Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2014, Article ID 368628, 10 pages
http://dx.doi.org/10.1155/2014/368628

Research Article

An Improved Fuzzy *c*-Means Clustering Algorithm Based on Shadowed Sets and PSO

Jian Zhang[1] and Ling Shen[2]

[1] *School of Mechanical Engineering, Tongji University, Shanghai 200092, China*
[2] *Precision Medical Device Department, University of Shanghai for Science and Technology, Shanghai 200093, China*

Correspondence should be addressed to Jian Zhang; jianzh@tongji.edu.cn

Received 8 June 2014; Revised 30 September 2014; Accepted 14 October 2014; Published 12 November 2014

Academic Editor: Daoqiang Zhang

To organize the wide variety of data sets automatically and acquire accurate classification, this paper presents a modified fuzzy *c*-means algorithm (SP-FCM) based on particle swarm optimization (PSO) and shadowed sets to perform feature clustering. SP-FCM introduces the global search property of PSO to deal with the problem of premature convergence of conventional fuzzy clustering, utilizes vagueness balance property of shadowed sets to handle overlapping among clusters, and models uncertainty in class boundaries. This new method uses Xie-Beni index as cluster validity and automatically finds the optimal cluster number within a specific range with cluster partitions that provide compact and well-separated clusters. Experiments show that the proposed approach significantly improves the clustering effect.

1. Introduction

Clustering is the process of assigning a homogeneous group of objects into subsets called clusters, so that objects in each cluster are more similar to each other than objects from different clusters based on the values of their attributes [1]. Clustering techniques have been studied extensively in data mining [2], pattern recognition [3], and machine learning [4].

Clustering algorithms can be generally grouped into two main classes, namely, supervised clustering and unsupervised clustering where the parameters of classifier are optimized. Many unsupervised clustering algorithms have been developed. One such algorithm is k-means, which assigns n objects to k clusters by minimizing the sum of squared Euclidean distance between the objects in each cluster to the cluster center. The main drawback of the k-means algorithm is that the result is sensitive to the selection of initial cluster centroids and may converge to local optima [5].

For handling those random distribution data sets, soft computing has been introduced in clustering [6], which exploits the tolerance for imprecision and uncertainty in order to achieve tractability and robustness. Fuzzy sets and rough sets have been incorporated in the c-means framework to develop the fuzzy c-means (FCM) [7] and rough c-means (RCM) [8] algorithms.

Fuzzy algorithms can assign data object partially to multiple clusters and handle overlapping partitions. The degree of membership in the fuzzy clusters depends on the closeness of the data object to the cluster centers. The most popular fuzzy clustering algorithm is FCM which is introduced by Bezdek [9] and now it is widely used. FCM is an effective algorithm, but the random selection in center points makes iterative process fall into the saddle points or local optimal solution easily. Furthermore, if the data sets contain severe noise points or if the data sets are high dimensional, such as bioinformatics [10], the alternating optimization often fails to find the global optimum. In these cases, the probability of finding the global optimum can be increased by stochastic methods such as evolutionary or swarm-based methods. Bezdek and Hathaway [11] optimized the hard c-means (HCM) model with a genetic algorithm. Runkler [12] introduced an ant colony optimization algorithm which explicitly minimizes the HCM and FCM cluster models. Al-Sultan and Selim [13] proposed the simulated annealing algorithm (SA) to overcome some of these limits and got promising results.

PSO is a population based optimization tool developed by Eberhart and Kennedy [14], which can be implemented and applied easily to solve various function optimization problems. Runkler and Katz [15] introduced two new methods for minimizing the reformulated objective functions of the FCM clustering model by PSO: PSO-V and PSO-U. In order to overcome the shortcomings of FCM, a PSO-based fuzzy clustering algorithm was discussed [16]; this algorithm uses the global search capacity of PSO to overcome the shortcomings of FCM. For finding more appropriate cluster centers, a generalized FCM optimized by PSO algorithm [17] was proposed.

Shadowed sets are considered as a conceptual and algorithmic bridge between rough sets and fuzzy sets, thereby incorporate the generic merits, and have been successfully used for unsupervised learning. Shadowed sets introduce (0, 1) interval to denote the belongingness of those clustering points, and the uncertainty among patterns lying in the shadowed region is efficiently handled in terms of membership. Thus, in order to disambiguate and capture the essence of a distribution, recently the concept of shadowed sets has been introduced [18], which can also raise the efficiency in the iteration process of the new prototypes by eliminating some "bad points" that have bad influence on cluster structure [19, 20]. Compared with FCM, the capability of shadowed c-means is enhanced when dealing with outlier [21].

Although lots of clustering algorithms based on FCM, PSO, or shadowed sets were proposed, most of them need to input the preestimated cluster number C. To obtain the desirable cluster partitions in a given data, commonly C is set manually, and this is a very subjective and somewhat arbitrary process. A number of approaches have been proposed to select the appropriate C. Bezdek et al. [22] suggested the rule of thumb $C \le N^{1/2}$ where the upper bound must be determined based on knowledge or applications about the data. Another approach is to use a cluster validity index as a measure criterion about the data partition, such as Davies-Bouldin (DB) [23], Xie-Beni (XB) [24], and Dunn [25] indices. These indices often follow the principle that the distance between objects in the same cluster should be as small as possible and the distance between objects in different clusters should be as large as possible. They have also been used to acquire the optimal number of clusters C according to their maximum or minimum value.

Therefore, we wish to find the best C in some range, obtain cluster partitions by considering compactness and intercluster separation, and reduce the sensitivity to initial values. Here, we propose a modified algorithm named as SP-FCM which integrates the merits of PSO and interleaves shadowed sets between stabilization iterations. And it can automatically estimate the optimal cluster number with a faster initialization than our previous approach.

The structure of the paper is as follows. Section 2 outlines all necessary prerequisites. In Section 3, a new clustering approach called SP-FCM is presented for automatically finding the optimal cluster number. Section 4 includes the results of experiments involving UCI data sets, yeast gene expression data sets, and real data set. In Section 5, main conclusions are covered.

2. Related Clustering Algorithms

In this section, we briefly describe some basic concepts of FCM, PSO, shadowed sets, and XB validity index and review the PSO-based clustering method.

