



Analysis of parameter selections for fuzzy *c*-means

Kuo-Lung Wu

Department of Information Management, Kun Shan University, Yunk-Kang, Tainan, Taiwan 71023, ROC

ARTICLE INFO

Article history:

Received 3 September 2010

Received in revised form

22 December 2010

Accepted 10 July 2011

Available online 20 July 2011

Keywords:

Fuzzy clustering

Fuzzy *c*-means

Cluster core

ABSTRACT

The weighting exponent m is called the fuzzifier that can influence the performance of fuzzy *c*-means (FCM). It is generally suggested that $m \in [1.5, 2.5]$. On the basis of a robust analysis of FCM, a new guideline for selecting the parameter m is proposed. We will show that a large m value will make FCM more robust to noise and outliers. However, considerably large m values that are greater than the theoretical upper bound will make the sample mean a unique optimizer. A simple and efficient method to avoid this unexpected case in fuzzy clustering is to assign a cluster core to each cluster. We will also discuss some clustering algorithms that extend FCM to contain the cluster cores in fuzzy clusters. For a large theoretical upper bound case, we suggest the implementation of the FCM with a suitable large m value. Otherwise, we suggest implementing the clustering methods with cluster cores. When the data set contains noise and outliers, the fuzzifier $m=4$ is recommended for both FCM and cluster-core-based methods in a large theoretical upper bound case.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

The fuzzy *c*-means (FCM) algorithm [1,2] is the best-known method for fuzzy clustering. Although FCM is a good clustering algorithm, many extensions to the FCM algorithm have been proposed in the literature. Overall, these extended types of FCM can be divided into two categories. One is to extend the dissimilarity measure $d(x_j, a_i)$ between the data point x_j and the cluster center a_i in the FCM objective function by replacing the Euclidean distance with the other types of metric measures (see Refs. [3–7]). The other is to extend the FCM objective function by adding a penalized term (see Refs. [8–12]).

On the other hand, another important factor that influences the effectiveness of FCM is the weighting exponent m , which has been well investigated by Pal and Bezdek [13] and Yu et al. [14]. Pal and Bezdek [13] suggested taking $m \in [1.5, 2.5]$, and Yu et al. [14] proposed a theoretical upper bound for m that can prevent the sample mean from being the unique optimizer of an FCM objective function. In this paper, we will discuss the robust properties of FCM and show that the parameter m influences the robustness of FCM. According to our analysis, we find that a large m value will make FCM more robust to noise and outliers. However, a considerably large m value will make the sample mean the unique optimizer (see Ref. [14]). We know that FCM extends the *k*-means in order to allow its membership function $\mu_{ij} \in [0, 1]$. However, almost no data points have a crisp membership $\mu_{ij} \in \{0, 1\}$. That is, in fact, FCM always produces fuzzy memberships μ_{ij} in the open interval $(0, 1)$. This is the main reason for the sample mean to be the unique optimizer of FCM

when the fuzzifier m is larger than the theoretical upper boundary. A simple and efficient method to avoid this special case in FCM is to assign a cluster core to each cluster.

Özdemir and Akarun [8] proposed a partition index maximization (PIM) algorithm by adding the partition coefficient (PC) [15] into the FCM objective function and successfully applied PIM for a color quantization of images. This modification can form a cluster core for each cluster, and data points inside the cluster core will have membership values of zero or one. Since the volumes of each cluster core in PIM are equal, Wu et al. [12] proposed a fuzzy compactness and separation (FCS) algorithm. The volumes of each cluster core generated by FCS are different. Moreover, Yang et al. [16] proposed the alpha-cut implemented fuzzy clustering algorithm (FCM α) whose cluster cores are generated by the α -cut concept. That is, if the membership value of the data point that belongs to one cluster is larger than a given value α , then the point will exactly belong to that cluster with a membership value of 1. The volumes of each cluster core in FCM α are also different. Although these algorithms can avoid the sample mean from being the unique optimizer, they do not directly optimize their objective function. We also give an example to illustrate that FCM can produce a better clustering result than these algorithms in a large theoretical upper bound case. We know that a good clustering method should achieve the robustness properties (see Refs. [17–19]). Therefore, in this paper, we will also discuss the robust properties of FCM, PIM, FCS, and FCM α .

This paper is organized as follows: In Section 2, we review the FCM and discuss the parameter selections and robust properties. In Section 3, we give a detailed discussion of cluster-core-based algorithms and give a unified definition to control their cluster core volumes. We also discuss the robust properties of these cluster-core-based algorithms. In Section 4, we use some real and numerical

E-mail address: klwu@mail.ksu.edu.tw

examples to discuss the parameter selection problems and the performance of FCM, PIM, FCS, and FCM α . Conclusions are presented in Section 5.

2. Parameter selections and robust analysis of fuzzy c-means

Let $X = \{x_1, \dots, x_n\}$ be a data set in an s -dimensional space R^s . Let c be a positive integer with $1 < c < n$. A partition of X into c parts can be presented by a mutually disjoint set $X_1, \dots, X_c$ such that $X_1 \cup \dots \cup X_c = X$, or equivalently by the indicator functions $\mu_1, \dots, \mu_c$ such that $\mu_{ij} = \mu_i(x_j) = 1$ if $x_j \in X_i$ and $\mu_{ij} = 0$ if $x_j \notin X_i$ for $i = 1, \dots, c$ and $j = 1, \dots, n$. The set of indicator functions $\{\mu_1, \dots, \mu_c\}$ is called a hard c -partition of clustering X into c clusters. Now, consider an extension to allow $\mu_{ij} = \mu_i(x_j) \in [0, 1]$ to be a membership function of fuzzy sets μ_i on X such that $\sum_{i=1}^c \mu_{ij} = 1$ for all x_j and $\{\mu_1, \dots, \mu_c\}$ is called a fuzzy c -partition. In this section, we will give a brief review of the best-known fuzzy c -means (FCM) clustering method and discuss the influence of the fuzzifier m on the robustness of FCM.

2.1. Fuzzy c-means clustering algorithm and parameter selections

In unsupervised learning clustering literatures, the fuzzy c -means (FCM) algorithm is a well-known fuzzy clustering method with the objective function

$$J_{FCM}(\mu, a) = \sum_{i=1}^c \sum_{j=1}^n \mu_{ij}^m d(x_j, a_i) = \sum_{i=1}^c \sum_{j=1}^n \mu_{ij}^m \|x_j - a_i\|^2. \quad (1)$$

The weighting exponent $m > 1$ is a fuzziness index, $\mu = \{\mu_1, \dots, \mu_c\}$ with $\mu_{ij} = \mu_i(x_j)$ is a fuzzy c -partition, and $a = \{a_1, \dots, a_c\}$ is a set of c cluster centers. The necessary conditions for a minimizer (μ, a) of J_{FCM} are the following update equations:

$$\mu_{ij} = \frac{\|x_j - a_i\|^{-2/(m-1)}}{\sum_{k=1}^c \|x_j - a_k\|^{-2/(m-1)}} \quad (2)$$

and

$$a_i = \frac{\sum_{j=1}^n \mu_{ij}^m x_j}{\sum_{j=1}^n \mu_{ij}^m}, \quad (3)$$

where $j = 1, \dots, n$, $i = 1, \dots, c$. Note that $d(x_j, a_i) = \|x_j - a_i\|^2$ is used the most. However, other types of metric $d(x_j, a_i)$ may be used for improving the usage and effectiveness of FCM (see Refs. [3–7]).

