A Feature-Reduction Fuzzy Clustering Algorithm Based on Feature-Weighted Entropy

Miin-Shen Yang and Yessica Nataliani

Abstract—**Fuzzy clustering algorithms generally treat data points with feature components under equal importance. However, there are various data sets with irrelevant features involved in clustering process that may cause bad performance for fuzzy clustering algorithms. That is, different feature components should take different importance. In this paper, we present a novel method for improving fuzzy clustering algorithms that can automatically compute individual feature weight, and simultaneously reduce these irrelevant feature components. In fuzzy clustering, the fuzzy c-means (FCM) algorithm is the best known. We first consider the FCM objective function with feature-weighted entropy, and construct a learning schema for parameters, and then reduce these irrelevant feature components. We call it a feature-reduction FCM (FRFCM). During FRFCM processes, a new procedure for eliminating irrelevant feature(s) with small weight(s) is created for feature reduction. The computational complexity of FRFCM is also analyzed. Some numerical and real data sets are used to compare FRFCM with various feature-weighted FCM methods in the literature. Experimental results and comparisons actually demonstrate these good aspects of FRFCM with its effectiveness and usefulness in practice.**

Index Terms—**Clustering; Fuzzy clustering; Fuzzy c-means (FCM); Entropy; Feature weights; Feature reduction; Feature-reduction FCM (FRFCM).**

I. Introduction

Cluster analysis is a useful tool for data analysis. It is a method for finding groups within data with the most similarity in the same cluster and the most dissimilarity between different clusters. Since Zadeh [1] proposed fuzzy set and introduced the idea of partial memberships described by membership functions, it was successfully applied in clustering by Ruspini [2]. He first proposed fuzzy c-partitions as a fuzzy approach to clustering in the 1970s. In fuzzy clustering, the fuzzy c-means (FCM) algorithm proposed by Dunn [3] and Bezdek [4] is the most well-known and used method. Fuzzy clustering has been widely studied and applied in a variety of substantive areas [4-14].

In general, FCM treats all feature components of data to be equally important. However, in most cases of data sets, there exist some irrelevant features that always affect clustering results during FCM processes and may cause FCM produce incorrect clustering results. In this case, embedding feature reduction behavior in FCM with feature weights can take big advantage for improving FCM. It is known that feature weights are in the interval [0,1], and so the more influence a feature is, the greater its weight should be. On the other way, the more irrelevant a feature to a data set is, the less its weight should be.

Feature-weighted techniques had been used for clustering algorithms. In clustering, k-means and fuzzy c-means (FCM) algorithms are the best known methods. Some variants of feature-weighted k-means and FCM had been proposed in the literature, such as weighted k-means (WKM) [15], entropy-weighted k-means (EWKM) [16], sparse k-means [17], weighted FCM using feature-weight learning (WFCM) [18], and feature-weighted FCM, called simultaneous clustering and attribute discrimination (SCAD1 and SCAD2) [19]. Although these feature-weighted clustering algorithms may improve the performance of k-means or FCM, they were not to consider a feature-reduction schema, except both WKM [15] and sparse k-means [17] where they considered feature selection techniques with dependence of parameter selections. It is known that, if irrelevant features are included in clustering processes, then they may cause more computational time and yield incorrect clustering results, especially for high dimensional data sets. In this paper, we propose a novel algorithm called a feature-reduction FCM (FRFCM) algorithm that can automatically compute different feature weights by considering the FCM objective function with feature-weighted entropy. Moreover, we create a feature-reduction schema to eliminate these irrelevant features with small weights such that the computational time can be decreased with better clustering performance.

To evaluate the performance of FRFCM, we use synthetic and real data sets to compare FRFCM with five leading algorithms: WKM, EWKM, SCAD2, WFCM, and FCM. The computational complexity is also analyzed. Experimental results and comparisons actually show that FRFCM has good aspects with feature-reduction behaviors and producing better clustering results. The contributions of the paper can be summarized as follows.

- We propose a new schema to improve FCM by weighting feature components and eliminating those small weights. This schema can decrease the number of features with feature reduction and also decrease the computational time.
- We propose a learning procedure for estimating those parameters used in the FRFCM objective function, such that the FRFCM algorithm will be free of parameter selections.

The rest of this paper is organized as follows. In Section II, we briefly review some related works of feature-weighted k-means and FCM clustering algorithms. Section III presents the proposed FRFCM algorithm with the learning procedure for estimating the parameter value of the objective function. Explanations using examples are also given. Experiments and

Manuscript received July 4, 2016; revised December 14, 2016; accepted March 30, 2017. Date of publication ??? ??, 2017; date of current version ??? ??, 2017. This work was supported in part by the Ministry of Science and technology (MOST) of Taiwan under Grant MOST-105-2118-M-033-004-MY2. Miin-Shen Yang is with the Department of Applied Mathematics, Chung Yuan Christian University, Chung-Li 32023, Taiwan, where Miin-Shen Yang is the Corresponding Author, e-mail: msyang@math.cycu.edu.tw. Yessica Nataliani is with the Department of Applied Mathematics, Chung Yuan Christian University, Chung-Li 32023, Taiwan, and the Department of Information Systems, Satya Wacana Christian University, Salatiga 50711, Indonesia.

comparisons of the proposed method with other existing methods using artificial and real data sets are shown in Section IV. Finally, conclusions are stated in Section V.

II. RELATED WORKS

In this section, we review these clustering algorithms proposed in the literature that consider feature weights, especially for k-means and FCM. We first summarize most notations used in this paper as follows: n = number of data points; c = number of clusters; d = number of feature components; m = fuzziness index; $\mathbf{U}=[\mu_{ik}]_{n\times c}$ with μ_{ik} as a fuzzy membership of the i^{th} point in the k^{th} cluster; $\mathbf{V}=[v_{kj}]_{c\times d}$ with v_{kj} as the k^{th} cluster center with the j^{th} feature component; $\mathbf{W}=[w_j]_{1\times d}$ with w_j as a feature weight of the j^{th} feature; $\mathbf{W}_M=[w_{kj}]_{c\times d}$ with w_{kj} as a feature weight of the j^{th} feature in k^{th} cluster, and $1\le i\le n$, $1\le k\le c$, $1\le j\le d$.

Huang et al. [15] proposed weighting k-means (WKM) as an extension of k-means by adding the calculation of variable weights during iterative processes. The WKM objective function is

$$J(\mathbf{U},\mathbf{V},\mathbf{W}_M)=\sum_{k=1}^{c}\sum_{i=1}^{n}\sum_{j=1}^{d}\mu_{ik}w_{kj}^{\beta}(x_{ij}-v_{kj})^2 \qquad (1)$$

where $\beta<0$ or $\beta>1$ is a power parameter for feature weights. They also considered removing unimportant variables by choosing variables with small weights for heart disease and Australian credit card data sets to obtain better results. Jing et al. [16] considered subspace clustering that is especially useful for high dimensional sparse data by using a feature-weighted approach. They proposed entropy-weighted k-means (EWKM) by adding weight entropy term, such that it can simultaneously minimize the within cluster dispersion and maximize the negative weight entropy. Since feature weights represent the probability of a dimensional contributing to clustering results, it is used to determine subsets of important dimensions in each cluster. The EWKM objective function is

$$J(\mathbf{U},\mathbf{V},\mathbf{W}_M)=\sum_{k=1}^{c}\sum_{i=1}^{n}\sum_{j=1}^{d}\mu_{ik}w_{kj}(x_{ij}-v_{kj})^2+\gamma\sum_{k=1}^{c}\sum_{j=1}^{d}w_{kj}\log w_{kj} \qquad (2)$$

where $\gamma\ge 0$ is a parameter to control the size of feature weights in each cluster. They applied EWKM to high dimensional sparse data, such as text clustering and business transaction data, where many attributes have zero-dimension.

Witten and Tibshirani [17] proposed a new framework for feature selection in k-means clustering, called sparse k-means. They used L_1 (lasso) to make some of feature weights become zero when the weights are smaller than a tuning parameter. In sparse k-means, the algorithm first runs k-means with fixed weights. After that, k-means is applied with fixed cluster centers, and then updates feature weights until convergence. Wang et al. [18] proposed feature-weighted learning method to improve the performance of FCM, called WFCM. They proposed the objective function as

$$J(\mathbf{U},\mathbf{V},\mathbf{W})=\sum_{k=1}^{c}\sum_{i=1}^{n}\sum_{j=1}^{d}\mu_{ik}^{m}w_j^2(x_{ij}-v_{kj})^2 \qquad (3)$$

where w_j is a feature weight by minimizing an evaluation function that follows from the procedure of Yeung and Wang [20].

Frigui and Nasraoui [19] proposed feature-weighted FCM method, called simultaneous clustering and attribute discrimination (SCAD1). The weights are different for each feature in each cluster with the following objective function

$$J(\mathbf{U},\mathbf{V},\mathbf{W}_M)=\sum_{k=1}^{c}\sum_{i=1}^{n}\sum_{j=1}^{d}\mu_{ik}^{m}w_{kj}(x_{ij}-v_{kj})^2+\sum_{k=1}^{c}\sum_{i=1}^{n}\delta_k w_{kj}^2 \qquad (4)$$

where δ_k represents the importance of weights in each cluster. Moreover, they also proposed another type of SCAD, called SCAD2, by adding a discriminant exponent q for feature weights with the objective function

$$J(\mathbf{U},\mathbf{V},\mathbf{W}_M)=\sum_{k=1}^{c}\sum_{i=1}^{n}\sum_{j=1}^{d}\mu_{ik}^{m}w_{kj}^{q}(x_{ij}-v_{kj})^2 \qquad (5)$$

Frigui and Nasraoui [19] demonstrated that both SCAD1 and SCAD2 have similar behavior and clustering results. For their further study in [19], they applied SCAD2 for clustering data with unknown number of clusters.

Although there are variants of feature-weighted clustering for k-means or FCM, most of them are generally used to improve the original algorithm, except that WKM [15] and sparse k-means [17] also considered feature selection techniques, but not a feature reduction. In this paper, we create a feature-reduction schema by adding feature-weighted entropy in the weighted FCM objective function. The proposed schema is novel and different from most of feature selection algorithms where the new schema for feature reduction is embedded in clustering processes. The proposed algorithm will update memberships, cluster centers, and feature weights, iteratively, until it is optimized. At the first iterative, all features are used and then features with small weights will be eliminated during each iterative with updating feature weights. Therefore, the number of features used in the next step during clustering processes will be less than the previous step until its convergence. We next present our proposed clustering algorithm.

III. THE FEATURE-REDUCTION FUZZY CLUSTERING ALGORITHM

Before we go on the proposed feature-reduction fuzzy clustering algorithm, we give a brief review of the fuzzy c-means (FCM) clustering algorithm. We also provide an example to demonstrate the behavior of FCM when all features are used during clustering processes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2017.2692203, IEEE Transactions on Fuzzy Systems

III-1 The FCM Clustering Algorithm

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a data set in a *d*-dimensional space $\mathbb{R}^d$. The FCM objective function is defined as follows [4,5]

$$J(\mathbf{U},\mathbf{V}) = \sum_{k=1}^{c}\sum_{i=1}^{n}\sum_{j=1}^{d}\mu_{ik}^{m}(x_{ij}-v_{kj})^2 \qquad (6)$$

subject to

$$\sum\nolimits_{k=1}^{c}\mu_{ik} = 1 \text{ with } \mu_{ik} \in [0,1],\ 1 \le i \le n,\ 1 \le k \le c \qquad (7)$$

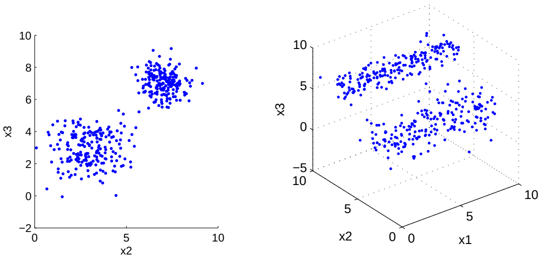
where the weighting exponent $1 < m < +\infty$ presents the degree of fuzziness. The FCM algorithm is iterated through the necessary conditions for minimizing $J(\mathbf{U},\mathbf{V})$, with the updating equations for membership function and cluster centers as follows: For $1 \le i \le n$ and $1 \le k \le c$,

$$\mu_{ik} = \left(\sum_{j=1}^{d}(x_{ij}-v_{kj})^2\right)^{-1/m-1} \Bigg/ \sum_{t=1}^{c}\left(\sum_{j=1}^{d}(x_{ij}-v_{tj})^2\right)^{-1/m-1} \qquad (8)$$

$$v_{kj} = \sum_{i=1}^{n}\mu_{ik}^{m}x_{ij} \Bigg/ \sum_{i=1}^{n}\mu_{ik}^{m},\ 1 \le k \le c \text{ and } 1 \le j \le d \qquad (9)$$

As we know, FCM treats all feature weights equally, no matter what relevant or irrelevant features are presented and it may induce incorrect clustering results as demonstrated in Example 1. For measuring clustering performance, accuracy rate (AR) with $\text{AR} = \frac{1}{n}\sum\nolimits_{k=1}^{c} n(c_k)$ is used, where $n(c_k)$ is the number of data points that obtained correct clustering for the cluster *k*, and *n* is the total number of data points. The larger AR is, the better clustering performance is.

Example 1 We use a two-cluster data set generated from a 2-component Gaussian mixture distribution $\sum\nolimits_{k=1}^{2}\alpha_k N(u_k,\Sigma_k)$ with a sample size 400 and the parameters $\alpha_1 = \alpha_2 = 0.5$, $u_1 = (3 \quad 3)^T$, $u_2 = (7 \quad 7)^T$, $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$, which is modified from Huang et al. [15] and Frigui and Nasraoui [19]. These 2-component points, namely x_2 and x_3, as shown in Fig. 1(a), will be stretched into the other component, namely x_1, by generating from uniform distribution through those 2-component points, as shown in Fig. 1(b). Furthermore, for this example, we add one more feature to the data set, namely x_4, such that the data points have four feature components. We set those features generated from the Gaussian mixture as the 2nd and 3rd features and generate the 1st and 4th features from uniform distributions over intervals [0, 10] and [0, 12], respectively. These different generation assignments are listed in Table 1. Uniform distributions make the 2-dimensional original data set into 4-dimensional data, but still keep the data points in the same clusters. Thus, we can say that two features with uniform distributions are two unimportant features. This is because the 4-dimensional data set can be projected into 2-dimensional original data set. Table 2 shows scatter plots for different combinations of any two features in the data set. It is shown that the 2nd and 3rd features present the original two-cluster data set. If we apply the FCM algorithm with all features, then the two clusters cannot be well separated (AR = 0.60), as shown in Fig. 2(a). It is because the 1st and 4th features distract the data grouping. However, the FCM algorithm with the 2nd and 3rd features produces a good clustering result with AR = 1.00, as shown in Fig. 2(b). We find that FCM algorithm is exactly influenced by irrelevant feature components, because it treats all feature components on the data set equally.

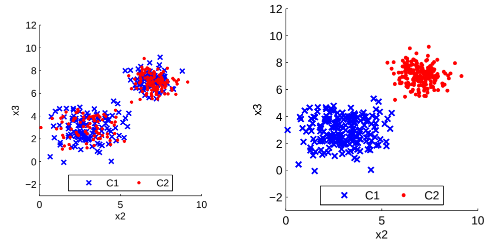


(a) (b)

Fig. 1 (a) 2-component points (x_2, x_3) generated from a Gaussian mixture model; (b) 3-component points (x_1, x_2, x_3), where x_2 and x_3 are obtained from (a) and x_1 is generated from uniform distribution.

Table 1 Different generation assignments

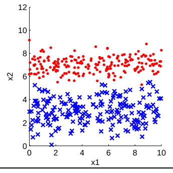
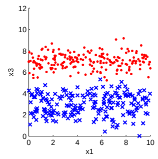
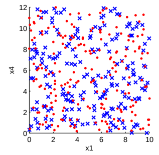
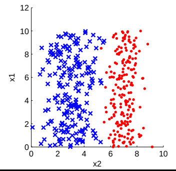
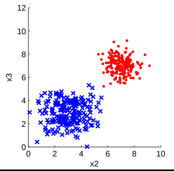
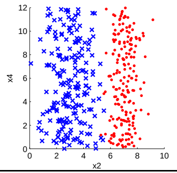
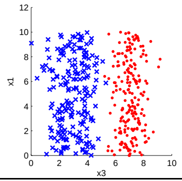
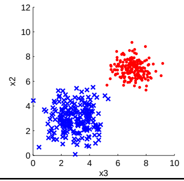
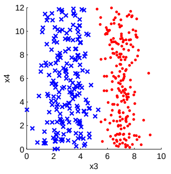
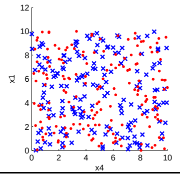
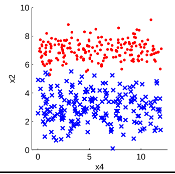
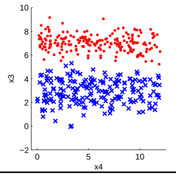
Clusters	1st feature	2nd and 3rd features	4th feature
C1	200 data generated from the uniform distribution over interval [0, 10]	200 data generated from Gaussian mixture distribution with $u = (3 \quad 3)^T$ and $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	200 data generated from the uniform distribution over interval [0, 12]
C2	200 data generated from the uniform distribution over interval [0, 10]	200 data generated from Gaussian mixture distribution with $u = (7 \quad 7)^T$ and $\Sigma = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$	200 data generated from the uniform distribution over interval [0, 12]



(a) (b)

Fig. 2 Clustering results of: (a) FCM with four features (x_1, x_2, x_3, x_4) (b) FCM with two features (x_2, x_3).