2.1. FCM. We define $X = \{x_1, \dots, x_N\}$ as the universe of a clustering data set, $B = \{\beta_1, \dots, \beta_C\}$ as the prototypes of C clusters, and $U = [u_{ij}]_{N \times C}$ as a fuzzy partition matrix, where $u_{ij} \in [0, 1]$ is the membership of x_i in a cluster with prototype β_j; $x_i, \beta_j \in R^P$, where P is the data dimensionality, $1 \le i \le N$, and $1 \le j \le C$. The FCM algorithm is derived by minimizing the objective function [22]

$$J_{\text{FCM}}(U, B, X) = \sum_{j=1}^{C} \sum_{i=1}^{N} u_{ij}^{m} d_{ij}^{2} \left(x_i, \beta_j\right), \qquad (1)$$

where $m > 1.0$ is the weighting exponent on each fuzzy membership and d_{ij} is the Euclidian distance from data vectors x_i to cluster center β_j. And

$$\begin{gathered} \sum_{j=1}^{C} u_{ij} = 1 \quad \forall i = 1, 2, \dots, N, \\ 0 < \sum_{i=1}^{N} u_{ij} < N \quad \forall j = 1, 2, \dots, C, \\ d_{ij} = \left\| x_i - \beta_j \right\|. \end{gathered} \qquad (2)$$

This produces the following update equations:

$$u_{ij} = \left(\sum_{k=1}^{C} \left(\frac{d\left(x_i, \beta_j\right)}{d\left(x_i, \beta_k\right)} \right)^{2/(m-1)} \right)^{-1}, \qquad (3)$$

$$\beta_j = \frac{\left(\sum_{i=1}^{N} \left(u_{ij}\right)^m x_i\right)}{\left(\sum_{i=1}^{N} \left(u_{ij}\right)^m\right)}. \qquad (4)$$

After computing the memberships of all the objects, the new prototypes of the clusters are calculated. The process stops when the prototypes stabilize. That is, the prototypes from the previous iteration are of close proximity to those generated in the current iteration, normally less than an error threshold.

2.2. PSO. PSO was originally introduced in terms of social and cognitive behavior of bird flocking and fish schooling. The potential solutions are called particles which fly through the problem space by following the current best particles. Each particle keeps track of its coordinates in the problem space which are associated with the best solution that has been achieved so far. The solution is evaluated by the fitness value, which is also stored. This value is called $p$best. Another best value that is tracked by the PSO is the best value, obtained so far by any particle in the swarm. The best value is a global best and is called $g$best. The search for the better positions follows the rule as

$$V(t+1) = wV(t) + c_1 r_1 (p\text{best}(t) - P(t)) + c_2 r_2 (g\text{best}(t) - P(t)), \qquad (5)$$
$$P(t+1) = P(t) + V(t+1),$$

where P and V are position and velocity vector of particle, respectively, w is inertia weight, c_1 and c_2 are positive constants, called acceleration coefficients which control the influence of $p$best and $g$best in search process, and r_1 and r_2 are random values in the range $[0, 1]$. The fitness value of each particle's position is determined by a fitness function, and PSO is usually executed with repeated application of (5) until a specified number of iterations have been exceeded or the velocity updates are close to zero over a number of iterations.

2.3. PSO-Based FCM. In this algorithm [26], each particle Part_l represents a cluster center vector, which is constructed as

$$\text{Part}_l = (P_{l1}, \ldots, P_{lj}, \ldots, P_{lC}), \qquad (6)$$

where l represents the lth particle, $l = 1, 2, \ldots L$, L is the number of particles, and $L < N$. P_{lj} is the jth cluster center of particle Part_l. Therefore, a swarm represents a number of candidates cluster center for the data vector. Each data vector belongs to a cluster according to its membership function and thus a fuzzy membership is assigned to each data vector. Each cluster has a cluster center per iteration and presents a solution which gives a vector of cluster centers. This method determines the position vector Part_l for every particle, updates it, and then changes the position of cluster center. And the fitness function for evaluating the generalized solutions is stated as

$$F(P) = \frac{1}{J_{\text{FCM}}}. \qquad (7)$$

The smaller is the J_{FCM}, the better is the clustering effect and the higher is the fitness function $F(P)$.

2.4. Shadowed Sets. Conventional uncertainty models like fuzzy sets tend to capture vagueness through membership values and associate precise numeric values of membership with vague concepts. By introducing α-cut [19], a fuzzy set can be converted into a classical set. Shadowed sets map each element of a given fuzzy set into 0, 1, and the unit interval $[0, 1]$, namely, excluded, included, and uncertain, respectively.

For constructing a shadowed set, Mitra et al. [21] proposed an optimization based on balance of vagueness. As elevating membership values of some regions to 1 and at the same time reducing membership values of some regions to 0, the uncertainty in these regions can be eliminated. To keep the balance of the total uncertainty regions, it needs to compensate these changes by the construction of uncertain regions, namely, shadowed sets that absorb the previous elimination of partial membership at low and high ranges. The shadowed sets are induced by fuzzy membership function in Figure 1.

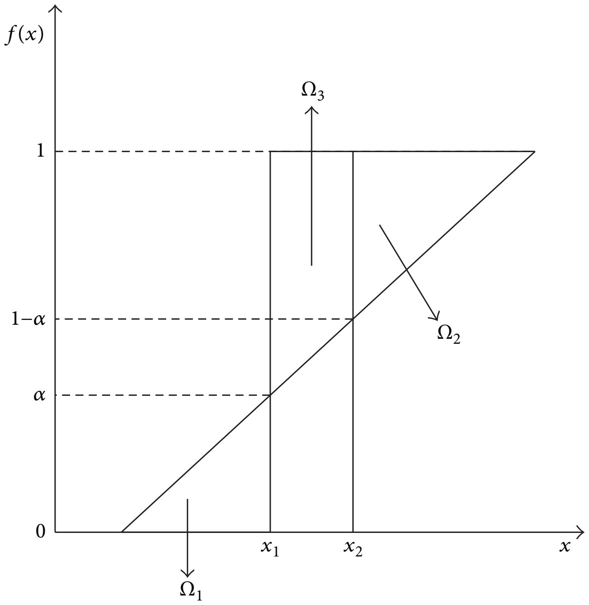


Figure 1: Shadowed sets induced by fuzzy function $f(x)$.

Here x denotes the objects; $f(x) \in [0, 1]$ is the continuous membership function of the objects belonging to a cluster. The symbol Ω_1 shows the reduction of membership, the symbol Ω_2 depicts the elevation of membership, and the symbol Ω_3 shows the formation of shadows. In order to balance the total uncertainty, the retention of balance translates into the following dependency:

$$\Omega_1 + \Omega_2 = \Omega_3. \qquad (8)$$

And the integral forms are given as

$$\Omega_1 = \int_{x: f(x) \leq \alpha} f(x)\,dx, \qquad \Omega_2 = \int_{x: f(x) \geq 1-\alpha} (1 - f(x))\,dx,$$
$$\Omega_3 = \int_{x: \alpha < f(x) < 1-\alpha} dx. \qquad (9)$$

The threshold of reducing and elevating is α and $1 - \alpha$ ($\alpha \in (0, 0.5)$). The optimal value of α should be acquired by translating it into the minimization of the following objective function:

$$O(\alpha) = |\Omega_1 + \Omega_2 - \Omega_3|. \qquad (10)$$

For a fuzzy set with discrete membership function, the balance equation is modified as

$$O\left(\alpha_j\right) = \left| \sum_{u_{ij} \le \alpha_j} u_{ij} + \sum_{u_{ij} \ge u_{j_{\max}} - \alpha_j} \left(u_{j_{\max}} - \alpha_j\right) - \operatorname{card}\left\{u_{ij} \mid \alpha_j < u_{ij} < u_{j_{\max}} - \alpha_j\right\} \right|. \tag{11}$$

In order to find the best α_j, it should satisfy the following optimal problem:

$$\alpha_j = \arg\min_{\alpha_j} O\left(\alpha_j\right), \tag{12}$$

where $u_{ij} \in [0, 1]$ is the membership of x_i in a cluster with prototype β_j; $u_{j_{\max}}$ and $u_{j_{\min}}$ denote the highest and lowest membership values to the jth cluster; and α_j is the threshold of the jth cluster. The range of feasible values of threshold α_j is $[u_{j_{\min}}, (u_{j_{\min}} + u_{j_{\max}})/2]$ [19].

