NORMALIZED CLUSTERING ALGORITHM BASED ON MAHALANOBIS DISTANCE

JENG-MING YIH
Center of General Education,
Min-Hwei College of Health Care Management
Tainan, Taiwan

YUAN-HORNG LIN
Department of Mathematics Education,
National Taichung University of Education
Taichung, Taiwan

Abstract—**FCM (fuzzy c-means algorithm) based on Euclidean distance function converges to a local minimum of the objective function, which can only be used to detect spherical structural clusters. The added fuzzy covariance matrices in their distance measure were not directly derived from the objective function. In this paper, an improved Normalized Clustering Algorithm Based on Mahalanobis distance by taking a new threshold value and a new convergent process is proposed.**

Index Terms—**Normalized, Mahalanobis distance, Clustering algorithm.**

I. Introduction and Motivation

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).

Fuzzy clustering 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.

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, 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.

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, common Mahalanobis distance and standardized Mahalanobis distance, respectively (FCM-M, and FCM-CM), [8-12,16] were proposed, and then, the fuzzy covariance matrices in the Mahalanobis distance can be directly derived by minimizing the objective function.

In this paper, not only replacing the common covariance matrix with the correlation matrix in the objective function in the FCM-CM algorithm but also replacing the threshold D

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

A new fuzzy clustering method, called the Fuzzy C-Means algorithm based on normalized Mahalanobis distance (FCM-NM), is proposed.

II. Literature Review

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.

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, GK algorithm needs added constraint of fuzzy covariance matrix, GG algorithm can only be used for the data with multivariate Gaussian distribution. A Fuzzy C-Means algorithm based on Mahalanobis distance (FCM-M) was proposed to improve those limitations of above two algorithms, but it is not stable enough when some of its covariance matrices are not equal. An improved Fuzzy C-Means algorithm based on Normalized Mahalanobis distance (FCM-NM) is proposed. The experimental results of two real data sets consistently show that the performance of our proposed FCM-NM algorithm is better than those of above algorithms.

A. GK 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 (1),(2),(3) and (4).

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

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

Where $V_i=\left|\Sigma_i\right|^{\frac{1}{p}}\Sigma_i^{-1}$ (3)

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

B. 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 (5)

$$P\left(X\mid\eta\right)=\sum_{i=1}^{c}P\left(X,\eta_i\right)=\sum_{i=1}^{c}P\left(\eta_i\right)P\left(X\mid\eta_i\right) \tag{5}$$

Gath and Geva (1989) [9] assumed that the normal distribution with expected value and covariance matrix is chosen for generating a datum with prior probability., satisfying

$$P\left(\underline{x}_j\mid\eta_i\right)=\frac{p_i}{\left(2\pi\right)^{\frac{p}{2}}\sqrt{\left|\Sigma_i\right|}}\exp\left[-\frac{1}{2}\left(\underline{x}_j-\underline{a}_i\right)'\Sigma_i^{-1}\left(\underline{x}_j-\underline{a}_i\right)\right] \tag{6}$$

C. FCM-M Algorithm

For improving the limitation of GK algorithm and GG algorithm, we added a regulating factor of covariance matrix, , to each class in the objective function, and deleted the constraint of the determinant of covariance matrices, in GK Algorithm as the objective function (1),(2),(3). We can obtain the objective function of Fuzzy C-Means based on adaptive Mahalanobis distance (FCM-M) as following [8-12];

$$J^{m}_{FCM-M}\left(U,A,\Sigma,X\right)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}d^{2}\left(\underline{x}_j,\underline{a}_i\right) \qquad (7)$$

