

A New Framework of Fuzzy Clustering Algorithm

Horng-Lin Shieh

Department of Electrical Engineering
St. John's University
Taipei, Taiwan
shieh@mail.sju.edu.tw

Abstract—In this paper, a novel data clustering algorithm based on the *subtractive clustering* (SC) algorithm and a new validity index are proposed. The SC algorithm is a simple method for data clustering; however, it has two problems which must be overcome. The first problem is such that the cluster centers found by SC are taken from data with the highest potential values, but that this data may not be the real cluster centers. The second problem is such that the cluster number generated by the SC algorithm is influenced by a predefined parameter. The proposed algorithm is based on distance relations between data and centers and is designed to ascertain the real centers generated by the SC algorithm. In addition, a novel robust cluster index is proposed to identify the real cluster number generated by SC algorithm.

Keywords— validity index, subtractive clustering (SC) algorithm, clustering algorithm

1. INTRODUCTION

In classical set theory, the membership of datum to a set is assigned using a value of 0 or 1. Let $\mu_A(x)$ denote the membership value of x belonging to cluster A (or set A). If one datum x belongs to cluster A , the membership value $\mu_A(x)$ is 1, otherwise, its value is 0; this clustering method is called hard partition. However, in practical applications, there are ambiguities found in the data set such that some data may or may not be partitioned into one cluster. In this situation, the membership value $\mu_A(x)$ of x is assigned to a range of $[0,1]$; this method is called *soft partition*. The fuzzy clustering algorithm based on fuzzy set theory is the most well known soft partition method. In the context of a fuzzy set, a value close to 1 for $\mu_A(x)$ denotes a high degree of membership, while a value close to 0 denotes a low degree of membership. Numerous successful applications of fuzzy clustering have been proposed, such as handwritten digit recognition [1], function approximation [2][3][4], and image processing [5].

There have been numerous fuzzy clustering methods proposed in the literature [6]–[11]. The fuzzy c-means algorithm (FCM) [7][8] has become the most popular fuzzy clustering approach for both theoretical and practical applications in recent decades. However, the FCM algorithm has two disadvantages: (1) the number of clusters has to be known in advance but, in many applications, this knowledge is often not available; (2) when the data training contains outliers, the resultant centers are always influenced by outliers. To overcome the first disadvantage, the results of the analyses pertaining to different numbers of clusters have to be compared to each other using a validity function in order to

ascertain an optimal cluster number. As regards the second disadvantage, when the training data contains noise/outliers, the results obtained by FCM are always affected by noise/outliers.

2. A New Clustering Algorithm

The *mountain method* (MM) proposed by Yager and Filev in 1994 [12] is an effective and simple method for identifying the cluster center of the sampling data. The mountain method divides data space into several intervals to construct grids and then evaluates its energies by substituting the distance from the grid to the data points into the objective exponential function in order to determine the approximated cluster centers at the grid with the maximum energy.

Let $X=\{x_1, x_2, \dots, x_n\}$ be a set of n vectors in a d -dimensional feature space, that is, $x_i \in \mathbb{R}^d$ is a vector, and x_{kj} denotes the j -th coordinate of the k -th data point, where $k = 1, 2, \dots, n$ and $j = 1, 2, \dots, d$.

The MM for data clustering is divided into three steps:

- Step1: first, the mountain function involves constructing a grid on the data space where the intersections of the grid lines are considered to be candidates for the cluster centers.
- Step2: the second step entails constructing a mountain function representing a data energy measurement. The value of the mountain function at grid node N_i is calculated by:

$$M(N_i) = \sum_{k=1}^n \exp(-\alpha \cdot d(x_k, N_i)), \quad (1)$$

where α is an application-specific positive constant and $d(x_k, N_i)$ is the distance between the data point x_k and the grid node N_i . The value of the mountain function can be seen to represent the potential ability of any of the grid nodes to be a cluster center. The higher mountain function value of the data is selected as the cluster center.