This approach considers all membership values with respect to a fixed cluster when updating the prototype of this cluster. The main merits of shadowed sets involve the optimization mechanism for choosing separate threshold and the reduction of the burden of plain numeric computations.

2.5. XB Clustering Validity Index. The clustering algorithms described above require prespecification of the number of clusters. The partition results are dependent on the choice of C. There exist validity indices to evaluate the goodness of clustering according to a given number of clusters; therefore, these validity indices can be used to acquire the optimal value of C [27].

The XB index presents a fuzzy-validity criterion based on a validity function which identifies overall compact and separate fuzzy c-partitions. This function depends upon the data set, geometric distance measure, and distance between cluster centroids and fuzzy partition, irrespective of any fuzzy algorithm used. For evaluating the goodness of the data partition, both cluster compactness and intercluster separation should be taken into account. For the FCM algorithm with $m = 2.0$, the Xie-Beni index can be shown to be

$$S_{\mathrm{XB}} = \frac{J_{\mathrm{FCM}}}{N d_{\min}^2}, \tag{13}$$

where $d_{\min} = \min_{i,j} \left\|\beta_i - \beta_j\right\|$ is the minimum distance between cluster centroids. The more separate the clusters, the larger the $d_{\min}$ and the smaller the S_{XB}.

3. Shadowed Sets-Based PSO-Fuzzy Clustering: SP-FCM

FCM strives to find C compact clusters in X where C is one of the specified parameters. But the process of selecting and adjusting C manually to obtain desirable cluster partitions in a given data set is very subjective and somewhat arbitrary. To seek the optimal cluster structure, C is always allowed to be overestimated [28], such that the distances between some clusters are not big enough or the membership values of some objects with different clusters are adjacent and ambiguous in a given data set. And, in this case, the modification of prototypes through long time iteration is meaningless.

The main subject of cluster validation is the evaluation of clustering results to find the partitioning that best fits the data set. Based on the foregoing algorithms, we wish to find cluster partitions that contain compact and well-separated clusters. In our algorithm C is also overestimated and the clusters compete for data membership. We can set $[C_{\min}, C_{\max}]$ as the reasonable range of cluster number based on the knowledge of the data. This provides a more transparent and tractable process of cluster number reduction. Considering the fuzzy partition matrix $U = [u_{ij}]_{N\times C}$, each column is comprised of the membership values of all feature vectors x_i with a single cluster center. Thus, an optimal threshold α_j ($j = 1, 2, \ldots C$) for each column should be found to create a harder partition by (12). The amount of data which are assigned membership value equal to 1 is identified as the cardinality of corresponding cluster. According to α_j, the cardinality of the jth column is

$$M_j = \operatorname{card}\left\{u_{ij} \mid u_{ij} \ge u_{j_{\max}} - \alpha_j\right\}. \tag{14}$$

Here, the threshold is not subjectively user-defined but it is established on the balance of uncertainty and can be adjusted automatically in the clustering process. This property of shadowed sets can be used to reduce the cluster number. In order to control the convergence speed, the decision to delete clusters can be based on some thresholds. Different threshold values should be set for different data sets depending on the cluster structure and size of data sets. Here, a threshold ε and attrition rate ρ ($0 < \rho < 1$) are set. The decision to delete clusters in SP-FCM is based solely on cluster cardinality and the threshold ε. If ε is too small, C is reduced more slowly and it may stop prematurely before the optimal cluster number is found. On the other hand, if ε is too large, C may be reduced too drastically. In our method, clusters whose cardinalities $M_j < \varepsilon$ are considered as "candidates" for removal. And we can remove up to $\lfloor \rho \times C \rfloor$ clusters having the lowest cardinality from the pool of candidates specified by ε. Limiting the number of clusters that can be removed at one time prevents C from being reduced too drastically when ε is set too high for a given data set. This would automatically estimate the best cluster number while also utilizing a faster, consistent, and repeatable initialization technique. For evaluating the goodness of the data partition, both cluster compactness and intercluster separation should be taken into account. Hence the XB index is adopted.

For each C in the range of $[C_{\min}, C_{\max}]$ a set of cluster validity indexes were calculated, where $C_{\max}$ is the initial cluster number which is set to be much larger than the expected cluster number. The partition matrix with C clusters with the best aggregate validity index is selected as the final cluster partition. The SP-FCM algorithm is summarized as in Algorithm 1.

```
(1) Initialize $C_{\max}$ and $C_{\min}$, let $C = C_{\max}$, the real number $m$, iteration counter $k = 0$, iteration counter $t = 0$, maximum
    iteration number $T$ of PSO.
(2) Initialize the population size $L$, the initial velocity of particles, the initial position of particles, $c_1$, $c_2$, $w$, the threshold $\varepsilon$
    and attrition rate $\rho$.
(3) DO {
        Repeat {
            (a) Update partition matrix $U(k)$for all particles by (3).
            (b) Calculate the cluster center for each particle by (4).
            (c) Calculate the fitness value for each particle by (7).
            (d) Calculate pbest for each particle.
            (e) Calculate gbest for the swarm.
            (f) Update the velocity and position of each particle by (5).
            (g) $t = t + 1$
        }
        Until PSO termination condition is met $(*)$
    (i) Calculate the optimal threshold $\alpha_j (j = 1, 2, \ldots, C)$ for each column of partition matrix $U(k)$ by (12),
        and relocate $u_{ij} (1 \le i \le N)$ of $j$th cluster according to $\alpha_j$
    (ii) Calculate cardinality $M_j$ for each cluster on the basis of the number of data whose membership value equal to 1 by (14),
        $1 \le j \le C$
    (iii) Remove all clusters whose $M_j < \varepsilon$ and $M_j$ is among $\lfloor \rho \times C \rfloor$ lowest cardinality
    (iv) Update cluster number $C$
    (v) Calculate cluster validity index $S_{\mathrm{XB}}(k)$ by (13)
    (vi) Update iteration counter $k = k + 1$
    }
While termination condition is not met $(**)$
    $(*)$ The termination condition of PSO in this method is $t \ge T$ (reach the maximum number of iterations) or the velocity
        updates are close to zero over a number of iterations.
    $(**)$ The algorithm can terminate under either of the following two conditions:
    (1) The prototype parameters in $B$ stabilize within some threshold $\delta$.
    (2) The number of clusters has reached the minimum limit $C_{\min}$.
```

Algorithm 1: SP-FCM.

