DE GRUYTER
OPEN

JAISCR, 2014, Vol. 4, No. 1, pp. 43 – 56
DOI 10.2478/jaiscr-2014-0024

PARTICLE SWARM OPTIMIZATION BASED FUZZY CLUSTERING APPROACH TO IDENTIFY OPTIMAL NUMBER OF CLUSTERS

Min Chen and Simone A. Ludwig
Department of Computer Science, North Dakota State University
Fargo, ND, USA

Abstract

Fuzzy clustering is a popular unsupervised learning method that is used in cluster analysis. Fuzzy clustering allows a data point to belong to two or more clusters. Fuzzy c-means is the most well-known method that is applied to cluster analysis, however, the shortcoming is that the number of clusters need to be predefined. This paper proposes a clustering approach based on Particle Swarm Optimization (PSO). This PSO approach determines the optimal number of clusters automatically with the help of a threshold vector. The algorithm first randomly partitions the data set within a preset number of clusters, and then uses a reconstruction criterion to evaluate the performance of the clustering results. The experiments conducted demonstrate that the proposed algorithm automatically finds the optimal number of clusters. Furthermore, to visualize the results principal component analysis projection, conventional Sammon mapping, and fuzzy Sammon mapping were used.

1 Introduction

Data mining is a field that is also referred to as exploratory data analysis, which is an analytic process designed to explore data. Data mining aims to search for consistent patterns or systematic relationships between variables within the data. The validation process then verifies the findings by applying the detected patterns to new subsets of data [1]. The application areas in which data mining is used for are astronomy, banking, customer relationship management, climate modeling, ecology, finance, life sciences, monitoring, manufacturing, network, retail, security, surveillance, web applications, etc. Data mining is a statistical analysis process that identifies clusters among a collection of data. The different tasks of data mining are classification, association, prediction, sequential pattern, and clustering [2].

Clustering analysis is one of the popular approaches and has been widely used in data mining, and is a process to identify groups or clusters based on some similarity measures. Most clustering algorithms can be categorized into two popular techniques known as hierarchical and partitional clustering. For hierarchical clustering, the output is a tree showing a sequence of clusters whereby each cluster being a partition of the data set. Hierarchical clustering on the other hand does not specify the number of clusters beforehand, and the output is independent of the initial condition. Furthermore, hierarchical clustering is static, i.e., the data points assigned to a cluster cannot be reassigned to another cluster. In addition, hierarchical clustering will fail to separate overlapping clusters due to the lack of information regarding the global shape or size of the clusters. Since partitional clustering requires a fixed number of clusters to be specified a priori, this is clearly a shortcoming. Usually objective functions such as square error function are used as a criterion in the optimization process during the data partitioning. Partitional clustering uses an iterative process to both optimize the cluster centers and the number of clusters. However, the challenge is to find the “optimum” number of clusters since it always requires prior knowledge about the data. To summarize, the advantages of hierarchical algorithms are the disadvantages of the partitional algorithms and vice versa.

The goal of clustering is to divide data points into homogeneous groups such that the data points in the same group are as similar as possible, and data points in different groups are as dissimilar as possible [3, 2]. The strength of clustering is documented in different areas such as pattern recognition [4], machine learning, image analysis [5], information retrieval, etc. Clustering methods can be categorized as either hard clustering [6, 7] or fuzzy clustering [8] depending on whether a data point belongs to a single cluster or several clusters with different membership degrees. Fuzzy set theory was proposed by Zadeh [9] in 1965. It is used to describe the membership degrees in fuzzy cluster analysis. Thus, each data point of a data set belongs to two or more clusters with a membership degree between 0 and 1. Due to the capability of handling uncertainty and vagueness, the potential of fuzzy clustering can be exploited to reveal the underlying structures in data with regard to similarities or dissimilarities among them [10].