Table 2 Scatter plots for different combinations of any two features, where x_1 and x_4 are generated from uniform distribution, while x_2 and x_3 are generated from a 2-component Gaussian mixture distribution

y-axis x-axis	1st feature (x_1)	2nd feature (x_2)	3rd feature (x_3)	4th feature (x_4)
1st feature (x_1)	-			
2nd feature (x_2)		-		
3rd feature (x_3)			-	
4th feature (x_4)				-

III-2 The Proposed Feature-Reduction FCM Clustering Algorithm

Since data sets may include some irrelevant feature components, feature selection becomes important in clustering algorithms, especially for high dimensional data. We propose a new schema to improve FCM by considering feature reduction with feature-weighted entropy, which is called a feature-reduction FCM (FRFCM). In this schema, each feature has its own weight that will be updated at each iterative. Afterwards, feature(s) with small weight(s) will be eliminated after some learning procedures. Let $\mathbf{X}=\{\mathbf{x}_1,\dots,\mathbf{x}_n\}$ be a d-dimensional data set and $\mathbf{W}=[w_j]_{1\times d}$ be with w_j as a feature weight of the j^{th} feature. The FRFCM objective function is considered as follows:

$$J(\mathbf{U},\mathbf{V},\mathbf{W})=\sum_{k=1}^{c}\sum_{i=1}^{n}\sum_{j=1}^{d}\mu_{ik}^{m}\delta_j w_j(x_{ij}-v_{kj})^2+\frac{n}{c}\sum_{j=1}^{d}\left(w_j\log\delta_j w_j\right) \tag{10}$$

subject to

$$\sum\nolimits_{k=1}^{c}\mu_{ik}=1, 0\le\mu_{ik}\le 1,\ \sum\nolimits_{j=1}^{d}w_j=1,\ 0\le w_j\le 1 \tag{11}$$

Note that δ_j is used to control feature weights. The learning procedure for δ_j will be demonstrated and explained later.

The FRFCM algorithm can be solved using three minimization steps. The first step is to fix $\mathbf{V}=\hat{\mathbf{V}}$ and $\mathbf{W}=\hat{\mathbf{W}}$, and then minimize $J(\mathbf{U},\hat{\mathbf{V}},\hat{\mathbf{W}})$ with respect to $\mathbf{U}$. Consider the Lagrangian function with

$$\tilde{J}(\mathbf{U},\mathbf{V},\mathbf{W})=\sum_{k=1}^{c}\sum_{i=1}^{n}\sum_{j=1}^{d}\mu_{ik}^{m}\delta_j w_j(x_{ij}-v_{kj})^2+\frac{n}{c}\sum_{j=1}^{d}\left(w_j\log\delta_j w_j\right)+ \lambda_1\left(\sum\nolimits_{k=1}^{c}\mu_{ik}-1\right)+\lambda_2\left(\sum\nolimits_{j=1}^{d}w_j-1\right) \tag{12}$$

Taking the partial derivative of the Lagrangian from (12) with respect to μ_{ik} and setting them to be zero, we have that

$$\frac{\partial\tilde{J}}{\partial\mu_{ik}}=\sum_{j=1}^{d}m\mu_{ik}^{m-1}\delta_j w_j(x_{ij}-v_{kj})^2+\lambda_1=0 \tag{13}$$

From (13), the updated equation for μ_{ik} can be obtained as follows

$$\mu_{ik}=\left(\sum_{j=1}^{d}\delta_j w_j(x_{ij}-v_{kj})^2\right)^{-1/m-1}\Bigg/\sum_{t=1}^{c}\left(\sum_{j=1}^{d}\delta_j w_j(x_{ij}-v_{tj})^2\right)^{-1/m-1} \tag{14}$$

The second step is to fix $\mathbf{U}=\hat{\mathbf{U}}$ and $\mathbf{W}=\hat{\mathbf{W}}$, and then minimize $J(\hat{\mathbf{U}},\mathbf{V},\hat{\mathbf{W}})$ from (12) with respect to $\mathbf{V}$. We have

$$\frac{\partial\tilde{J}}{\partial v_{kj}}=-2\sum_{i=1}^{n}\mu_{ik}^{m}\delta_j w_j(x_{ij}-v_{kj})=0 \tag{15}$$

From (15), the updated equation for v_{kj} can be obtained using

$$v_{kj}=\sum_{i=1}^{n}\mu_{ik}^{m}x_{ij}\Bigg/\sum_{i=1}^{n}\mu_{ik}^{m} \tag{16}$$

The third step is to fix $\mathbf{U}=\hat{\mathbf{U}}$ and $\mathbf{V}=\hat{\mathbf{V}}$, and then minimize $J(\hat{\mathbf{U}},\hat{\mathbf{V}},\mathbf{W})$ with respect to $\mathbf{W}$. From (12) we have

$$\frac{\partial\tilde{J}}{\partial w_j}=\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^{m}\delta_j(x_{ij}-v_{kj})^2+\frac{n}{c}\left(\log\delta_j w_j+1\right)+\lambda_2=0 \tag{17}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2017.2692203, IEEE Transactions on Fuzzy Systems

From (17), the updated equation for w_j can be obtained as follows

$$w_j = \frac{\dfrac{1}{\delta_j}\exp\left(\dfrac{-c\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^{m}\delta_j(x_{ij}-v_{kj})^2}{n}\right)}{\sum_{p=1}^{d}\dfrac{1}{\delta_p}\exp\left(\dfrac{-c\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^{m}\delta_p(x_{ip}-v_{kp})^2}{n}\right)} \qquad (18)$$

Furthermore, in order to retain the constraint $\sum_{j=1}^{d} w_j = 1$, we adjust w_j by

$$w_j{}' = w_j \Big/ \sum_{p=1}^{d(new)} w_p \qquad (19)$$

Let us recall the second term, $\frac{n}{c}\sum_{j=1}^{d} w_j \log\delta_j w_j$, in the FRFCM objective function (10). We explain why we use the constant n/c to handle effects of the term $\sum_{j=1}^{d} w_j \log\delta_j w_j$, which is also used in Eq. (18) as an updating equation for feature weights. As seen in Eq. (18), if the term $\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^{m}\delta_j(x_{ij}-v_{kj})^2$ is too large, then the numerator $\exp\left(-\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^{m}\delta_j(x_{ij}-v_{kj})^2\right)$ will become too small as close to a zero value. We need to avoid this case for preventing too many feature weights to be discarded during this updating step. On the other hand, if the term $\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^{m}\delta_j(x_{ij}-v_{kj})^2$ is too small, then the numerator $\exp\left(-\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^{m}\delta_j(x_{ij}-v_{kj})^2\right)$ will be closed to one so that it is difficult for the feature(s) to be discarded during the updating step. We also need to avoid this case. In this sense, we need to put a suitable constant to control it. In the FRFCM clustering algorithm, one goal is to cluster a data set (with n data points) into c clusters. The numbers n and c are the two commonly given constants. We can use the constant n/c to control the term $\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^{m}\delta_j(x_{ij}-v_{kj})^2$.

Another problem is how to estimate the value of δ_j in Eqs. (14) and (18). There are two terms in the FRFCM objective function. The first term is the sum of feature-weighted distances between data points and cluster centers, which is minimized when the distance between points and centers is small. The second term is a variant of the feature-weight entropy $\sum_{j=1}^{d} w_j \log w_j$, i.e., $\sum_{j=1}^{d} w_j \log\delta_j w_j$. Because the δ_j in $\sum_{k=1}^{c}\sum_{i=1}^{n}\sum_{j=1}^{d}\mu_{ik}^{m}\delta_j w_j(x_{ij}-v_{kj})^2$ and $\sum_{j=1}^{d} w_j \log\delta_j w_j$ are used for controlling the variants of feature weights, the choice of δ_j is important. To estimate the value of δ_j, we next propose a learning procedure.

In probability theory and statistics, standard deviation and variance are used to measure the dispersion of data. Another measurement as an index of dispersion is a well-known variance-to-mean-ratio (VMR), defined as $\text{VMR} = \sigma^2/\mu$ [21]. VMR can be used to observe a dispersed or clustered data set. Smaller dispersion means the data set would be closer to the cluster center, while larger dispersion means the data set is far from the cluster center. Because we need to retain features which have small dispersion and then discard those which have large dispersion, we consider the reciprocal of VMR, i.e., mean-to-variance ratio (MVR), which has been defined in [22], to be used in our algorithm.

Table 3 Computation of mean and variance of each feature in the data set of Example 1

	x_1	x_2	x_3	x_4
Mean of the data set	4.881	5.007	4.992	5.655
Variance of the data set	8.573	4.885	4.944	11.883
Variance-to-mean ratio (VMR)	1.756	0.976	0.990	2.101
Mean-to-variance ratio (MVR)	0.569	1.025	1.010	0.476

From Example 1, it is clear that the 1st and 4th features are the unimportant features. Thus, we want to drive those features to have small weights, so that they can be discarded during the clustering process. We borrow the idea of MVR in our algorithm. Table 3 presents the mean, variance, VMR, and MVR of the data set from Example 1. As shown in Table 3, VMR cannot produce small weights for the 1st and 4th features compared to the 2nd and 3rd features, while MVR can produce small weights. We find that the term $\left(\text{mean}(x)/\text{var}(x)\right)_j$ for the feature j in the FRFCM algorithm can actually handle the dispersion between clusters in the data set. Therefore, we use $\left(\text{mean}(x)/\text{var}(x)\right)_j$ to estimate δ_j. That is, we consider the estimate for δ_j as follows:

$$\delta_j = \left(\frac{\text{mean}(x)}{\text{var}(x)}\right)_j \qquad (20)$$

To create a feature-reduction schema in the FRFCM algorithm, we need to select these unimportant feature(s) (i.e., small weights) during clustering processes. In our construction, we use a threshold to determine which feature(s) will be selected and then discarded. We know that the data set has n data points in which each data point has d feature components with the constraint $\sum_{j=1}^{d} w_j = 1$ of feature weights. If d is large, then the threshold for feature reduction is intuitively chosen as $1/d$. However, we expect our feature reduction algorithm to be fitted for most data sets, even for small d. In this sense, the data number n should be considered as another factor. We know that $1/d = 1/\sqrt{d^2} = 1/\sqrt{dd}$. For a balance between small and large d, we replace one d with the data number n so that it becomes $1/\sqrt{nd}$. Therefore, we consider $1/\sqrt{nd}$ as a suitable threshold for discarding these unimportant feature(s) in the FRFCM clustering algorithm. Thus, the proposed FRFCM algorithm can be summarized as follows.

FRFCM Algorithm

Fix $\varepsilon > 0$. Give cluster number c, randomly initialize cluster center $\mathbf{V}^{(0)}$, randomly initialize feature weight $\mathbf{W}^{(0)}$ (user may define $\mathbf{W}^{(0)} = [w_j]_{1\times d}, j = 1,\ldots,d, w_j = 1/d$), and set $t = 0$.

```
Step 1: Calculate $\delta_j$ using data points $\mathbf{X}$ by Eq. (20).
Step 2: Compute membership function $\mathbf{U}^{(t)}$ using $\delta_j$, $\mathbf{V}^{(t-1)}$,
        and $\mathbf{W}^{(t-1)}$ by Eq. (14).
Step 3: Update cluster center $\mathbf{V}^{(t)}$ using $\mathbf{U}^{(t)}$ by Eq. (16).
Step 4: Update $\mathbf{W}^{(t)}$ using $\delta_j$, $\mathbf{U}^{(t)}$ and $\mathbf{V}^{(t)}$ by Eq. (18).
Step 5: Discard total $d_r$ number of these $j$ feature
        components for $\mathbf{W}^{(t)}$, if $\mathbf{W}^{(t)} \le 1/\sqrt{nd}$, and set
        $d^{(new)} = d - d_r$.
Step 6: Adjust $\mathbf{W}^{(t)}$ by Eq. (19).
Step 7: If $\left\| \|\mathbf{W}^{(t)}\| - \|\mathbf{W}^{(t-1)}\| \right\| < \varepsilon$, then Stop;
        Else set $t = t+1$, $d = d^{(new)}$ and go back to Step 1.
```

Example 1 (cont.) We continue Example 1 by implementing the FRFCM algorithm for the data set with equal feature-weights as the initialization $\mathbf{W}^{(0)}$. At the first iteration, we find that the weights of the 1st and 4th features become very small. It is good because the 1st and 4th features are originally unimportant features. After two iterations, FRFCM clearly demonstrates the 2nd and 3rd features as important features. This feature reduction behavior is shown in Table 4, while the clustering result with AR = 1.00 is shown in Fig. 3. We demonstrate the performance of the FRFCM algorithm. It is well known that the final clustering results of FCM depend on initializations of cluster centers. In FRFCM, besides initial cluster centers, initial weights may also influence clustering results. Thus, we start by using different initial cluster centers with fixed feature weights (defined by $\mathbf{W} = [0.250 \;\; 0.250 \;\; 0.250 \;\; 0.250]$). The final feature weights with ARs are shown in Table 5. From these results, it is shown that the proposed FRFCM is quite robust to different initializations of cluster centers with good clustering results (average of AR = 0.999) and gets almost the same final cluster centers. We also apply FRFCM with different initial feature weights and fixed cluster centers. The final feature weights with their ARs are listed in Table 6. As presented, the final feature weights and ARs from 10 different initial feature weights are almost the same with average AR = 0.998. Thus, the FRFCM is also robust to different initialization of feature weights.

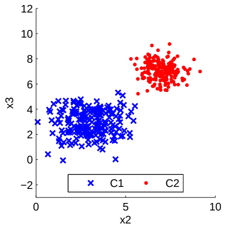


Fig. 3 Clustering result of Example 1 using FRFCM.

Table 4 Feature reduction behavior from 4 features to 2 features by FRFCM for Example 1 with initial $\mathbf{W}^{(0)} = [0.25 \;\; 0.25 \;\; 0.25 \;\; 0.25]$

	Feature weights			
	x_1	x_2	x_3	x_4
Initialization	0.250	0.250	0.250	0.250
Iteration 1	0.028	0.438	0.520	0.014
Iteration 2	0.001	0.478	0.522	-
Iteration 3	-	0.484	0.516	-
Iteration 4	-	0.485	0.515	-
Iteration 5	-	0.486	0.515	-
Iteration 6	-	0.486	0.514	-
Iteration 7	-	0.486	0.514	-

Table 5 Final cluster centers and ARs with different initial cluster centers using FRFCM

	Initial cluster centers		Final cluster centers		AR
	1st cluster	2nd cluster	1st cluster	2nd cluster	
1	(3.378, 3.549, 2.929,3.850)	(3.464, 6.402, 7.560, 4.523)	(2.936, 2.913)	(7.001, 7.010)	1.000
2	(6.897, 2.190, 4.765, 1.107)	(9.419, 7.168, 7.737, 7.206)	(2.943, 2.927)	(6.996, 7.002)	1.000
3	(9.745, 7.223, 7.655, 4.516)	(3.934, 2.069, 2.964, 9.459)	(2.939, 2.919)	(6.999, 7.007)	1.000
4	(2.175, 2.437, 2.500, 11.956)	(7.704, 7.770, 6.276, 2.304)	(2.957, 2.938)	(6.982, 6.988)	0.995
5	(8.363, 8.279, 6.371, 9.356)	(8.058, 2.322, 3.871, 2.451)	(2.943, 2.937)	(6.998, 7.003)	1.000
6	(8.644, 0.888, 3.896, 9.399)	(10.050, 7.402, 7.028, 1.376)	(2.944, 2.927)	(6.995, 7.002)	1.000
7	(6.517, 5.787, 5.330, 11.980)	(1.999, 2.629, 3.552, 10.640)	(2.937, 2.918)	(6.998, 7.006)	1.000
8	(1.488, 7.693, 7.309, 6.102)	(3.213, 2.776, 3.963, 2.317)	(2.936, 2.917)	(7.001, 7.008)	1.000
9	(3.159, 4.145, 4.382, 3.301)	(6.483, 7.981, 6.984, 3.368)	(2.943, 2.923)	(6.999, 7.006)	1.000
10	(4.107, 2.469, 4.681, 9.797)	(3.931, 7.284, 6.606, 8.213)	(2.929, 2.916)	(6.999, 7.005)	0.998

Table 6 Final feature weights and ARs with different initial feature weights using FRFCM

	Initial features weights				Final features weights				AR
	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	
1	0.250	0.250	0.250	0.250	-	0.484	0.516	-	1.000

2	0.290	0.500	0.001	0.210	-	0.489	0.511	-	0.998
3	0.301	0.018	0.380	0.301	-	0.482	0.518	-	1.000
4	0.267	0.344	0.141	0.248	-	0.485	0.515	-	1.000
5	0.302	0.171	0.304	0.223	-	0.483	0.517	-	1.000
6	0.100	0.393	0.093	0.414	-	0.486	0.514	-	1.000
7	0.428	0.159	0.393	0.020	-	0.472	0.528	-	0.998
8	0.038	0.387	0.217	0.359	-	0.482	0.518	-	0.995
9	0.269	0.299	0.268	0.164	-	0.478	0.572	-	0.995
10	0.009	0.440	0434	0.117	-	0.476	0.524	-	0.998

III-3 Convergence Theorems for the FRFCM Clustering Algorithm

From Table 4, since the 1st and 4th features had been reduced after two iterations, the FRFCM objective function is almost minimized since then. This is demonstrated in Fig. 4. We run the first step for the FRFCM algorithm to update the cluster centers and membership matrix. After that, new feature weights are computed and features with small weights are discarded during clustering processes. After adjusting points and cluster centers with new feature components, the first term is restarted again using the new feature weights, cluster centers, and membership matrix. This process is continued until the objective function is minimized. Each point in Fig. 4 represents the values of the objective function at its corresponding iteration of the FRFCM algorithm, from initial feature weights until the final feature weights are obtained. The FRFCM converges after seven iterations.