Here, if $\lfloor \rho \times C \rfloor$ is equal to 0, we can let it to be 1. This means that the cluster with the lowest cardinality may be removed. The initial $C_{\max}$ cluster prototypes can be initialized using exemplars from data points selected by $\beta_j = x_{\lfloor (N/C_{\max}) j \rfloor}$. After termination, the B and U from $C \in [C_{\min}, C_{\max}]$ with the best cluster validity index S_{XB} are selected as the final cluster prototype and partition.

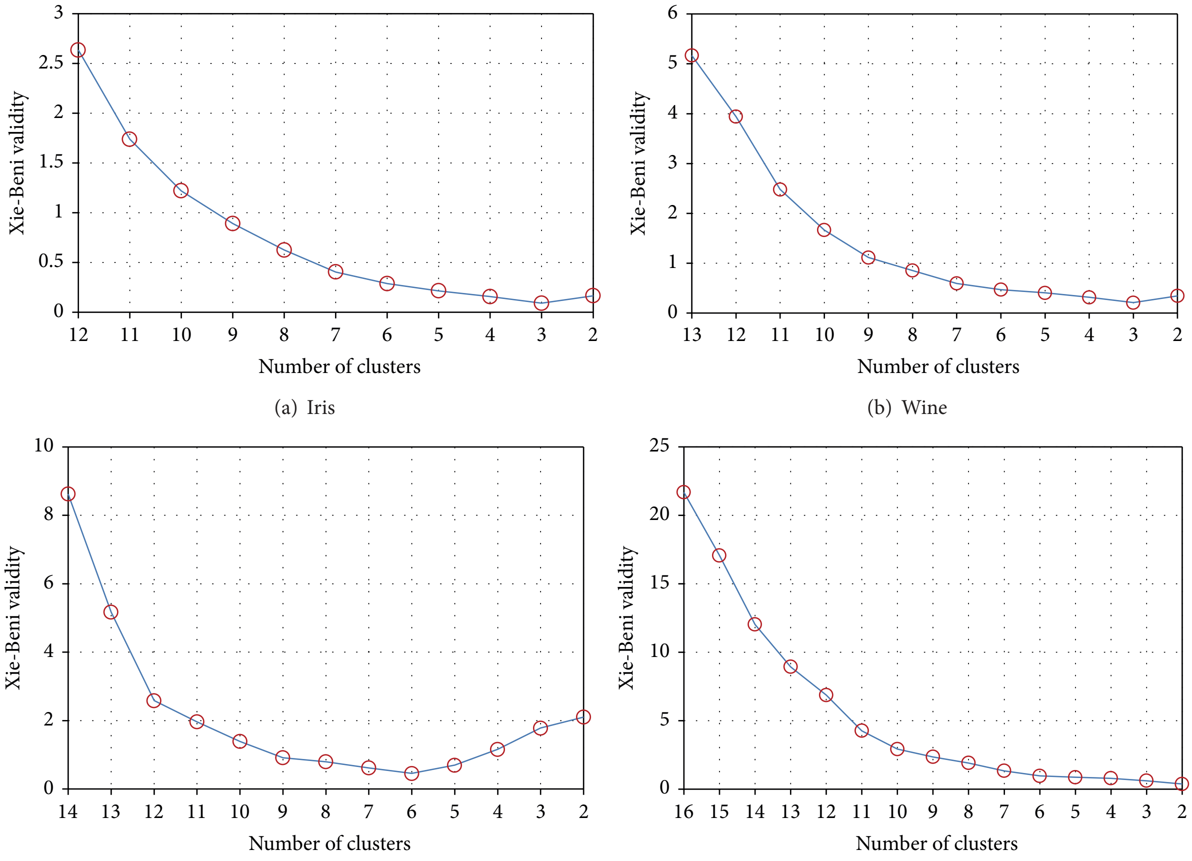
4. Experimental Results

In this section, the performance of FCM, RCM, shadowed c-means (SCM) [21], shadowed rough c-means (SRCM) [19], and SP-FCM algorithms is presented on four UCI datasets, four yeast gene expression datasets, and real data. For evaluating the convergence effect, the fundamental criterion can be described as follows: the distance between different objects in the same cluster should be as close as possible; the distance between different objects in different cluster should be as far as possible. Here we use DB index and Dunn index to evaluate the clustering effect. For a given data set and C value, the higher the similarity values within the clusters and the intercluster separation, the lower the DB index value. A good clustering procedure should make the value of DB index as low as possible. Reversely, higher values of the Dunn index indicate better clustering in the sense that the clusters are well separated and relatively compact. The details of experiments are mentioned below.

4.1. UCI Data Set. In our experiments, totally four UCI data sets are used, including 4-dimensional Iris, 13-dimensional Wine, 10-dimensional Glass, and 34-dimensional Ionosphere. There are 3 clusters in data set of Iris, each of which has 50 data patterns; 3 clusters in data set of Wine, which have 50, 60, and 68 data patterns; 6 clusters in data set of Glass, which have 30, 35, 40, 42, 36, and 31 separately; and 2 clusters in data set of Ionosphere, which have 226 and 125 data patterns. The validity indices of each method are compared in Table 1. SP-FCM can identify compact groups compared to other algorithms when given the cluster number C. It can also be seen that SRCM and SP-FCM have more obvious advantages than FCM, RCM, and SCM. SP-FCM performs slightly better than SRCM in most cases due to the global search ability which enables it to converge to an optimum or near optimum solutions. Moreover, shadowed set- and rough set-based clustering methods, namely, SP-FCM, SRCM, RCM, and SCM, perform better than FCM. It implies that the partition of approximation regions can reveal the nature of data structure and only the lower bound and boundary region of each cluster have positive contribution in the process of updating the prototypes.

Table 1: Performance of FCM, RCM, SCM, SRCM, and SP-FCM on four UCI data sets.