- Step 3: the third step involves an iterative selection of the next cluster centers and a sequential deconstruction of the mountain function value. In each iterative process, the grid node that has the largest energy is selected as the cluster center and the energies of the remaining grid nodes must be eliminated in order to identify the next one. Let N_k^* , $k=1, 2, \dots, c_{\max}$, denote the grid node that

has the maximal value of the mountain function at the k -th iteration. The maximal value of the mountain function M_k^* is denoted as:

$$M_k^* = \max_i M_k(N_i) \quad (2)$$

The new modified value of the mountain function is defined as follows:

$$M_{k+1}(N_i) = M_k(N_i) - M_k^* \times e^{-\beta d(N_i - N_k^*)} \quad (3)$$

Where β is a predefined positive constant. The MM continually produces new cluster centers and reduces the value of the mountain function until the level of the current maximum M_k^* compared with the original maximum M_1^* becomes too low.

Although the mountain method is a simple method for data clustering and has robust properties against outliers and noise, it has a notable problem that must be overcome. That is, because the mountain function calculates its energies on the grid nodes, the grid nodes and computational loads increase exponentially as the dimensions increase. In 1994, Chiu [13] modified the mountain method to construct a potential function for calculating the clustering center of the sampling data. The proposed method is also called the subtractive clustering (SC) algorithm. In the SC algorithm, the data points are likened to energy sources. The potential generated by each feature point has a peak value at the location of the data point which decreases rapidly at any point away from the data point [14]. The potential of each data point x_i is defined as:

$$P(x_i) = \sum_{k=1}^n \exp(-\alpha \cdot d(x_i, x_k)), \quad (4)$$

where $\alpha = \frac{4}{r_a}$, and r_a is a positive constant that is effectively the radius defining a neighborhood. In Equation (4), the total potential of any data point x_i is defined as the sum of the potentials contributed by each data point at x_k . It is reasonable to assume that the peaks of the total potential function correspond to cluster prototypes and that the valleys correspond to the decision boundaries between the clusters [14].

After the potential of each data point is obtained, the data point with the highest potential is selected as the first cluster center. Assume x_1^* is the first cluster center and its potential value is p_1^* . Then, the potential of each data point can be revised by:

$$P(x_i) = P(x_i) - p_1^* e^{-\beta d(x_i, x_1^*)}, \quad (5)$$

where $\beta = \frac{4}{r_b}$, and r_b is a positive constant which is effectively the radius defining a neighborhood that will have measurable reductions in potential. In [13], a good choice for r_b is $1.5 r_a$. To find the next cluster center, x_2^* , the new revised

potential is maximized and the effects of this cluster center are again removed. The process is repeated until a given threshold for the potential is obtained so that:

$$\frac{p_j^*}{p_1^*} < \delta. \quad (6)$$

In the above equation, δ is a given threshold. The choice of δ is an important factor affecting the clustering results; if δ is too large, then too few data points will be accepted as cluster centers and if δ is too small, then too many cluster centers will be generated [13].

There are two problems that must be overcome: (1) the SC algorithm assumes that the clustering center is x_i^* obtained by the i^{th} iteration. However, the x_i^* may be a rough cluster center which does not represent the actual cluster center; (2) there is no index to indicate what the optimal number of clusters for a data set is. For overcoming the first problem, this paper proposes a new algorithm based on distance measurements to identify the actual cluster centers. For overcoming second problem, this paper proposes a new validity index for the SC algorithm to indicate the goodness of the partitions in various cluster numbers. The proposed validity index combines the compactness in each cluster, the separation between clusters, and the correct ratio of data being properly partitioned into clusters to whole data set.

Let x_i^* , $i = 1, 2, \dots, c_{\max}$ are the cluster centers obtained by the SC algorithm, the distance relation between the centers x_i^* and data x_j is defined by the following equation:

$$D_{ij} = \exp(-\alpha \|x_j - x_i^*\|^2), i = 1, 2, \dots, c_{\max}, j = 1, 2, \dots, n, \quad (7)$$