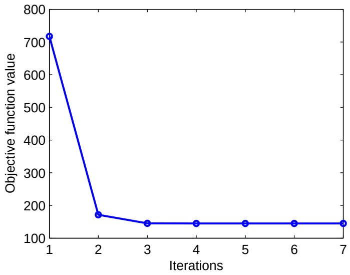


Fig. 4 Convergence plot of FRFCM

We next provide convergence theorems for the FRFCM clustering algorithm that a FRFCM convergent subsequence can tend to optimal solutions. Zangwill's convergence theorem [23] and bordered Hessian matrix [24] will be applied to our convergence proving. We mention that this way had been used in Yang and Tian [25]. Originally, Zangwill defined a point-to-set map with $T:V\to P(V)$, where $P(V)$ represents the power set of V and a closed point-to-set map must be defined. However, the FRFCM algorithm here is a point-to-point map and the "closed" property is exactly "continuity" for the case of the point-to-point map. Thus, the Zangwill's convergence theorem is given as follows.

Zangwill's Convergence Theorem [23]: Let the point-to-point map $T:V\to P(V)$ generate a sequence $\{z_k\}_{k=0}^{\infty}$ by $z_{k+1}=T(z_k)$. Let a solution set $\Omega\in V$ be given and suppose that:

1. There is a continuous function $Z:V\to R$ such that, if $z\notin\Omega$, $Z(T(z))<Z(z)$, and if $z\notin\Omega$, $Z(T(z))\le Z(z)$.
2. The map T is continuous on $V\setminus\Omega$.
3. All points z_k are contained in a compact set $S\subseteq V$.

Then the limit of any convergent subsequence shall be in the solution set Ω and $Z(z_k)$ will monotonically converge to $Z(z)$ for some $z\in\Omega$.

Set $M_{fcn}=\left\{\mathbf{U}=[\mu_{ik}]_{n\times c}\middle|\sum_{k=1}^{c}\mu_{ik}=1,\mu_{ik}\ge 0\right\}$ and $M_w=\left\{\mathbf{W}=[w_j]_{d\times 1}\middle|\sum_{j=1}^{d}w_j=1,w_j\ge 0\right\}$, and $\mathbf{V}=(\mathbf{v}_1,\dots,\mathbf{v}_c)$.

Let Ω_{FRFCM} be a solution set for the FRFCM algorithm, defined as

$$\Omega_{FRFCM}=\left\{(\mathbf{U}^*,\mathbf{V}^*,\mathbf{W}^*)\middle|\begin{array}{l}\forall\mathbf{U}\in M_{fcn},\mathbf{U}\ne\mathbf{U}^*,J(\mathbf{U}^*,\mathbf{V}^*,\mathbf{W}^*)<J(\mathbf{U},\mathbf{V}^*,\mathbf{W}^*);\\ \forall\mathbf{V}\ne\mathbf{V}^*,J(\mathbf{U}^*,\mathbf{V}^*,\mathbf{W}^*)<J(\mathbf{U}^*,\mathbf{V},\mathbf{W}^*);\\ \forall\mathbf{W}\in M_w,\mathbf{W}\ne\mathbf{W}^*,J(\mathbf{U}^*,\mathbf{V}^*,\mathbf{W}^*)<J(\mathbf{U}^*,\mathbf{V}^*,\mathbf{W});\\ \text{Discard } d_r\ (0\le d_r<d) \text{ number of these } j \text{ feature}\\ \text{components for } \mathbf{W}^* \text{ if } w_j^*<1\big/\sqrt{nd}\,;\ d^{(new)}=d-d_r;\\ w_{j'}^*=w_{j'}^*\big/\sum_{p'=1}^{d^{(new)}}w_{p'}^*;\ d=d^{(new)}\end{array}\right\}$$

$$\mathbf{U}^*=[\mu_{ik}^*]_{n\times c},\ \mu_{ik}^*=\left(\sum_{j=1}^{d}\delta_j w_j(x_{ij}-v_{kj})^2\right)^{-1/m-1}\Bigg/\sum_{t=1}^{c}\left(\sum_{j=1}^{d}\delta_j w_j(x_{ij}-v_{tj})^2\right)^{-1/m-1},$$

$$\mathbf{V}^*=(\mathbf{v}_1^*,\dots,\mathbf{v}_c^*),\qquad v_{kj}^*=\sum_{i=1}^{n}\mu_{ik}^m x_{ij}\Bigg/\sum_{i=1}^{n}\mu_{ik}^m,\qquad \mathbf{W}^*=[w_j^*]_{d\times 1},$$

$$w_j^*=\frac{1}{\delta_j}\exp\left(\frac{-c\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^m\delta_j(x_{ij}-v_{kj})^2}{n}\right)\Bigg/\sum_{p=1}^{d}\frac{1}{\delta_p}\exp\left(\frac{-c\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^m\delta_p(x_{ip}-v_{kp})^2}{n}\right).$$

Let $E:(\mathbb{R}^d)^c\times M_w\to M_{fcn}$ with $E(\mathbf{V},\mathbf{W})=\mathbf{U}=[\mu_{ik}]_{n\times c}$, where μ_{ik} is calculated by

$$\mu_{ik}=\left(\sum_{j=1}^{d}\delta_j w_j(x_{ij}-v_{kj})^2\right)^{-1/m-1}\Bigg/\sum_{t=1}^{c}\left(\sum_{j=1}^{d}\delta_j w_j(x_{ij}-v_{tj})^2\right)^{-1/m-1}.$$

Let $F:M_{fcn}\to(\mathbb{R}^d)^c$, $F(\mathbf{U})=\mathbf{V}=(\mathbf{v}_1,\dots,\mathbf{v}_c)$, where v_{kj} is calculated by $v_{kj}=\sum_{i=1}^{n}\mu_{ik}^m x_{ij}\Big/\sum_{i=1}^{n}\mu_{ik}^m$. Let $G:M_{fcn}\times(\mathbb{R}^d)^c\to M_w$, $G(\mathbf{U},\mathbf{V})=\mathbf{W}=[w_j]_{d\times 1}$, where

w_j is calculated by Eq. (18). The FRFCM operator can be defined as follows.

Definition 1 The FRFCM operator $T : M_{fcn} \times (\mathbb{R}^d)^c \times M_w \to M_{fcn} \times (\mathbb{R}^d)^c \times M_w$ is defined by $T = A_2 \circ A_1$ where $A_1 : M_{fcn} \times (\mathbb{R}^d)^c \times M_w \to M_{fcn} \times (\mathbb{R}^d)^c$ with $A_1(\mathbf{U},\mathbf{V},\mathbf{W}) = E(\mathbf{V},\mathbf{W})$ and $A_2 : M_{fcn} \to M_{fcn} \times (\mathbb{R}^d)^c \times M_w$ with $A_2(\mathbf{U}) = \left(\mathbf{U}, F(\mathbf{U}), G(\mathbf{U}, F(\mathbf{U}))\right)$. Thus, we have

$$T(\mathbf{U},\mathbf{V},\mathbf{W}) = (A_2 \circ A_1)(\mathbf{U},\mathbf{V},\mathbf{W}) = A_2\left(A_1(\mathbf{U},\mathbf{V},\mathbf{W})\right) = A_2\left(E(\mathbf{V},\mathbf{W})\right)$$
$$= \left(E(\mathbf{V},\mathbf{W}), F\left(E(\mathbf{V},\mathbf{W})\right), G\left(E(\mathbf{V},\mathbf{W}), F\left(E(\mathbf{V},\mathbf{W})\right)\right)\right) = (\mathbf{U}^*,\mathbf{V}^*,\mathbf{W}^*)$$

where $\mathbf{U}^* = E(\mathbf{V},\mathbf{W})$, $\mathbf{V}^* = F\left(E(\mathbf{V},\mathbf{W})\right) = F(\mathbf{U}^*)$, and $\mathbf{W}^* = G\left(E(\mathbf{V},\mathbf{W}), F\left(E(\mathbf{V},\mathbf{W})\right)\right) = G(\mathbf{U}^*,\mathbf{V}^*)$.

In general, the sufficient and necessary condition for a strict minimizer of an objective function is to analyze the Jacobian matrix and the Hessian matrix. However, if some constraints are considered, Lagrange's multipliers in addition to a bordered Hessian matrix must be assessed as follows.

Theorem 1 (Lagrange's theorem [24]). Let functions $f : D_f \to R$, $D_f \subseteq \mathbb{R}^n$, and $g_i : D_{g_i} \to R$, $D_{g_i} \subseteq \mathbb{R}^n$, $i = 1,\ldots,t$, $t < n$, be continuously partially differentiable and let $x^0 = (x_1^0,\ldots,x_n^0) \in D_f$ be a local extreme point of the function f subject to the constraints $g_i(x_1,\ldots,x_n) = 0$, $i = 1,\ldots,t$. Let $L(x;\lambda) = f(x_1,\ldots,x_n) + \sum_{i=1}^{t} \lambda_i g_i(x_1,\ldots,x_n)$ and

$$|J| = \begin{vmatrix} \frac{\partial g_1(x)}{\partial x_1} & \cdots & \frac{\partial g_1(x)}{\partial x_t} \\ \vdots & & \vdots \\ \frac{\partial g_t(x)}{\partial x_1} & \cdots & \frac{\partial g_t(x)}{\partial x_t} \end{vmatrix} \neq 0$$

at the point x^0. Then, we have that the gradient of $L(x;\lambda)$ at the point (x^0,λ^0) is 0, i.e., $\nabla L(x^0,\lambda^0) = 0$.

Theorem 2 (local sufficient conditions [24]). Let functions $f : D_f \to R$, $D_f \subseteq \mathbb{R}^n$, and $g_i : D_{g_i} \to R$, $D_{g_i} \subseteq \mathbb{R}^n$, $i = 1,\ldots,t$, $t < n$, be twice continuously partially differentiable and let $(x^0;\lambda^0)$ with $x^0 \in D_f$ be a solution of the system $\nabla L(x^0;\lambda^0) = 0$. Let

$$H_L(x,\lambda) = \begin{pmatrix} 0 & \cdots & 0 & \frac{\partial^2 L}{\partial\lambda_1 \partial x_1} & \cdots & \frac{\partial^2 L}{\partial\lambda_1 \partial x_n} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & \frac{\partial^2 L}{\partial\lambda_t \partial x_1} & \cdots & \frac{\partial^2 L}{\partial\lambda_t \partial x_n} \\ \frac{\partial^2 L}{\partial x_1 \partial\lambda_1} & \cdots & \frac{\partial^2 L}{\partial x_1 \partial\lambda_t} & \frac{\partial^2 L}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 L}{\partial x_1 \partial x_n} \\ \vdots & & \vdots & \vdots & & \vdots \\ \frac{\partial^2 L}{\partial x_n \partial\lambda_1} & \cdots & \frac{\partial^2 L}{\partial x_n \partial\lambda_t} & \frac{\partial^2 L}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 L}{\partial x_n \partial x_n} \end{pmatrix}$$

be the bordered Hessian and consider its leading principle minors $\left|\bar{H}_r(x^0;\lambda^0)\right|$ of the order $r = 2t+1, 2t+2,\ldots,n+t$ at point $(x^0;\lambda^0)$. Therefore, the following expressions can be derived:

1. If all leading principle minors $\left|\bar{H}_r(x^0;\lambda^0)\right|$, $2t+1 \le r \le n+t$, have the sign $(-1)^t$, then $x^0 = (x_1^0,\ldots,x_n^0)$ is a local minimum point of function f subject to the constraints $g_i(x) = 0$, $i = 1,\ldots,t$.
2. If the signs of all leading principle minors $\left|\bar{H}_r(x^0;\lambda^0)\right|$, $2t+1 \le r \le n+t$, are alternated, the sign of $\left|\bar{H}_{n+t}(x^0;\lambda^0)\right| = \left|\bar{H}_L(x^0;\lambda^0)\right|$ being that of $(-1)^n$, then $x^0 = (x_1^0,\ldots,x_n^0)$ is a local maximum point of function f subject to the constraints $g_i(x) = 0$, $i = 1,\ldots,t$.
3. If neither the conditions of (1) nor those of (2) are satisfied, then x^0 is not a local extreme point of function f subject to the constraints $g_i(x) = 0$, $i = 1,\ldots,t$. Here, the case in which one or several leading principal minors have a value of zero is not considered a violation of condition (1) or (2).

Lemma 1 If $\mathbf{U} = \hat{\mathbf{U}}$ and $\mathbf{W} = \hat{\mathbf{W}}$ are fixed, then $J(\hat{\mathbf{U}},\mathbf{V},\hat{\mathbf{W}})$ is minimized at $\mathbf{V}^* = (\mathbf{v}_1^*,\ldots,\mathbf{v}_c^*)$ if and only if $v_{kj}^* = \sum_{i=1}^{n} \mu_{ik}^m x_{ij} \Big/ \sum_{i=1}^{n} \mu_{ik}^m$, $\forall k = 1,\ldots,c$, $\forall j = 1,\ldots,d$.

Lemma 2 If $\mathbf{V} = \hat{\mathbf{V}}$ and $\mathbf{W} = \hat{\mathbf{W}}$ are fixed, then $J(\mathbf{U},\hat{\mathbf{V}},\hat{\mathbf{W}})$ subject to $\sum_{k=1}^{c} \mu_{ik} = 1$ is locally minimized at $\mathbf{U}^* = [\mu_{ik}^*]_{n\times c}$ if and only if

$$\mu_{ik}^* = \left(\sum_{j=1}^{d} \delta_j w_j (x_{ij} - v_{kj})^2\right)^{-1/_{m-1}} \Bigg/ \sum_{t=1}^{c} \left(\sum_{j=1}^{d} \delta_j w_j (x_{ij} - v_{tj})^2\right)^{-1/_{m-1}} \quad \forall i,k.$$

Lemma 3 If $\mathbf{U} = \hat{\mathbf{U}}$ and $\mathbf{V} = \hat{\mathbf{V}}$ are fixed, then $J(\hat{\mathbf{U}},\hat{\mathbf{V}},\mathbf{W})$ subject to $\sum_{j=1}^{d} w_j = 1$ is minimized at $\mathbf{W}^* = [w_j^*]_{d\times 1}$ if and only if

$$w_j^* = \frac{\frac{1}{\delta_j} \exp\left(\frac{-c\sum_{k=1}^{c}\sum_{i=1}^{n} \mu_{ik}^m \delta_j (x_{ij} - v_{kj})^2}{n}\right)}{\sum_{p=1}^{d} \frac{1}{\delta_p} \exp\left(\frac{-c\sum_{k=1}^{c}\sum_{i=1}^{n} \mu_{ik}^m \delta_p (x_{ip} - v_{kp})^2}{n}\right)} \quad \forall j.$$

Lemma 4 J is continuous on $M_{fcn} \times (\mathbb{R}^d)^c \times M_w$.

Lemma 5 Let Ω_{FRFCM} be the solution set of J . We have that $J\left(T(\mathbf{U},\mathbf{V},\mathbf{W})\right) = J(\mathbf{U}^*,\mathbf{V}^*,\mathbf{W}^*) < J(\mathbf{U},\mathbf{V},\mathbf{W})$ for any $(\mathbf{U},\mathbf{V},\mathbf{W}) \notin \Omega_{FRFCM}$.

Lemma 6 The FRFCM operator T is continuous on $M_{fcn} \times (\mathbb{R}^d)^c \times M_w$.

