\varepsilon > 0$ (such as $\varepsilon = 0.001$). Randomly choose the initial membership

$u^{(0)}{}_{ij}$, $i = 1,2,...,c,\ j = 1,2,...,n$, such that

$$\sum_{1 \le i \le c} u^{(0)}{}_{ij} = 1,\ j = 1,2,...,n \tag{3}$$

Step 2: Find

$$\underline{a}_i^{(k)} = \left[\sum_{j=1}^{n} \left[\mu_{ij}^{(k-1)} \right]^m \right]^{-1} \sum_{j=1}^{n} \left[\mu_{ij}^{(k-1)} \right]^m \underline{x}_j, \quad i = 1,2,...,c \tag{4}$$

$$\mu_{ij}^{(k)} = \left[\sum_{l=1}^{c} \left[\frac{\left(\underline{x}_j - \underline{a}_i^{(k)} \right)' \left(\underline{x}_j - \underline{a}_i^{(k)} \right)}{\left(\underline{x}_j - \underline{a}_l^{(k)} \right)' \left(\underline{x}_j - \underline{a}_l^{(k)} \right)} \right]^{\frac{1}{m-1}} \right]^{-1} \tag{5}$$

Step 3: Increment k; until $\frac{1}{c} \sum_{i=1}^{c} \left\| \underline{a}_i^{(k)} - \underline{a}_i^{(k-1)} \right\|^2 < \varepsilon$ (6)

III. GK ALGORITHM

The reason that FCM can only work well for spherical shaped clusters because in the objective function the distances between data points to the centers of the clusters are calculated by Euclidian distances. To overcome the above drawback, we could try to extend the distance measure to Mahalanobis distance (MD). However, Krishnapuram and Kim (1999) [2] pointed out that the Mahalanobis distance can not be used directly in clustering algorithm.

Gustafson and Kessel (1979) extended the Euclidian distances of the standard FCM by employing an adaptive norm, in order to detect clusters of different geometrical shape without changing the clusters' sizes in one data set. The objective function of GK algorithm is given in Equation (7),(8),(9) and (10).

$$J_{GK}^{m}(U, A, V, X) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^{m} d^2 \left(\underline{x}_j, \underline{a}_i \right) \tag{7}$$

$$d^2 \left(\underline{x}_j, \underline{a}_i \right) = \left\| \underline{x}_j - \underline{a}_i \right\|_{V_i}^2 = \left(\underline{x}_j - \underline{a}_i \right)' V_i \left(\underline{x}_j - \underline{a}_i \right) \tag{8}$$

Where $V_i = \left| \Sigma_i \right|^{\frac{1}{p}} \Sigma_i^{-1}$ (9)

and

$$\Sigma_i = \left[\sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^{m} \right]^{-1} \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^{m} \left(\underline{x}_j - \underline{a}_i \right) \left(\underline{x}_j - \underline{a}_i \right)' \tag{10}$$

The steps of the GK algorithm are listed as follows:

Step 1: Determining the number of cluster; c and m-value (let m=2), given converge error, $\varepsilon > 0$ (such as $\varepsilon = 0.001$).

Randomly choose the initial membership

$u^{(0)}{}_{ij}$, $i = 1,2,...,c,\ j = 1,2,...,n$, such that

$$\sum_{1 \le i \le c} u^{(0)}{}_{ij} = 1,\ j = 1,2,...,n \tag{11}$$

Step 2: Find

$$\underline{a}_i^{(k)} = \left[\sum_{j=1}^{n} \left(\mu_{ij}^{(k-1)} \right)^m \right]^{-1} \sum_{j=1}^{n} \left(\mu_{ij}^{(k-1)} \right)^m \underline{x}_j, \quad i = 1,2,...,c, \quad k = 1,2,.. \tag{12}$$

$$\Sigma_i^{(k)} = \left[\sum_{i=1}^{c} \sum_{j=1}^{n} \left[\mu_{ij}^{(k-1)} \right]^m \right]^{-1} \sum_{i=1}^{c} \sum_{j=1}^{n} \left[\mu_{ij}^{(k-1)} \right]^m \left(\underline{x}_j - \underline{a}_i^{(k)} \right) \left(\underline{x}_j - \underline{a}_i^{(k)} \right)' \tag{13}$$