where $\|x_j - x_i^*\|$ represents the Euclidean distance between x_j and x_i^* and where α is a constant defined in Equation (4). The value of D_{ij} indicates the distance relationship between x_j and x_i^* . In other words, for a cluster center x_i^* , any data x_j that has high value of D_{ij} indicates that this data x_j is close to the cluster i . Set N_ϵ denotes the number of data which has $D_{ij} \geq \epsilon$, where $\epsilon \geq 0$ is a constant. In this paper, ϵ is set to 0.5. The new center of cluster i is defined by the following equation:

$$v_i = \frac{\sum_{k=1}^{N_\epsilon} x_k^i}{N_\epsilon}, \quad (8)$$

where x_k^i are the data which have $D_{ik} \geq \epsilon$. In Equation (8), v_i is the mean of data close to cluster i . In each iterative epoch of the SC algorithm, when a new rough cluster center x_i^* is found, Equation (8) is adopted to evaluate the real cluster center; the potential of the data point can be revised by:

$$P(x_k^i) = P(x_k^i) - p_i^* e^{-\beta d(x_k^i, v_i)}. \quad (9)$$

The differences between the proposed method and the original SC algorithm are such that: (1) the cluster centers obtained by the proposed method provide a more accurate real cluster center than the cluster centers obtained by the SC algorithm; and (2), in each iterative epoch of the proposed method, the revised potential of Equation (9) uses relative data, x_k^i , that is close to the cluster center i , while the revised potential of the SC algorithm in Equation (5) uses all sampling data. That is to say that the proposed method requires less computation of data than SC algorithm.

3. A ROBUST VALIDITY INDEX

A *validity index* is a function which assigns to the output of the clustering algorithm a value to measure the qualities of the partitions provided by the output. The cluster number of SC algorithm is determined in Equation (6) but it is not used for evaluating the fitness of partitions obtained through fuzzy clustering. The validity index for finding an optimal c , denoted c^* , which can completely describe the data structure, has become the most studied topic in cluster validity. Several popular validity indexes are reviewed in [15][16].

The partition coefficient (PC) index proposed by Bezdek [17] in 1974 was the first index for the FCM algorithm. let $X = \{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^d$ be a data set in a d -dimensional feature space and μ_{ik} be the membership value of data x_i belonging to cluster k . The PC index is defined as:

$$PC(c) = \frac{1}{n} \sum_{k=1}^c \sum_{i=1}^n (\mu_{ik})^2, \quad (10)$$

and $\sum_{k=1}^c \mu_{ik} = 1, \quad \frac{1}{c} \leq PC(c) \leq 1.$

As regards a clustering algorithm for the partitioning of a data set, the optimal cluster number occurs when the PC (c) is at maximal value. The core idea of the PC index is such that the optimal cluster number occurs when the data are most unambiguously partitioned into clusters. The disadvantage of the PC index is that it only considers the fuzzy membership degree μ_{ik} for each cluster without considering the structure of the clusters.

The other index proposed by Bezdek is the partition entropy (PE) index [17][18]. The PE index is defined as:

$$PE(c) = -\frac{1}{n} \sum_{k=1}^c \sum_{i=1}^n \mu_{ik} \log_a(\mu_{ik}), \quad (11)$$

where a is the base of the logarithm and the range of the PE(c) is $[0, \log_a(c)]$. For finding an optimal c^* , we solve $\min_{2 \leq c \leq c_{\max}} PE(c)$ to produce the best clustering performance for a data set. The disadvantage of the PE index is analogous to the PC index, i.e., the PE index will take the minimum value on every hard partition. These two indexes only evaluate the fuzziness and do not consider the data structure of the clusters.

The Xie and Beni (XB) index [19], combined compactness and separation properties, is defined as follows:

$$XB(c) = \frac{\sum_{k=1}^c \sum_{i=1}^n \mu_{ik}^2 \|x_i - v_k\|^2}{n \times \min_{i \neq j} (\|v_i - v_j\|)}. \quad (12)$$

Where v_i and v_j represent the centers of cluster i and j , respectively, while $\|v_i - v_j\|$ represents the Euclidean distance between v_i and v_j . In Equation (12), the numerator indicates the compactness of the fuzzy partition while the denominator indicates the strength of the separation between clusters [20]. A small value in the numerator of the XB index represents a high level of compactness, while a high value in the denominator denotes good separation between clusters. Hence, the lowest $XB(c)$, $2 \leq c \leq c_{\max}$ indicates the optimal cluster number for a data set X . However, the separation measurement of the XB index adopts the minimal distance between cluster centers; when cluster number c is close to the candidate of the data set, the XB index decreases to the lowest value.