One of the widely used methods in fuzzy clustering is Fuzzy C-Means clustering (FCM) [11]. FCM attempts to partition a data set into a collection of c fuzzy groups. The algorithm finds a cluster center in each group such that the intra-distance within the group is minimized, and the inter-distance between each group is maximized. Most of the fuzzy clustering methods that have been applied recently use an extension of the FCM algorithm. As we have mentioned before, partitional clustering suffers from the following drawbacks:

- The number of clusters needs to be pre-specified, and prior knowledge or ground truth is required of the data.
- Most data points in overlapping areas cannot be categorized correctly.

In order to address these two shortcomings, we proposed a fuzzy c-means clustering approach using a Particle Swarm Optimization (PSO) approach that is applied to clustering analysis. The remainder of this paper is organized as follows: In Section 2, fuzzy c-means and PSO are introduced. The proposed algorithm is described in Section 3, and a list of validity indices is given in Section 4. The experimental results and analysis is described in Section 5, and the paper is concluded in Section 6.

2 Related Work

CM was first developed by [12] in 1973, and was extended by [11] in 1981. Since then, FCM is one of the best fuzzy clustering methods. Many different variants of FCM have been introduced. For example, the Gustafson-Kessel (GK) algorithm [7] is a fuzzy clustering technique which can estimate local covariance to partition data into subsets, which can be well fitted with linear sub-models. However, since considering a general structure of the covariance matrix can have a substantial effect on the modeling approach, the Gath-Geva algorithm [13] was proposed to overcome this shortcoming. Another algorithm, called Fuzzy C-Varieties (FCV) [14] clustering algorithm, is a fuzzy clustering method for which the prototype of each cluster is represented as a multi-dimensional linear vector. The approach is similar to cluster analysis, however, it uses the statistical method of principal component analysis for the clustering task. Another algorithm, referred to as generalized FCM algorithm, is presented in [15], in which setting of the algorithm parameters is being done automatically.

Related work lists many evolutionary computation methods that have been applied for clustering. For example, a hybrid technique based on combining the k-means algorithm and Nelder-Mead simplex search was applied for cluster analysis in [16]. Another algorithm based on the combination of Genetic Algorithm (GA), k-means and logarithmic regression expectation maximization was introduced in [17]. In [18], a k-means algorithm that performs correct clustering without preassigning the exact number of clusters was proposed. A genetic k-means algorithm for cluster analysis was introduced in [19]. In [20], a GA based method to solve the clustering problem and experiment on synthetic and real life data sets to evaluate the performance was proposed. A GA algorithm that exchanges neighboring centers for k-means clustering has been introduced in [21]. A combination of evolutionary algorithm with an ant colony algorithm for the clustering problem was presented in [21, 22]. A clonal selection based method has been combined with FCM in [23].

PSO has also been applied to data clustering. In particular, two methods called PSO-V and PSOU are introduced in [24], whereby a reformulated objective function of fuzzy c-means is minimized by the PSO algorithm for the cluster analysis task. Another PSO-based fuzzy clustering algorithm is introduced to overcome the shortcomings of FCM in [25]. An ant colony clustering algorithm is applied for solving the clustering task in [26]. The algorithm uses the global pheromone update and heuristic information to find clustering solutions. In [27], a genetic fuzzy K-modes algorithm for clustering categorical data is proposed, which uses a genetic algorithm to obtain the global optimal clustering solution. A hybrid data clustering algorithm that uses the merits of PSO and K-harmonic means is proposed in [28]. The hybrid algorithm helps to escape from local optima, and thus overcomes the problem of slow convergence of the PSO algorithm. A hybrid evolutionary algorithm, called FAPSO-ACO-K, is introduced in [29]. The hybrid algorithm combines PSO, ACO and k-means applied to cluster analysis. Another method for dynamic parameter adaptation in PSO is proposed in [30]. The proposed algorithm uses fuzzy logic to improve the convergence and diversity of the swarm in PSO.