$$V_i^{(k)} = \left| \Sigma_i^{(k)} \right|^{\frac{1}{p}} \left[\Sigma_i^{(k)} \right]^{-1} \tag{14}$$

$$\mu_{ij}^{(k)} = \left[\sum_{l=1}^{c} \left[\frac{\left(\underline{x}_j - \underline{a}_i^{(k)} \right)' V_i^{(k)} \left(\underline{x}_j - \underline{a}_i^{(k)} \right)}{\left(\underline{x}_j - \underline{a}_l^{(k)} \right)' V_l^{(k)} \left(\underline{x}_j - \underline{a}_l^{(k)} \right)} \right]^{\frac{1}{m-1}} \right]^{-1} \tag{15}$$

Step 3: Increment k; until $\frac{1}{c} \sum_{i=1}^{c} \left\| \underline{a}_i^{(k)} - \underline{a}_i^{(k-1)} \right\|^2 < \varepsilon$

IV. GG ALGORITHM

Gath-Geva (GG) fuzzy clustering algorithm is an extension of Gustafson-Kessel (GK) fuzzy clustering algorithm, and also takes the size and density of clusters for classification (Hoppner et al, 1999)[7], Hence, it has better behaviors for irregular features.

Probabilistic interpretation of GG clustering is shown by Equation (16)

$$P(X \mid \eta) = \sum_{i=1}^{c} P(X, \eta_i) = \sum_{i=1}^{c} P(\eta_i) P(X \mid \eta_i) \tag{16}$$

Gath and Geva (1989) [9] assumed that the normal distribution N_i with expected value $\underline{a}_i$ and covariance matrix Σ_i is chosen for generating a datum with prior probability. p_i, satisfying

$$P\left(\underline{x}_j \mid \eta_i\right)=\frac{p_i}{(2\pi)^{\frac{p}{2}}\sqrt{|\Sigma_i|}}\exp\left[-\frac{1}{2}\left(\underline{x}_j-\underline{a}_i\right)'\Sigma_i^{-1}\left(\underline{x}_j-\underline{a}_i\right)\right] \quad (17)$$

Where $X=\left[\underline{x}_1,\underline{x}_2,...,\underline{x}_n\right]$, $\underline{x}_j\in R^p$, $j=1,2,...,n$ is the data matrix, the covariance matrix of cluster i is $\Sigma_i\in R^{p\times p}$. P is the number of dimension of data, c is the number of clusters, $d^2\left(\underline{x}_j,\underline{a}_i\right)$ for GG algorithm is chosen to be indirectly proportional to Equation (17) which is the posterior probability (likelihood) function. A small distance means a high probability, and a large distance means a low probability for membership. GG algorithm is based on minimization of the sum of weighted square distances between the data and the cluster centers of the objective function in Equation (18)

$$J_{GG}^m\left(U,A,\Sigma,X\right)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^m d^2\left(\underline{x}_j,\underline{a}_i\right) \quad (18)$$

Conditions for probablity clusters partiton are

$$m\in[1,\infty);U=\left[\mu_{ij}\right]_{c\times n};\mu_{ij}\in\left[0,1\right],i=1,2,...,c,\ j=1,2,...,n$$

$$\sum_{i=1}^{c}\mu_{ij}=1,\ j=1,2,...,n,\ 0<\sum_{j=1}^{n}\mu_{ij}<n,\ i=1,2,...,c \quad (19)$$

$$d^2\left(\underline{x}_j,\underline{a}_i\right)=\frac{1}{P\left(\underline{x}_j\mid\eta_i\right)}=\frac{(2\pi)^{\frac{p}{2}}\sqrt{|\Sigma_i|}_i}{p_i}\exp\left[\frac{1}{2}\left(\underline{x}_j-\underline{a}_i\right)'\Sigma_i^{-1}\left(\underline{x}_j-\underline{a}_i\right)\right] \quad (20)$$

Minimizing the objective function respect to all parameters in Equation (18), with the constraint (19), we can obtain the following GG algorithm.