The Fukuyama and Sugeno (FS) index [21] is another index which combines the properties of compactness and separation measurements. The FS index is defined as follows:

$$FS(c) = \sum_{i=1}^n \sum_{k=1}^c (\mu_{ik})^m \|x_i - v_k\|^2 - \sum_{k=1}^c \left\{ \left[\sum_{i=1}^n (\mu_{ik})^2 \right] \cdot \|v_k - \bar{v}\| \right\}, \quad (13)$$

where $1 < m < \infty$, and $\bar{v} = \frac{1}{c} \sum_{i=1}^c v_i$ represents the mean of the cluster centroids. In Equation (13), the first term represents the geometrical compactness of the clusters and the second term indicates the separation between the clusters. When m is 2, the first term of the FS index equals the numerator of the XB index. The second term of the FS index is used for the separation measurements. The optimal cluster number is obtained by solving $\min_{2 \leq c \leq n-1} FC(c)$.

The I-index was proposed by Maulik and Bandyopadhyay [22] in 2002. The I-index is defined as:

$$I(c) = \left(\frac{1}{c} \times \frac{E_1}{E_c} \times D_c \right)^2. \quad (14)$$

Where

$$E_c = \sum_{i=1}^n \sum_{j=1}^c \mu_{ij} \|x_i - v_j\|, \quad (15)$$

and

$$D_c = \max_{i,j=1}^c \|v_i - v_j\|. \quad (16)$$

To find the optimal c^* , the authors tried to solve the $\max_{2 \leq c \leq c_{\max}} I(c)$ in order to obtain the best clustering performance for the data set.

In real applications, data domains are often affected by noises and outliers. The resultant partitions obtained by clustering algorithms are consistently affected by those variables and, surprisingly, there has been little discussion in

the literature about the influence of noises and outliers on validity indexes. In this paper, a novel robust validity index is proposed to indicate the adequacy of partitions in a data set with regard to noise and outliers.

Let $X = \{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^d$ be a data set in a d -dimensional feature space and $W = \{w_{ik} \mid 0 \leq w_{ik} \leq 1\}$, $1 \leq i \leq n, 1 \leq k \leq c$ be a matrix, where w_{ik} is the membership degree of x_i belonging to cluster k , and $v = \{v_1, v_2, \dots, v_c\}$ is a set of the cluster centers, where $v_k \in \mathbb{R}^d$. The w_{ik} is defined by:

$$w_{ik} = \exp\left(-\frac{\|x_i - v_k\|^2}{\rho^2}\right), \quad (17)$$

where $\|x_i - v_k\|$ represents the Euclidean distance between x_i and v_k , and ρ is the width of the Gaussian function. In the proposed method, $\rho = \sqrt{\alpha}$, where α is defined by Equation (4). Let F represent a fuzzy set of data x_i belonging to cluster k . Hence, F can be characterized by Equation (17). An α -cut of a fuzzy set F is a crisp set F_α that contains all the elements in X which have membership values in F greater than or equal to α , that is:

$$F_\alpha = \{x_i \in X \mid w_{ik} \geq \alpha\} \quad (18)$$

This paper assumes that $\alpha = 0.5$ and that Γ represents the number of data satisfying $w_{ik} \geq \alpha$, that is:

$$\Gamma = |F_{0.5}|, \quad (19)$$

where $|\cdot|$ is the number of elements in a set. When a data x_i closes to cluster k , then w_{ik} is greater than or equal to 0.5 and x_i can be partitioned into cluster k . Hence, this paper defines a parameter τ named *partitioned index* to indicate the ratio of data that has been properly partitioned to all data. Thus, τ is defined as:

$$\tau = \frac{\Gamma}{|X|}. \quad (20)$$

For indicating the fitness of the number of partitions in a data set with noises and outliers, this paper combines compactness, separation, and τ properties to propose a new validity index, $CS\alpha$, as follows:

$$CS\alpha(c) = \left(\frac{1}{\tau \times D_{\min}}\right) \times com. \quad (21)$$