Conditions for FCM-M are

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

$$\sum_{i=1}^{c}\mu_{ij}=1,\ j=1,2,...,n,\ 0<\sum_{j=1}^{n}\mu_{ij}<n,\ i=1,2,...,c \qquad (8)$$

$$d^{2}\left(\underline{x}_j,\underline{a}_i\right)=\begin{cases}\left(\underline{x}_j-\underline{a}_i\right)'\Sigma_i^{-1}\left(\underline{x}_j-\underline{a}_i\right)-\ln\left|\Sigma_i^{-1}\right| & if\left(\underline{x}_j-\underline{a}_i\right)'\Sigma_i^{-1}\left(\underline{x}_j-\underline{a}_i\right)-\ln\left|\Sigma_i^{-1}\right|\geq 0\\ 0 & if\left(\underline{x}_j-\underline{a}_i\right)'\Sigma_i^{-1}\left(\underline{x}_j-\underline{a}_i\right)-\ln\left|\Sigma_i^{-1}\right|<0\end{cases} \qquad (9)$$

Minimizing the objective function respect to all parameters in Equation (7), with the constraint (8), (9) 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

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

$$\sum_{1\leq i\leq c}u^{(0)}_{ij}=1,\ j=1,2,...,n \qquad (10)$$

$$\underline{a}_i^{(0)}=\left[\sum_{j=1}^{n}\mu_{ij}^{(0)}\right]^{-1}\sum_{j=1}^{n}\mu_{ij}^{(0)}\underline{x}_j,\quad i=1,2,...,c \qquad (11)$$

$$D=\sum_{i=1}^{c}\sum_{j=1}^{n}\left[\mu_{ij}^{(0)}\right]^{m}\left[\left(\underline{x}_j-\underline{a}_i^{(0)}\right)'\left(\underline{x}_j-\underline{a}_i^{(0)}\right)\right]>0 \qquad (12)$$

$$\Sigma_i^{(0)}=\left[\sum_{j=1}^{n}\mu_{ij}^{(0)}\right]^{-1}\sum_{j=1}^{n}\mu_{ij}^{(0)}\left(\underline{x}_j-\underline{a}_i^{(0)}\right)\left(\underline{x}_j-\underline{a}_i^{(0)}\right)'\quad i=1,2,...,c \qquad (13)$$

$$if\quad\left|\Sigma_i^{(0)}\right|>D,\ \ or\quad\left|\Sigma_i^{(0)}\right|<\frac{1}{D}\quad then\ \ \Sigma_i^{(0)}=I \qquad (14)$$

Step 2: Find

$$\underline{a}_i^{(k)}=\left[\sum_{j=1}^{n}\left(\mu_{ij}^{(k-1)}\right)^{m}\right]^{-1}\sum_{j=1}^{n}\left(\mu_{ij}^{(k-1)}\right)^{m}\underline{x}_j,\quad i=1,2,...,c,\quad k=1,2,.. \qquad (15)$$

$$\Sigma_i^{(k)}=\frac{\sum_{j=1}^{n}\left[\mu_{ij}^{(k-1)}\right]^{m}\left(\underline{x}_j-\underline{a}_i^{(k)}\right)\left(\underline{x}_j-\underline{a}_i^{(k)}\right)'}{\sum_{j=1}^{n}\left[\mu_{ij}^{(k-1)}\right]^{m}}, \qquad (16)$$

$$if\quad\left|\Sigma_i^{(k)}\right|>D,\ \ or\quad\left|\Sigma_i^{(k)}\right|<\frac{1}{D}\quad then\ \ \Sigma_i^{(k)}=I \qquad (17)$$

$$\mu_{ij}^{(k)}=\left[\sum_{s=1}^{c}\left[\frac{\left(\underline{x}_j-\underline{a}_i^{(k)}\right)'\left[\Sigma_i^{-1}\right]^{(k)}\left(\underline{x}_j-\underline{a}_i^{(k)}\right)-\ln\left|\left[\Sigma_i^{-1}\right]\right|^{(k)}}{\left(\underline{x}_j-\underline{a}_s^{(k)}\right)'\left[\Sigma_s^{-1}\right]^{(k)}\left(\underline{x}_j-\underline{a}_s^{(k)}\right)-\ln\left|\left[\Sigma_s^{-1}\right]\right|^{(k)}}\right]^{\frac{1}{m-1}}\right]^{-1} \qquad (18)$$

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].