Lemma 7 Let $[\mathrm{conv}(\mathbf{X})]^c$ be the c-fold Cartesian product of the convex hull of $\mathbf{X}$, and let $\left(E(\mathbf{V}^{(0)},\mathbf{W}^{(0)}),\mathbf{V}^{(0)},\mathbf{W}^{(0)}\right)$ be the starting point of iteration with T . Then, $T^{(t)}\left(E(\mathbf{V}^{(0)},\mathbf{W}^{(0)}),\mathbf{V}^{(0)},\mathbf{W}^{(0)}\right)\in M_{fcn}\times[\mathrm{conv}(\mathbf{X})]^c\times M_w$ is compact in $M_{fcn}\times(\mathbb{R}^d)^c\times M_w$.

These proofs of Lemmas 1-7 are shown in Appendix.

According to Lemmas 4-7 by assessing the condition of Zangwill's convergence theorem, we obtain Theorem 3 of convergence theorem for the FRFCM clustering algorithm.

Theorem 3 Let $\mathbf{X}=\{\mathbf{x}_1,\dots,\mathbf{x}_n\}$ be bounded in $\mathbb{R}^d$ with the FRFCM objective function $J(\mathbf{U},\mathbf{V},\mathbf{W})$ subject to $\sum_{k=1}^{c}\mu_{ik}=1$ and $\sum_{j=1}^{d}w_j=1$. Let $T:M_{fcn}\times(\mathbb{R}^d)^c\times M_w\to M_{fcn}\times(\mathbb{R}^d)^c\times M_w$ be the FRFCM operator as defined in Definition 1. Then, for any FRFCM convergent subsequence $T^{(t_j)}\left(E(\mathbf{V}^{(0)},\mathbf{W}^{(0)}),\mathbf{V}^{(0)},\mathbf{W}^{(0)}\right)$ will tend to the optimal solution $(\mathbf{U}^*,\mathbf{V}^*,\mathbf{W}^*)$ in Ω_{FRFCM} , and the FRFCM sequence $T^{(t)}\left(E(\mathbf{V}^{(0)},\mathbf{W}^{(0)}),\mathbf{V}^{(0)},\mathbf{W}^{(0)}\right)$ will monotonically converge to the optimal solution $(\mathbf{U}^*,\mathbf{V}^*,\mathbf{W}^*)$ in Ω_{FRFCM}.

IV. COMPARISONS AND EXPERIMENTAL RESULTS

In this section, some comparisons and experiments using synthetic and real data sets are presented. We compare the FRFCM algorithm with FCM [3,4], WKM [15], EWKM [16], WFCM [18], and SCAD2 [19]. For these experimental comparisons, all methods use the same initial cluster centers and the same initial feature weights. For the exponent m in FRFCM, SCAD2, WFCM, and FCM, $m=2$ is used. These comparisons and experimental results demonstrate the effectiveness and usefulness of the proposed FRFCM algorithm.

Example 2 In this example, a three-cluster data set with 1000 data points generated from the Gaussian mixture distribution $\sum_{k=1}^{3}\alpha_k N(u_k,\Sigma_k)$ with parameters $\alpha_1=\alpha_2=\alpha_3=1/3$, $u_1=\begin{pmatrix}0 & 2\end{pmatrix}^T$, $u_2=\begin{pmatrix}4 & 7\end{pmatrix}^T$, $u_3=\begin{pmatrix}8 & 12\end{pmatrix}^T$, $\Sigma_1=\Sigma_3=\begin{pmatrix}10 & 0.01\\ 0.01 & 0.01\end{pmatrix}$, $\Sigma_2=\begin{pmatrix}4 & 0\\ 0 & 1\end{pmatrix}$, as shown in Fig. 5(a). The clustering result of FRFCM, as shown in Fig. 5(c), gives a perfectly separated clustering result with AR = 1.000. The FRFCM algorithm detects that the 2nd feature (y-axis) gives a larger weight than the 1st feature (x-axis). In fact, if we discard the 1st feature and only use the 2nd feature, we still get the same clustering result as shown in Fig. 5(c). Using WKM, EWKM, SCAD2, and WFCM, all of them also get good clustering results with ARs 1.000, 0.998, 1.000, 0.999, respectively. However, the clustering result using FCM, as shown in Fig. 5(b), cannot cluster the data set so well with AR = 0.933.

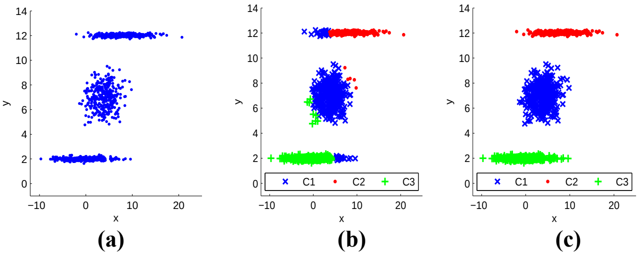


Fig. 5 (a) Original data set; Clustering results using (b) FCM; (c) FRFCM, WKM, EWKM, SCAD2, and WFCM

To further compare the performance of the proposed FRFCM with WKM, EWKM, SCAD2, WFCM, and FCM, we consider the obtained cluster center from algorithms as an estimate of the mean value of Gaussian distribution in Example 1(cont.) and Example 2. In each Example, we generate 30 samples from the Gaussian mixture distribution. We use the criterion of the mean squared error (MSE) to evaluate the accuracy. The MSE is defined as $MSE=\sum_{t=1}^{30}SE_t\big/30$, where the squared error (SE) for each sample t is defined as $SE_t=\sum_{k=1}^{c}\left\|\hat{u}_k-u_k\right\|\big/c$ in which $\hat{u}_k$ is the kth cluster center from each algorithm. The MSE results are shown in Table 7. We find that the proposed FRFCM has the smallest MSE among WKM, EWKM, SCAD2, WFCM, and FCM. This indicates that the FRFCM presents well for estimating Gaussian means.

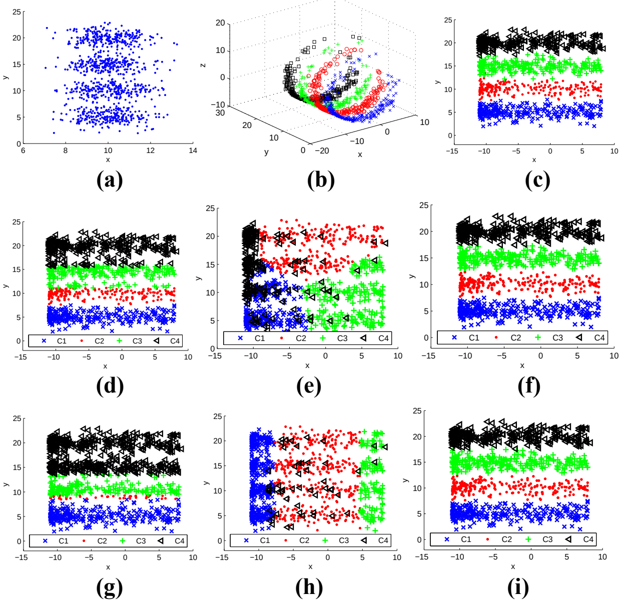
Table 7 MSE values for each algorithm

Data set	FRFCM	WKM	EWKM	SCAD2	WFCM	FCM
Example 1 (cont.)	**0.005**	**0.005**	0.008	0.006	0.148	4.745
Example 2	**0.006**	0.132	0.141	0.142	0.153	0.422

Example 3 In this example, a four-cluster data set with 1000 sample points are generated from the Gaussian mixture distribution $\sum_{k=1}^{4}\alpha_k N(u_k,\Sigma_k)$ with parameters $\alpha_k=1/4,\ \forall k$, $u_1=\begin{pmatrix}10 & 5\end{pmatrix}^T$, $u_2=\begin{pmatrix}10 & 10\end{pmatrix}^T$, $u_3=\begin{pmatrix}10 & 15\end{pmatrix}^T$, $u_4=\begin{pmatrix}10 & 20\end{pmatrix}^T$, and $\Sigma_k=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$, $\forall k$, as shown in Fig. 6(a). This data set is converted from 2-dimensional into 3-dimensional with a mapping function $(x,y)\to(x,y,z)=(x\cos x,\ y,\ x\sin x)$, as shown in Fig. 6(b). The projection of the 3-dimensional data set into the xy-plane is shown in Fig. 6(c). We implement the FRFCM algorithm for this data set. The clustering result in xy-plane is shown in Fig. 6(d) where FRFCM actually detects that the data set just depends on the 2nd feature. The clustering results in the xy-plane with FCM, WKM, EWKM, SCAD2, and WFCM are shown in Figs. 6(e), 6(f), 6(g), 6(h), and 6(i) respectively. The ARs for the FRFCM, FCM, WKM, EWKM, SCAD2, and WFCM algorithms are 0.992, 0.407, 0.992, 0.691, 0.232, and 0992 respectively. It is obviously that the proposed FRFCM gets the best clustering result, as shown in Fig. 6(d), but EWKM and SCAD2 are not good enough, because they just compute the feature weights in each cluster

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2017.2692203, IEEE Transactions on Fuzzy Systems

and do not select which features are important. While WKM and WFCM obtain the 2nd feature as the largest feature weight with $\mathbf{W}=[0.223\ 0.545\ 0.232]$ and $\mathbf{W}=[0.002\ 0.996\ 0.002]$, respectively.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 6 (a) Original 2-dimension data set; (b) Mapping into 3-dimension with marker and color; (c) Projection into *xy*-plane; Clustering results in the *xy*-plane using: (d) FRFCM; (e) FCM; (f) WKM; (g) EWKM; (h) SCAD2; (i) WFCM

Table 8 ARs of the Iris data set with FCM using different combinations of features

Feature used (SL, SW, PL, PW)	AR
(1, 1, 1, 1)	0.893
(1, 1, 1, 0)	0.887
(1, 1, 0, 1)	0.827
(1, 1, 0, 0)	0.807
(1, 0, 1, 1)	0.907
(1, 0, 1, 0)	0.880
(1, 0, 0, 1)	0.827
(1, 0, 0, 0)	0.707
(0, 1, 1, 1)	0.940
(0, 1, 1, 0)	0.927
(0, 1, 0, 1)	0.927
(0, 1, 0, 0)	0.427
(0, 0, 1, 1)	0.947
(0, 0, 1, 0)	0.933
(0, 0, 0, 1)	0.960

Example 4 In this example, we consider the well-used real data set, Iris data, that contain 150 data points with four attributes, i.e., sepal length (SL, in cm), sepal width (SW, in cm), petal length (PL, in cm), and petal width (PW, in cm). The Iris data set has three clusters, i.e., setosa, versicolor, and virginica. Table 8 shows the ARs of FCM using different features. By using all features, FCM gives the AR of 0.893 (16 incorrect data of 150 data), while using the features PL and PW, FCM obtains the second largest AR with 0.947. Moreover, using only feature PW gives the AR of 0.960 that is the largest. This means, using FCM, feature PW is the most important feature and PL is the second most important feature for the Iris data set.

We also implement the WKM, EWKM, SCAD2, and WFCM algorithms for the Iris data set. As depicted in Table 9, WKM shows that the 1st feature has the smallest weight and the 4th feature has the largest weight. EWKM and SCAD2 algorithms show that the 2nd and 4th features have large weights for clusters 1 and 2, while for cluster 3, the 3rd and 4th features have large feature weights than the 1st and 2nd features. While WFCM gives the 3rd and 4th features as the features with large weights. Together with FCM, thus, we can say that the 3rd and 4th features are more important than the 1st and 2nd features. The ARs of WKM, EWKM, SCAD2, and WFCM are 0.953, 0.960, 0.960 and 0.953, respectively.

Table 9 Feature weights using WKM, EWKM, SCAD2, and WFCM for the Iris data set

		Final feature weights			
		1st (SL)	2nd (SW)	3rd (PL)	4th (PW)
WKM		0.177	0.251	0.203	0.369
EWKM	Cluster 1	0.013	0.248	0.017	0.722
	Cluster 2	0.006	0.385	0.013	0.596
	Cluster 3	0.079	0.056	0.367	0.498
SCAD2	Cluster 1	0.168	0.291	0.154	0.387
	Cluster 2	0.162	0.334	0.164	0.340
	Cluster 3	0.144	0.134	0.256	0.466
WFCM		0.023	0.117	0.431	0.429

We next use FRFCM for the Iris data set with equal feature-weight initial $\mathbf{W}^{(0)}$. The feature reduction behavior from FRFCM for the Iris data set is shown in Table 10. After the first iteration, the 3rd and 4th features give larger feature weights than the 1st and 2nd features. Since the 1st and 2nd features have very small feature weights, these two features are discarded during clustering processes. FRFCM finally retains the two features, i.e., the 3rd feature (PL) and the 4th feature (PW), after three iterations with a very good clustering result of AR = 0.973 (i.e., 4 incorrect data of 150 data).

Table 10 Feature reduction behavior for the Iris data set using FRFCM

	Updating feature weights			
	1st (SL)	2nd (SW)	3rd (PL)	4th (PW)
Initialization (equal feature-weight)	0.250	0.250	0.250	0.250
Iteration 1	0.006	0.005	0.457	0.532
Iteration 2	-	-	0.565	0.435
Iteration 3	-	-	0.565	0.435

Beside the above synthetic data sets and the Iris dataset, we next use more real data sets which contain plant data (seeds, soybean), disease data (pima Indians, thyroid, bupa), cancer data (breast cancer, colon cancer, ovariance cancer), handwritten data (USPS), text data (basehock), and gene data (SMK-CAN-187), taken from UCI data repository [26] and Kent Ridge Biomedical Data Set [27].

For comparisons, except using the criterion of AR, we also consider the other well-used clustering performance measures. Let *C* is the set of original clusters in data set and *C** is the set of clusters obtained by the clustering algorithm. For a pair of points (x_i, x_j), *a* is the number of pairs if both points belong to the same cluster in *C* and *C**, *b* is the number of pairs if both points belong to the same cluster in *C* and different clusters in *C**, *c* is the number of pairs if both points belong to two different clusters in *C* and the same cluster in *C**, and *d* is the number of pairs if both points belong to the two different clusters in *C* and *C**. In 1971, Rand [28] proposed objective criteria for the evaluation of clustering methods, known as the Rand Index (RI). Up to now, RI had popularly used for measuring similarity between two clustering partitions. It was widely applied to various areas [29-33]. The RI is defined by $\mathrm{RI} = (a+d)/(a+b+c+d)$, and so the larger RI is, the better clustering performance is. Another clustering measurement is Jaccard Index (JI) proposed by Jaccard [34]. The JI is defined as $\mathrm{JI} = a/(a+b+c)$. The larger JI is, the better clustering performance is. The RI and JI were extended by Yeh and Yang [35]. In our experiments, we also use normalized mutual information (NMI) [36] to evaluate the clustering performance. MI is often used to evaluate the accuracy of clustering results. In clustering evaluation, it measures how much information the presence/absence of a term contributes to making the correct classification decision. The normalized MI (NMI), whose value is always a number between 0 and 1, is used to compare clustering with different numbers of clusters. NMI is defined as $\mathrm{NMI} = \dfrac{I(X:Y)}{[H(X)+H(Y)]/2}$, where $H(X)$ is an entropy of X and $I(X:Y)$ is a mutual information between $H(X)$ and $H(Y)$. The larger NMI is, the better clustering performance is.

Example 5 In this example, we consider some real data sets, which their properties are given in Table 11. We compare AR, RI, JI, and NMI of FRFCM with FCM, WKM, EWKM, SCAD2, and WFCM. The comparisons using different initial cluster centers with fixed initial feature weights are shown in Table 12, while comparisons using different initial feature weights with fixed initial cluster centers are presented in Table 13. From the worst, average, and the best ARs, RIs, JIs, and NMIs obtained from all algorithms, the proposed FRFCM algorithm actually presents better results.