The steps of the GG algorithm are listed as follows:

Step 1: Determining the number of cluster; c and m-value (let m=2), given converge error, $\varepsilon>0$ (such as $\varepsilon=0.001$). Randomly choose the initial membership

$u^{(0)}{}_{ij}$, $i=1,2,...,c$, $j=1,2,...,n$, such that

$$\sum_{1\le i\le c}u^{(0)}{}_{ij}=1,\ j=1,2,...,n \quad (21)$$

Step 2: Find

$$\underline{a}_i^{(k)}=\left[\sum_{j=1}^{n}\left(\mu_{ij}^{(k-1)}\right)^m\right]^{-1}\sum_{j=1}^{n}\left(\mu_{ij}^{(k-1)}\right)^m\underline{x}_j,\quad i=1,2,...,c,\quad k=1,2,... \quad (22)$$

$$\Sigma_i^{(k)}=\left[\sum_{j=1}^{n}\left[\mu_{ij}^{(k-1)}\right]^m\right]^{-1}\sum_{j=1}^{n}\left[\mu_{ij}^{(k-1)}\right]^m\left(\underline{x}_j-\underline{a}_i^{(k)}\right)\left(\underline{x}_j-\underline{a}_i^{(k)}\right)', \quad (23)$$

$$P_i^{(k)}=\left[\sum_{l=1}^{c}\sum_{j=1}^{n}\left(\mu_{lj}^{(k-1)}\right)^m\right]^{-1}\sum_{j=1}^{n}\left(\mu_{ij}^{(k-1)}\right)^m,\quad i=1,2,...,c,\quad k=1,2,.. \quad (24)$$

$$\mu_{ij}^{(k)}=\left[\sum_{l=1}^{c}\left[\frac{d^2\left(\underline{x}_j,\underline{a}_i^{(k)}\right)}{d^2\left(\underline{x}_j,\underline{a}_l^{(k)}\right)}\right]^{\frac{1}{m-1}}\right]^{-1} \quad (25)$$

Where

$$d^2\left(\underline{x}_j,\underline{a}_i^{(k)}\right)=\frac{(2\pi)^{\frac{p}{2}}\sqrt{|\Sigma_i^{(k)}|}_i}{p_i^{(k)}}\exp\left[\frac{1}{2}\left(\underline{x}_j-\underline{a}_i^{(k)}\right)'\left[\Sigma_i^{(k)}\right]^{-1}\left(\underline{x}_j-\underline{a}_i^{(k)}\right)\right]$$

Step 3: Increment k; until $\frac{1}{c}\sum_{i=1}^{c}\left\|\underline{a}_i^{(k)}-\underline{a}_i^{(k-1)}\right\|^2<\varepsilon$

V. FCM-M ALGORITHM

For improving the limitation of GK algorithm and GG algorithm, we added a regulating factor of covariance matrix, $-\ln\left|+\Sigma_i^{-1}\right|$, to each class in the objective function, and deleted the constraint of the determinant of covariance matrices, in GK Algorithm as the objective function (7),(8),(9). We can obtain the objective function of Fuzzy C-Means based on adaptive Mahalanobis distance (FCM-M) as following [8,12];

$$J_{FCM-M}^m\left(U,A,\Sigma,X\right)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^m d^2\left(\underline{x}_j,\underline{a}_i\right) \quad (26)$$

Conditions for FCM-M are

$$m\in[1,\infty);U=\left[\mu_{ij}\right]_{c\times n};\mu_{ij}\in\left[0,1\right],i=1,2,...,c,\ j=1,2,...,n$$

$$\sum_{i=1}^{c}\mu_{ij}=1,\ j=1,2,...,n,\ 0<\sum_{j=1}^{n}\mu_{ij}<n,\ i=1,2,...,c \quad (27)$$

$$d^2\left(\underline{x}_j,\underline{a}_i\right)=\begin{cases}\left(\underline{x}_j-\underline{a}_i\right)'\Sigma_i^{-1}\left(\underline{x}_j-\underline{a}_i\right)-\ln\left|\Sigma_i^{-1}\right| & if\left(\underline{x}_j-\underline{a}_i\right)'\Sigma_i^{-1}\left(\underline{x}_j-\underline{a}_i\right)-\ln\left|\Sigma_i^{-1}\right|\ge 0\\ 0 & if\left(\underline{x}_j-\underline{a}_i\right)'\Sigma_i^{-1}\left(\underline{x}_j-\underline{a}_i\right)-\ln\left|\Sigma_i^{-1}\right|<0\end{cases} \quad (28)$$

Minimizing the objective function respect to all parameters in Equation (26), with the constraint (27), (28) we can obtain the following FCM-M algorithm;

The steps of the FCM-M are listed as follows [8].