D. 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:

$$J^{m}_{FCM-CM}\left(U,A,\Sigma,X\right)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}d^{2}\left(\underline{x}_j,\underline{a}_i\right) \qquad (19)$$

Conditions for FCM-CM are

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

$$\sum_{i=1}^{c}\mu_{ij}=1,\ j=1,2,...,n,\ 0<\sum_{j=1}^{n}\mu_{ij}<n,\ i=1,2,...,c \qquad (20)$$

$$d^{2}\left(\underline{x}_j,\underline{a}_i\right)=\begin{cases}\left(\underline{x}_j-\underline{a}_i\right)'\Sigma^{-1}\left(\underline{x}_j-\underline{a}_i\right)-\ln\left|\Sigma^{-1}\right| & if\left(\underline{x}_j-\underline{a}_i\right)'\Sigma^{-1}\left(\underline{x}_j-\underline{a}_i\right)-\ln\left|\Sigma^{-1}\right|\geq 0\\ 0 & if\left(\underline{x}_j-\underline{a}_i\right)'\Sigma^{-1}\left(\underline{x}_j-\underline{a}_i\right)-\ln\left|\Sigma^{-1}\right|<0\end{cases} \qquad (21)$$

Minimizing the objective function respect to all parameters in Equation (19) with the constraint (21), 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

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

$$\sum_{1\leq i\leq c}u^{(0)}_{ij}=1,\ j=1,2,...,n \qquad (22)$$

$$\underline{a}_i^{(0)}=\left[\sum_{j=1}^{n}\mu_{ij}^{(0)}\right]^{-1}\sum_{j=1}^{n}\mu_{ij}^{(0)}\underline{x}_j,\quad i=1,2,...,c,\quad k=1,2,.. \qquad (23)$$

$$D=\sum_{i=1}^{c}\sum_{j=1}^{n}\left[\mu_{ij}^{(0)}\right]^{m}\left[\left(\underline{x}_j-\underline{a}_i^{(0)}\right)'\left(\underline{x}_j-\underline{a}_i^{(0)}\right)\right]>0 \qquad (24)$$

$$\Sigma^{(0)}=\left[\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{(0)}\right]^{-1}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{(0)}\left(\underline{x}_j-\underline{a}_i^{(0)}\right)\left(\underline{x}_j-\underline{a}_i^{(0)}\right)' \qquad (25)$$

$$if\quad\left|\Sigma^{(0)}\right|>D,\ \ or\quad\left|\Sigma^{(0)}\right|<\frac{1}{D}\quad then\ \ \Sigma^{(0)}=I \qquad (26)$$

Step 2: Find

$$\underline{a}_i^{(k)}=\left[\sum_{j=1}^{n}\left(\mu_{ij}^{(k-1)}\right)^{m}\right]^{-1}\sum_{j=1}^{n}\left(\mu_{ij}^{(k-1)}\right)^{m}\underline{x}_j,\quad i=1,2,...,c,\quad k=1,2,.. \qquad (27)$$

$$\Sigma^{(k)}=\left[\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{(k-1)}\right]^{-1}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{(k-1)}\left(\underline{x}_j-\underline{a}_i^{(k)}\right)\left(\underline{x}_j-\underline{a}_i^{(k)}\right)' \qquad (28)$$

$$if\quad\left|\Sigma^{(k)}\right|>D,\ \ or\quad\left|\Sigma^{(k)}\right|<\frac{1}{D}\quad then\ \ \Sigma^{(k)}=I \qquad (29)$$

$$\mu_{ij}^{(k)}=\left[\sum_{s=1}^{c}\left[\frac{\left(\underline{x}_j-\underline{a}_i^{(k)}\right)'\left[\Sigma^{-1}\right]^{(k)}\left(\underline{x}_j-\underline{a}_i^{(k)}\right)-\ln\left\|\Sigma^{-1}\right\|^{(k)}}{\left(\underline{x}_j-\underline{a}_s^{(k)}\right)'\left[\Sigma^{-1}\right]^{(k)}\left(\underline{x}_j-\underline{a}_s^{(k)}\right)-\ln\left\|\Sigma^{-1}\right\|^{(k)}}\right]^{\frac{1}{m-1}}\right]^{-1} \qquad (30)$$

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].