Different indices	Algorithm	Data sets			
		Iris	Wine	Glass	Ionosphere
DB index	FCM	0.7642	0.8803	2.2971	2.0587
	RCM	0.6875	0.5692	1.9635	1.5434
	SCM	0.6862	0.5327	1.8495	1.4763
	SRCM	0.6613	0.4436	1.5804	1.3971
	SP-FCM	0.6574	0.4328	1.5237	1.4066
Dunn index	FCM	2.3106	2.5834	0.1142	0.8381
	RCM	2.7119	2.8157	0.2637	1.0233
	SCM	2.4801	2.7992	0.3150	1.0319
	SRCM	3.0874	3.1342	0.5108	1.1924
	SP-FCM	3.3254	3.1764	0.4921	1.2605



(a) Iris

(b) Wine

(c) Glass

(d) Ionosphere

Figure 2: XB validity index of four UCI data sets with cluster number C.

As usual, the number of clusters is implied by the nature of the problem. Here, with the shadowed sets involved, one can anticipate that the optimal number of clusters could be found. The fuzzification coefficient m can be optimized; however, it is common to assume a fixed value of 2.0, which associates with the form of the membership functions of the generated clusters. For testing the SP-FCM algorithm, the rule $C \le N^{1/2}$ is adopted, and the range of the expected cluster number can be set as (1) Iris, $[C_{\min} = 2, C_{\max} = 12]$; (2) Wine, $[C_{\min} = 2, C_{\max} = 13]$; (3) Glass, $[C_{\min} = 2, C_{\max} = 14]$; (4) Ionosphere, $[C_{\min} = 2, C_{\max} = 16]$. The swarm size is set as $L = 20$, the maximum iteration number of PSO $T = 50$, and, for cluster reduction, the cluster cardinality threshold $\varepsilon = 10$ and the attrition rate $\rho = 0.1$. In each cycle, we get the distribution of every cluster, remove part of them according to their cardinality, and calculate the XB index, and the cluster number C varies from $C_{\max}$ to $C_{\min}$. After ending the circulation, the partition with the lowest value is selected as the final result. Figure 2 presents the validity indices in the process of generating the optimal cluster number. Smaller

Table 2: Performance of FCM, RCM, SCM, SRCM, and SP-FCM on four yeast expression data sets.

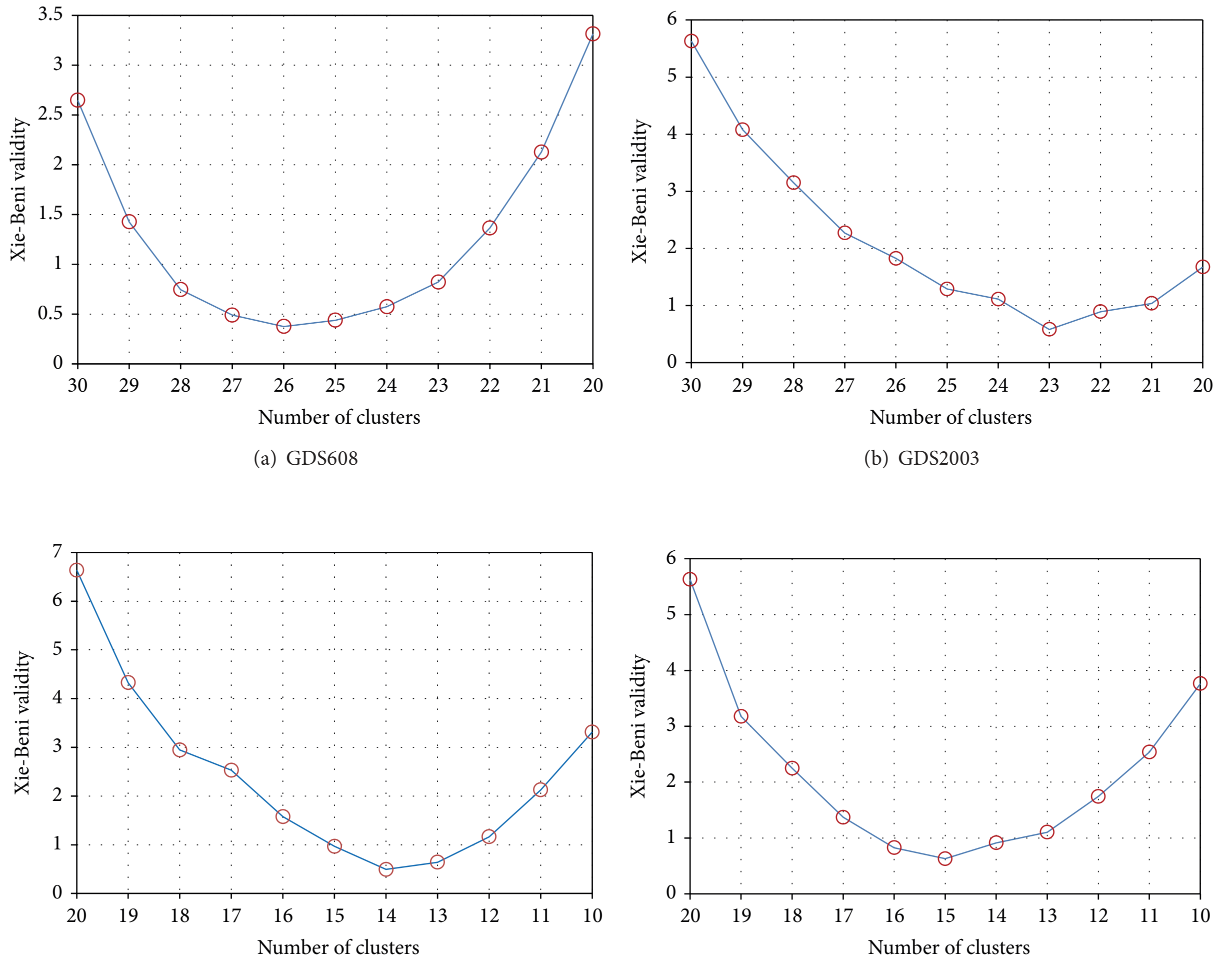
Different indices	Algorithm	Data sets			
		GDS608	GDS2003	GDS2267	GDS2712
DB index	FCM	2.0861	2.4671	1.5916	1.9526
	RCM	1.6109	2.2104	1.0274	1.2058
	SCM	1.5938	2.1346	0.8946	1.0965
	SRCM	1.3274	1.9523	0.7438	0.7326
	SP-FCM	1.2958	1.8946	0.7962	0.6843
Dunn index	FCM	0.2647	0.2976	0.4208	0.3519
	RCM	0.3789	0.3981	0.7164	0.6074
	SCM	0.3865	0.3775	0.8439	0.6207
	SRCM	0.5126	0.4953	0.9759	0.8113
	SP-FCM	0.5407	0.5026	0.9182	0.8049

values indicate more compact and well-separated clusters. The validity indices reach their minimum value at $C = 3$, 3, 6, and 2 separately, which correspond to the final cluster prototype and the best partition. Through the shadowed sets and PSO approaches, the influence of each boundary region to the formation of the prototypes and the clusters can be properly resolved. Although more computing time is required to run SP-FCM, the reasonable result can be acquired for processing the overlapping and vagueness data patterns.