Where,

$$D_{\min} = \min_{i \neq j} (\|v_i - v_j\|), \quad (22)$$

$$com = \left(\frac{1}{n^2} \sum_{k=1}^c \sum_{i=1}^n \mu_{ik}^2 \|x_i - v_k\|^2 \times \sum_{k=1}^c \sum_{i=1}^n \mu_{ik} \|x_i - v_k\|\right), \quad (23)$$

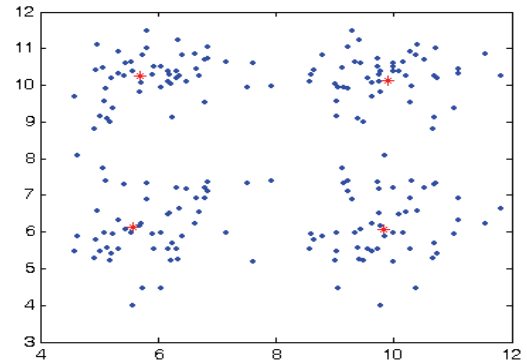
$$\text{and, } \sum_{k=1}^c \mu_{ik} = 1. \quad (24)$$

To find the optimal number of clusters, this paper solves the $\min_{2 \leq c \leq c_{\max}} CS\alpha$. The $CS\alpha$ is subject to three factors: $\frac{1}{\tau}$, $\frac{1}{D_{\min}}$, and com .

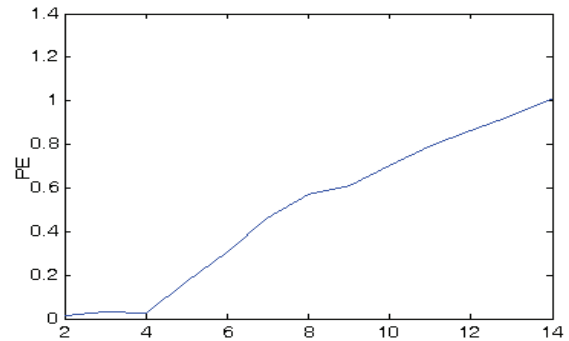
4. Experimental Results

In this section, two data sets are used to show that the validity index proposed in this paper outperforms other indexes proposed in the literature. For all experiments undertaken in this paper, it is reasonable to regard the square root of n as the upper limit of the optimal number of clusters [23], that is $c_{\max} = \sqrt{n}$, where n is the candidate of data set.

Example 1: this example is used to test the performance of various indexes for a data set without noise. Fig. 1(a) shows a data set made up of four clusters and the stars denote the cluster centers obtained by the proposed method. Each cluster has fifty data points, hence, the $c_{\max} = \sqrt{200} = 14$. Figs. 1(b)-(f) show the results obtained by PE, XB, FS, I-index and $CS\alpha$, respectively. The results show that the XB, FS, I-index and $CS\alpha$ indexes can correctly indicate the cluster number of this data set, but the PE index indicate that the optimal numbers of clusters is two. This result shows that the XB, FS, I-index and $CS\alpha$ indexes perform equally as well, but that PE did not have good performance.



(a)



(b)