Step 1: Determining the number of cluster; c and m-value (let m=2), given converge error, $\varepsilon>0$ (such as $\varepsilon=0.001$).

Randomly choose the initial membership

$u^{(0)}{}_{ij}$, $i=1,2,...,c$, $j=1,2,...,n$, such that

$$\sum_{1\le i\le c}u^{(0)}{}_{ij}=1,\ j=1,2,...,n \quad (29)$$

$$\underline{a}_i^{(0)}=\left[\sum_{j=1}^{n}\mu_{ij}^{(0)}\right]^{-1}\sum_{j=1}^{n}\mu_{ij}^{(0)}\underline{x}_j,\quad i=1,2,...,c \quad (30)$$

$$D=\sum_{i=1}^{c}\sum_{j=1}^{n}\left[\mu_{ij}^{(0)}\right]^m\left[\left(\underline{x}_j-\underline{a}_i^{(0)}\right)'\left(\underline{x}_j-\underline{a}_i^{(0)}\right)\right]>0 \quad (31)$$

$$\Sigma_i^{(0)} = \left[\sum_{j=1}^{n}\mu_{ij}^{(0)}\right]^{-1}\sum_{j=1}^{n}\mu_{ij}^{(0)}\left(\underline{x}_j-\underline{a}_i^{(0)}\right)\left(\underline{x}_j-\underline{a}_i^{(0)}\right)' \quad i=1,2,...,c \tag{32}$$

$$if \quad \left|\Sigma_i^{(0)}\right| > D, \ or \quad \left|\Sigma_i^{(0)}\right| < \frac{1}{D} \quad then \ \Sigma_i^{(0)} = I \tag{33}$$

Step 2: Find

$$\underline{a}_i^{(k)} = \left[\sum_{j=1}^{n}\left(\mu_{ij}^{(k-1)}\right)^m\right]^{-1}\sum_{j=1}^{n}\left(\mu_{ij}^{(k-1)}\right)^m \underline{x}_j, \quad i=1,2,...,c, \quad k=1,2,.. \tag{34}$$

$$\Sigma_i^{(k)} = \frac{\sum_{j=1}^{n}\left[\mu_{ij}^{(k-1)}\right]^m\left(\underline{x}_j-\underline{a}_i^{(k)}\right)\left(\underline{x}_j-\underline{a}_i^{(k)}\right)'}{\sum_{j=1}^{n}\left[\mu_{ij}^{(k-1)}\right]^m}, \tag{35}$$

$$if \quad \left|\Sigma_i^{(k)}\right| > D, \ or \quad \left|\Sigma_i^{(k)}\right| < \frac{1}{D} \quad then \ \Sigma_i^{(k)} = I \tag{36}$$

$$\mu_{ij}^{(k)} = \left[\sum_{s=1}^{c}\left[\frac{\left(\underline{x}_j-\underline{a}_i^{(k)}\right)'\left[\Sigma_i^{-1}\right]^{(k)}\left(\underline{x}_j-\underline{a}_i^{(k)}\right)-\ln\left|\left[\Sigma_i^{-1}\right]^{(k)}\right|}{\left(\underline{x}_j-\underline{a}_s^{(k)}\right)'\left[\Sigma_s^{-1}\right]^{(k)}\left(\underline{x}_j-\underline{a}_s^{(k)}\right)-\ln\left|\left[\Sigma_s^{-1}\right]^{(k)}\right|}\right]^{\frac{1}{m-1}}\right]^{-1} \tag{37}$$

Step 3: Increment k; until $\frac{1}{c}\sum_{i=1}^{c}\left\|\underline{a}_i^{(k)}-\underline{a}_i^{(k-1)}\right\|^2 < \varepsilon$.