Note that a widely unified approach to generalizing FCM had been studied in [3]. On the other hand, another important factor that influences the effectiveness of FCM is the fuzziness index m , which has previously been thoroughly investigated in Pal and Bezdek [13] and Yu et al. [14].

The weighting exponent m is called the fuzzifier; it can influence the clustering performance of FCM. The influence of the weighting exponent m on the FCM membership function is shown in Figs. 1(a)–(d). These figures are produced by assuming that there are only two clusters with centers 0 and 2. The curves with different m values are the membership functions that belong to the cluster with center 0. When $m = 1$, the FCM will reduce to the traditional hard c -means. When m tends to infinity, $\mu_{ij} = 1/c$ for all i, j , and the sample mean will be a unique optimizer of the FCM objective function. In fact, this situation may occur for any specified m value, and Yu et al. [14] proposed a theoretical upper bound for m that can prevent the sample mean from being the unique optimizer of the FCM objective function. The rule is that $\forall i, a_i = \bar{x}$ is stable for FCM if $\lambda_{\max}(C_X) < 0.5$ and $m \geq (1 - 2\lambda_{\max}(C_X))^{-1}$, where $C_X = \sum_{j=1}^n (x_j - \bar{x})(x_j - \bar{x})^T / n \|x_j - \bar{x}\|^2$ and $\lambda_{\max}(C_X)$ is the maximum eigenvalue of the matrix C_X . Therefore, for FCM, we should set $m < (1 - 2\lambda_{\max}(C_X))^{-1}$ if $\lambda_{\max}(C_X) < 0.5$. If $\lambda_{\max}(C_X) \geq 0.5$, we can take m to be any positive value; further, Pal and Bezdek [13] have suggested that $m \in [1.5, 2.5]$.

Table 1 (refer to Yu et al. [14]) shows the upper limit of m for some data sets obtained from the UCI Repository of Machine Learning Databases. For the data sets with $\lambda_{\max}(C_X) < 0.5$, the sample mean will be the unique optimizer of FCM when the fuzzifier $m \geq (1 - 2\lambda_{\max}(C_X))^{-1}$ (the upper limit). In the Iris data set [17,18], since the maximum eigenvalue of the matrix C_X is larger than 0.5, the upper limit of m for the Iris data set is positive infinity. Yang and Wu [11] also confirmed a part of these results. We now give a simple example to demonstrate the abovementioned properties.

We implement the Normal-4 data set to test the influence of m on the performances of FCM. Pal and Bezdek [13] proposed the Normal-4 data set, which is a 4-dimensional data set with the sample size $n = 800$; each of the four clusters contains 200 points. The population mean vectors are $c_1 = (3, 0, 0, 0)$, $c_2 = (0, 3, 0, 0)$, $c_3 = (0, 0, 3, 0)$, and $c_4 = (0, 0, 0, 3)$. The covariance matrix for each population is the identity matrix I_4 . We randomly generate 100 Normal-4 data sets and implement the FCM algorithm for each one of the data sets with the parameters $m = 1.5, 2, 2.5, 3$, and 3.5 . We then calculate the average MSE and the average number of iterations for these 100 randomly

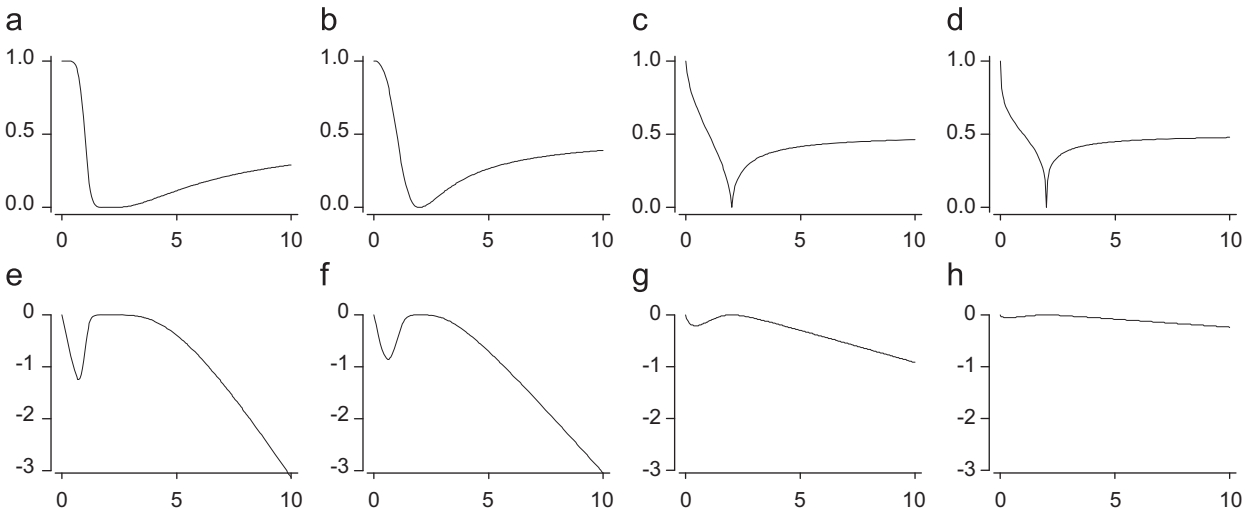


Fig. 1. The membership functions and ϕ functions of FCM. (a), (b), (c), and (d) are membership functions with different fuzzifier $m = 1.5, 2, 4$, and 6 , respectively. (e), (f), (g), and (h) are ϕ functions with different fuzzifier $m = 1.5, 2, 4$, and 6 , respectively. The membership function and ϕ function are calculated by Eqs. (2) and (5) and the x -axis denotes the location of an observation x_j .

generated Normal-4 data sets. The mean squared error (MSE) is calculated by $(1/4) \sum_{i=1}^4 \|a_i - c_i\|^2$, where a_i is the cluster center output from the FCM algorithm. The results are shown in Table 2. Yu et al. [14] concluded that the theoretical valid m for a random Normal-4 data set should not be greater than 2.6; the sample mean in this case is an optimizer with approximately 50% probability.

Table 1

The upper limit of m for the data sets.

Data set	Number of samples	Number of features	Number of clusters	$\lambda_{\max}$ (Cx)	Upper limit of m
Vowel	990	10	11	0.2189	1.7787
Waveform	5000	21	3	0.3272	2.8935
Glass	214	9	6	0.3424	3.1726
Iris	150	4	3	0.6652	$+\infty$

Table 2

Mean square error (MSE) and number of iterations of 100 randomly generated normal-4 data sets.

	MSE			Number of iterations		
	Best case	Average	Worst case	Best case	Average	Worst case
$m=1.5$	0.05791	0.14554	0.33201	10	14.93	23
$m=2$	0.16153	0.45912	0.91060	13	20.20	35
$m=2.5$	3.87754	5.95604	6.52174	50	50.00	50
$m=3$	6.67529	6.74813	6.79683	36	42.24	50
$m=3.5$	6.73916	6.75495	6.78513	27	30.80	36

In Table 2, when $m=2.5$, the worst case is when the sample mean is the unique optimizer with $MSE=6.52$. When $m=3$ and 3.5, the FCM always produces the unique optimizer $\bar{x}$. These results are coincident to Yu et al. [14]. Although a considerably large m may cause some problems in FCM, a suitably large m value can make FCM more robust to noise and outliers. We will discuss this property in the next subsection.