E. FCM-NM Algorithm

In this paper, not only z-score normalizing for each feature in the objective function in the FCM-CM algorithm, but also replacing the threshold D where

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

with the determinant value of the crisp correlation matrix, and then, the new fuzzy clustering method, called the Fuzzy C-Means algorithm based on normalized Mahalanobis distance (FCM-NM) is proposed. We can obtain the objective function of FCM-NM as following:

$$J^{m}_{FCM-NM}\left(U,A,R,Z\right)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}d^{2}\left(z_{j},\underline{a}_{i}\right) \qquad (32)$$

$$X=\left[\underline{x}_{1},\underline{x}_{2},\ldots,\underline{x}_{n}\right],\ \underline{x}_{j}\in R^{p},\ j=1,2,\ldots,n \qquad (33)$$

$$\underline{z}_{j}=\left(z_{1j},z_{2j},\ldots,z_{pj}\right)',\ z_{tj}=\frac{x_{tj}-\overline{x}_{t}}{s_{t}},\ j=1,2,\ldots,n,\ t=1,2,\ldots,p \qquad (34)$$

$$\overline{x}_{t}=\frac{1}{n}\sum_{j=1}^{n}x_{tj},\ s_{t}=\frac{1}{n}\sum_{j=1}^{n}\left(x_{tj}-\overline{x}_{t}\right)^{2},\quad t=1,2,\ldots,p \qquad (35)$$

Conditions for FCM-NM are

$$\begin{gathered}m\in[1,\infty);\ U=\left[\mu_{ij}\right]_{c\times n};\ \mu_{ij}\in\left[0,1\right],\ i=1,2,\ldots,c,\ j=1,2,\ldots,n\\ \sum_{i=1}^{c}\mu_{ij}=1,\ j=1,2,\ldots,n,\ 0<\sum_{j=1}^{n}\mu_{ij}<n,\ i=1,2,\ldots,c\end{gathered} \qquad (36)$$

$$d^{2}\left(\underline{z}_{j},\underline{a}_{i}\right)=\begin{cases}\left(\underline{z}_{j}-\underline{a}_{i}\right)'R^{-1}\left(\underline{z}_{j}-\underline{a}_{i}\right)-\ln\left|\Sigma^{-1}\right| & if\left(\underline{z}_{j}-\underline{a}_{i}\right)'R^{-1}\left(\underline{z}_{j}-\underline{a}_{i}\right)-\ln\left|R^{-1}\right|\geq 0\\ 0 & if\left(\underline{z}_{j}-\underline{a}_{i}\right)'R^{-1}\left(\underline{z}_{j}-\underline{a}_{i}\right)-\ln\left|R^{-1}\right|<0\end{cases} \qquad (37)$$

The steps of the FCM-NM are listed as follows
Step 1: Determining the number of cluster; c, m-value (let m=2), and the threshold $|R|$ as follows;

$$|R|=|\frac{1}{n}\sum_{j=1}^{n}\left(\underline{z}_{j}-\underline{a}_{t}\right)\left(\underline{z}_{j}-\underline{a}_{t}\right)'| \qquad (38)$$