Note that FCM is a special case of FCM-M, when covariance matrices equal to identity matrices [8].

VI. FCM-CM ALGORITHM

For improving the stability of the clustering results, we replace all of the covariance matrices with the same common covariance matrix in the objective function in the FCM-M algorithm, and then, an improve fuzzy clustering method, called the Fuzzy C-Means algorithm based on common Mahalanobis distance (FCM-CM) is proposed. We can obtain the objective function of FCM-CM as following:

$$J_{FCM-CM}^{m}\left(U,A,\Sigma,X\right)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}d^2\left(\underline{x}_j,\underline{a}_i\right) \tag{38}$$

Conditions for FCM-CM are

$$m\in[1,\infty);U=\left[\mu_{ij}\right]_{c\times n};\mu_{ij}\in\left[0,1\right], i=1,2,...,c, \ j=1,2,...,n$$
$$\sum_{i=1}^{c}\mu_{ij}=1, \ j=1,2,...,n, \ 0<\sum_{j=1}^{n}\mu_{ij}<n, i=1,2,...,c \tag{39}$$

$$d^2\left(\underline{x}_j,\underline{a}_i\right)=\begin{cases}\left(\underline{x}_j-\underline{a}_i\right)'\Sigma^{-1}\left(\underline{x}_j-\underline{a}_i\right)-\ln\left|\Sigma^{-1}\right| & if \left(\underline{x}_j-\underline{a}_i\right)'\Sigma^{-1}\left(\underline{x}_j-\underline{a}_i\right)-\ln\left|\Sigma^{-1}\right|\geq 0\\ 0 & if \left(\underline{x}_j-\underline{a}_i\right)'\Sigma^{-1}\left(\underline{x}_j-\underline{a}_i\right)-\ln\left|\Sigma^{-1}\right|<0\end{cases} \tag{40}$$

Minimizing the objective function respect to all parameters in Equation (40) with the constraint (39), we can obtain the following FCM-CM algorithm.

The steps of the FCM-CM are listed as follows [12]

Step 1: Determining the number of cluster; c and m-value (let m=2), given converge error, $\varepsilon > 0$ (such as $\varepsilon = 0.001$).

Randomly choose the initial membership

$u^{(0)}_{ij}, i=1,2,...,c, \ j=1,2,...,n$, such that

$$\sum_{1\le i\le c}u^{(0)}_{ij}=1, \ j=1,2,...,n \tag{41}$$

$$\underline{a}_i^{(0)} = \left[\sum_{j=1}^{n}\mu_{ij}^{(0)}\right]^{-1}\sum_{j=1}^{n}\mu_{ij}^{(0)}\underline{x}_j, \quad i=1,2,...,c, \quad k=1,2,.. \tag{42}$$

$$D=\sum_{i=1}^{c}\sum_{j=1}^{n}\left[\mu_{ij}^{(0)}\right]^m\left[\left(\underline{x}_j-\underline{a}_i^{(0)}\right)'\left(\underline{x}_j-\underline{a}_i^{(0)}\right)\right]>0 \tag{43}$$

$$\Sigma^{(0)} = \left[\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{(0)}\right]^{-1}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{(0)}\left(\underline{x}_j-\underline{a}_i^{(0)}\right)\left(\underline{x}_j-\underline{a}_i^{(0)}\right)' \tag{44}$$

$$if \quad \left|\Sigma^{(0)}\right| > D, \ or \quad \left|\Sigma^{(0)}\right| < \frac{1}{D} \quad then \ \Sigma^{(0)} = I \tag{45}$$

Step 2: Find

$$\underline{a}_i^{(k)} = \left[\sum_{j=1}^{n}\left(\mu_{ij}^{(k-1)}\right)^m\right]^{-1}\sum_{j=1}^{n}\left(\mu_{ij}^{(k-1)}\right)^m \underline{x}_j, \quad i=1,2,...,c, \quad k=1,2,.. \tag{46}$$

$$\Sigma^{(k)} = \left[\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{(k-1)}\right]^{-1}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{(k-1)}\left(\underline{x}_j-\underline{a}_i^{(k)}\right)\left(\underline{x}_j-\underline{a}_i^{(k)}\right)' \tag{47}$$