4.2. Yeast Gene Expression Data Set. There are four yeast gene expression data sets used in the experiments, including GDS608, GDS2003, GDS2267, and GDS2712 downloaded from Gene Expression Omnibus. The number of classes and samples of GDS608 is 26 and 6303; for GDS2003, the number of classes and samples is 23 and 5617, for GDS2267 is 14 and 9275, and for GDS2712 is 15 and 9275. Table 2 presents the validity indices of different methods after the cluster number C was given. The SP-FCM and SRCM obtain the same effect and perform better than other clustering algorithms. The improvement can be attributed to the fact that the global search capacity of PSO is conducive to finding more appropriate cluster centers while escaping from local optima.

For getting the optimum C automatically, we let $m = 2.0$, $c_1 = 1.49$, $c_2 = 1.49$, and $w = 0.72$, and the rule $C \leq N^{1/2}$ is adopted. The swarm size is set as $L = 20$, the maximum iteration number of PSO is $T = 80$, and, for cluster reduction, the range of the expected cluster number, the cluster cardinality threshold ε, and the attrition rate ρ can be set as (1) GDS608, $[C_{\min} = 20, C_{\max} = 80]$, $\varepsilon = 20$, $\rho = 0.05$; (2) GDS2003, $[C_{\min} = 20, C_{\max} = 75]$, $\varepsilon = 20$, $\rho = 0.05$; (3) GDS2267, $[C_{\min} = 10, C_{\max} = 96]$, $\varepsilon = 20$, $\rho = 0.08$; (4) GDS2712, $[C_{\min} = 10, C_{\max} = 96]$, $\varepsilon = 20$, $\rho = 0.08$. In each cycle, we get the distribution of every cluster, remove part of them according to their cardinality, and calculate the XB index, and the cluster number C varies from $C_{\max}$ to $C_{\min}$. The partition with the lowest value is selected as the final result after the loop is ended. As seen in Figure 3, for GDS608, at the beginning the cluster number decreases at a faster rate, it takes 26 iterations to reduce the cluster number from $C = 80$ to $C = 30$ and 4 iterations from $C = 30$ to $C = 26$, and the XB index begins to increase when the cluster number $C < 26$. For GDS2003, it takes 24 iterations to reduce the cluster number from $C = 75$ to $C = 30$ and 7 iterations from $C = 30$ to $C = 23$, and the XB index begins to increase when the cluster number $C < 23$. For GDS2267, it takes 23 iterations to reduce the cluster number from $C = 96$ to $C = 20$ and 6 iterations from $C = 20$ to $C = 14$, and the XB index begins to increase when the cluster number $C < 14$. For GDS2712, it takes 23 iterations to reduce the cluster number from $C = 96$ to $C = 20$ and 5 iterations from $C = 20$ to $C = 15$, and the XB index begins to increase when the cluster number $C < 15$. Here, the advantages of fuzzy sets, PSO, and shadowed sets are integrated in the SP-FCM and make this algorithm applicable to deal with overlapping partitions, the uncertainty, and vagueness arising from the boundary regions, and the optimization process in the shadowed sets makes this method robust to outliers, so that the approximation regions of each cluster can be determined accurately and the obtained prototypes approach the desired locations.

4.3. Real Data. In this experiment totally 10 different packages are tested. Each package is represented by 100 frames captured from different angles by camera, and each frame is extracted SIFT feature points which are used for training a recognition system. Figure 4 shows some images with their SIFT keypoints. And this data set is comprised of 248150 descriptors. We let $m = 2.0$, $c_1 = 1.49$, $c_2 = 1.49$, $w = 0.72$, $L = 20$, $\varepsilon = 30$, and $\rho = 0.01$ for the SP-FCM and choose the reasonable range $[C_{\min} = 200, C_{\max} = 360]$ according to the category amount of packages and distribution of keypoints in each image. Eighty iterations of PSO are run on each given C to produce the cluster prototype B and partition matrix U as the starting point for the shadowed sets. Longer PSO stabilization is needed to obtain more stable cluster partitions.



(a) GDS608

(b) GDS2003

(c) GDS2267

(d) GDS2712

Figure 3: XB validity index of four yeast gene expression data sets with cluster number C.

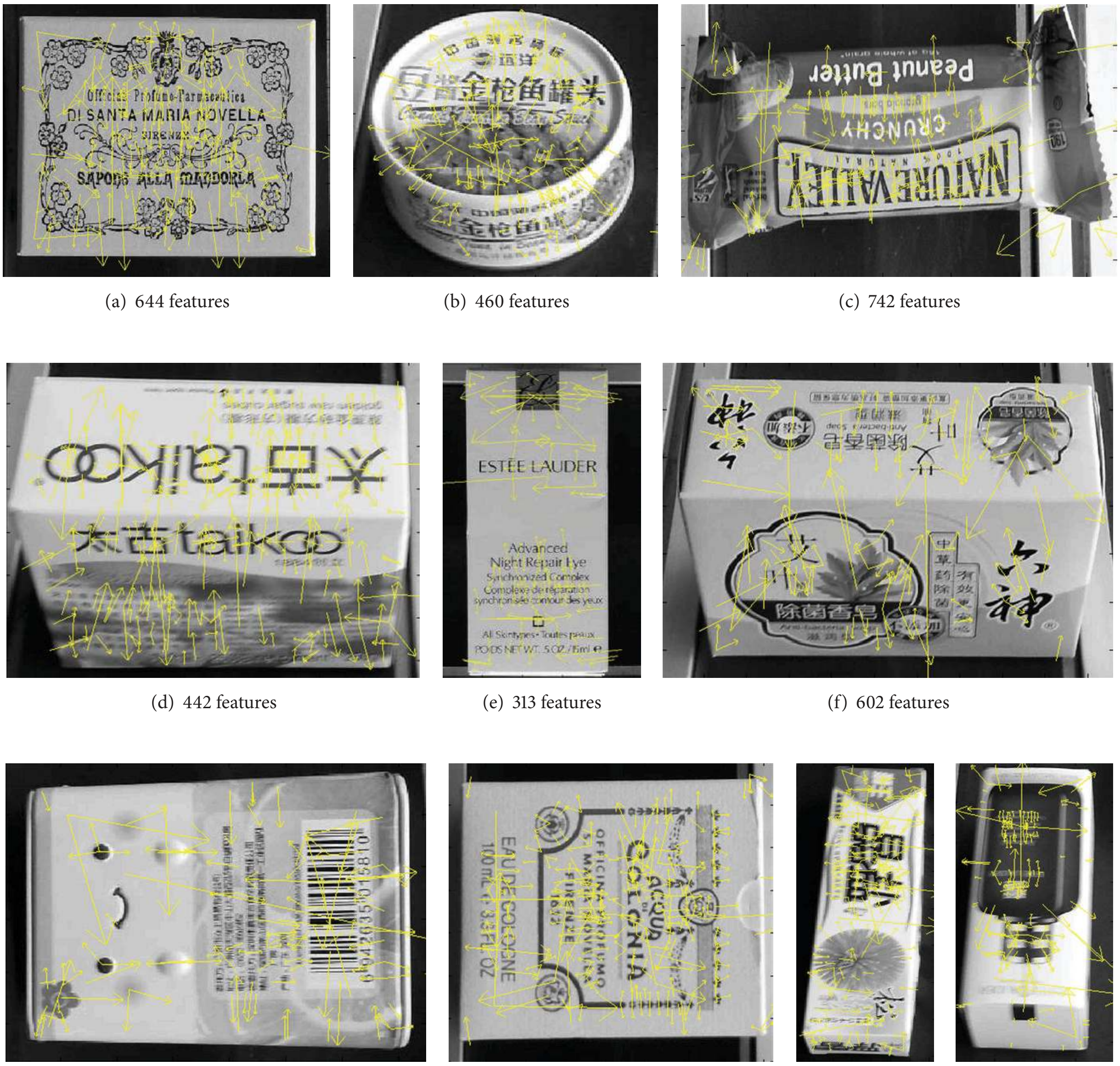
Table 3: Performance of FCM, RCM, SCM, SRCM, and SP-FCM on package datasets.