2.2. Robust analysis of fuzzy c-means

The influence curve (IC) can help us to assess the relative influence of an individual observation toward the value of an estimate. The influence function of an M-estimator is proportional to its φ function [17]. If the influence function of an estimator is unbounded, an outlier might lead to problems when the φ function is used for denoting the degree of the influence. Let the loss between the data point x_j and the cluster center a_i be

$$\rho_{FCM}(x_j - a_i) = \mu_{ij}^m \|x_j - a_i\|^2 \quad (4)$$

and

$$\varphi_{FCM}(x_j - a_i) = \frac{d}{da_i} \rho(x_j - a_i) = -2\mu_{ij}^m (x_j - a_i). \quad (5)$$

By solving the equation $\sum_{j=1}^n \varphi_{FCM}(x_j - a_i) = 0$, we have the result shown in Eq. (3). Therefore, the FCM cluster center estimate is an M-estimator with the loss function (4) and φ function (5). Refer to Figs. 1(a)–(d), the corresponding φ functions with different m are illustrated in Figs. 1(e)–(h). The influences of adding a point will become very small when m is large. That is, FCM can be robust to

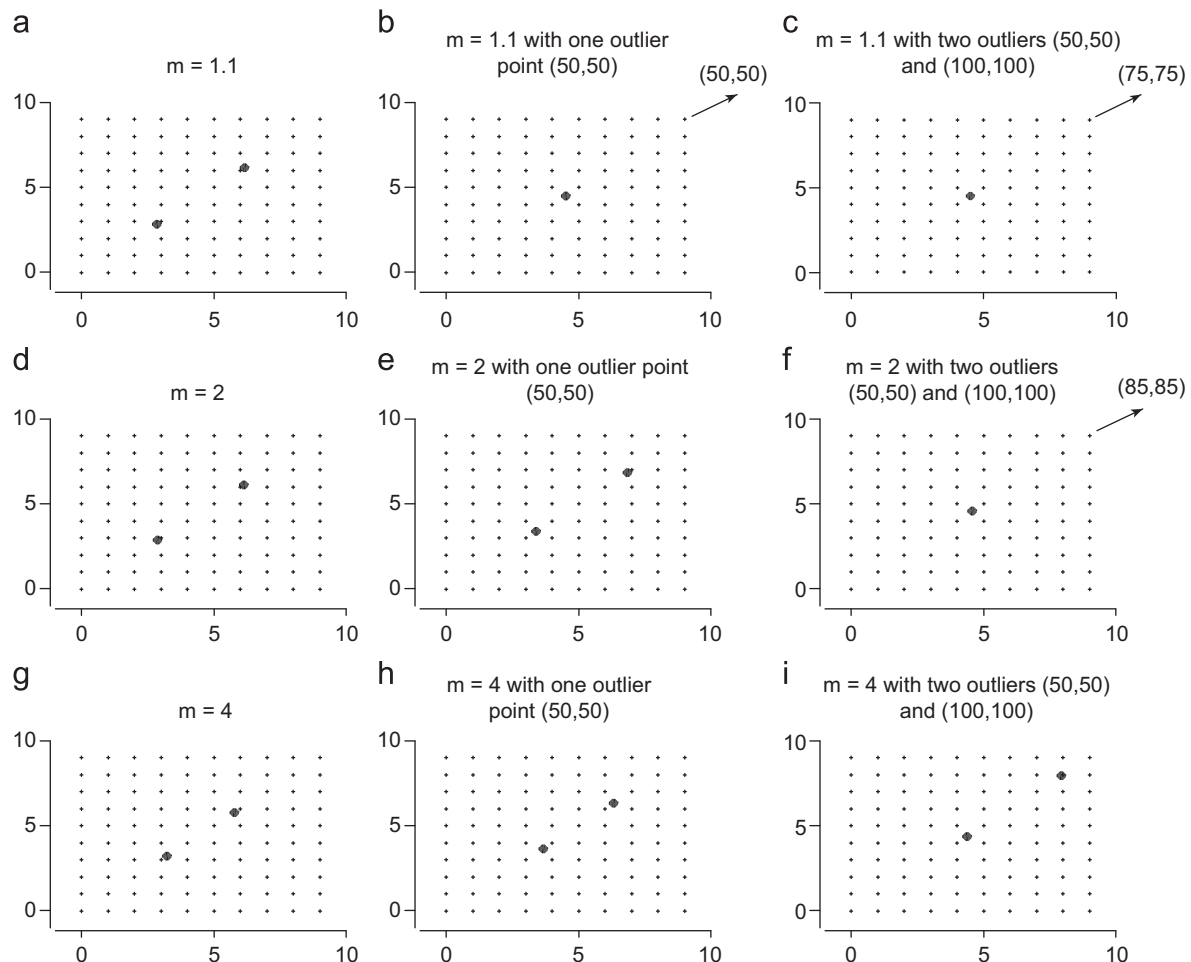


Fig. 2. Clustering results of FCM for the grid data.

noise and outliers with a large m value. Fig. 2 is a simple example to illustrate this phenomenon. This is an artificial data set with grid points. We implement FCM with different m values, and the results are shown in Figs. 2(a), (d), and (g). We then add an outlier point to the coordinate (50,50), and the results are shown in Figs. 2(b), (e) and (h). In this case, the result of FCM is influenced by this outlier point when m is small with $m=1.1$. When $m=2$ and 4, the results of FCM are robust to this outlier. Moreover, we add one more outlier point to the coordinate (100,100), and the results are shown in Figs. 2(c), (f), and (i). Only a large m value with $m=4$ can avoid the influence of these two outliers. These results are coincident to the phenomenon illustrated in Fig. 1. This example shows the robust properties of FCM when m becomes large.

Note that the cluster center update Eq. (3) can be seen as a weighted mean for the data set $\{x_1, \dots, x_n\}$. The weight of the data point x_j is

$$\frac{\mu_{ij}^m}{\sum_{j=1}^n \mu_{ij}^m}, \quad (6)$$

which is proportional to μ_{ij}^m . For a noise or outlier point x_j , the weight is proportional to $(1/c)^m$. In a large m case, the weight will tend to zero, and hence, the influence of noise and outliers on the cluster centers will become very small. This can explain why FCM is robust to noise and outliers when m becomes large.

When m tends to infinity, the weight and φ function of a finite point x_j will tend to 0. However, in a real application, we never observe an infinity data point. We then have $\lim_{m \rightarrow \infty} \varphi(x_j - a_i) = 0$ for a real data point. That is, a very large m value will make FCM very robust. However, this is not a good guideline for selecting m in FCM. Although FCM becomes very robust in a large m case, the membership value for each data point will be very close to $1/c$, and the sample mean may become the unique optimizer. Fig. 1(d) also shows that the membership values for the data points become approximately 0.5 when $m=6$.

Note that the largest membership of a noise or outlier point is approximately 0.5; Fig. 3 shows the weights of 0.5^m with $m=1.1, 1.5, 2, 3, 4, \dots, 10$. When $m \geq 5$, the weights are very close to zero. Here, we suggest implementing FCM with $m \in [1.5, 4]$ under the restriction that m is smaller than the theoretical upper bound proposed by Yu et al. [14].