The high computational cost and the slow convergence rate severely limit the use of PSO on clustering analysis. For these reasons, a chaotic map PSO with an accelerated convergence rate strategy was introduced in [31]. The algorithm works by adopting chaotic maps and adaptive action to avoid local minima. In [32], a hybrid fuzzy clustering method based on FCM and FPSO is proposed to overcome the shortcomings of PSO. Another modified version of PSO, known as Multi-Elitist PSO (MEPSO), is proposed in [33]. This approach solves the hard clustering problem by automatically determining the optimal number of clusters. This approach shows that PSO is guaranteed to solve clustering problems automatically.

This paper addresses the shortcoming of the FCM algorithm, which is the predefined cluster count. A clustering approach based on PSO is proposed whose aim it is to automatically determine the optimal number of clusters using a threshold vector. The algorithm partitions the data set randomly (within a preset maximum number of clusters) and uses a reconstruction criterion to evaluate the performance of the clustering results. This paper is an extended version of [34].

3 Fuzzy C-Means and Particle Swarm Optimization

3.1 Fuzzy C-means Clustering

Fuzzy clustering is a method of clustering that allows one piece of data to belong to two or more clusters. The FCM algorithm is an iterative partition clustering technique that was first introduced by Dunn [12], and was then extended by Bezdek [11]. FCM uses a standard least squared error model that generalizes an earlier and very popular non-fuzzy c-means model that produces hard clusters of the data. An optimal c partition is iteratively produced by minimizing the weighted within group sum of squared error objective function

$$J = \sum_{i=1}^{n} \sum_{j=1}^{c} (u_{ij})^m d^2(y_i, c_j) \tag{1}$$

where $Y = [y_1, y_2, ..., y_n]$ is the data set in a d-dimensional vector space, n is the number of data items, c is the number of clusters that is defined by the user where $2 \leq c \leq n$, u_{ij} is the degree of membership of y_i in the j^{th} cluster, m is a weighted exponent on each fuzzy membership, c_j is the center of cluster j, $d^2(x_i, c_j)$ is a square distance measure between object y_i and cluster c_j.

An optimal solution with c partitions can be obtained via an iterative process which is as follows:

- Input(c, m, ε, data)
- Initialize the fuzzy partition matrix $U = [u_{ij}]$
- Iteration starts and set t=1
- Calculate the c cluster centers with U^t:
$$c_i = \frac{\sum_{i=1}^{n} (u_{ij})^m y_i}{\sum_{i=1}^{n} (u_{ij})^m} \tag{2}$$
- Calculate the membership U^{t+1} using:
$$u_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{d_{ij}}{d_{kj}})^{\frac{2}{(m-1)}}} \tag{3}$$

- If the stopping criteria is not met, $t = t + 1$ and go to Step 4)

3.2 Particle Swarm Optimization

PSO was designed and introduced by Eberhart and Kennedy [35]. PSO is a population-based search algorithm that simulates the choreography of a bird flock. Each individual, called particle, within the swarm is represented by a vector in a multidimensional search space. A velocity vector is assigned to each particle to determine the next movement of the particle. Each particle updates its velocity based on the current velocity, best personal position it has explored so far, and the global best position explored by the swarm

The velocity and position of the particle at the next iteration is updated as

$$V_i(t+1) = wV_i(t) + c_1 r_1 (X_i^l(t) - X_i(t)) + c_2 r_2 (X^g - X_i(t)) \quad (4)$$

$$X_i(t+1) = X_i(t) + V_i(t+1) \quad (5)$$

for the i^{th} particle, where w is the inertia weight, $V_i(t)$ is the previous velocity at iteration t of i^{th} particle, c_1 and c_2 are coefficients. Generally, r_1 and r_2 are random numbers between 0 and 1. $(X_i^l(t) - X_i(t))$ is the difference between the local best X_i^l of the i^{th} particle and the previous position $X_i(t)$. Similarly, $(X^g - X_i(t))$ is the difference between the global best X^g and the previous position $X_i(t)$.