Table 11 Real data sets

Data set	# of instances	# of features	# of clusters
Thyroid	215	5	3
Bupa	345	6	2
Seeds	210	7	3
Breast cancer	699	8	2
Pima Indians	768	8	2
Soybean	47	21	4
USPS	4000	256	10
Colon cancer	62	2000	2
Ovariance cancer	216	4000	2
Basehock	1993	4862	2
SMK-CAN-187	187	19993	2

Table 12 The worst/average/the best ARs, RIs, JIs, and NMIs using different initial cluster centers with fixed initial feature weights

Data set		FRFCM	WKM	EWKM	SCAD2	WFCM	FCM
Iris	AR	**0.947**/**0.961**/**0.973**	0.880/0.924/0.960	0.840/0.903/0.960	**0.947**/0.956/0.960	**0.947**/0.950/0.953	0.893/0.893/0.893
	RI	**0.934**/**0.951**/**0.966**	0.868/0.910/0.950	0.837/0.892/0.950	**0.934**/0.946/0.950	**0.934**/0.934/0.934	0.880/0.880/0.880
	JI	**0.819**/**0.861**/**0.901**	0.673/0.769/0.892	0.622/0.730/0.858	0.818/0.847/0.858	0.818/0.818/0.818	0.654/0.654/0.654
	NMI	**0.833**/**0.870**/**0.901**	0.744/0.812/0.864	0.715/0.793/0.864	0.832/0.856/0.864	0.831/0.831/0.831	0.749/0.749/0.749
Thyroid	AR	**0.865**/**0.881**/0.907	0.628/0.747/0.861	0.423/0.493/0.661	0.563/0.869/**0.954**	0.754/0.854/**0.954**	0.791/0.791/0.791
	RI	**0.794**/0.815/0.857	0.591/0.681/0.792	0.503/0.530/0.589	0.569/**0.842**/**0.924**	0.682/0.803/0.924	0.719/0.719/0.719
	JI	**0.675**/0.711/0.765	0.403/0.518/0.675	0.280/0.314/0.446	0.367/**0.758**/0.868	0.505/0.686/**0.868**	0.550/0.550/0.550
	NMI	**0.499**/0.551/0.610	0.251/0.321/0.495	0.110/0.160/0.268	0.252/**0.668**/**0.795**	0.273/0.530/0.786	0.344/0.344/0.344
Bupa	AR	0.551/0.551/0.551	0.504/0.545/0.588	0.502/0.559/**0.597**	0.542/0.535/0.513	0.525/0.525/0.525	**0.559**/**0.560**/0.563
	RI	**0.531**/**0.531**/**0.531**	0.499/0.503/0.514	0.499/0.506/0.518	0.499/0.502/0.503	0.506/0.506/0.506	0.500/0.500/0.500
	JI	0.343/0.343/0.343	0.339/0.409/0.447	0.336/0.371/**0.480**	0.344/0.379/0.411	0.360/0.361/0.362	**0.429**/**0.429**/0.429
	NMI	**0.011**/**0.011**/0.011	0.000/0.002/0.018	0.000/0.008/**0.018**	0.000/0.001/0.003	0.004/0.005/0.006	0.007/0.007/0.007
Seeds	AR	0.862/**0.895**/**0.919**	0.857/0.875/0.909	0.714/0.809/0.891	0.857/0.869/0.881	**0.895**/**0.895**/0.895	**0.895**/**0.895**/0.895
	RI	0.842/**0.876**/**0.900**	0.830/0.852/0.891	0.731/0.801/0.872	0.829/0.844/0.859	0.873/0.873/0.873	**0.874**/0.874/0.874
	JI	0.616/**0.685**/**0.737**	0.594/0.635/0.716	0.475/0.573/0.677	0.582/0.620/0.651	0.678/0.678/0.678	**0.682**/0.682/0.682
	NMI	0.675/**0.697**/**0.725**	0.603/0.652/0.722	0.524/0.613/0.696	0.601/0.627/0.656	0.676/0.676/0.676	**0.695**/0.695/0.695
Breast cancer	AR	0.927/**0.947**/0.953	0.924/0.944/0.955	0.876/0.944/**0.957**	0.790/0.813/0.920	**0.938**/0.938/0.938	0.936/0.936/0.936
	RI	0.865/**0.902**/**0.919**	0.860/0.894/0.915	0.782/0.895/0.918	0.667/0.701/0.852	**0.884**/0.884/0.884	0.879/0.879/0.879
	JI	**0.821**/**0.840**/0.848	0.779/0.827/0.857	0.782/0.835/**0.861**	0.583/0.616/0.769	0.812/0.812/0.812	0.805/0.805/0.805
	NMI	0.619/**0.687**/0.708	0.604/0.675/0.723	0.426/0.676/**0.729**	0.252/0.306/0.590	0.642/0.642/0.642	**0.653**/0.653/0.653
Pima Indians	AR	**0.921**/**0.954**/**1.000**	0.615/0.758/**1.000**	0.595/0.669/**1.000**	0.617/0.939/**1.000**	0.720/0.720/0.720	0.659/0.659/0.659
	RI	**0.854**/**0.915**/**1.000**	0.526/0.693/**1.000**	0.517/0.587/**1.000**	0.527/0.878/**1.000**	0.548/0.548/0.548	0.5500.550/0.550
	JI	**0.756**/**0.853**/**1.000**	0.376/0.621/**1.000**	0.376/0.481/**1.000**	0.406/0.842/**1.000**	0.455/0.455/0.455	0.442/0.442/0.442

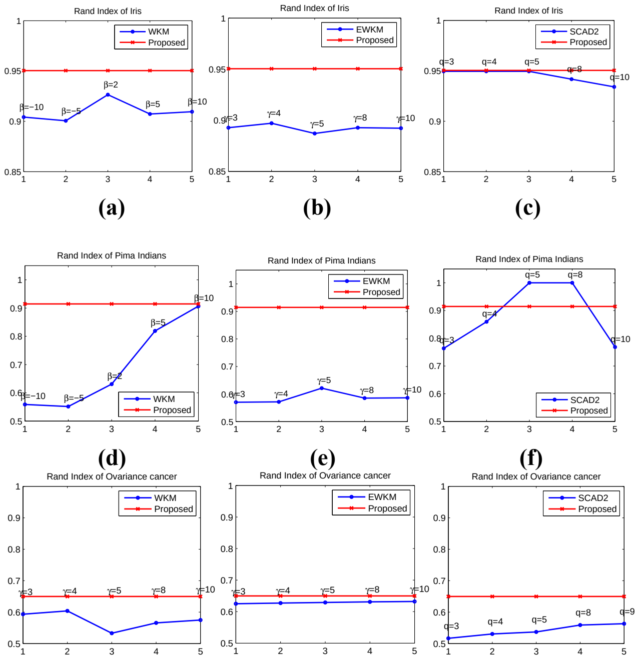
	NMI	**0.664**/**0.784**/**1.000**	0.001/0.344/**1.000**	0.015/0.140/**1.000**	0.014/0.737/**1.000**	0.119/0.119/0.119	0.034/0.034/0.034
Soybean	AR	**0.787**/**0.887**/0.979	0.681/0.770/0.830	0.702/0.791/**1.000**	0.575/0.575/0.575	**0.787**/0.872/0.894	0.723/0.736/0.787
	RI	**0.843**/**0.910**/0.843	0.699/0.825/0.859	0.809/0.857/**1.000**	0.593/0.593/0.593	0.841/0.893/0.906	0.832/0.834/0.843
	JI	0.518/**0.742**/0.910	0.251/0.502/0.627	0.459/0.601/**1.000**	0.381/0.381/0.381	**0.597**/0.661/0.676	0.489/0.527/0.615
	NMI	**0.783**/**0.888**/0.944	0.364/0.694/0.826	0.692/0.813/**1.000**	0.618/0.618/0.618	0.774/0.801/0.807	0.716/0.742/0.849
USPS	AR	**0.436**/**0.447**/0.464	0.260/0.420/**0.527**	0.214/0.344/0.443	0.128/0.276/0.335	0.313/0.393/0.402	0.314/0.393/0.402
	RI	**0.842**/**0.855**/0.868	0.622/0.847/**0.893**	0.426/0.693/0.842	0.509/0.613/0.716	0.701/0.702/0.702	0.702/0.702/0.702
	JI	**0.297**/**0.302**/0.315	0.129/0.267/**0.350**	0.121/0.185/0.272	0.128/0.154/0.181	0.218/0.218/0.218	0.224/0.224/0.224
	NMI	**0.373**/0.393/0.410	0.211/**0.460**/**0.553**	0.146/0.319/0.480	0.178/0.259/0.321	0.359/0.359/0.359	0.381/0.381/0.381
Colon cancer	AR	0.513/**0.597**/**0.613**	0.500/0.532/0.597	0.500/0.562/**0.613**	**0.548**/0.548/0.548	**0.548**/0.563/0.565	**0.548**/0.548/0.548
	RI	**0.511**/**0.515**/**0.518**	0.492/0.495/0.511	0.492/0.503/**0.518**	0.497/0.497/0.497	0.497/0.500/0.501	0.497/0.497/0.497
	JI	**0.491**/**0.492**/**0.493**	0.338/0.344/0.358	0.338/0.403/**0.493**	0.344/0.344/0.344	0.349/0.349/0.349	0.344/0.344/0.344
	NMI	**0.005**/**0.027**/**0.062**	0.001/0.004/0.021	0.001/0.016/**0.062**	0.013/0.013/0.013	0.013/0.023/0.024	0.013/0.013/0.013
Ovariance cancer	AR	**0.750**/**0.775**/0.796	0.621/0.699/0.759	0.690/0.754/**0.810**	0.597/0.645/0.681	0.741/0.741/0.741	0.713/0.713/0.713
	RI	**0.623**/**0.650**/0.674	0.527/0.579/0.633	0.570/0.629/**0.691**	0.517/0.515/0.563	0.614/0.614/0.614	0.589/0.589/0.589
	JI	**0.474**/**0.496**/0.518	0.362/0.424/0.470	0.445/0.488/**0.544**	0.357/0.379/0.398	0.457/0.457/0.457	0.431/0.431/0.431
	NMI	**0.297**/**0.324**/**0.351**	0.039/0.163/0.273	0.149/0.266/0.323	0.037/0.079/0.114	0.235/0.235/0.235	0.182/0.182/0.182
Basehock	AR	**0.555**/**0.588**/**0.635**	0.501/0.508/0.551	0.501/0.515/0.616	0.501/0.505/0.554	0.537/0.537/0.537	0.536/0.536/0.536
	RI	**0.506**/**0.516**/**0.536**	0.499/0.501/0.505	0.500/0.504/0.535	0.499/0.501/0.506	0.503/0.503/0.503	0.502/0.502/0.502
	JI	0.396/0.412/0.450	0.366/0.482/0.499	0.340/0.479/0.500	**0.476**/**0.497**/**0.500**	0.361/0.361/0.361	0.359/0.359/0.359
	NMI	**0.021**/**0.042**/0.093	0.000/0.014/0.048	0.001/0.023/**0.138**	0.000/0.009/0.098	0.005/0.005/0.005	0.005/0.005/0.005
SMK-CAN-187	AR	0.604/**0.611**/0.621	0.503/0.563/0.631	0.503/0.570/**0.642**	0.556/0.559/0.562	**0.609**/0.609/0.609	**0.609**/0.609/0.609
	RI	0.519/**0.522**/0.527	0.497/0.510/0.532	0.497/0.513/**0.538**	0.503/0.504/0.505	**0.522**/0.522/0.522	**0.522**/0.522/0.522
	JI	0.349/0.354/0.359	0.345/0.371/**0.398**	0.349/**0.373**/0.396	0.347/0.348/0.350	**0.359**/0.359/0.359	**0.359**/0.359/0.359
	NMI	0.031/**0.036**/0.042	0.001/0.022/0.064	0.000/0.026/**0.069**	0.008/0.010/0.011	**0.035**/0.035/0.035	**0.035**/0.035/0.035

Table 13 The worst/average/the best ARs, RIs, JIs, and NMIs using different initial feature weights with fixed initial cluster centers

Data set		FRFCM	WKM	EWKM	SCAD2	WFCM
Iris	AR	**0.953**/**0.964**/**0.973**	0.873/0.925/0.960	0.887/0.924/0.960	0.947/0.956/0.960	0.807/0.893/0.947
	RI	**0.942**/**0.953**/**0.966**	0.862/0.912/0.949	0.874/0.911/0.949	0.934/0.945/0.949	0.811/0.882/0.934
	JI	**0.837**/**0.867**/**0.901**	0.662/0.769/0.858	0.684/0.768/0.857	0.818/0.846/0.857	0.553/0.701/0.818
	NMI	**0.850**/**0.879**/**0.901**	0.737/0.808/0.871	0.751/0.809/0.864	0.832/0.854/0.864	0.610/0.746/0.837
Thyroid	AR	**0.735**/0.854/0.912	0.721/0.752/0.842	0.405/0.590/0.739	0.563/**0.869**/**0.953**	0.651/0.764/**0.954**
	RI	**0.657**/0.796/0.857	0.647/0.675/0.774	0.498/0.563/0.670	0.569/**0.843**/**0.924**	0.571/0.688/0.924
	JI	**0.545**/0.700/0.771	0.474/0.511/0.648	0.278/0.402/0.537	0.367/**0.750**/**0.868**	0.391/0.528/0.868
	NMI	**0.381**/0.544/0.629	0.282/0.305/0.409	0.049/0.138/0.271	0.252/**0.667**/**0.795**	0.241/0.362/0.786
Bupa	AR	**0.551**/0.551/0.551	0.501/0.544/0.551	0.510/**0.553**/**0.609**	0.513/0.535/0.542	0.516/0.540/0.559
	RI	**0.504**/0.504/0.504	0.501/0.503/0.504	0.499/**0.505**/**0.522**	0.499/0.501/0.503	0.499/0.502/0.506
	JI	0.343/0.343/0.343	0.343/0.407/0.441	0.342/0.344/0.373	0.343/0.379/0.411	**0.359**/**0.409**/**0.447**
	NMI	**0.011**/0.011/0.011	0.000/0.001/0.011	0.001/**0.011**/**0.031**	0.000/0.001/0.003	0.000/0.004/0.014
Seeds	AR	**0.886**/**0.901**/**0.914**	0.547/0.806/0.881	0.705/0.812/0.891	0.857/0.869/0.881	0.857/0.883/**0.914**
	RI	**0.866**/**0.882**/**0.896**	0.626/0.806/0.862	0.747/0.819/0.872	0.829/0.844/0.860	0.839/0.864/0.895
	JI	**0.665**/**0.698**/**0.728**	0.288/0.565/0.656	0.534/0.601/0.677	0.592/0.620/0.650	0.613/0.661/0.726
	NMI	**0.675**/**0.700**/0.719	0.192/0.563/0.692	0.583/0.633/0.682	0.601/0.627/0.658	0.661/0.687/**0.727**
Breast cancer	AR	**0.923**/0.936/**0.958**	0.765/**0.943**/0.954	0.718/0.876/0.957	0.790/0.867/0.920	0.907/0.932/0.953
	RI	**0.857**/0.880/**0.921**	0.640/**0.893**/0.912	0.595/0.801/0.918	0.667/0.776/0.852	0.831/0.873/0.907
	JI	**0.776**/0.807/**0.864**	0.492/**0.825**/0.853	0.525/0.734/0.861	0.583/0.692/0.769	0.744/0.798/0.846
	NMI	**0.599**/0.649/**0.737**	0.355/**0.677**/0.717	0.159/0.512/0.729	0.252/0.45/0.590	0.555/0.636/0.706
Pima Indians	AR	**0.959**/**0.979**/**1.000**	0.665/0.888/**1.000**	0.602/0.637/**1.000**	0.617/0.899/**1.000**	0.650/0.676/0.729
	RI	**0.922**/**0.960**/**1.000**	0.554/0.846/**1.000**	0.520/0.543/**1.000**	0.527/0.874/**1.000**	0.544/0.564/0.604
	JI	**0.865**/**0.929**/**1.000**	0.416/0.799/**1.000**	0.308/0.426/**1.000**	0.406/0.841/**1.000**	0.424/0.442/0.462
	NMI	**0.786**/**0.878**/**1.000**	0.043/0.682/**1.000**	0.016/0.040/**1.000**	0.014/0.737/**1.000**	0.031/0.060/0.131
Soybean	AR	**0.808**/**0.946**/0.979	0.681/0.743/0.809	0.638/0.774/**1.000**	0.574/0.574/0.574	0.723/0.804/0.957
	RI	**0.850**/**0.950**/0.976	0.790/0.831/0.850	0.757/0.846/**1.000**	0.593/0.593/0.593	0.827/0.866/0.967
	JI	**0.533**/**0.832**/0.910	0.399/0.493/0.533	0.406/0.586/**1.000**	0.381/0.381/0.381	0.479/0.583/0.876
	NMI	**0.748**/**0.922**/0.944	0.564/0.710/0.748	0.590/0.799/**1.000**	0.618/0.618/0.618	0.680/0.769/0.918
USPS	AR	**0.458**/**0.463**/0.467	0.237/0.386/**0.556**	0.220/0.307/0.380	0.271/0.305/0.335	0.397/0.401/0.404
	RI	**0.865**/**0.867**/0.868	0.664/0.834/**0.903**	0.441/0.661/0.762	0.528/0.624/0.667	0.698/0.702/0.705
	JI	**0.314**/**0.316**/0.317	0.124/0.247/**0.393**	0.115/0.160/0.207	0.132/0.157/0.175	0.218/0.223/0.229