3. Algorithms with cluster cores

According to our analysis, we find that a large m value will make FCM more robust to noise and outliers. However, a considerably large m value may make the sample mean the optimizer. We know that FCM extends the k -means in order to allow its membership function $\mu_{ij} \in [0, 1]$. However, almost no data

points have a crisp membership $\mu_{ij} \in \{0, 1\}$. That is, in fact, FCM always produces fuzzy memberships μ_{ij} in the open interval $(0, 1)$. This is the main reason for the sample mean to be the unique optimizer of FCM when the fuzzifier m is larger than the theoretical upper boundary. A simple and efficient method to avoid this special case in FCM is to assign a cluster core to each cluster. Here, we will review some algorithms with cluster cores and discuss their robust properties.

3.1. Partition index maximization (PIM)

Özdemir and Akarun [8] proposed an alternative clustering algorithm, called the partition index maximization (PIM), for the color quantization of images. PIM modified the distance measure of FCM and proposed the objective function

$$J_{PIM}(\mu, a) = \sum_{i=1}^c \sum_{j=1}^n \mu_{ij}^m d(x_j, a_i) = \sum_{i=1}^c \sum_{j=1}^n \mu_{ij}^m (\|x_j - a_i\|^2 - \beta). \quad (7)$$

The objective of PIM is to minimize the FCM objective function and simultaneously maximize the partition coefficient (PC) [12]. We can rewrite Eq. (7) as

$$J_{PIM}(\mu, a) = \sum_{i=1}^c \sum_{j=1}^n \mu_{ij}^m \|x_j - a_i\|^2 - \beta \sum_{i=1}^c \sum_{j=1}^n \mu_{ij}^m = J_{FCM}(\mu, a) - \beta \sum_{i=1}^c \sum_{j=1}^n \mu_{ij}^m. \quad (8)$$

The update equations for $J_{PIM}(\mu, a)$ are

$$a_i = \frac{\sum_{j=1}^n \mu_{ij}^m x_j}{\sum_{j=1}^n \mu_{ij}^m} \quad (9)$$

and

$$\mu_{ij} = \begin{cases} \frac{(\|x_j - a_i\|^2 - \beta)^{-1/(m-1)}}{\sum_{k=1}^c (\|x_j - a_k\|^2 - \beta)^{-1/(m-1)}} & \text{if } \|x_j - a_i\|^2 > \beta, \\ 1 & \text{if } \|x_j - a_i\|^2 \leq \beta. \end{cases} \quad (10)$$

For a given data point x_j in PIM, if $\|x_j - a_i\|^2 \leq \beta$, then $\mu_{ij} = 1$ and $\mu_{ij} = 0$ for all $i \neq j$. That is, if the distance between the data points and the i th cluster center is smaller than β , these data points will then exactly belong to the i th cluster with the membership value one. Each cluster in PIM will have a crisp boundary such that all data points inside this boundary will have a crisp membership value $\mu_{ij} \in \{0, 1\}$ and other data points outside this boundary will have fuzzy membership values $\mu_{ij} \in [0, 1]$. Each crisp boundary will form a hyper-ball for the corresponding cluster and can be seen as a cluster core. The parameter β will determine the volume of each cluster core and the general condition that any two of these c cluster cores do not overlap should be

$$\beta = (\gamma/4) \min_{i \neq j} \|a_i - a_j\|^2, \quad 0 \leq \gamma < 1. \quad (11)$$

Each volume of the cluster core will be a monotone increasing function of γ with the maximum volume when γ approaches 1 and with the minimum volume when γ equals 0. The volumes of each cluster core in PIM are equal. In the next subsection, we will review the algorithm that uses the cluster cores with different volumes.

3.2. Fuzzy compactness and separation (FCS)

Fukuyama and Sugeno [22] and Sugeno and Yasukawa [23] proposed an extended FCM objective function by adding a penalized term based on cluster prototypes a_i with

$$J^{FS}(\mu, a) = \sum_{j=1}^n \sum_{i=1}^c [(\mu_{ij})^m d^2(x_j, a_i) - (\mu_{ij})^m d^2(a_i, \bar{x})], \quad (12)$$

where $d^2(x_j, a_i) = \|x_j - a_i\|^2$. They used $FS(c) = J^{FS}(\mu, a)$ as an index for the cluster validity problem. A relatively small $FS(c)$ will lead to good fuzzy clustering results. However, this $FS(c) = J^{FS}(\mu, a)$ index

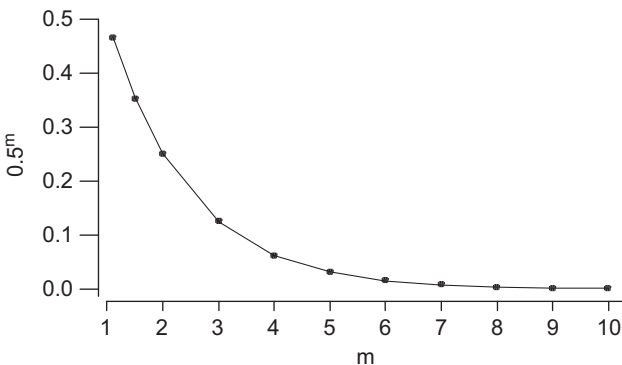


Fig. 3. The weights of the noise or outlier points.

cannot be a clustering objective function because there do not exist update equations for the cluster center a_i by differentiating $FS(c)=F^S(\mu,a)$ with respect to a_i .

Wu et al. [12] proposed a fuzzy compactness and separation (FCS) algorithm, which is a modification of the FS validity index. The objective function is defined as

$$J_{FCS}(\mu,a) = \sum_{i=1}^c \sum_{j=1}^n \mu_{ij}^m d(x_j, a_i) = \sum_{j=1}^n \sum_{i=1}^c \mu_{ij}^m (\|x_j - a_i\|^2 - \eta_i \|a_i - \bar{x}\|^2) \\ = J_{FCM}(\mu,a) - \sum_{j=1}^n \sum_{i=1}^c \mu_{ij}^m \eta_i \|a_i - \bar{x}\|^2 \quad (13)$$

where the parameter $\eta_i \geq 0$. The update equations for $J_{FCS}(\mu,a)$ are as follows:

$$a_i = \frac{\sum_{j=1}^n \mu_{ij}^m x_j - \eta_i \sum_{j=1}^n \mu_{ij}^m \bar{x}}{\sum_{j=1}^n \mu_{ij}^m - \eta_i \sum_{j=1}^n \mu_{ij}^m} \quad (14)$$

and

$$\mu_{ij} = \begin{cases} \frac{(\|x_j - a_i\|^2 - \eta_i \|a_i - \bar{x}\|^2)^{-1/(m-1)}}{\sum_{k=1}^c (\|x_j - a_k\|^2 - \eta_k \|a_k - \bar{x}\|^2)^{-1/(m-1)}} & \text{if } \|x_j - a_i\|^2 > \eta_i \|a_i - \bar{x}\|^2, \\ 1 & \text{if } \|x_j - a_i\|^2 \leq \eta_i \|a_i - \bar{x}\|^2. \end{cases} \quad (15)$$

For a given data point x_j in FCS, if $\|x_j - a_i\|^2 \leq \eta_i \|a_i - \bar{x}\|^2$, then $\mu_{ij} = 1$ and $\mu_{i'j} = 0$ for all $i' \neq i$. That is, if the distance between the data points and the i th cluster center is smaller than $\eta_i \|a_i - \bar{x}\|^2$, these data points will belong exactly to the i th cluster with the membership value of one. Each cluster in FCS will have a crisp boundary such that all data points inside this boundary have a crisp membership value $\mu_{ij} \in \{0,1\}$ and the other data points outside this boundary have fuzzy membership values $\mu_{ij} \in [0,1]$. The parameter η_i will determine the volume of each cluster core and the general condition that any two of these c cluster cores do not overlap should be

$$\eta_i = \frac{(\gamma/4) \min_{i \neq i'} \|a_i - a_{i'}\|^2}{\max_k \|a_k - \bar{x}\|^2}, \quad 0 \leq \gamma < 1. \quad (16)$$

Each volume of a cluster core will be a monotone increasing function of γ with the maximum volume when γ approaches 1 and with the minimum volume when γ equals 0.