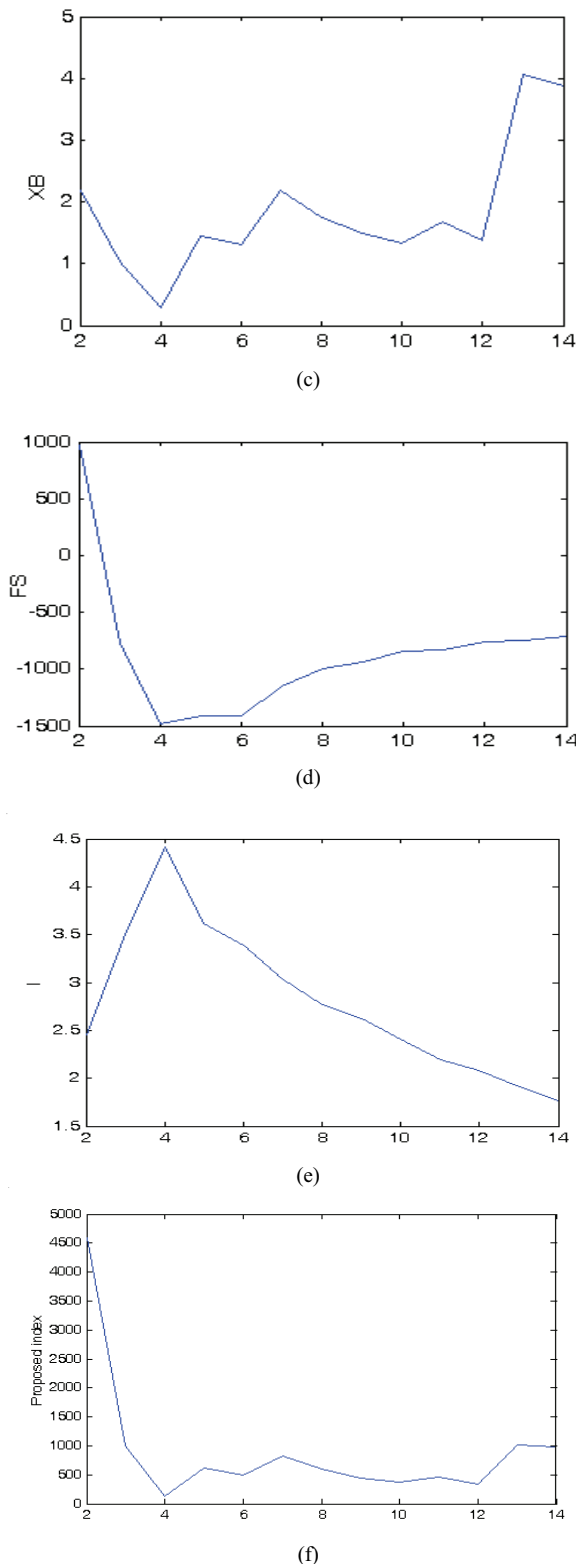


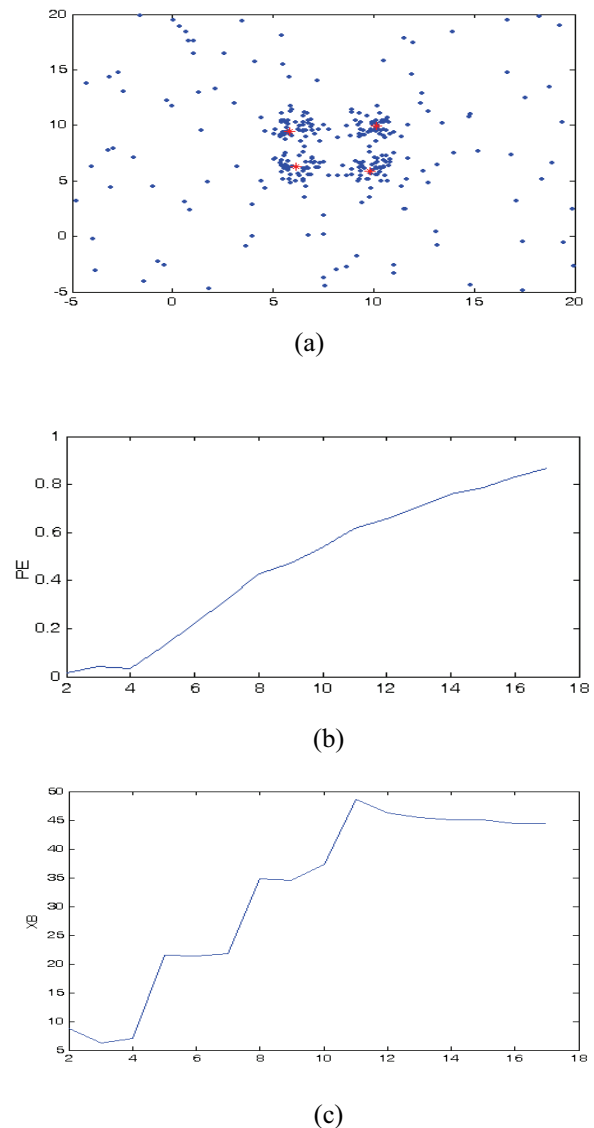
Fig.1: A data set contains four clusters.