Different indices	Algorithms				
	FCM	RCM	SCM	SRCM	SP-FCM
DB index	184.569	159.671	143.194	124.038	107.826
Dunn index	92.647	116.298	125.376	169.422	167.313

Within each cluster, the optimal α_j decides the cardinality and realizes cluster reduction, and XB index is calculated. Each C-partition is ranked using this index and selected as the final output by the smallest index value that indicates the best compact and well-separated clusters. At the beginning, the cluster number decreases at a faster speed; it takes 26 iterations to reduce the cluster number from $C = 360$ to $C = 289$ and 20 iterations from $C = 289$ to $C = 267$. The XB index increases at a relatively faster rate when the cluster number $C < 267$. Figure 5 shows the XB index for $C \in [267, 289]$. The index reaches its minimum value at $C = 276$ that means the best partition for this data set is 276 clusters. Table 3 exhibits the comparative analysis of convergence effect. As expected, SP-FCM can provide sound results for the real data; the performance is assessed by those validity indices.

5. Conclusions

This paper presents a modified fuzzy c-means algorithm based on the particle swarm optimization and shadowed sets to perform unsupervised feature clustering. This algorithm called SP-FCM utilizes the global search property of PSO and vagueness balance property of shadowed sets, such that it can estimate the optimal cluster number as it runs through its alternating optimization process. SP-FCM as a randomized based approach has the capability to alleviate the problems faced by FCM, which has some demerits of initialization and falling in local minima. Moreover, this algorithm avoids the subjective and somewhat arbitrary trials to estimate the appropriate value of cluster number, and it enhances this capability to find the optimal cluster number within a specific



(a) 644 features (b) 460 features (c) 742 features

(d) 442 features (e) 313 features (f) 602 features

(g) 373 features (h) 724 features (i) 539 features (j) 124 features

Figure 4: Ten package images with SFIT features.

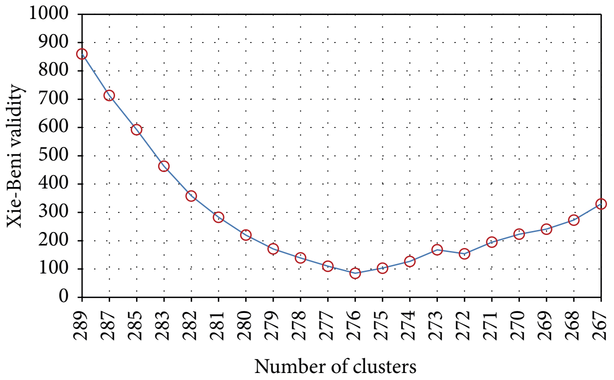


Figure 5: XB validity index of bag data set with cluster number C.

range using cluster validity measures as indicators. The use of XB validity index allows the algorithm to find the optimum cluster number with cluster partitions that provide compact and well-separated clusters. From the experiments, we have shown that the SP-FCM algorithm produces good results with reference to DB and Dunn indices, especially to the high dimension and large data cases.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (Grant no. 61105089, NSFC).

References

[1] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[2] H.-P. Kriegel, P. Kröger, and A. Zimek, "Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering," *ACM Transactions on Knowledge Discovery from Data*, vol. 3, no. 1, article 1, 2009.

[3] P. Melin and O. Castillo, "A review on type-2 fuzzy logic applications in clustering, classification and pattern recognition," *Applied Soft Computing Journal*, vol. 21, pp. 568–577, 2014.

[4] M. Yuwono, S. W. Su, B. D. Moulton, and H. T. Nguyen, "Data clustering using variants of rapid centroid estimation," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 3, pp. 366–377, 2014.

[5] S. C. Satapathy, G. Pradhan, S. Pattnaik, J. V. R. Murthy, and P. V. G. D. P. Reddy, "Performance comparisons of PSO based clustering," *InterJRI Computer Science and Networking*, vol. 1, no. 1, pp. 18–23, 2009.

[6] G. Peters and R. Weber, "Dynamic clustering with soft computing," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 226–236, 2012.

[7] 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.

[8] P. Maji and S. Paul, "Robust rough-fuzzy C-means algorithm: design and applications in coding and non-coding RNA expression data clustering," *Fundamenta Informaticae*, vol. 124, no. 1-2, pp. 153–174, 2013.

[9] J. Bezdek, *Fuzzy mathematics in pattern classification [Ph.D. thesis]*, Cornell University, New york, NY, USA, 1974.

[10] V. Olman, F. Mao, H. Wu, and Y. Xu, "Parallel clustering algorithm for large data sets with applications in bioinformatics," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, no. 2, pp. 344–352, 2009.

[11] J. C. Bezdek and R. J. Hathaway, "Optimization of fuzzy clustering criteria using genetic algorithms," in *Proceedings of the 1st IEEE Conference on Evolutionary Computation*, pp. 589–594, June 1994.

[12] T. A. Runkler, "Ant colony optimization of clustering models," *International Journal of Intelligent Systems*, vol. 20, no. 12, pp. 1233–1251, 2005.

[13] K. S. Al-Sultan and S. Z. Selim, "A global algorithm for the Fuzzy Clustering problem," *Pattern Recognition*, vol. 26, no. 9, pp. 1357–1361, 1993.

[14] R. Eberhart and J. Kennedy, "New optimizer using particle swarm theory," in *Proceedings of the 6th International Symposium on Micro Machine and Human Science*, pp. 39–43, Nagoya, Japan, October 1995.