3.3. Alpha-cut implemented fuzzy clustering algorithms (FCM α)

According to the concepts of the cluster core, Yang et al. [16] proposed the alpha-cut implemented fuzzy clustering algorithm (FCM α). We mention that PIM and FCS generate the cluster cores such that if the distances between the data points and the i th cluster center are smaller than a threshold, these data points will exactly belong to the i th cluster with a membership value of 1. In fact, this cluster core concept can be extended in a more general manner such that if the membership value μ_{ij} of the data point x_j in the i th cluster is larger than a given value α , then the point x_j will belong exactly to the i th cluster with the membership value of 1 and $\mu_{i'j} = 0$ for all $i' \neq i$. A simple rule to guarantee that no two of these c cluster cores overlap is $0.5 < \alpha \leq 1$. These types of algorithms implemented by α -cut FCM algorithms are called FCM α . The cluster cores generated by FCM α can be determined by a relative distance measure with

$$\mu_{ij} = \frac{\|x_j - a_i\|^{-2/(m-1)}}{\sum_{k=1}^c \|x_j - a_k\|^{-2/(m-1)}} > \alpha, \quad (17)$$

which is equivalent to

$$\|x_j - a_i\|^2 < \left(\alpha \sum_{k=1}^c \|x_j - a_k\|^{-2/(m-1)} \right)^{-(m-1)}. \quad (18)$$

Hence, the update equations for FCM α are

$$a_i = \frac{\sum_{j=1}^n \mu_{ij}^m x_j}{\sum_{j=1}^n \mu_{ij}^m} \quad (19)$$

and

$$\mu_{ij} = \begin{cases} 1 & \text{if } \|x_j - a_i\|^2 < \left(\alpha \sum_{k=1}^c \|x_j - a_k\|^{-2/(m-1)} \right)^{-(m-1)}, \\ \frac{(\|x_j - a_i\|^{-2/(m-1)})}{\sum_{k=1}^c (\|x_j - a_k\|^{-2/(m-1)})} & \text{otherwise} \end{cases} \quad (20)$$

For a given data point x_j in FCM α , if $\|x_j - a_i\|^2 < \left(\alpha \sum_{k=1}^c \|x_j - a_k\|^{-2/(m-1)} \right)^{-(m-1)}$, then $\mu_{ij} = 1$ and $\mu_{i'j} = 0$ for all $i' \neq i$. Each cluster in FCM α will have a crisp boundary such that all data points inside this boundary have a crisp membership value $\mu_{ij} \in \{0,1\}$ and the other data points outside this boundary have fuzzy membership values $\mu_{ij} \in [0,1]$. The parameter α will determine the volume of each cluster core, and the general condition that any two of these c cluster cores do not overlap should be $0.5 < \alpha \leq 1$. If we take

$$\alpha = 1 - 0.5\gamma, \quad 0 \leq \gamma < 1, \quad (21)$$

then each volume of the cluster core will be a monotone increasing function of γ with the maximum volume when γ approaches 1 and with the minimum volume when γ equals 0.

It has been shown that the above-reviewed clustering algorithms with a large fuzzifier m ($m=4$) can be more robust to noise and outliers than FCM with $m=2$ (see Refs. [12,16,24]). However, according to our analysis in Section 2, we find that a large m value will make FCM more robust to noise and outliers such as PIM, FCS, and FCM α . Since the φ function of FCM has been discussed in Section 2.2, we now give the φ functions of these cluster-core-based algorithms.

3.4. φ Function of PIM, FCS, and FCM α

We now define the loss function of PIM as

$$\rho_{PIM}(x_j - a_i) = \mu_{ij}^m (\|x_j - a_i\|^2 - \beta) \quad (22)$$

and

$$\varphi_{PIM}(x_j - a_i) = \begin{cases} -2\mu_{ij}^m (x_j - a_i) & \text{if } x_j \text{ outside the cluster core,} \\ -2(x_j - a_i) & \text{if } x_j \text{ inside the cluster core} \end{cases} \quad (23)$$

Equivalently, we can define the loss function of FCS as

$$\rho_{FCS}(x_j - a_i) = \mu_{ij}^m (\|x_j - a_i\|^2 - \eta_i \|a_i - \bar{x}\|^2) \quad (24)$$

and we have

$$\varphi_{FCS}(x_j - a_i) = \begin{cases} -2\mu_{ij}^m (x_j - a_i) - \eta_i \mu_{ij}^m (a_i - \bar{x}) & \text{if } x_j \text{ outside the cluster core,} \\ -2(x_j - a_i) - \eta_i (a_i - \bar{x}) & \text{if } x_j \text{ inside the cluster core} \end{cases} \quad (25)$$

Since the loss function of FCM α is equivalent to the FCM, we can conclude that the φ function of FCM α is equivalent to the φ function of PIM. That is

$$\varphi_{FCM\alpha}(x_j - a_i) = \varphi_{PIM}(x_j - a_i). \quad (26)$$

Note that the φ function can help us to assess the relative influence of an individual observation on the value of an estimate. Since the φ functions of FCM, PIM, FCS, and FCM α are monotone with $(x_j - a_i)$, the influence of adding a noise or an outlier point on FCM will approximate to adding a noise or outlier point on PIM, FCS, and FCM α . That is, the robust properties of FCM, PIM, FCS, and FCM α are similar and will be illustrated in the next section.

In this section, we review some cluster-core-based algorithms. The volumes of the cluster cores are controlled by the specified

parameters; we give a unified definition with parameter γ to control their volumes as shown in Eqs. (11), (16), and (21). When $\gamma=0$, these cluster-core-based algorithms are equivalent to FCM. When γ approaches 1, these algorithms are implemented with the largest cluster core volumes.

4. Examples

We will give some numerical examples to compare the performances of FCM, PIM, FCS, and FCM α . The first is the Vowel data set [25] that has $n=990$ points in a 10-dimensional space with 11 clusters. The second real data set is the Iris data set [20,21] that has $n=150$ points in a 4-dimensional space that represents three clusters, each with 50 points. The third data set is the grid data illustrated in Fig. 2. In order to sufficiently use all the information, we will normalize the real data sets, including the Iris and Vowel data. Suppose we have a data set $X=\{x_1, \dots, x_n\}$ in an s -dimensional space with each $x_j=\{x_{j1}, \dots, x_{js}\}$, we normalize the data by replacing x_{jk} with x'_{jk} as

$$x'_{jk} = \frac{x_{jk} - \sum_{l=1}^n x_{lk}/n}{\sqrt{\sum_{l=1}^n (x_{lk} - \sum_{l=1}^n x_{lk}/n)^2 / (n-1)}}, \quad (27)$$

where $k=1, \dots, s$ and $j=1, \dots, n$. After normalization, each characteristic of the real data set will have common sample mean and dispersion measures.