4 Proposed Approach

The proposed algorithm is based on PSO and FCM. The particle encoding, velocity encoding, decoding and clustering validation are described separately below followed by the procedures of the proposed algorithm.

4.1 Particle Encoding

A particle is a $2 \times k$ matrix, where k is the maximum number of clusters that is predefined. The first row represents the centers. Each value in the second row controls the activation of each center in the first row.

$$X_i = \begin{pmatrix} x_{1,1}^i & x_{1,2}^i & \cdots & x_{1,k}^i \\ t_{2,1}^i & t_{2,2}^i & \cdots & t_{2,k}^i \end{pmatrix} \quad (6)$$

where $x_{1,k}^i$ represents the i^{th} particle's position in cluster k. $x_{1,k}^i$ should be in the range of $[x_{min}, x_{max}]$. $t_{2,k}^i$ is the i^{th} particle's threshold value in the range of $[0,1]$. If the threshold value is greater than 0.5, the center is activated, otherwise it is deactivated.

4.2 Velocity Encoding

The velocity matrix has to have the same dimension as the position matrix with a range. Suppose we set the range as $[v_{min}, v_{max}]$, all values of the velocity matrix have to be between v_{min} and v_{max}. Thus, the i^{th} velocity is denoted as:

$$V_i = \begin{pmatrix} v_{x1,1}^i & v_{x1,2}^i & \cdots & v_{x1,k}^i \\ v_{t2,1}^i & v_{t2,2}^i & \cdots & v_{t2,k}^i \end{pmatrix} \quad (7)$$

Similarly, k is the maximum number of clusters. The first row represents the velocity of the centers, and the second row represents the velocity of the threshold values.

4.3 Decoding

$Y = (y_1, y_2, .., y_n)$ is the data set with d dimensions. The cluster centers can be decoded as $C = (c_1, c_2, ...c_k)$ using Equation 2.

4.4 Clustering Validation Techniques

The aim of clustering validation is to evaluate the clustering results by finding the best partition that fits the underlying data. Thus, cluster validity is used to quantitatively evaluate the results of clustering algorithms. Compactness and separation two widely considered criteria for measuring the quality of the partitioning of a data set into different numbers of clusters. Conventional approaches use an iterative approach by choosing different input values, and they select the best validity measure to determine the ”optimum” number of clusters. A list of validity indices for fuzzy clustering is listed below.

4.4.1 Dunn's Index (DI)

The Dunn's Index is proposed to identify the compactness and separation of the clusters. The function that uses to calculate the result of the clustering is as follow

$$DI = \min\{ \min_{j \in c, i \neq j} \{ \frac{\min_{x \in C_i, y \in C_j} d(x,y)}{\max_{k \in c}\{\max_{x,y \in C} d(x,y)\}} \} \} \tag{8}$$

where $d(x,y)$ is the distance of the two cluster centers. DI takes its minimum value when the cluster structure is optimal.

4.4.2 Weighted Inter-Intra (Wint) Index

The weighted inter-intra (Wint) measure is introduced by Strehl [42] in 2002. It compares the compactness of the data to its separation.

$$Wint = (1 - \frac{2c}{n}) \cdot (1 - \frac{\sum_i \frac{1}{n-|c_i|} \sum_{j \neq i} inter(c_i, c_j)}{\sum_i \frac{2}{|c_i|-1} intra(c_i)}) \tag{9}$$

where $intra(c_i)$ is the average intra-distance within cluster i. $inter(c_i, c_j)$ is the average inter-distance between cluster i and cluster j. Wint obtains its maximum value when the cluster structure is optimal.

4.4.3 Least Squared Error (SE) Index

The weighted within cluster sum of squared error function is used [36]:

$$J_m = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^m ||y_i - c_j||^2 \tag{10}$$

where y_i is the i^{th} data point with d dimensions. c_j is the value of the j^{th} cluster, and $||y_i - c_j||$ is the Euclidean distance between y_i and c_j. J_m takes its minimum value when the cluster structure is best.