Where $$0\leq|R|\leq 1, \qquad (39)$$

And $$\underline{a}_{t}=\frac{1}{n}\sum_{j=1}^{n}\underline{z}_{j} \qquad (40)$$

Randomly choose the initial membership

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

$$\sum_{1\leq i\leq c}u^{(0)}{}_{ij}=1,\ j=1,2,\ldots,n \qquad (41)$$

$$\underline{a}_{i}^{(0)}=\left[\sum_{j=1}^{n}\mu_{ij}^{(0)}\right]^{-1}\sum_{j=1}^{n}\mu_{ij}^{(0)}\underline{z}_{j},\quad i=1,2,\ldots,c,\quad k=1,2,.. \qquad (42)$$

$$R^{(0)}=\left[\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{(0)}\right]^{-1}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{(0)}\left(\underline{z}_{j}-\underline{a}_{i}^{(0)}\right)\left(\underline{z}_{j}-\underline{a}_{i}^{(0)}\right)' \qquad (43)$$

$$if\qquad\left|R^{(0)}\right|<|R|\quad then\ \ R^{(0)}=I \qquad (44)$$

Step 2: Find

$$\underline{a}_{i}^{(k)}=\left[\sum_{j=1}^{n}\left(\mu_{ij}^{(k-1)}\right)^{m}\right]^{-1}\sum_{j=1}^{n}\left(\mu_{ij}^{(k-1)}\right)^{m}\underline{z}_{j},\quad i=1,2,\ldots,c,\quad k=1,2,.. \qquad (45)$$

$$R^{(k)}=\left[\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{(k-1)}\right]^{-1}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{(k-1)}\left(\underline{z}_{j}-\underline{a}_{i}^{(k)}\right)\left(\underline{z}_{j}-\underline{a}_{i}^{(k)}\right)' \qquad (46)$$

$$if\qquad\left|R^{(k)}\right|<|R|\quad then\ \ R^{(k)}=I \qquad (47)$$

$$\mu_{ij}^{(k)}=\left[\sum_{s=1}^{c}\left[\frac{\left(\underline{z}_{j}-\underline{a}_{i}^{(k)}\right)'\left[R^{-1}\right]^{(k)}\left(\underline{z}_{j}-\underline{a}_{i}^{(k)}\right)-\ln\left|\left[R^{-1}\right]\right|^{(k)}}{\left(\underline{z}_{j}-\underline{a}_{s}^{(k)}\right)'\left[R^{-1}\right]^{(k)}\left(\underline{z}_{j}-\underline{a}_{s}^{(k)}\right)-\ln\left|\left[R^{-1}\right]\right|^{(k)}}\right]^{\frac{1}{m-1}}\right]^{-1} \qquad (48)$$

Step 3: Increment k; until

$$\sum_{i=1}^{c}\left\|\underline{a}_{i}^{(k-1)}-\underline{a}_{i}^{(k-2)}\right\|^{2}\geq\sum_{i=1}^{c}\left\|\underline{a}_{i}^{(k)}-\underline{a}_{i}^{(k-1)}\right\|^{2}\geq\ldots\geq\sum_{i=1}^{c}\left\|\underline{a}_{i}^{(k+9)}-\underline{a}_{i}^{(k+8)}\right\|^{2},$$

Step 4: Classification strategy;

If $\arg\max_{1\leq i\leq c}u_{ij}^{(k)}=t$ then x_{j} is assigned to cluster t.

Note that the threshold, $|R|$, of FCM-NM is a dynamic value rather than a constant, and the convergent process is different from all of before mentioned algorithms[16].

F. Clustering Accuracy

In [17], C. Ding, T. Li, and W. Ping, use the clustering accuracy, as follows,

$$A_{c}=\frac{1}{n}\max\sum_{C_{s},L_{t}}T\left(C_{s},L_{t}\right) \qquad (49)$$

where n is the number of objects in the data set, C_{s} is the s-th cluster and L_{t} is the t-th class, $T\left(C_{s},L_{t}\right)$ is the number of objects which belong to class t and are assigned to cluster s. Accuracy computes the maximum sum of $T\left(C_{s},L_{t}\right)$ for all pairs of clusters and these pairs have no overlaps. Accuracy, A_{c}, is the percentage of the points that were correctly recovered in a clustering result. Generally, the grater the accuracy values the better the cluster performance.