Example 1. We have mentioned that the use of the cluster cores in FCM can avoid the sample mean from being the optimizer when the fuzzifier m is larger than the theoretical upper bound proposed by Yu et al. [14]. The theoretical upper bound for the Vowel data set is 1.7787. We now implement the PIM, FCS, and FCM α for the Vowel data with parameter $\gamma=0, 0.2, 0.4, 0.6, 0.8$, and 0.99 . The results of different m values with $m=1.1, 1.5$, and 2 are shown in Fig. 4(a), (b), and (c), respectively. We use the non-fuzziness index (NFI) values [13] to compare the performances of the clustering results. If the membership values of each data point are close to 0 or 1, the NFI index will then be close to 1 and if the membership values of each

data points are close to $1/c$, the NFI index will then be close to 0. The results of FCM are equivalent to the case with $\gamma=0$. Fig. 4 shows that the NFI index seems to be a monotone increasing function of parameter γ and the use of the cluster cores in FCM can avoid the sample mean from being the optimizer when the fuzzifier m is larger than the theoretical upper bound 1.7787. In this example, we show the advantage of using the cluster cores in FCM. However, if the data set has a large theoretical upper bound, the use of the cluster cores may not result in a better performance than FCM, as shown in the next two examples.

Example 2. In this example, we will implement the Iris data set; the error counts will be calculated after clustering. Note that the error counts obtained by FCM for this data set are approximately 16 in many literatures. According to the above example, the parameter $\gamma=0.99$ will be adopted for the cluster-core-based algorithms. Fig. 5 shows the error counts for the algorithms over a set of specified m with $m=1.1, 1.5, 2, 3, 4, \dots, 40$. Since the theoretical upper bound for the Iris data is ∞ , the clustering results will become stable when m becomes large and the clustering error counts of FCM, PIM, FCS, and FCM α converge to 13, 15, 14, and 13, as shown in Fig. 5(a), (b), (c), and (d), respectively. In general, it is suggested that the fuzzifier m in FCM and many of its extensions not be assigned a considerably large value. However, this example considers the problem of selecting the fuzzifier m from a different point of view.

Note that the aim of this example is not to find the best fuzzifier m . We want to show that the use of cluster cores may not always result in a better performance than FCM. Table 3 shows some summarized information, including the average error count and the standard deviation of the error count. The average error count of FCM is 13.22, which is the smallest among these algorithms. Although the average error count of FCM α is considerably close to FCM, FCM α still has a large standard deviation. This means that the performance of FCM α as compared to FCM is relatively unstable. Both PIM and FCS have relatively stable error counts with a small standard deviation, but the average error

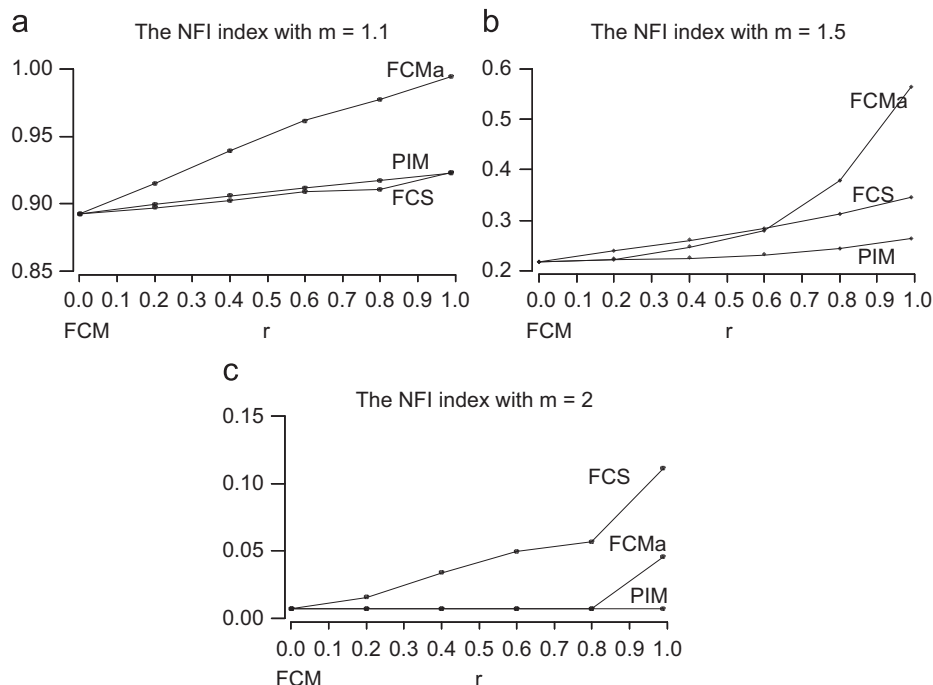


Fig. 4. The NFI index values for the vowel data set. The parameter $\gamma=0, 0.2, 0.4, 0.6, 0.8$, and 0.99 are used in PIM, FCS, and FCM α . The results of different m values with $m=1.1, 1.5$, and 2 are shown in (a), (b), and (c), respectively. The results of FCM are equivalent to the case with $\gamma=0$.

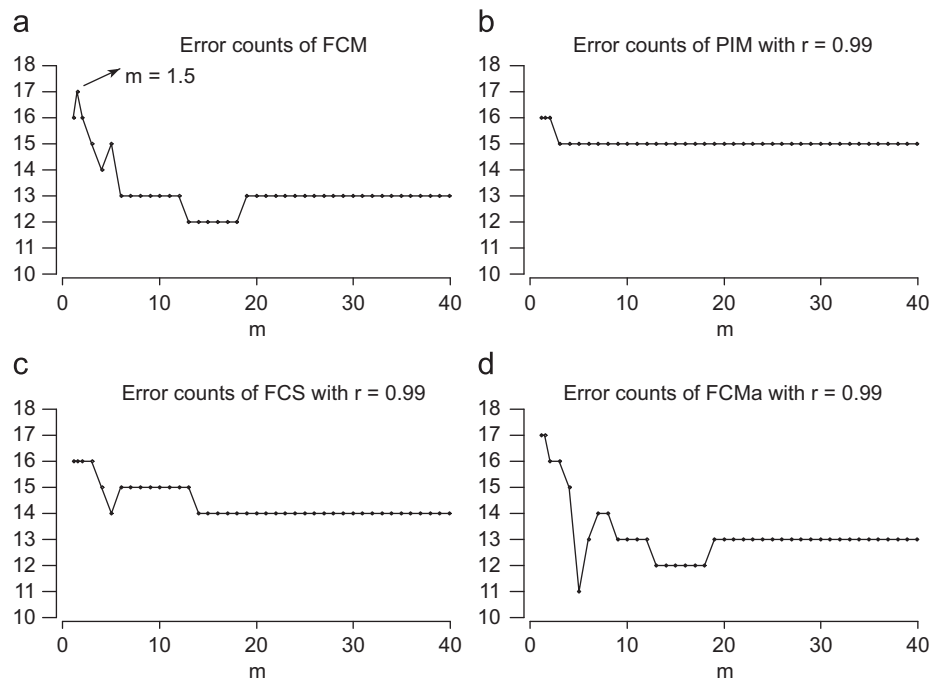


Fig. 5. The results of the iris data set. The parameter $\gamma=0.99$ is used in PIM, FCS, and FCM α .