4.4.4 Partition Coefficient (PC) Index

The partition coefficient (PC) is defined as [11]

$$PC = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^2 \tag{11}$$

PC obtains its maximum value when the cluster structure is optimal.

4.4.5 Partition Entropy (PE) Index

The partition entropy was defined as [14]

$$PE = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij} log_b(u_{ij}) \tag{12}$$

where b is the logarithmic base. PE gets its minimum value when the cluster structure is optimal.

4.4.6 Modified Partition Coefficient (MPC) Index

Modification of the PC index, which can reduce the monotonic tendency, is proposed by Dave in 1996 [37].

$$MPC = 1 - \frac{c}{c-1}(1 - PC) \tag{13}$$

where c is the number of cluster. An optimal cluster number is found by maximizing MPC to produce a best clustering performance for a data set.

4.4.7 Fukuyama and Sugeno (FS) Index

Fukuyama and Sugeno proposed a validity function in 1989 [38]. It is defined as

$$FS = \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij}^m ||x_i - c_j|| - \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij}^m ||c_j - \bar{c}|| \tag{14}$$

where $\bar{c} = \sum_{j=1}^{c} c_j / c$. It measures the separation. The first term equals to J_m which is the least squared error. It measures the compactness. The best clustering performance for a data set is found by maximizing the value of FS.

4.4.8 Xie-Beni (XB) Index

Xie and Beni proposed a validity function in 1991 [39], and later it was modified by Bezdek in 1995 [40].

$$XB = \frac{J_m}{n \times \min_{i \neq j} ||z_i - z_j||^2} \tag{15}$$

XB reaches its minimum value when the cluster structure is optimal.

4.4.9 Partition Coefficient and Exponential Separation (PCAES) Index

The partition coefficient and exponential separation (PCAES) index [41] is defined as

$$PCAES = \sum_{i=1}^{n} \sum_{j=1}^{c} \frac{(u_{ij})^2}{u_M} - \sum_{k=1}^{c} exp(-\min_{k \neq i} ||z_i - z_k||^2 / \beta_T) \tag{16}$$

where $u_M = \min_{1 \leq j \leq c}\{\sum_{i=1}^{n} u_{ij}^2\}$ and $\beta_T = (\sum_{j=1}^{c} ||z_j - \bar{z}||^2)/c$. $\bar{z} = \sum_{i=1}^{n}(y_i/n)$. PCAES takes its maximum value when the cluster structure is optimal.

The procedure of the proposed algorithm is as follows:

Input: data set $Y = [y_1, y_2, ..., y_n]$, number of cluster c, fuzzification coefficient m.

Output: a $n \times c$ partition matrix U and corresponding centers.

- Randomly initialize a swarm
- Iteration starts and set t=1
- Update the velocity of each particle using Equation 4
- Update the position of each particle using Equation 5
- Update the personal best and global best
- Calculate the partition matrix U
- If the stopping criterion is not met, $t = t + 1$ and go back to Step 3)
- The partition matrix U of the global best is used to reconstruct the original data
- Calculate the reconstruction error. In order to use a consistent method to evaluate the eight different indices, the reconstruction criterion (RC) [43] is used. The reconstruction criterion uses the cluster prototypes and partition matrix to “reconstruct” the original data vectors. The reconstructed version of the original data vectors, $\hat{Y} = [\hat{y}_1, \hat{y}_2, ..., \hat{y}_n]$, is calculated as

$$\hat{y}_i = \frac{\sum_{j=1}^{c} u_{ij}^m c_j}{\sum_{j=1}^{c} u_{ij}^m} \tag{17}$$

  Once the reconstruction has been finished, the squared error of the reconstruction vectors and original vectors are evaluated using Equation 18.

$$E = \sum_{i=1}^{n} ||\hat{y}_i - y_i||^2 \tag{18}$$

- Select the partition matrix and centers corresponding to the minimum reconstruction error.