Threshold D In this paper, not only z-score normalizing for each feature in the objective function in the FCM-CM algorithm, but also replacing the threshold D where

$$D=\sum_{i=1}^{c}\sum_{j=1}^{n}\left[\mu_{ij}^{(0)}\right]^{m}\left[\left(\underline{x}_{j}-\underline{a}_{i}^{(0)}\right)'\left(\underline{x}_{j}-\underline{a}_{i}^{(0)}\right)\right]>0 \qquad (50)$$

III. Empirical Analysis

The data set 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 1.

TABLE I. The Details of The Used Datasets

Datasets	Attributes	Classes	Sample number
Iris	**4**	**3**	**150**
Wdbc	**30**	**2**	**569**

The performances of FCM, GK, GG, FCM-M, FCM-CM, FCM-SM, and FCM-NM 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].

TABLE II. The Accuracies of Five Algorithms

Algorithms	Iris	Wdbc
GK	**0.9000**	**0.7404**
GG	**0.7649**	**0.7767**
FCM-M	**0.9000**	**0.7978**
FCM-CM	**0. 9279**	**0.9172**
FCM-NM	**0. 9299**	**0.9183**

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, GK algorithm needs added constraint of fuzzy covariance matrix, GG algorithm can only be used for the data with multivariate Gaussian distribution. A Fuzzy C-Means algorithm based on Mahalanobis distance (FCM-M) was proposed to improve those limitations of above two algorithms, but it is not stable enough when some of its covariance matrices are not equal. An improved Fuzzy C-Means algorithm based on Normalized Mahalanobis distance (FCM-NM) is proposed. The experimental results of two real data sets consistently show that the performance of our proposed FCM-NM algorithm is better than those of the FCM algorithms. In this paper, each cluster of data can easily describe features of knowledge structures[18,19].

The Mean clustering Accuracies of 100 different initial value sets of GK, GG, FCM-M, FCM-CM, and FCM-NM for these two Datasets were shown in TABLE II. From this table, we can find that the performance of GG algorithm always worse than FCM-NM for above two datasets. Although the performance of GK algorithm is better than which of GG algorithm in Iris dataset, but the performance of the former is worse than which of the later in Wdbc dataset. The performances of our proposed three algorithms, FCM-M, FCM-CM, and FCM-NM are simultaneously better than which of GK and GG algorithm in two datasets. In other words, our proposed two algorithms, FCM-CM, and FCM-NM are better than GG algorithm and GK algorithm. Among our proposed two algorithms, the new algorithm, FCM-NM, has the best performance. In a word, FCM-NM algorithm is better than others.

IV. Conclusions

Clustering technique plays an important role in data analysis and interpretation. Fuzzy clustering is a branch in clustering analysis and it is widely used in the pattern recognition field. 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) pointed out that the Mahalanobis distance can not be used directly in clustering algorithm. Gustafson-Kessel (GK) clustering algorithm and Gath-Geva (GG) clustering algorithm 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. 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 Mahalanobis distance (FCM-M),was proposed, and then For improving the stability of the FCM-M clustering results, Replace all of the covariance matrices with the same common covariance matrix in the objective function in the FCM-M algorithm.

Proper clustering number will be decided in advance and one student will be randomly selected from each cluster to describe features of knowledge structures. The proper number of cluster is 3 as Iris.

In this paper, we use the best performance of clustering Algorithm FCM-CM 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 widely used in the pattern recognition field. Hence each cluster of data can easily describe features of knowledge structures. Manage the knowledge structures of Mathematics Concepts to construct the model of features in the pattern recognition completely[20,21,22].

The well-known FCM is based on Euclidean distance function[23,24], 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. three improved Fuzzy C-Means algorithm based on different Mahalanobis distance, called FCM-M, FCM-CM, and FCM-SM were proposed by our previous works. In this paper, a further improved Fuzzy C-Means algorithm based on a normalized Mahalanobis distance (FCM-NM) by taking a new convergent process is proposed.

The experimental results of two real data sets show that our proposed new algorithms have the best performance.

Acknowledgment

This paper is partially supported by Min-Hwei College of Health Care Management of Taiwan, The project's title is Construct Remedial Instruction's Interface for the Sample of Nursing Background to Learning Foundational Mathematics in Longitudinal Study (NSC 103).