Table 3

The summarized results of the Iris data set. The parameter $\gamma=0.99$ is used in PIM, FCS, and FCM α .

	FCM	PIM	FCS	FCM α
Error counts with $m=2$	16	16	16	16
Error counts with $m=4$	14	15	15	15
Average error count over a set of specified m	13.22	15.07	14.42	13.24
Standard deviation of error count over a set of specified m	1.084	0.264	0.670	1.261

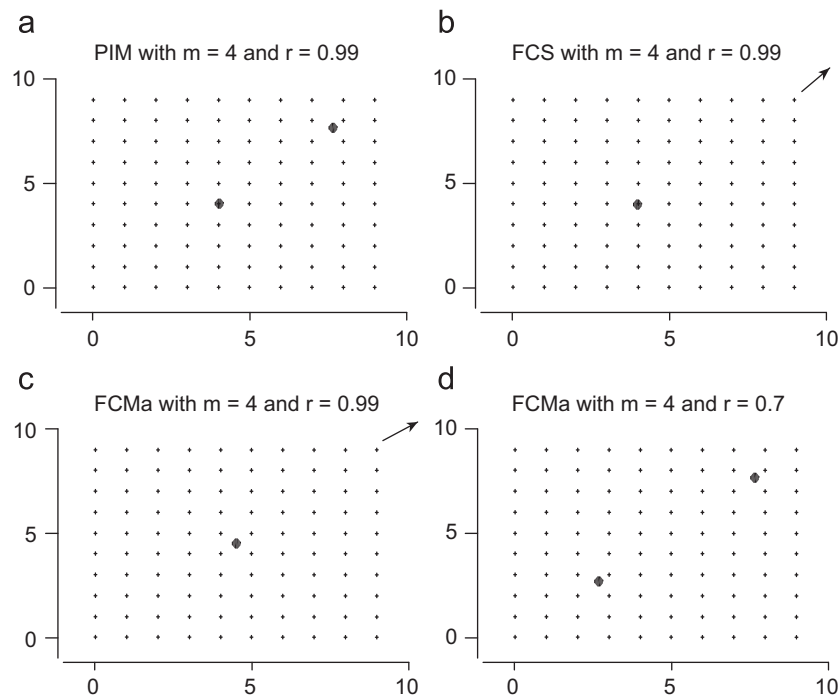


Fig. 6. Clustering results of PIM, FCS and FCM α for the grid data set with two outliers (50,50) and (100,100): (a) The result of PIM with $m=4$ and $\gamma=0.99$; (b) the result of FCS with $m=4$ and $\gamma=0.99$; (c) the result of FCM α with $m=4$ and $\gamma=0.99$, and (d) the result of FCM α with $m=4$ and $\gamma=0.7$.

count is relatively large. FCM exhibits a relatively good clustering performance in this example, and we suggest implementing FCM with $m=4$ in a large theoretical upper bound case. In the next example, we will compare the performance of these algorithms when the fuzzifier $m=4$.

Example 3. Fig. 6 shows the clustering results of PIM, FCS, and FCM α for the grid data set with two outliers (50,50) and (100,100). Fig. 6(a) shows the result of PIM with $m=4$ and $\gamma=0.99$, which is similar to the result of FCM, as shown in Fig. 2(i). The cluster centers obtained by FCS with $m=4$ and $\gamma=0.99$ are influenced by the outliers, as shown in Fig. 6(b). In our simulations, the results of FCS are not as expected when the data set does not contain good cluster structures. Note that the results of FCM α with $\gamma=1$ are equivalent to the k -means when the cluster number $c=2$. Hence, the result shown in Fig. 6(c) is similar to the result of FCM with $m=1.1$, as shown in Fig. 2(c). In this case, Yang et al. [16] suggested specifying the parameter $\alpha=0.65$ (i.e., $\gamma=0.7$). Fig. 6(d) shows the result of FCM α with $m=4$ and $\gamma=0.7$, which is also similar to the result of FCM, as shown in Fig. 2(i). Although these cluster-core-based algorithms have similar robust properties with FCM, their performances still depend on the specified parameter γ .

Example 4. This example contains three randomly generated data sets denoted by data 1, data 2, and data 3. Each cluster in these data sets consist of 100 data points generated from a Normal distribution. The cluster numbers of data 1, data 2, and data 3 are $c=2$, $c=3$, and $c=4$, respectively. We also add 25% Uniform noise points to each data set. Table 4 shows the information about data 1, data 2, and data 3. The corresponding Histograms of these data sets are illustrated in Fig. 7. Under the same initial values and stopping

Table 4

The randomly generated three data sets. Each cluster in these data sets consist of 100 data points generated from a Normal distribution. We also add 25% Uniform noise points to each data set. The sample size of data 1, data 2, and data 3 are 250, 375, and 500, respectively.

Data 1, c=2				
Cluster 1	Cluster 2	Noise		
$n=100$ $N(0,1)$	$n=100$ $N(10,1)$	$n=50$ $U[0,10]$		
Data 2, c=3				
Cluster 1	Cluster 2	Cluster 3	Noise	
$n=100$ $N(0,1)$	$n=100$ $N(10,1)$	$n=100$ $N(20,1)$	$n=75$ $U[0,20]$	
Data 3, c=4				
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Noise
$n=100$ $N(0,1)$	$n=100$ $N(10,1)$	$n=100$ $N(20,1)$	$n=100$ $N(30,1)$	$n=100$ $U[0,30]$

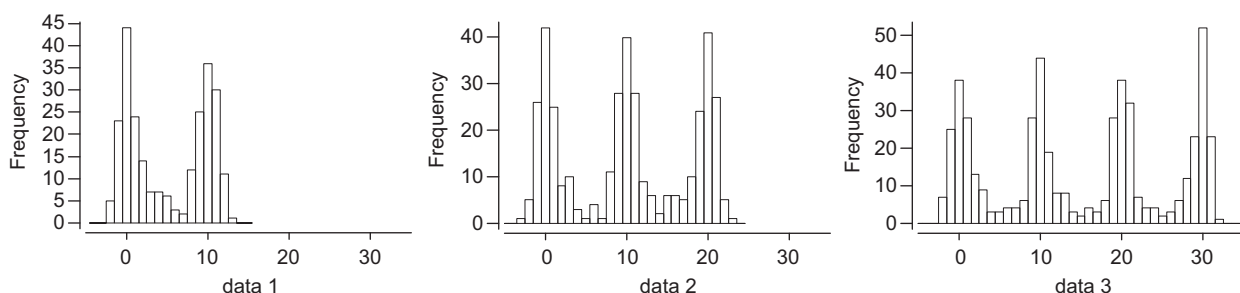


Fig. 7. Histograms of three randomly generated data sets.

conditions, we then implement FCM, PIM, FCS, and FCM α to estimate the cluster centers of these data sets. The performance of the algorithms are evaluated by the MSE values with $\sum_{i=1}^c (a_i - \theta_i)^2 / c$ where θ_i denotes the true cluster center. Table 5 shows the average MSE values of 100 repeated MSE measures obtained by different combinations of fuzzifier m and clustering algorithms. For example, the average MSE of 100 repeated MSE of data 1 estimated by FCM with $m=2$ is 0.1670. From the results in Table 5, FCM with a suitable large m with $m=4$ always performs well than other clustering methods, except the FCM α with $m=4$ in data 3. Note that, the “–” symbol means the algorithm cannot converge. When the data set has a large theoretical upper bound, the use of the cluster cores may not result in a better performance than FCM and FCM with $m=4$ always performs well than FCM with $m=2$. In this example, the cluster-core-based algorithms are implemented with a large cluster core volumes ($\gamma=0.99$) and hence some noise points are inside the cluster cores. Although we can try to find a suitable γ to improve the performance of the cluster-core-based algorithms, this example reveal the superiority of using FCM with a suitable large m value.