Example 2: In this experiment, 100 noises are added to the data set used in Example 1 to test the proposed clustering and validity index. Fig. 2(a) shows the data set containing three clusters and noises and stars denote the cluster centers obtained by the proposed clustering algorithm. The

performances of various indexes for this data set are shown in Figs. 2(b)-(f). The results show that only the proposed index can correctly indicate the optimal number of clusters. The PE, XB, FS, and I-index obtain the optimal number of clusters at two clusters.

5. Conclusions

In this paper, a novel data clustering algorithm based on the SC algorithm and a new validity index are proposed. The proposed clustering algorithm adopts the distance relations between data points and cluster centers obtained by using the SC algorithm to evaluate real cluster centers. In addition, a new validity index composed of compactness, separation, and a partitioned index is proposed to measure the fitness of various cluster numbers. Two examples show that the proposed clustering algorithm can obtain correctly the cluster centers, and the proposed index can indicate correctly the cluster numbers of data set with noises.



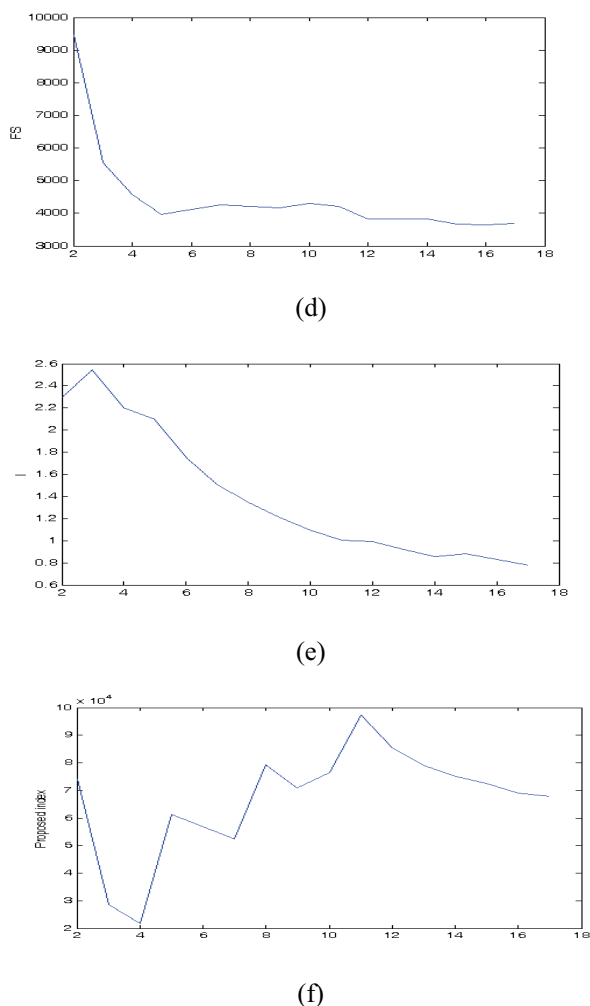


Fig. 2: A data set containing 4 clusters with 100 noises

References

- [1] S. Chatzis, and T. Varvarigou, "Factor Analysis Latent Subspace Modeling and Robust Fuzzy Clustering Using t-Distributions," IEEE Trans. on Fuzzy Systems, vol. 17, no. 3, pp.505-517, 2009.
- [2] H.-L. Shieh, Y.-K. Yang, P.-L. Chang, J.-T. Jeng, "Robust Neural-Fuzzy Method for Function Approximation," Expert Systems with Applications, 36, pp.6903-6913, 2009.
- [3] H.-L. Shieh, Y. K. Yang and C.-N. Lee, "A Robust Approach of Fuzzy Clustering on Data Sets with Noise and Outliers", Cybernetics and Systems: An International Journal, Vol.37, No.7, pp.755-778, 2006
- [4] H.-L. Shieh, Y.-K. Yang, P.-L. Chang, J.-T. Jeng, "Robust Neural-Fuzzy Method for Function Approximation," Expert Systems with Applications, 36, pp.6903-6913, 2009.
- [5] U. Maulik, and I. Saha, "Automatic Fuzzy Clustering Using Modified Differential Evolution for Image Classification," IEEE Trans. on Geoscience and Remote sensing, vol. 48, no. 9, pp. 3502-3510, 2010.
- [6] M. S. Yang and K.-L. Wu, "A Similarity-Based Robust Clustering Method," IEEE Trans. On Pattern Analysis and machine Intelligence, Vol. 25, No. 4, pp. 434-448, April, 2004.