$$if \quad \left|\Sigma^{(k)}\right| > D, \ or \quad \left|\Sigma^{(k)}\right| < \frac{1}{D} \quad then \ \Sigma^{(k)} = I \tag{48}$$

$$\mu_{ij}^{(k)} = \left[\sum_{s=1}^{c}\left[\frac{\left(\underline{x}_j-\underline{a}_i^{(k)}\right)'\left[\Sigma^{-1}\right]^{(k)}\left(\underline{x}_j-\underline{a}_i^{(k)}\right)-\ln\left|\left[\Sigma^{-1}\right]^{(k)}\right|}{\left(\underline{x}_j-\underline{a}_s^{(k)}\right)'\left[\Sigma^{-1}\right]^{(k)}\left(\underline{x}_j-\underline{a}_s^{(k)}\right)-\ln\left|\left[\Sigma^{-1}\right]^{(k)}\right|}\right]^{\frac{1}{m-1}}\right]^{-1} \tag{49}$$

Step 3: Increment k; until $\frac{1}{c}\sum_{i=1}^{c}\left\|\underline{a}_i^{(k)}-\underline{a}_i^{(k-1)}\right\|^2 < \varepsilon$.

Note that FCM is a special case of FCM-CM, when covariance matrices equal to identity matrices [12].

VII. FCM-SM ALGORITHM

For normalizing each feature in the objective function in the FCM-CM algorithm, and then, the all covariance matrices become the corresponding correlation matrices, and then, the new fuzzy clustering method, called the Fuzzy C-Means algorithm based on standard Mahalanobis distance (FCM-SM) is proposed. We can obtain the objective function of FCM-SM as following:

$$J_{FCM-SM}^{m}\left(U,A,R,Z\right)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}d^2\left(z_j,\underline{a}_i\right) \tag{50}$$

Conditions for FCM-CM are

$$m \in [1,\infty); U = \left[\mu_{ij}\right]_{c\times n}; \mu_{ij} \in [0,1], i=1,2,...,c, j=1,2,...,n$$
$$\sum_{i=1}^{c}\mu_{ij}=1,\ j=1,2,...,n,\ 0<\sum_{j=1}^{n}\mu_{ij}<n,\ i=1,2,...,c \tag{51}$$

$$d^2\left(\underline{z}_j,\underline{a}_i\right)=\begin{cases}\left(\underline{z}_j-\underline{a}_i\right)' R^{-1}\left(\underline{z}_j-\underline{a}_i\right)-\ln\left|\Sigma^{-1}\right| & if\ \left(\underline{z}_j-\underline{a}_i\right)' R^{-1}\left(\underline{z}_j-\underline{a}_i\right)-\ln\left|R^{-1}\right|\ge 0\\ 0 & if\ \left(\underline{z}_j-\underline{a}_i\right)' R^{-1}\left(\underline{z}_j-\underline{a}_i\right)-\ln\left|R^{-1}\right|<0\end{cases} \tag{52}$$

The steps of the FCM-SM are listed as follows

Step 1: Determining the number of cluster; c and m-value (let m=2), given converge error, $\varepsilon > 0$ (such as $\varepsilon = 0.001$).

Randomly choose the initial membership

$u^{(0)}{}_{ij}$, $i=1,2,...,c,\ j=1,2,...,n$, such that

$$\sum_{1\le i\le c} u^{(0)}{}_{ij}=1,\ j=1,2,...,n \tag{53}$$

$$\underline{a}_i^{(0)}=\left[\sum_{j=1}^{n}\mu_{ij}^{(0)}\right]^{-1}\sum_{j=1}^{n}\mu_{ij}^{(0)}\underline{z}_j,\quad i=1,2,...,c,\quad k=1,2,.. \tag{54}$$