5. Conclusions

In this paper, we discussed the parameter selections of FCM from a robustness point of view and found that a large m value will make FCM more robust to noise and outliers. We suggest implementing FCM with $m \in [1.5, 4]$. However, considerably large m values that are greater than the theoretical upper bound will make the sample mean the unique optimizer. To avoid this unexpected case, we can adopt the cluster core concept in FCM. We then reviewed some cluster-core-based algorithms and discussed these algorithms in detail. Example 1 showed the results of the Vowel data set whose theoretical upper bound of m is 1.7787; the advantages of the cluster-core-based algorithms were evident in this example. However, if the data set has a large theoretical upper bound, the use of the cluster cores may not result in a better

Table 5

The average MSE values of 100 repeated MSE measures obtained by the clustering algorithms. The parameter $\gamma=0.99$ is used in PIM, FCS and FCM α .

	FCM	PIM	FCS	FCM α
Data 1				
$m=2$	0.1607	0.2643	–	0.2633
$m=4$	0.0893	0.2630	–	0.2563
Data 2				
$m=2$	0.0756	0.1315	4.6704	0.1264
$m=4$	0.0358	0.1279	5.4639	0.0432
Data 3				
$m=2$	0.0526	0.0959	1.1585	0.0882
$m=4$	0.0258	0.0944	1.4259	0.0186

performance than FCM, as illustrated in Examples 2, 3, and 4. For a large theoretical upper bound case, we suggest implementing the FCM with a suitable large m value. Otherwise, we suggest implementing the clustering methods with cluster cores. When the data set contains noise and outliers, the fuzzifier $m=4$ is recommended for both FCM and cluster-core-based methods in a large theoretical upper bound case.

Acknowledgments

The author is grateful to the anonymous referees for their critical and constructive comments and suggestions to improve the presentation of the paper. This work was supported in part by the National Science Council of Taiwan, under Kuo-Lung Wu's Grant: NSC-99-2118-M-168-001.

References

- [1] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithm, Plenum Press, 1981.
- [2] M.S. Yang, A survey of fuzzy clustering, Math. Comput. Modelling 18 (1993) 1–16.
- [3] J. Yu, General c -means clustering model, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1197–1211.
- [4] K.L. Wu, M.S. Yang, Alternative c -means clustering algorithms, Pattern Recognition 35 (2002) 2267–2278.
- [5] J. Leski, Towards a robust fuzzy clustering, Fuzzy Sets Syst. 137 (2003) 215–233.
- [6] H. Frigui, R. Krishnapuram, A robust competitive clustering algorithm with applications in computer vision, IEEE Trans. Pattern Anal. Mach. Intell. 21 (1999) 450–465.
- [7] J. Yu, M.S. Yang, Optimality test for generalized FCM and its application to parameter selection, IEEE Trans. Fuzzy Syst. 13 (2005) 164–176.
- [8] D. Özdemir, L. Akarun, A fuzzy algorithm for color quantization of images, Pattern Recognition 35 (2002) 1785–1791.
- [9] D. Özdemir, L. Akarun, Fuzzy algorithms for combined quantization and dithering, IEEE Trans. Image Process. 10 (2001) 923–931.
- [10] R. Krishnapuram, J.M. Keller, A possibilistic approach to clustering, IEEE Trans. Fuzzy Syst. 1 (1993) 98–110.
- [11] M.S. Yang, K.L. Wu, Unsupervised possibilistic clustering, Pattern Recognition 39 (2006) 5–21.
- [12] K.L. Wu, J. Yu, M.S. Yang, A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality tests, Pattern Recognition Lett. 26 (2005) 639–652.
- [13] N.R. Pal, J.C. Bezdek, On cluster validity for fuzzy c -means model, IEEE Trans. Fuzzy Syst. 1 (1995) 370–379.
- [14] J. Yu, Q. Cheng, H. Huang, Analysis of the weighting exponent in the FCM, IEEE Trans. Syst. Man, Cybern. B 34 (2004) 634–639.
- [15] J.C. Bezdek, Cluster validity with fuzzy sets, J. Cybern. 3 (1974) 58–73.
- [16] M.S. Yang, K.L. Wu, J.N. Hsieh, J. Yu, Alpha-cut implemented fuzzy clustering algorithms and switching regressions, IEEE Trans. Syst. Man Cybern. B 38 (2008) 588–603.
- [17] P.J. Huber, Robust Statistics, Wiley, 1981.
- [18] R.N. Dave, R. Krishnapuram, Robust clustering methods: a unified view, IEEE Trans. Fuzzy Syst. 5 (1997) 270–293.
- [19] M.S. Yang, K.L. Wu, A similarity-based robust clustering method, IEEE Trans. Pattern Anal. Mach. Intell. 26 (2004) 434–448.
- [20] E. Anderson, The Irises of the gaspe peninsula, Bull. Am. Iris Soc. 59 (1935) 2–5.
- [21] J.C. Bezdek, J.M. Keller, R. Krishnapuram, L.I. Kuncheva, N.R. Pal, Will the Iris data please stand up? IEEE Trans. Fuzzy Syst. 7 (1999) 368–369.
- [22] Y. Fukuyama, M. Sugeno, A new method of choosing the number of clusters for fuzzy c -means method, in: Proceedings of the 5th Fuzzy System Symposium, 1989, pp. 247–250 (in Japanese).
- [23] M. Sugeno, T. Yasukawa, A fuzzy-logic-based approach to qualitative modeling, IEEE Trans. Fuzzy Syst. 1 (1993) 7–31.
- [24] K.L. Wu, An analysis of partition index maximization algorithm, in: IEEE International Conference on Fuzzy Systems, FUZZ-IEEE, ICC Jeju, Jeju Island, Korea, August 20–24, 2009, pp. 1785–1790.
- [25] C.L. Blake, C.J. Merz, UCI repository of machine learning databases, a huge collection of artificial and real-world data sets (1998). Available from: <<http://www.ics.uci.edu/~mllearn/MLRepository.html>>.

KUO-LUNG WU received his BS degree in mathematics in 1997, the MS and PhD degrees in applied mathematics in 2000 and 2003, all from the Chung Yuan Christian University, Chungli, Taiwan. In 2003–2006, he was an Assistant Professor in the Department of Information Management at Kun Shan University of Technology, Tainan, Taiwan. He has been an Associate Professor at the same university. He is a member of the Phi Tau Phi Scholastic Honor Society of Taiwan. His research interests include fuzzy theorem, cluster analysis, pattern recognition and neural networks.