	NMI	**0.394**/**0.399**/0.405	0.152/0.411/**0.588**	0.185/0.282/0.393	0.208/0.262/0.321	0.359/0.376/0.402
Colon cancer	AR	**0.613**/**0.613**/0.613	0.500/0.530/**0.645**	0.516/0.555/0.613	0.548/0.548/0.548	0.516/0.544/0.564
	RI	**0.518**/**0.518**/0.518	0.492/0.496/**0.535**	0.492/0.503/0.518	0.496/0.496/0.496	0.492/0.496/0.500
	JI	**0.491**/**0.491**/**0.491**	0.338/0.345/0.382	0.356/0.410/**0.491**	0.344/0.344/0.344	0.340/0.343/0.349
	NMI	0.005/0.005/0.005	0.0000/0.005/**0.050**	0.001/0.003/0.004	**0.013**/**0.013**/0.013	0.003/0.011/0.023
Ovariance cancer	AR	**0.782**/**0.785**/0.792	0.602/0.676/**0.796**	0.560/0.673/0.773	0.597/0.645/0.681	0.704/0.720/0.745
	RI	**0.658**/**0.661**/0.669	0.517/0.564/**0.674**	0.505/0.566/0.648	0.517/0.542/0.563	0.581/0.596/0.619
	JI	**0.502**/**0.505**/0.512	0.350/0.403/**0.538**	0.458/0.489/0.513	0.357/0.379/0.398	0.420/0.437/0.460
	NMI	**0.330**/**0.334**/**0.344**	0.026/0.105/0.306	0.027/0.161/0.262	0.037/0.077/0.114	0.157/0.193/0.241
Basehock	AR	**0.585**/**0.590**/0.599	0.500/0.510/**0.619**	0.501/0.532/0.581	0.501/0.521/0.554	0.515/0.532/0.551
	RI	**0.514**/**0.516**/0.520	0.499/0.501/**0.528**	0.500/0.503/0.513	0.500/0.502/0.506	0.500/0.502/0.505
	JI	0.397/0.399/0.401	0.434/**0.493**/**0.500**	0.436/0.465/0.491	**0.476**/0.491/**0.500**	0.342/0.355/0.364
	NMI	**0.033**/0.036/0.044	0.000/0.021/**0.169**	0.000/0.024/0.068	0.000/**0.037**/0.098	0.001/0.004/0.009
SMK-CAN-187	AR	**0.615**/**0.619**/0.620	0.508/0.544/**0.658**	0.508/0.516/0.524	0.556/0.559/0.562	0.610/0.610/0.610
	RI	**0.524**/**0.526**/0.526	0.497/0.504/**0.547**	0.497/0.498/0.498	0.504/0.504/0.505	0.521/0.521/0.521
	JI	0.357/0.358/0.359	0.343/0.371/**0.447**	**0.375**/**0.378**/0.393	0.347/0.348/0.350	0.359/0.359/0.359
	NMI	**0.038**/**0.041**/0.042	0.000/0.009/**0.072**	0.000/0.001/0.002	0.008/0.009/0.010	0.034/0.034/0.034

Since WKM, EWKM, and SCAD2 need extra parameter value setting, we make more comparisons of FRFCM with WKM, EWKM, and SCAD2 for Iris, pima Indians, and ovariance cancer under different parameter values. These results are shown in Fig. 7. Different parameter values can obtain different clustering results. Overall, the proposed FRFCM, not necessary to have parameter value setting, almost has better clustering results with large RI.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 7 Rand index comparisons of FRFCM with WKM, EWKM, and SCAD2 for Iris, pima Indians, and ovariance cancer under different parameter values

Example 6 In this example, we use Olivetti Research Laboratory (ORL) database of faces contains of 400 different images from 40 individuals, where each individual has 10 different images [37]. The images were taken at different times, different condition of lighting, different variations of face expression (open/closed eyes, smiling/not smiling), and different facial details (glasses/no glasses). All images were taken against a dark homogeneous background with a tolerance for some tilting and rotation of the face. We use 100 face images from 10 individuals of ORL database with 1024 attributes, as shown in Fig. 8. We implement FRFCM for the 100 face images and get its AR of 0.420, while WKM, EWKM, SCAD2, WFCM, and FCM have 0.340, 0.410, 0.200, 0.200, and 0.200 of AR, respectively.

Fig. 8 100 face images of ORL database

In all above experiments, we had assumed the number c of clusters as a known number. However, it is usually unknown in clustering. Therefore, to check the effectiveness of FRFCM, SCAD2, WFCM, and FCM, we also use cluster validity indices, such as partition coefficient (PC) proposed by Bezdek [38] and Dunn [39], partition entropy (PE) proposed by Bezdek [40], XB index proposed by Xie and Beni [41], and C-index (CI) proposed by Hubert and Levin [42]. Table 14 shows the estimated optimal cluster numbers obtained from these cluster validity indices for real data sets using FRFCM, SCAD2, WFCM, and FCM. From Table 14, we see that FRFCM seems to be better for finding the true cluster number c than the other fuzzy clustering algorithms, SCAD2, WFCM, and FCM for these cluster validity indices.

We next analyze the computational complexities for the FRFCM, FCM, WKM, EWKM, SCAD2, and WFCM algorithms. FRFCM algorithm is divided into three parts: (1)

Compute the membership partition, μ_{ik} , which needs $O(nc^2d)$; (2) Update cluster center, v_k , which needs $O(nc)$; and (3) Update the weight w_j , which needs $O(ncd^2)$. Because the notation of big O (i.e., $O(\cdot)$) only considers the upper bound on the growth rate of the function, the total computational complexity for the FRFCM algorithm is $O(nc^2d+ncd^2)$, where n is the number of data, c is the number of clusters, and d is the dimension of data points. While for FCM, it needs $O(nc^2d)$, WKM needs $O(ncd^2)$, EWKM needs $O(ncd^2)$, SCAD2 needs $O(nc^2d+ncd^2)$, and WFCM needs $O(nc^2d)$.

Finally, we demonstrate that the iteration time of FRFCM will generally decrease rapidly after several iterations for most data sets. This situation occurs because the features will be discarded during iterations, so that the number d of features will also decrease. The final numbers of features obtained from FRFCM for different real data sets are shown in Table 15. In Table 15, we also present the speed of clustering with the total running time from each algorithm for real data sets. As shown, we can see that, if the data set has more feature numbers, FRFCM will give much less running time than the other algorithms, especially for data sets with high-dimensional features. To further demonstrate this phenomenon, we also consider the computation time in seconds per iteration for these data sets, i.e., Gaussian mixture model with $d=4$ of Example 1, Iris data set with $d=4$, and ovariance cancer with $d=4000$. As shown in Fig. 9, at the first iteration, FRFCM needs more times than the second iteration. Start from the third iteration, time needed for clustering processes reduces up to 90% from the first iteration. This means that the computation time will be decreased rapidly after some iteration.

Table 14 Optimal cluster numbers obtained from validity indices for real data sets under different clustering algorithms

Data set	True c	Validity indices	Optimal c*			
			FRFCM	SCAD2	WFCM	FCM
Iris	3	PC	2	2	2	2
		PE	2	2	2	2
		XB	2	2	2	2
		CI	2	2	2	2
Thyroid	3	PC	**3**	2	2	2
		PE	**3**	2	2	2
		XB	**3**	2	**3**	**3**
		CI	6	**3**	5	5
Seeds	3	PC	**3**	2	2	2
		PE	2	2	2	2
		XB	**3**	2	**3**	2
		CI	**3**	4	**3**	4
Breast cancer	2	PC	**2**	**2**	**2**	**2**
		PE	**2**	6	**2**	**2**
		XB	**2**	**2**	**2**	**2**
		CI	6	**2**	4	3
Pima Indians	2	PC	**2**	**2**	**2**	**2**
		PE	**2**	**2**	**2**	**2**
		XB	**2**	4	**2**	**2**
		CI	3	4	3	6
Ovariance cancer	2	PC	**2**	**2**	**2**	**2**
		PE	**2**	**2**	**2**	**2**
		XB	**2**	**2**	**2**	**2**
		CI	**2**	**2**	5	5

Table 15 Numbers of original and final features obtained from FRFCM and total running time (in seconds) for real data sets of each algorithm

Real data sets	Original d	Final d by FRFCM	Total running time (seconds)					
			FRFCM	WKM	EWKM	SCAD2	WFCM	FCM
Iris	4	2	0.102	0.080	0.088	0.154	0.164	0.081
Thyroid	5	3	0.099	0.083	0.084	0.197	0.404	0.344
Bupa	6	1	0.100	0.087	0.077	0.178	0.511	0.112
Seeds	7	5	0.110	0.095	0.081	0.183	0.222	0.087
Breast cancer	8	7	0.112	0.103	0.085	0.197	2.452	0.099
Pima indians	8	3	0.114	0.095	0.090	0.218	10.881	0.159
Soybean	21	9	0.098	0.078	0.079	0.149	0.140	0.095
USPS	256	26	17.589	17.224	19.520	105.129	355.297	29.444
Colon cancer	2000	2	0.107	0.110	0.173	0.381	0.487	0.154
Ovariance cancer	4000	71	0.234	0.249	0.322	2.918	7.142	0.648
Basehock	4862	517	8.929	9.849	20.051	105.390	1761.603	25.188
SMK-CAN-187	19993	51	0.804	0.821	1.566	19.169	31.713	4.056
ORL	1024	29	0.130	0.122	0.252	5.301	1.113	0.659

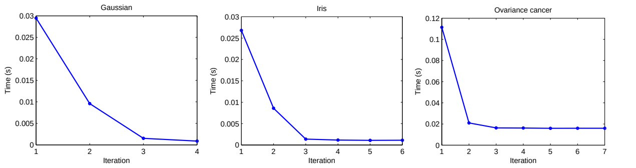


Fig. 9 Plots of per iteration time for different data sets: (a) Gaussian mixture of Example 1; (b) Iris data set; (c) Ovariance cancer data set

V. CONCLUSIONS AND DISCUSSION

In this paper, we proposed the feature-reduction FCM (FRFCM) clustering algorithm that can reduce feature dimension automatically, and also produce good clustering results. The FRFCM algorithm computes a new weight for each feature by adding feature-weight entropy in the FRFCM objective function. These new weights are then used to update the memberships and cluster centers for the data set during iterations. The proposed FRFCM algorithm is not only improving FCM performance, but also able to select important features by weighting them and reduce feature dimension by discarding unimportant features. Experimental results and comparisons actually demonstrate the effectiveness and good aspects of the proposed FRFCM algorithm. We mention that FRFCM is similar as FCM and other feature-weighted FCM in which the fuzziness index *m* has impact on clustering results, even though *m* = 2 had been commonly used. In this sense, a procedure for finding an optimal *m* should be useful, but it is actually a difficult problem. In our future work, we will consider an approach based on Markov chain Monte Carlo (MCMC) and simulated annealing for finding an approximately optimal *m* value of FCM, FRFCM and other feature-weighted FCM in real applications. On the other hand, a theoretical upper bound for *m* may offer additional information on selecting *m*. That is, we may consider the Jacobian matrix approach proposed by Chaomurilige et al. [43] for finding a theoretical upper bound of *m* in the FRFCM and other feature-weighted FCM clustering algorithms.

Most of feature-weighted FCM algorithms, including FRFCM, are necessary to assign a number of clusters a priori. We aware that there are variants of cluster validity indices for fuzzy clustering (see [30-32, 40-42, 44]) to validate if the clustering results accurately present the actual structure of data. Although there are many validity indices to be used for finding an optimal number of clusters, it is a separated procedure, but not an embedding part of the algorithm. In our future work, we are also interested in constructing a framework for the FRFCM algorithm such that it can be free of membership initializations and the fuzziness index *m* with automatically finding an optimal number of clusters. Furthermore, the FRFCM algorithm is only for clustering numerical data. Since some real data sets also contain categorical attributes or mixed attributes (numeric and categorical), we will consider extending FRFCM, such that it can handle these categorical or mixed attribute data sets.

Acknowledgements

The authors would like to thank the anonymous referees for their helpful comments in improving the presentation of this paper.

Appendix

Proof of Lemma 1: Recall that

$$J(\mathbf{U},\mathbf{V},\mathbf{W})=\sum_{k=1}^{c}\sum_{i=1}^{n}\sum_{j=1}^{d}\mu_{ik}^{m}\delta_j w_j (x_{ij}-v_{kj})^2+\frac{n}{c}\sum_{j=1}^{d}\left(w_j \log \delta_j w_j\right).$$

With the gradient of J w.r.t. v_{kj}, we have $\frac{\partial \tilde{J}}{\partial v_{kj}}=-2\sum_{i=1}^{n}\mu_{ik}^{m}\delta_j w_j (x_{ij}-v_{kj})=0$, and then $v_{kj}=\sum_{i=1}^{n}\mu_{ik}^{m}x_{ij}\Big/\sum_{i=1}^{n}\mu_{ik}^{m}$, $\forall k,j$. Thus, we proved the "only if" condition. The proof of the "if" condition is as follows. If $\mathbf{U}=\hat{\mathbf{U}}$ and $\mathbf{W}=\hat{\mathbf{W}}$ are fixed, then we have $\frac{\partial \tilde{J}}{\partial v_{kj}}=-2\sum_{i=1}^{n}\mu_{ik}^{m}\delta_j w_j (x_{ij}-v_{kj})$ and $\frac{\partial \tilde{J}}{\partial v_{kj}\partial v_{lj}}=2\beta_{kl}\mathbf{I}_d\sum_{i=1}^{n}\mu_{ik}^{m}\delta_j w_j$, where β_{kl} is Kronecker index with $\beta_{kl}=\begin{cases}1, & \text{if } k=l\\ 0, & \text{if } k\neq l\end{cases}$. The Hessian matrix of $J(\hat{\mathbf{U}},\mathbf{V},\hat{\mathbf{W}})$ w.r.t. v_{kj} is $2\times diag\left(\mathbf{I}_d\sum_{i=1}^{n}\mu_{i1}^{m}\delta_j w_j,\mathbf{I}_d\sum_{i=1}^{n}\mu_{i2}^{m}\delta_j w_j,\ldots,\mathbf{I}_d\sum_{i=1}^{n}\mu_{ic}^{m}\delta_j w_j\right)$ and obviously, the Hessian matrix is positive definite. That is, $J(\hat{\mathbf{U}},\mathbf{V},\hat{\mathbf{W}})$ is minimized at $\mathbf{V}^*=(\mathbf{v}_1^*,\ldots,\mathbf{v}_c^*)$ with $v_{kj}^*=\sum_{i=1}^{n}\mu_{ik}^{m}x_{ij}\Big/\sum_{i=1}^{n}\mu_{ik}^{m}$, $\forall k,j$. ■

Proof of Lemma 2: The minimization is $\min J(\mathbf{U},\hat{\mathbf{V}},\hat{\mathbf{W}})=\min\left(\sum_{k=1}^{c}\sum_{i=1}^{n}\sum_{j=1}^{d}\mu_{ik}^{m}\delta_j w_j (x_{ij}-v_{kj})^2+\frac{n}{c}\sum_{j=1}^{d}\left(w_j \log \delta_j w_j\right)\right)$ subject to $\sum_{k=1}^{c}\mu_{ik}=1$, $\forall i=1,\ldots,n$. Since for all i, the constraints $\sum_{k=1}^{c}\mu_{ik}=1$ are all the same, i.e., $g_i(x_1,\ldots,x_n)=\cdots=g_n(x_1,\ldots,x_n)$, we may only consider a fixed i. Thus, $\forall i$, let the Lagrangian function be $L_1=\sum_{k=1}^{c}\sum_{i=1}^{n}\sum_{j=1}^{d}\mu_{ik}^{m}\delta_j w_j (x_{ij}-v_{kj})^2+\lambda_1\left(\sum_{k=1}^{c}\mu_{ik}-1\right)$, where λ_1 is a Lagrangian multiplier. With the gradient of L_1 w.r.t. μ_{ik} and λ_1, we have $\frac{\partial L_1}{\partial \mu_{ik}}=\sum_{j=1}^{d}m\mu_{ik}^{m-1}\delta_j w_j (x_{ij}-v_{kj})^2+\lambda_1=0$ and $\frac{\partial L_1}{\partial \lambda_1}=\sum_{k=1}^{c}\mu_{ik}-1=0$. This implies that

$$\mu_{ik}^*=\left(\sum_{j=1}^{d}\delta_j w_j (x_{ij}-v_{kj})^2\right)^{-1/m-1}\Bigg/\sum_{t=1}^{c}\left(\sum_{j=1}^{d}\delta_j w_j (x_{ij}-v_{tj})^2\right)^{-1/m-1},$$

$\lambda_1^*=-m\mu_{ik}^{m-1}\sum_{j=1}^{d}\delta_j w_j (x_{ij}-v_{kj})^2$ and the "only if" condition is proved. For the proof of the "if" condition, we follow Theorem 2. If $\mathbf{V}=\hat{\mathbf{V}}$ and $\mathbf{W}=\hat{\mathbf{W}}$ are fixed, then

$\frac{\partial L_1}{\partial \mu_{ik}} = \sum_{j=1}^{d} m \mu_{ik}^{m-1} \delta_j w_j (x_{ij} - v_{kj})^2 + \lambda_1$, $\frac{\partial^2 L_1}{\partial \mu_{ik} \partial \mu_{rl}} = \beta_{ir} \beta_{kl} m(m-1) \mu_{ik}^{m-2} \sum_{j=1}^{d} \delta_j w_j (x_{ij} - v_{kj})^2$ and $\frac{\partial^2 L_1}{\partial \mu_{ik} \partial \lambda_1} = \frac{\partial^2 L_1}{\partial \lambda_1 \partial \mu_{ik}} = 1$. Thus, the bordered Hessian matrix w.r.t. μ_{ik} and λ_1 is

$$H_{L_1}(\mu_{ik}, \lambda_1) = \begin{pmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & \frac{\partial^2 L_1}{\partial \mu_{i1} \partial \mu_{i1}} & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & \frac{\partial^2 L_1}{\partial \mu_{ic} \partial \mu_{ic}} \end{pmatrix}.$$