- [7] R. J. Hathaway, J. C. Bezdek, and Y. Hu, "Generalized Fuzzy c-Means Clustering Strategies Using L_p Norm Distances," IEEE Trans. on Fuzzy Set, vol. 8, no. 5, pp.576~582, 2000.
- [8] J.C. Bezdek, "A convergence theorem for the fuzzy ISODATA clustering algorithms," IEEE Trans. on Pattern Analysis and Machine Intelligence, pp.1-8, 1980.
- [9] G. Zouridakis, B. H. Jansen and N. N. Boutros, "A Fuzzy Clustering Approach to EP Estimation," IEEE Tran. On Biomedical Engineering,

Vol. 44, no. 8, pp.673-680, 1977.

- [10] C. Borgelt, "Accelerating fuzzy clustering," Information Sciences, 179, pp.3985-3997, 2009.
- [11] S.-M. Chen, Y.-C. Chang, "Multi-variable fuzzy forecasting based on fuzzy clustering," Information Sciences 180, pp.4772-4783, 2010.
- [12] R. R. Yager and D. P. Filev, "Approximate Clustering Via the Mountain Method," IEEE Trans. on Systems, Man and Cybernetics, vol. 24, no. 8, pp.1279-1284, 1994.
- [13] S. Chiu, "Fuzzy model identification based on cluster estimation," Journal of Intelligent & Fuzzy Systems, 2, pp.267-278, 1994.
- [14] R. N. Dave and R. Krishnapuram, "Robust Clustering Methods: A Unified View," IEEE Trans. on Fuzzy System, vol. 5, no. 2, pp.270-293, 1997.
- [15] K. L. Wu, M. S. Yang, "A cluster validity index for fuzzy clustering," Pattern Recognition Letters, 26, pp.1275-1291, 2005.
- [16] H.-L. Shieh, C.-C. Kuo, "A Novel Validity Index for the Subtractive Clustering Algorithm," to be appeared in International Journal of Pattern Recognition and Artificial Intelligence.
- [17] J. C. Bezdek, "Numerical Taxonomy with Fuzzy Sets," Journal of Mathematical Biology, 1, pp.57-71, 1974.
- [18] J. C. Bezdek, "Cluster validity with fuzzy sets, "Cybernetics and Systems, 3, pp.58-73, 1973.
- [19] X. L. Xie and G. Beni, "A Validity Measure for Fuzzy Clustering," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.13 pp. 841-847, 1991.
- [20] W. Wang and Y. Zhanga, "On fuzzy cluster validity indices," Fuzzy Sets and Systems, 158, pp.2095-2117, 2007.
- [21] Y. Fukuyama and M. Sugeno, "new method of choosing the number of clusters for the fuzzy c-means method." In Proc. 5th Fuzzy Systems Symposium, pp. 247-250, 1989, Japanese.
- [22] U. Maulik, and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," IEEE Trans. on Pattern Analysis Machine Intelligence, vol. 24, no. 12, pp. 1650-1654, 2002.
- [23] J.C. Bezdek and N.R. Pal, "Some new indexes of cluster validity," IEEE Trans. Systems, Man Cybernetics. Part B: Cybernetics, 28(3), pp.301-15, 1998.