Where

$$\underline{z}_j=\left(z_{1j},z_{2j},...,z_{pj}\right)',\ z_{tj}=\frac{x_{tj}-\overline{x}_t}{s_t},\ j=1,2,...,n,\ t=1,2,...,p \tag{55}$$

$$\overline{x}_t=\frac{1}{n}\sum_{j=1}^{n}x_{tj},\ s_t=\frac{1}{n}\sum_{j=1}^{n}\left(x_{tj}-\overline{x}_t\right)^2,\quad t=1,2,\ldots,p \tag{56}$$

$$R^{(0)}=\left[\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{(0)}\right]^{-1}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{(0)}\left(\underline{z}_j-\underline{a}_i^{(0)}\right)\left(\underline{z}_j-\underline{a}_i^{(0)}\right)' \tag{57}$$

$$if\quad \left|R^{(0)}\right|<\frac{1}{100}\quad then\ \ R^{(0)}=I \tag{58}$$

Step 2: Find

$$\underline{a}_i^{(k)}=\left[\sum_{j=1}^{n}\left(\mu_{ij}^{(k-1)}\right)^m\right]^{-1}\sum_{j=1}^{n}\left(\mu_{ij}^{(k-1)}\right)^m\underline{z}_j,\quad i=1,2,...,c,\quad k=1,2,.. \tag{59}$$

$$R^{(k)}=\left[\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{(k-1)}\right]^{-1}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{(k-1)}\left(\underline{z}_j-\underline{a}_i^{(k)}\right)\left(\underline{z}_j-\underline{a}_i^{(k)}\right)' \tag{60}$$

$$if\quad \left|R^{(k)}\right|<\frac{1}{100}\quad then\ \ R^{(k)}=I \tag{61}$$

$$\mu_{ij}^{(k)}=\left[\sum_{s=1}^{c}\left[\frac{\left(\underline{z}_j-\underline{a}_i^{(k)}\right)'\left[R^{-1}\right]^{(k)}\left(\underline{z}_j-\underline{a}_i^{(k)}\right)-\ln\left|\left[R^{-1}\right]\right|^{(k)}}{\left(\underline{z}_j-\underline{a}_s^{(k)}\right)'\left[R^{-1}\right]^{(k)}\left(\underline{z}_j-\underline{a}_s^{(k)}\right)-\ln\left|\left[R^{-1}\right]\right|^{(k)}}\right]^{\frac{1}{m-1}}\right]^{-1} \tag{62}$$

Step 3: Increment k; until $\frac{1}{c}\sum_{i=1}^{c}\left\|\underline{a}_i^{(k)}-\underline{a}_i^{(k-1)}\right\|^2<\varepsilon$.

Note that FCM is a special case of FCM-SM, when each correlation matrix equal to identity matrix.

VIII. Empirical Analysis

Three data sets from the University of California at Irvine (UCI) Machine Learning Repository [13,14] are used in the empirical study, The information about the data is shown in TABLE I.

The performances of FCM, GK, GG, FCM-M, FCM-CM, and FCM-SM all with the fuzzifier m=2, are compared in these experiments. The results of FCM, GK, and GG are obtained by applying the Matlab toolbox developed by [15].

The Mean clustering Accuracies of 100 initial value sets of FCM, GK, GG, FCM-M, FCM-CM, and FCM-SM for these three Datasets were shown in TABLE II.

TABLE I.
The Details of The Used Datasets

Datasets	Attributes	Classes	Sample number
Iris	4	3	150
Wdbc	30	2	569
Dermatology	34	6	366

TABLE II.
The Accuracies of Six Algorithms

	Iris	Wdbc	Dermatology
FCM	0.8400	0.9139	0.5132
GK	0.9000	0.7404	0.4796
GG	0.7173	0.7767	0.3602
FCM-M	0.8482	0.9170	0.6509
FCM-CM	0.8501	0.9172	0.6611
FCM-SM	0.8520	0.9172	0.6611

From TABLE II, we can find that GG algorithm always has the worst performance in three datasets. Although the performance of GK algorithm is better than which of FCM algorithm in Iris dataset, but the performance of the former is worse than which of the later in Wdbc dataset and Dermatology dataset. The performances of our proposed three algorithms, FCM-M, FCM-CM and FCM-SM are simultaneously better than which of FCM algorithm in three datasets. In other words, our proposed three algorithms, FCM-M, FCM-CM and FCM-SM are better than FCM algorithm, GG algorithm and GK algorithm. Among our proposed three algorithms, the new algorithm, FCM-SM, has the best performance. In a word, FCM-SM algorithm is a better one than others.