Then, all leading principle minors are checked with

$$\left|\bar{H}_3(\mu_i^*, \lambda_1^*)\right| = \begin{vmatrix} 0 & 1 & 1 \\ 1 & m(m-1)\mu_{i1}^{m-2} \sum_{j=1}^{d} \delta_j w_j (x_{ij} - v_{1j})^2 & 0 \\ 1 & 0 & m(m-1)\mu_{i2}^{m-2} \sum_{j=1}^{d} \delta_j w_j (x_{ij} - v_{2j})^2 \end{vmatrix}_{\mu_i = \mu_i^*, \lambda_1 = \lambda_1^*}$$

$$= -\left(m(m-1)\mu_{i1}^{m-2} \sum_{j=1}^{d} \delta_j w_j (x_{ij} - v_{1j})^2 + m(m-1)\mu_{i2}^{m-2} \sum_{j=1}^{d} \delta_j w_j (x_{ij} - v_{2j})^2 \right)_{\mu_i = \mu_i^*, \lambda_1 = \lambda_1^*} < 0,$$

$$\left|\bar{H}_4(\mu_i^*, \lambda_1^*)\right| = -\left(\sum_{k=1}^{3} \prod_{\substack{l=1 \\ l \neq k}}^{3} m(m-1)\mu_{il}^{m-2} \sum_{j=1}^{d} \delta_j w_j (x_{ij} - v_{lj})^2 \right)_{\mu_i = \mu_i^*, \lambda_1 = \lambda_1^*} < 0$$

, … and

$$\left|\bar{H}_{c+1}(\mu_i^*, \lambda_1^*)\right| = -\left(\sum_{k=1}^{c} \prod_{\substack{l=1 \\ l \neq k}}^{c} m(m-1)\mu_{il}^{m-2} \sum_{j=1}^{d} \delta_j w_j (x_{ij} - v_{lj})^2 \right)_{\mu_i = \mu_i^*, \lambda_1 = \lambda_1^*} < 0.$$

Thus, by Theorem 2, $J(\mathbf{U}, \hat{\mathbf{V}}, \hat{\mathbf{W}})$ subject to $\sum_{k=1}^{c} \mu_{ik} = 1$ is locally minimized at $\mathbf{U}^* = [\mu_{ik}^*]_{n \times c}$ with, , $\forall i, k$

$$\mu_{ik}^* = \left(\sum_{j=1}^{d} \delta_j w_j (x_{ij} - v_{kj})^2 \right)^{-1/m-1} \Bigg/ \sum_{t=1}^{c} \left(\sum_{j=1}^{d} \delta_j w_j (x_{ij} - v_{tj})^2 \right)^{-1/m-1}. \qquad ■$$

Proof of Lemma 3: The updating function of $\mathbf{W}^*$ in the FRFCM algorithm is separated into two cases, that is, without reduction case, i.e., $w_j^* \geq 1/\sqrt{nd}$ and with reduction case, i.e., $w_j^* < 1/\sqrt{nd}$. For the without reduction case, the Lagrangian is $L_2 = \sum_{k=1}^{c} \sum_{i=1}^{n} \sum_{j=1}^{d} \mu_{ik}^m \delta_j w_j (x_{ij} - v_{kj})^2 + \frac{n}{c} \sum_{j=1}^{d} \left(w_j \log \delta_j w_j \right) + \lambda_2 \left(\sum_{j=1}^{d} w_j - 1 \right)$ where λ_2 is a Lagrangian multiplier. With the gradient of L_2 w.r.t. w_j and λ_2, we have $\frac{\partial L_2}{\partial w_j} = \sum_{k=1}^{c} \sum_{i=1}^{n} \mu_{ik}^m \delta_j (x_{ij} - v_{kj})^2 + \frac{n}{c} \left(\log \delta_j w_j + 1 \right) + \lambda_2 = 0$ and $\frac{\partial L_2}{\partial \lambda_2} = \sum_{j=1}^{d} w_j - 1 = 0$. Thus, we have

$$w_j^* = \frac{1}{\delta_j} \exp\left(\frac{-c \sum_{k=1}^{c} \sum_{i=1}^{n} \mu_{ik}^m \delta_j (x_{ij} - v_{kj})^2}{n} \right) \Bigg/ \sum_{p=1}^{d} \frac{1}{\delta_p} \exp\left(\frac{-c \sum_{k=1}^{c} \sum_{i=1}^{n} \mu_{ik}^m \delta_p (x_{ip} - v_{kp})^2}{n} \right)$$

and $\lambda_2 = -\sum_{k=1}^{c} \sum_{i=1}^{n} \mu_{ik}^m \delta_j (x_{ij} - v_{kj})^2 - \frac{n}{c} \left(\log \delta_j w_j + 1 \right)$, and so the "only if" condition is proved. For the proof of the "if" condition, we follow Theorem 2 as follows. If $\mathbf{U} = \hat{\mathbf{U}}$ and $\mathbf{V} = \hat{\mathbf{V}}$ are fixed, then $\frac{\partial L_2}{\partial w_j} = \sum_{k=1}^{c} \sum_{i=1}^{n} \mu_{ik}^m \delta_j (x_{ij} - v_{kj})^2 + \frac{n}{c} \left(\log \delta_j w_j + 1 \right) + \lambda_2$, $\frac{\partial^2 L_2}{\partial w_j \partial w_p} = \beta_{jp} \frac{n}{c w_j}$ and $\frac{\partial^2 L_2}{\partial w_j \partial \lambda_2} = \frac{\partial^2 L_2}{\partial \lambda_2 \partial w_j} = 1$. Thus, the bordered Hessian matrix w.r.t. w and λ_2 is

$$H_{L_2}^*(w, \lambda_2) = \begin{pmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & \frac{\partial^2 L_2}{\partial w_1 \partial w_1} & 0 & \cdots & 0 \\ 1 & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & 0 & & \cdots & \frac{\partial^2 L_2}{\partial w_d \partial w_d} \end{pmatrix}.$$

Note that we only have one constraint, i.e., $t = 1$, so $(-1)^1 = -1 < 0$. Next, all leading principle minors are checked as follows:

$$\left|\bar{H}_3(w^*, \lambda_2^*)\right| = \begin{vmatrix} 0 & 1 & 1 \\ 1 & \frac{n}{c w_1} & 0 \\ 1 & 0 & \frac{n}{c w_2} \end{vmatrix}_{w = w^*, \lambda_2 = \lambda_2^*} = -\left(\frac{n}{c w_1} + \frac{n}{c w_2} \right)_{w = w^*, \lambda_2 = \lambda_2^*} < 0$$

$$\left|\bar{H}_4(w^*, \lambda_2^*)\right| = \begin{vmatrix} 0 & 1 & 1 & 1 \\ 1 & \frac{n}{c w_1} & 0 & 0 \\ 1 & 0 & \frac{n}{c w_2} & 0 \\ 1 & 0 & 0 & \frac{n}{c w_3} \end{vmatrix}_{w = w^*, \lambda_2 = \lambda_2^*}$$

$$= -\left(\frac{n^2}{c^2 w_2 w_3} + \frac{n^2}{c^2 w_1 w_3} + \frac{n^2}{c^2 w_1 w_2} \right)_{w = w^*, \lambda_2 = \lambda_2^*} < 0, \ldots \text{ and}$$

$$\left|\bar{H}_{d+1}(w^*, \lambda_2^*)\right| = \begin{vmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & \frac{n}{c w_1} & 0 & \cdots & 0 \\ 1 & 0 & \frac{n}{c w_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & \frac{n}{c w_d} \end{vmatrix}_{w = w^*, \lambda_2 = \lambda_2^*}$$

$$= -\left(\sum_{j=1}^{d} \prod_{\substack{p=1 \\ p \neq j}}^{d} \frac{n^2}{c^2 w_p^2} \right)_{w = w^*, \lambda_2 = \lambda_2^*} < 0$$

Thus, by Theorem 2, $J(\hat{\mathbf{U}}, \hat{\mathbf{V}}, \mathbf{W})$ subject to $\sum_{j=1}^{d} w_j = 1$ is minimized at $\mathbf{W}^* = [w_j^*]$ with , $\forall j$

$$w_j^* = \frac{1}{\delta_j}\exp\left(\frac{-c\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^m\delta_j(x_{ij}-v_{kj})^2}{n}\right)\Bigg/\sum_{p=1}^{d}\frac{1}{\delta_p}\exp\left(\frac{-c\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^m\delta_p(x_{ip}-v_{kp})^2}{n}\right)$$

The second case is for the reduction case. In this case, the algorithm discards d_r ($0 \le d_r < d$) of some j feature components for $\mathbf{W}^*$, if $w_j^* < 1\big/\sqrt{nd}$. After the reduction process, we set $d^{(new)} = d - d_r$ and have an adjustment for $\mathbf{W}^*$, with $w_{j'}^* = w_{j'}^*\Big/\sum_{p'=1}^{d^{(new)}} w_{p'}^*$, to keep $\sum_{j'=1}^{d^{(new)}} w_{j'} = 1$. Thus the Lagrangian becomes to be

$$L_2 = \sum_{k=1}^{c}\sum_{i=1}^{n}\sum_{j'=1}^{d^{(new)}}\mu_{ik}^m\delta_{j'}w_{j'}(x_{ij'}-v_{kj'})^2 + \frac{n}{c}\sum_{j'=1}^{d^{(new)}}\left(w_{j'}\log\delta_{j'}w_{j'}\right) + \lambda_2\left(\sum_{j'=1}^{d^{(new)}} w_{j'} - 1\right)$$

where λ_2 is a Lagrangian multiplier. For the proof of the "if and only if" condition for L_2, subject to $\sum_{j'=1}^{d^{(new)}} w_{j'} = 1$ is minimized at $\mathbf{W}^* = [w_{j'}^*]_{d^{(new)}\times 1}$ with

$$w_{j'}^* = \frac{1}{\delta_{j'}}\exp\left(\frac{-c\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^m\delta_{j'}(x_{ij'}-v_{kj'})^2}{n}\right)\Bigg/\sum_{p'=1}^{d^{(new)}}\frac{1}{\delta_{p'}}\exp\left(\frac{-c\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^m\delta_{p'}(x_{ip'}-v_{kp'})^2}{n}\right)$$

We can obtain the similar Jacobian matrix and bordered Hessian matrix with gradients of L_2 w.r.t. $w_{j'}$ and λ_2 , where only the dimension is changed from d to $d^{(new)}$. Thus, we also have all leading principle minors are checked with

$$\left|\bar{H}_{d^{(new)}+1}(w^*,\lambda_2^*)\right| = \begin{vmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & \frac{n}{cw_{1'}} & 0 & \cdots & 0 \\ 1 & 0 & \frac{n}{cw_{2'}} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & \frac{n}{cw_{d^{(new)}}} \end{vmatrix}_{w=w^*,\lambda_2=\lambda_2^*} = -\left(\sum_{j'=1}^{d^{(new)}}\prod_{\substack{p'=1\\p'\ne j'}}^{d^{(new)}}\frac{n^2}{c^2w_{p'}^2}\right)_{w=w^*,\lambda_2=\lambda_2^*} < 0$$

Thus, by Theorem 2, $J(\hat{\mathbf{U}},\hat{\mathbf{V}},\mathbf{W})$ subject to $\sum_{j'=1}^{d^{(new)}} w_{j'} = 1$ is minimized at $\mathbf{W}^* = [w_{j'}^*]$ with, $\forall j'$

$$w_{j'}^* = \frac{\frac{1}{\delta_{j'}}\exp\left(\frac{-c\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^m\delta_{j'}(x_{ij'}-v_{kj'})^2}{n}\right)}{\sum_{p'=1}^{d^{(new)}}\frac{1}{\delta_{p'}}\exp\left(\frac{-c\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^m\delta_{p'}(x_{ip'}-v_{kp'})^2}{n}\right)}. \quad \blacksquare$$

Proof of Lemma 4: Since $\left\{v_{kj} \to \left\|x_{ij}-v_{kj}\right\|^2\right\}$, $\left\{\mu_{ik} \to \mu_{ik}^m\right\}$, and $\left\{w_j \to w_j\log\delta_j w_j\right\}$ are continuous, the sum of products of $\left\{v_{kj} \to \left\|x_{ij}-v_{kj}\right\|^2\right\}$ and $\left\{\mu_{ik} \to \mu_{ik}^m\right\}$ is continuous, and the sum of products of $\left\{\mu_{ik} \to \mu_{ik}^m\right\}$ and $\left\{w_j \to w_j\log\delta_j w_j\right\}$ is also continuous. Therefore, J is continuous on $M_{fcn}\times(\mathbb{R}^d)^c\times M_w$. ■

Proof of Lemma 5: Let $(\mathbf{U},\mathbf{V},\mathbf{W}) \notin \Omega_{FRFCM}$. Then $J\left(T(\mathbf{U},\mathbf{V},\mathbf{W})\right) = J\left(A_2\circ A_1(\mathbf{U},\mathbf{V},\mathbf{W})\right) = J(\mathbf{U}^*,\mathbf{V}^*,\mathbf{W}^*) < J(\mathbf{U},\mathbf{V}^*,\mathbf{W}^*)$ by Lemma 2, and then $< J(\mathbf{U},\mathbf{V},\mathbf{W}^*)$ by Lemma 1, and finally $< J(\mathbf{U},\mathbf{V},\mathbf{W})$ by Lemma 3. That is, $J\left(T(\mathbf{U},\mathbf{V},\mathbf{W})\right) = J(\mathbf{U}^*,\mathbf{V}^*,\mathbf{W}^*) < J(\mathbf{U},\mathbf{V},\mathbf{W})$ for any $(\mathbf{U},\mathbf{V},\mathbf{W}) \notin \Omega_{FRFCM}$. ■

Proof of Lemma 6: We have that $A_1(\mathbf{U},\mathbf{V},\mathbf{W}) = E(\mathbf{V},\mathbf{W}) = \left(E_{11}(\mathbf{V},\mathbf{W}),E_{21}(\mathbf{V},\mathbf{W}),\ldots,E_{nc}(\mathbf{V},\mathbf{W})\right)$, where

$$E_{ik}(\mathbf{V},\mathbf{W}) = \frac{\left(\sum_{j=1}^{d}\delta_j w_j(x_{ij}-v_{kj})^2\right)^{-1/m-1}}{\sum_{t=1}^{c}\left(\sum_{j=1}^{d}\delta_j w_j(x_{ij}-v_{tj})^2\right)^{-1/m-1}} = \mu_{ik}$$

. Since $\left\{v_{kj} \to \left\|x_{ij}-v_{kj}\right\|^2\right\}$, $\left\{w_j \to w_j\log\delta_j w_j\right\}$, and $\left\{\sum_{j=1}^{d}\delta_j w_j(x_{ij}-v_{kj})^2 \to \left(\sum_{j=1}^{d}\delta_j w_j(x_{ij}-v_{kj})^2\right)^{-1/m-1}\right\}$ are continuous, and the sum of continuous functions is continuous, $E_{ik}(\mathbf{V},\mathbf{W}) = \mu_{ik}$, the quotient of two continuous, is also continuous. Therefore, $E(\mathbf{V},\mathbf{W}) = \left(E_{11}(\mathbf{V},\mathbf{W}),E_{21}(\mathbf{V},\mathbf{W}),\ldots,E_{nc}(\mathbf{V},\mathbf{W})\right)$ is continuous. Furthermore, $A_2(\mathbf{U}) = \left(\mathbf{U},F(\mathbf{U}),G(\mathbf{U},F(\mathbf{U}))\right)$ where $F(\mathbf{U}) = \left(F_{11}(\mathbf{U}),\ldots,F_{cd}(\mathbf{U})\right)$, $G(\mathbf{U},\mathbf{V}) = \left(G_1(\mathbf{U},\mathbf{V}),\ldots,G_d(\mathbf{U},\mathbf{V})\right)$ with $F_{kj}(\mathbf{U}) = \sum_{i=1}^{n}\mu_{ik}^m x_{ij}\Big/\sum_{i=1}^{n}\mu_{ik}^m$ and

$$G_j(\mathbf{U},\mathbf{V}) = \frac{1}{\delta_j}\exp\left(\frac{-c\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^m\delta_j(x_{ij}-v_{kj})^2}{n}\right)\Bigg/\sum_{p=1}^{d}\frac{1}{\delta_p}\exp\left(\frac{-c\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}^m\delta_p(x_{ip}-v_{kp})^2}{n}\right).$$

Since $\left\{\mu_{ik} \to \mu_{ik}^m\right\}$ and $\left\{\mu_{ik} \to \mu_{ik}^m x_{ij}\right\}$ are continuous, and the sum of continuous function is continuous, $F_{kj}(\mathbf{U})$ and $G_j(\mathbf{U},\mathbf{V})$ with the quotient of two continuous is also continuous. Therefore, $F(\mathbf{U}) = \left(F_{11}(\mathbf{U}),\ldots,F_{cd}(\mathbf{U})\right)$ and $G(\mathbf{U},\mathbf{V}) = \left(G_1(\mathbf{U},\mathbf{V}),\ldots,G_d(\mathbf{U},\mathbf{V})\right)$ are continuous. Thus, $T = A_2\circ A_1$ is continuous on $M_{fcn}\times(\mathbb{R}^d)^c\times M_w$. ■

Proof of Lemma 7: Let $\left(E(\mathbf{V}^{(0)},\mathbf{W}^{(0)}),\mathbf{V}^{(0)},\mathbf{W}^{(0)}\right)$ be the starting point of iteration with T , where $E(\mathbf{V}^{(0)},\mathbf{W}^{(0)}) = \left(E_{11}(\mathbf{V}^{(0)},\mathbf{W}^{(0)}),E_{21}(\mathbf{V}^{(0)},\mathbf{W}^{(0)}),\ldots,E_{nc}(\mathbf{V}^{(0)},\mathbf{W}^{(0)})\right)$ with

$$E_{ik}(\mathbf{V}^{(0)},\mathbf{W}^{(0)}) = \frac{\left(\sum_{j=1}^{d}\delta_j w_j^{(0)}(x_{ij}-v_{kj}^{(0)})^2\right)^{-1/m-1}}{\sum_{t=1}^{c}\left(\sum_{j=1}^{d}\delta_j w_j^{(0)}(x_{ij}-v_{tj}^{(0)})^2\right)^{-1/m-1}} = \mu_{ik}^{(0)}.$$

Then $v_{kj}^{(1)} = F_{kj}(\mathbf{U}^{(0)}) = \sum_{i=1}^{n}(\mu_{ik}^{(0)})^m x_{ij}\Big/\sum_{i=1}^{n}(\mu_{ik}^{(0)})^m$. Let $d_{ik} = (\mu_{ik}^{(0)})^m\Big/\sum_{q=1}^{n}(\mu_{qk}^{(0)})^m, \forall k$. Thus, $0 \le d_{ik} \le 1, \forall i,k$ and

$v_{kj}^{(1)} = \sum_{i=1}^{n} d_{ik} x_{ij}$ with $\sum_{i=1}^{n} d_{ik} = \sum_{i=1}^{n} \frac{(\mu_{ik}^{(0)})^m}{\sum_{q=1}^{n} (\mu_{qk}^{(0)})^m} = \frac{\sum_{i=1}^{n} (\mu_{ik}^{(0)})^m}{\sum_{q=1}^{n} (\mu_{qk}^{(0)})^m} = 1$.