References

[1] J. C. Bezdek, Pattern recognition with fuzzy objective function algorithms (Plenum press, 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, 10(3), 2008, 168-173.

[3] R. Krishnapuram and J. Kim, A note on the Gustafson-Kessel and adaptive fuzzy clustering algorithm, IEEE Transactions on Fuzzy Systems, 7(4), 1999, 453-461.

[4] D. E. Gustafson and W. C. Kessel, Fuzzy Clustering with a Fuzzy Covariance Matrix, Proc. IEEE Conf. Decision Contr. San Diego, CA, 1979, 761-766.

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

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

[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-means algorithm based on Mahalanobis distances and better initial values. Proceedings of the 10th Joint Conference& 12th International Conference on Fuzzy Theory & Technology. 1, 2007, 1398-1404.

[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, .7(7), 2007, 3939-3944.

[10] H.-C. Liu, J.-M. Yih, D.-B. Wu and S.-W. Liu, Fuzzy C-means algorithm based on "complete" Mahalanobis cistances. Proceedings of International conference on Machine Learning and Cybernetics, 7(6), 2008, 3569-3574.

[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. (in press).

[12] H.-C. Liu, J.-M. Yih, W.-C. Lin and D.-B. Wu, Fuzzy C-mans algorithm based on common Mahalanobis dstances. Journal of Multiple Valued Logic & Soft Computing, 15, 2009, 581-595.

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

[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/MLReposi tory.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.

[16] H.-C. Liu, B.-C. Jeng, J.-M. Yih, and Y.-K. Yu, Fuzzy C-means algorithm based on standard mahalanobis distances. Proceedings of the 2009 International Symposium on Information Processing (ISIP'09), ISBN 928-952-5726-02-2. 2009, 422-427.

[17] C. Ding, T. Li, W. Ping, On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing, Computational Statistics and Data Analysis 52, 2008, 3913-3927.

[18] H.-C. Liu, B.-C. Jeng, J.-M. Yih, and Y.-K. Yu, Fuzzy C-means algorithm based on standard mahalanobis distances. Proceedings of the 2009 International Symposium on Information Processing (ISIP'09), ISBN 928-952-5726-02-2.(2009). pp.422-427.

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

[20] Hasanzadeh R. P. R., Moradi M. H. and Sadeghi S. H. H., Fuzzy clustering to the detection of defects from nondestructive testing, 3rd International Conference: Sciences of Electronic Technologies of Information and Telecommun ication(2005). March 27-31, Tunisia.

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

[22] Jeng-Ming Yih, Yuan-Horng Lin, Hsiang-Chuan Liu (2008). FCM Algorithm Besed on Unsupervised Mahalanobis Distances with Better Initial Values and Separable Criterion. Proceedings of The 8th WSEAS International Conference on APPLIED COMPUTER SCIENCE (ACS'08), pp. 326-331, ISSN: 1790-5109/ ISBN: 978-960-474-028-4. [Venice, Italy, November 21-23, 2008].

[23] Hsiang-Chuan Liu, Jeng-Ming Yih, Der-Bang Wu, Shin-Wu Liu (2008). Fuzzy Possibility C-Mean Clustering Algorithms Based on Complete Mahalanobis Distances. 2008 International Conference on Wavelet Analysis and Pattern Recognition. pp. 50-55, ISBN: 978-1-4244-2239-5. [The Mira Hotel, Hong Kong, August 30-31, 2008].

[24] Jeng-Ming Yih, Yuan-Horng Lin, Hsiang-Chuan Liu (2008). FCM Algorithm Besed on Unsupervised Mahalanobis Distances with Better Initial Values and Separable Criterion. Proceedings of The 8th WSEAS International Conference on APPLIED COMPUTER SCIENCE (ACS'08), pp. 326-331, ISSN: 1790-5109/ ISBN: 978-960-474-028-4. EI 級論文 [Venice, Italy, November 21-23, 2008].