[15] T. A. Runkler and C. Katz, "Fuzzy clustering by particle swarm optimization," in *Proceedings of the IEEE International Conference on Fuzzy Systems*, pp. 601–608, can, July 2006.

[16] C.-F. Juang, C.-M. Hsiao, and C.-H. Hsu, "Hierarchical cluster-based multispecies particle-swarm optimization for fuzzy-system optimization," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 1, pp. 14–26, 2010.

[17] H. Izakian and A. Abraham, "Fuzzy C-means and fuzzy swarm for fuzzy clustering problem," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1835–1838, 2011.

[18] W. Pedrycz, "Shadowed sets: representing and processing fuzzy sets," *IEEE Transactions on Systems, Man, and Cybernetics B: Cybernetics*, vol. 28, no. 1, pp. 103–109, 1998.

[19] J. Zhou, W. Pedrycz, and D. Miao, "Shadowed sets in the characterization of rough-fuzzy clustering," *Pattern Recognition*, vol. 44, no. 8, pp. 1738–1749, 2011.

[20] W. Pedrycz, "From fuzzy sets to shadowed sets: interpretation and computing," *International Journal of Intelligent Systems*, vol. 24, no. 1, pp. 48–61, 2009.

[21] S. Mitra, W. Pedrycz, and B. Barman, "Shadowed c-means: integrating fuzzy and rough clustering," *Pattern Recognition*, vol. 43, no. 4, pp. 1282–1291, 2010.

[22] J. C. Bezdek, J. M. Keller, R. Krishnapuram, and N. R. Pal, *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, vol. 4 of *The Handbooks of Fuzzy Sets*, Springer, New York, NY, USA, 1999.

[23] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, pp. 224–227, 1978.

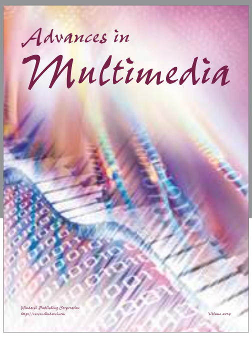
[24] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841–847, 1991.

[25] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 28, no. 3, pp. 301–315, 1998.

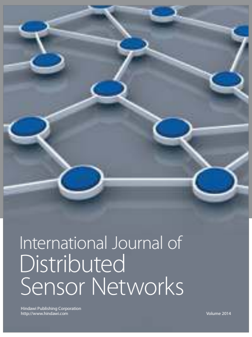
[26] S. C. Satapathy, S. K. Patnaik, C. D. P. Dash, and S. Sahoo, "Data clustering using modified fuzzy-PSO (MFPSO)," in *Proceedings of the 5th Multi-Disciplinary Int. Workshop on Artificial Intelligence*, Hyderabad, India, 2011.

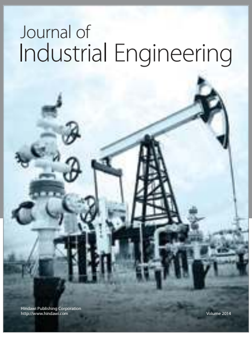
[27] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern Recognition*, vol. 37, no. 3, pp. 487–501, 2004.

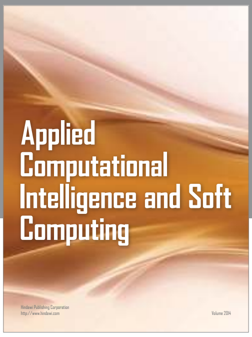
[28] R. Krishnapuram and J. M. Keller, "The possibilistic C-means algorithm: insights and recommendations," *IEEE Transactions on Fuzzy Systems*, vol. 4, no. 3, pp. 385–393, 1996.

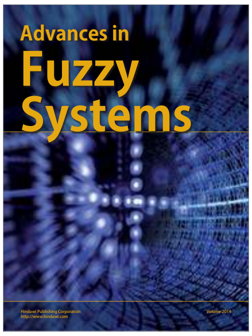

Advances in
Multimedia

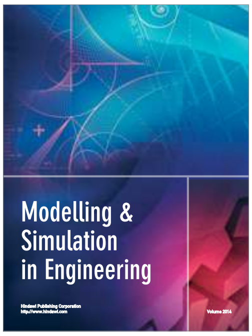
The Scientific
World Journal

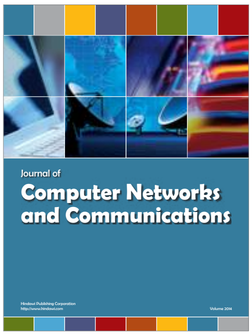

International Journal of
Distributed
Sensor Networks

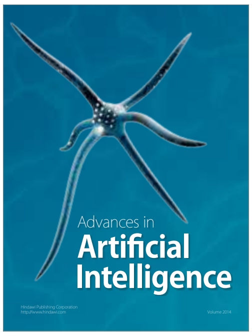

Journal of
Industrial Engineering

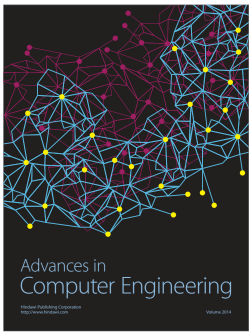

Applied
Computational
Intelligence and Soft
Computing


Advances in
Fuzzy
Systems

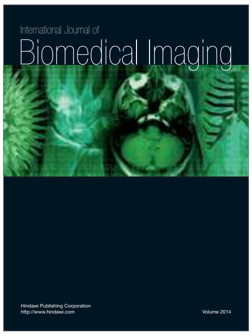

Modelling &
Simulation
in Engineering

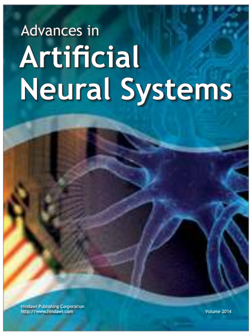

Journal of
Computer Networks
and Communications

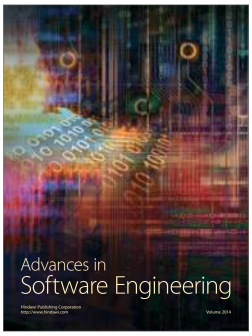

Advances in
Artificial
Intelligence


Advances in
Computer Engineering

International Journal of
Computer Games
Technology

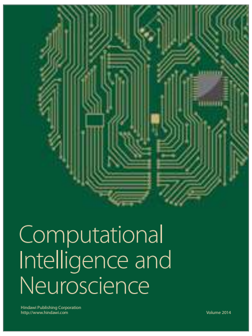

International Journal of
Biomedical Imaging


Advances in
Artificial
Neural Systems


Advances in
Software Engineering

Journal of
Robotics

Advances in
Human-Computer
Interaction


Computational
Intelligence and
Neuroscience

International Journal of
Reconfigurable
Computing

Journal of
Electrical and Computer
Engineering