Therefore, $v_{kj}^{(1)} \in [\mathrm{conv}(\mathbf{X})]$ and $\mathbf{V}^{(1)} \in [\mathrm{conv}(\mathbf{X})]^c$. Continuing recursively, $\mathbf{V}^{(t)} \in [\mathrm{conv}(\mathbf{X})]^c, \forall t \geq 1$. Obviously,

$$w_j^{(1)} = G_j(\mathbf{U}^{(0)}, \mathbf{V}^{(1)}) = \frac{\frac{1}{\delta_j} \exp\left(\frac{-c \sum_{k=1}^{c} \sum_{i=1}^{n} (\mu_{ik}^{(0)})^m \delta_j (x_{ij} - v_{kj}^{(1)})^2}{n} \right)}{\sum_{p=1}^{d} \frac{1}{\delta_p} \exp\left(\frac{-c \sum_{k=1}^{c} \sum_{i=1}^{n} (\mu_{ik}^{(0)})^m \delta_p (x_{ip} - v_{kp}^{(1)})^2}{n} \right)}$$

and $w^{(1)} \in M_w$,

$$\mu_{ik}^{(1)} = E_{ik}(\mathbf{V}^{(1)}, \mathbf{W}^{(1)}) = \frac{\left(\sum_{j=1}^{d} \delta_j w_j^{(1)} (x_{ij} - v_{kj}^{(1)})^2 \right)^{-\frac{1}{m-1}}}{\sum_{t=1}^{c} \left(\sum_{j=1}^{d} \delta_j w_j^{(1)} (x_{ij} - v_{tj}^{(1)})^2 \right)^{-\frac{1}{m-1}}} \in M_{fcn}$$

and $\mathbf{U}^{(1)} \in M_{fcn}$. Continuing recursively, $\mathbf{W}^{(t)} \in M_w$ and $\mathbf{U}^{(t)} \in M_{fcn}, \forall t \geq 1$. Thus, $(T)^{(t)} \left(E(\mathbf{V}^{(0)}, \mathbf{W}^{(0)}), \mathbf{V}^{(0)}, \mathbf{W}^{(0)} \right) \in M_{fcn} \times [\mathrm{conv}(\mathbf{X})]^c \times M_w, \forall t.$ We next prove that $M_{fcn} \times [\mathrm{conv}(\mathbf{X})]^c \times M_w$ is compact in $M_{fcn} \times (\mathbb{R}^d)^c \times M_w$. Since $\mathbf{X}$ is finite, each $\mathbf{x}_i \in \mathbf{X}$ has finite components. Therefore, the diameter of $\mathbf{X}$, that is equal to diameter of $\mathrm{conv}(\mathbf{X})$, is bounded. Since $\mathrm{conv}(\mathbf{X})$ is the convex hull of finitely many generators $\mathbf{x}_i$, it is closed. Thus, $\mathrm{conv}(\mathbf{X})$ is bounded and closed in $\mathbb{R}^d$, and so $\mathrm{conv}(\mathbf{X})$ is compact. Based on the generalized Heine-Borel theorem, we have that $[\mathrm{conv}(\mathbf{X})]^c$ is also compact. For claiming that M_{fcn} and M_w are compact, let us consider $M_{hcn} = \left\{ \mathbf{U} = [\mu_{ik}]_{n \times c} \middle| \sum_{k=1}^{c} \mu_{ik} = 1, \mu_{ik} \in \{0,1\} \right\}$ and $M_{w0} = \left\{ \mathbf{W} = [w_j]_{d \times 1} \middle| \sum_{j=1}^{d} w_j = 1, w_j \in \{0,1\} \right\}$. An argument is similar as that every respect $(M_{fcn} = \mathrm{conv}(M_{hcn})\ \&\ M_w = \mathrm{conv}(M_{w0}))$ given by Ball and Hall [45] establishes compactness of M_{fcn} and M_w . Thus, we have $M_{fcn} \times [\mathrm{conv}(\mathbf{X})]^c \times M_w$ is compact. Furthermore, for feature-reduction process, the dimensions of $\mathbf{X}$ and $\mathbf{V}$ are reduced to $d^{(new)}$ with $\mathbf{X}^{(new)} = \{\mathbf{x}_1^{(new)}, \ldots, \mathbf{x}_n^{(new)}\}$ and $\mathbf{V}^{(new)} = \{\mathbf{v}_1^{(new)}, \ldots, \mathbf{v}_n^{(new)}\}$. Since $\mathbf{X}$ and $\mathbf{V}$ are bounded, $\mathbf{X}^{(new)}$ and $\mathbf{V}^{(new)}$ are also bounded. Similarly, we can claim that $M_{fcn} \times [\mathrm{conv}(\mathbf{X}^{(new)})]^c \times M_w$ is compact in $M_{fcn} \times (\mathbb{R}^{d^{(new)}})^c \times M_w$. After replacing d with $d^{(new)}$, then $(T)^{(t)} \left(E(\mathbf{V}^{(0)}, \mathbf{W}^{(0)}), \mathbf{V}^{(0)}, \mathbf{W}^{(0)} \right) \in M_{fcn} \times [\mathrm{conv}(\mathbf{X})]^c \times M_w$ is compact in $M_{fcn} \times (\mathbb{R}^d)^c \times M_w$. ■

REFERENCES

[1] L. Zadeh, "Fuzzy Sets," *Information and Control,* vol. 8, no. 3, pp. 338-353, 1965.

[2] E. H. Ruspini, "A new approach to clustering," *Information and Control,* vol. 15, no. 1, pp. 22-32, 1969.

[3] J. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics,* vol. 3, no. 3, pp. 32-57, 1973.

[4] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Massachusetts: Kluwer Academic Publishers, 1981.

[5] M. S. Yang, "A survey of fuzzy clustering," *Mathematical and Computer Modelling,* vol. 18, no. 11, pp. 1-16, 1993.

[6] A. Baraldi and P. Blonda, "A survey of fuzzy clustering algorithms for pattern recognition-part I and II," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 29, no. 6, pp. 778-801, 1999.

[7] F. Höppner, F. Klawonn, R. Kruse and T. Runkler, Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition, New York: Wiley, 1999.

[8] M.S. Yang, W.L. Hung and F.J. Cheng, "Mixed-variable fuzzy clustering approach to part family and machine cell formation for GT applications," *International Journal of Production Economics*, vol. 103, no. 1, pp. 185-198, 2006.

[9] M. S. Yang, K. L. Wu, J. N. Hsieh and J. Yu, "Alpha-cut implemented fuzzy clustering algorithms and switching regressions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B,* vol. 38, no. 3, pp. 588-603, 2008.

[10] L. F. S. Coletta, L. Vendramin, E. R. Hruschka, R. J. G. B. Campello and W. Pedrycz, "Collaborative fuzzy clustering algorithms: Some refinements and design guidelines," *IEEE Transactions on Fuzzy Systems,* vol. 20, no. 3, pp. 444-462, 2012.

[11] M. Gong, L. Su, M. Jia and W. Chen, "Fuzzy clustering with a modified MRF energy function for change detection in synthetic aperture radar images," *IEEE Transactions on Fuzzy Systems,* vol. 22, no. 1, pp. 98-109, 2014.

[12] S. T. Chang, K. P. Lu and M. S. Yang, "Fuzzy change-point algorithms for regression models," *IEEE Transactions on Fuzzy Systems,* vol. 23, no. 6, pp. 2343-2357, 2015.

[13] P. Fazendeiro and J. V. D. Oliveira, "Observer-Biased Fuzzy Clustering," *IEEE Transactions on Fuzzy*

Systems, vol. 23, no. 1, pp. 85-97, 2015.

[14] Z. Deng, Y. Jiang, F.L. Chung, H. Ishibuchi, K. Choi and S. Wang, "Transfer prototype-based fuzzy clustering," *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 5, pp. 1210-1232, 2016.

[15] J. Z. Huang, M. K. Ng, H. Rong and Z. Li, "Automated Variable Weighting in k-Means Type Clustering," *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 27, no. 5, pp. 657-668, 2005.

[16] L. Jing, M. K. Ng and J. Z. Huang, "An Entropy Weighting k-Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data," *IEEE Transactions on Knowledge and Data Engineering,* vol. 19, no. 8, pp. 1026-1041, 2007.

[17] D. M. Witten and R. Tibshirani, "A Framework for Feature Selection in Clustering," *Journal of the American Statistical Association,* vol. 105, no. 490, pp. 713-726, 2010.

[18] X. Wang, Y. Wang and L. Wang, "Improving fuzzy c-means clustering based on feature weight learning," *Pattern Recognition Letters,* vol. 25, no. 10, pp. 1123-1132, 2004.

[19] H. Frigui and O. Nasraoui, "Unsupervised learning of prototypes and attribute weights," *Pattern Recognition,* vol. 37, no. 3, pp. 567-581, 2004.

[20] D. Yeung and X. Wang, "Improving performance of similarity-based clustering by feature weight learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, no. 4, pp. 556-561, 2002.

[21] D. R. Cox and P. A. Lewis, The Statistical Analysis of Series of Events, London: Methuen & Co., 1966.

[22] Z. Bai, K. Wang and W. K. Wong, "Mean-Variance Ratio Test, A Complement to Coefficient of Variation Test and Sharpe Ratio Test," *Statistics & Probability Letters,* vol. 81, no. 8, pp. 1078-1085, 2011.

[23] W. I. Zangwill, Nonlinear programming: a unified approach, Englewood Cliffs, NJ: Prentice Hall, 1969.

[24] F. Werner and Y. N. Sotskov, Mathematics of Economics and Business, London and New York: Routledge, Taylor & Francis Group, 2006.

[25] M. S. Yang and Y. C. Tian, "Bias-correction fuzzy clustering algorithms," *Information Sciences,* vol. 309, pp. 138-162, 2015.

[26] https://archive.ics.uci.edu/ml/datasets.html, UCI data set.

[27] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack and A. Levine, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," in *Proceedings of the National Academy of Sciences of the USA*, 1999.

[28] W.M. Rand, Objective criteria for the evaluation of clustering methods, *J. Amer. Stat. Assoc.*, vol. 66, pp. 846-850, 1971.

[29] E.B. Fowlkes and C.L. Mallows, "A method for comparing two hierarchical clusterings," *J. Amer. Stat. Assoc.*, vol. 78, pp. 553-569, 1983.

[30] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, pp. 193-218, 1985.

[31] Y. Yin and K. Yasuda, "Similarity coefficient methods applied to the cell formation problem: A taxonomy and review," *Int. J. Production Economics*, vol. 101, pp. 329-352, 2006.

[32] D. T. Anderson, J. C. Bezdek, M. Popescu and J. M. Keller, "Comparing fuzzy, probabilistic, and possibilistic partitions," *IEEE Transactions on Fuzzy Systems,* vol. 18, no. 5, pp. 906-918, 2010.

[33] C.C. Yeh and M.S. Yang, Evaluation measures for cluster ensembles based on a fuzzy generalized Rand index. Applied Soft Computing, 2017, DOI: 10.1016/j.asoc.2017.03.030. (Accepted & in press)

[34] P. Jaccard, "Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines". *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 241-272, 1901.

[35] C.C. Yeh and M.S. Yang, A generalization of Rand and Jaccard indices with its fuzzy extension, International Journal of Fuzzy Systems, vol. 18, pp. 1008–1018, 2016.

[36] C.H. Coombs, R.M. Dawes, A. Tversky, Mathematical Psychology: An Elementary Introduction. Englewood Cliffs, NJ: Prentice-Hall, 1970.

[37] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html. Olivetti face data set.

[38] J. C. Bezdek, "Numerical taxonomy with fuzzy sets," *Journal of Mathematical Biology,* vol. 1, no. 1, pp. 57-71, 1974.

[39] J. Dunn, "Indices of partition fuzziness and the detection of clusters in large data sets," in *M. Gupta and G. Saridis (Eds.), Fuzzy Automata and Decision Processes*, New York, Elsevier, 1976, pp. 271-284.

[40] J. C. Bezdek, "Cluster validity with fuzzy sets," *Journal of Cybernetics,* vol. 3, no. 3, pp. 58-73, 1974.

[41] X. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 13, no. 8, pp. 841-847, 1991.

[42] L. J. Hubert and J. R. Levin, "A general statistical framework for assessing categorical clustering in free recall," *Psychological Bulletin,* vol. 83, no. 6, pp. 1072-1080, 1976.

[43] Chaomurilige, J. Yu, and M. S. Yang, Analysis of parameter selection for Gustafson-Kessel fuzzy clustering using Jacobian matrix, *IEEE Transactions on Fuzzy Systems*, vol. 23, no. 6, pp. 2329−2342, 2015.

[44] C. H. Wu, C. S. Ouyang, L. W. Chen and L. W. Lu, "A new fuzzy clustering validity index with a median factor for centroid-based clustering," *IEEE Transactions on Fuzzy Systems,* vol. 23, no. 3, pp. 701-718, 2015.

[45] G. H. Ball and D. J. Hall, "A clustering technique for summarizing multivariate data," *Systems Research and Behavioral Science,* vol. 12, no. 2, pp. 153-155, 1967.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2017.2692203, IEEE Transactions on Fuzzy Systems

Miin-Shen Yang received the BS degree in mathematics from the Chung Yuan Christian University, Chung-Li, Taiwan, in 1977, the MS degree in applied mathematics from the National Chiao-Tung University, Hsinchu, Taiwan, in 1980, and the PhD degree in statistics from the University of South Carolina, Columbia, USA, in 1989.

In 1989, he joined the faculty of the Department of Mathematics in the Chung Yuan Christian University (CYCU) as an Associate Professor, where, since 1994, he has been a Professor. From 1997 to 1998, he was a Visiting Professor with the Department of Industrial Engineering, University of Washington, Seattle, USA. During 2001-2005, he was the Chairman of the Department of Applied Mathematics in CYCU. During 2012-2016, he was the Director of Chaplain's Office in CYCU. Since 2012, he has been a Distinguished Professor of the Department of Applied Mathematics in CYCU. His research interests include applications of statistics, control charts, fuzzy clustering, pattern recognition, and machine learning.

Dr. Yang was an Associate Editor of the IEEE Transactions on Fuzzy Systems (2005-2011), and is an Associate Editor of the Applied Computational Intelligence & Soft Computing, and Editor-in-Chief of Advances in Computational Research.

Yessica Nataliani received B.S. degree in mathematics from Gadjah Mada University, Yogyakarta, Indonesia, in 2004 and M.S. degree in computer science at the same university in 2006. She is currently a Ph.D. student at Department of Applied Mathematics, Chung Yuan Christian University, Taiwan. Her research interests include cluster analysis and pattern recognition.