IX. Conclusions

The well-known FCM is based on Euclidean distance function, which can only be used to detect spherical structural clusters. GK algorithm and GG algorithm were developed to detect non-spherical structural clusters. However, the former needs added constraint of fuzzy covariance matrix, the later can only be used for the data with multivariate Gaussian distribution. Two improved

Fuzzy C-Means algorithm based on different Mahalanobis distance, called FCM-M and FCM-CM were proposed by our previous works, In this paper, A improved Fuzzy C-Means algorithm based on a standard Mahalanobis distance (FCM-SM) is proposed The experimental results of three real data sets show that our proposed new algorithm has the better performance.

ACKNOWLEDGMENT

This paper is partially supported by the National Science Council grant (NSC 97-2410-H-468-014).

REFERENCES

[1] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum press*, pp. 65-70, N.Y. 1981.

[2] C.-H. Li, W.-C. Huang, B.-C. Kuo and C.-C. Hung "A Novel Fuzzy Weighted C-Method for Image Classification," International Journal of Fuzzy Systems, vol. 10, no. 3, pp.168-173, September 2008

[3] R. Krishnapuram and J. Kim, A note on the Gustafson-Kessel and adaptive fuzzy clustering algorithm, *IEEE Transactions on Fuzzy Systems*, vol. 7 no. 4 August 1999.

[4] D. E. Gustafson and W. C. Kessel, *Proc. IEEE Conf. Decision Contr. San Diego, CA*, 761 ,1979.

[5] Gath, and A. B. Geva, Unsupervised optimal fuzzy clustering, *IEEE Trans. Pattern Anal. Machine Intell.* vol.11, pp.773-781, 1989.

[6] J. C. Dunn, A fuzzy relative of the isodata process and its use in detecting compact, well-separated clusters *J. Cybern.* vol.3, vol.3, pp. 32-57, 1973.

[7] F. Höppner, F. Klawonn, R. Kruse, T. Runkler, Fuzzy Cluster Analysis, John Wiley and Sons, 1999

[8] H.-C. Liu, J.-M. Yih and S.-W. Liu, Fuzzy C-Mean Algorithm Based On Mahalanobis Distances and Better Initial Values. *Proceedings of the 10th Joint Conference& 12th International Conference on Fuzzy Theory & Technology.* vol.1, p.1398-1404, 2007.

[9] H.-C. Liu, J.-M. Yih, T.-W. Sheu and S.-W. Liu, A new Fuzzy Possibility Clustering Algorithms Based On Unsupervised Mahalanobis Distances, *Proceedings of International conference on Machine Learning and Cybernetics*, vol.7, no.7, pp.3939-3944, 2007.

[10] H.-C. Liu, J.-M. Yih, D.-B. Wu and S.-W. Liu, Fuzzy C-Mean Algorithm Based on "Complete" Mahalanobis Distances. *Proceedings of International conference on Machine Learning and Cybernetics*, vol.7, no.6, pp.3569-3574, 2008.

[11] H.-C. Liu, J.-M. Yih, W.-C. Lin and T.-S. Liu, Fuzzy C-Means Algorithm Based on PSO and Mahalanobis Distances. *International Journal of Innovative Computing, Information and Control*, ISSN 1349-4198, January 21 2009.(Accepted)

[12] H.-C. Liu, J.-M. Yih, W.-C. Lin and D.-B. Wu, Fuzzy C-Means Algorithm Based on common Mahalanobis Distances. *Journal of Multiple Valued Logic & Soft Computing*, pp1-13, July 10 2009.(Accepted)

[13] R. A. Fisher, The use of multiple measurements in taxonomic problems. *Annals of Eugenics*. vol.7, pp.179-188, 1936.

[14] C, Blake, E. Keogh, C. J. Merz, UCI Repository of Machine Learning Databases. Dept. Inf. Comput. Sci., Univ. California, Irvine, CA, 1998. From
http://www.ics.uci.edu/~mlearn/MLRepository.html

[15] B. Balasko, J. Abonyi and B. Feil " Fuzzy Clustering and. Data Analysis Toolbox For Use with Matlab" From
http://www.mathworks.com/matlabcentral/fileexchange/7473