5 Experiments and Results

In this section, the experimental setup, datasets and experimental study are described in detail.

5.1 Experimental Setup

The experiments are implemented and evaluated on an ASUS desktop (Intel(R) Dual Core I3 CPU @3.07 GHz, 3.07 GHz) Matlab Version 7.13. All measurements of the proposed algorithm are executed 30 times and the average is taken. The parameters required for the proposed algorithm are listed in Table 1.

Table 1: Parameters and their values of the proposed algorithm.

Parameter	Value
Maximum number of cluster	10
Maximum iteration	50
Swarm size	25
Maximum run	30
Fuzzification coefficient (m)	2

5.2 Datasets

The experiments are conducted on a number of datasets taken from the UCI repository [44], and synthetic data sets were generated using Matlab. The datasets are described in Table 2.

Table 2: Datasets used for the experiments.

Data Set	Dimensions	Instances	Classes
Pinwheel	2	1000	2
Transfusion	4	748	2
Haberman	3	306	2
Breast-W	9	699	2
Jain	2	373	2
Thyroid	5	215	2
Iris	4	150	3
DIM032	32	320	5
DIM064	64	320	5
DIM128	128	320	5
DIM256	256	320	5

5.3 Experimental Study

5.3.1 Use of Synthetic Data

In order to investigate the clustering performance with different numbers of clusters, we use a synthetic data set, named pinwheel, to test the clustering performance using K-means [46], K-medoid [47], FCM, Gustafson-Kessel (GK) and our proposed algorithm (FPSO). K-means is one of the unsupervised learning methods that uses an iterative refinement technique. The number of desired cluster, *k*, is defined in advance. K-medoid is another unsupervised learning method related to the K-means algorithm. Similarly, the K-medoid classifies the data set into *k* clusters. However, K-medoid is more robust to noise and outliers as compared to K-means. Unlike K-means, the medoid is defined as the data point whose average dissimilarity within the cluster is minimal. K-means and K-medoid are traditional hard clustering techniques, while FCM, GK and FPSO are soft clustering techniques. The nine validity indices listed in Equations 8-16 are used.

The cluster performance of the pinwheel data set is displayed in Figure 1. The first figure is the original data set. As can be seen, the cluster centers are different using Kmeans, K-medoid, FCM, GK and FPSO, respectively.

Figure 2 shows the performance of the pinwheel data set using the K-means algorithm. Figure 3 shows the performance of the pinwheel data set using the K-medoid algorithm. Since K-means and K-medoid are algorithms using hard partitioning, the DI index, Wint index and SE index are used for validation.

In Figure 4, the performance of the pinwheel data set using FCM is given. Figure 5 shows the performance of the pinwheel data set using the GK algorithm. The performance of the proposed algorithm (FPSO) is displayed in Figure 6. The correct number of clusters found for the nine indices are listed in Table 3. The correct cluster number for the pinwheel data set is 2. The correct number of clusters found measuring DI using the five different algorithms are 5, 5, 7, 10, and 7, respectively. The correct number of clusters found applying Wint are consistent with the correct cluster number. The correct number of clusters found by SE is consistent with 10. As the number of clusters increases, the SE values decrease. PC, PE and MPC using the FCM, GK and FPSO algorithms find the correct cluster number. Measuring FS using the FCM, GK and FPSO algorithm are similar to SE. As the number of clusters increases, the FS values decrease. The correct number of clusters found applying XB using the FCM, GK and FPSO algorithms are 10, 7, and 10, respectively. The correct number of clusters found measuring PCAES using the FCM, GK and FPSO algorithms are 4, 10, and 8, respectively. Overall, the Wint, PC, PE and MPC indices, which outperform the other indices, find the correct number of clusters.

5.3.2 Use of Real-World Data

In this section, we investigate the behavior of the clustering results using nine different validity measures listed.

In Table 4, the reconstruction errors of the transfusion data set, where c ranges from 2 to 9, have been calculated using the proposed algorithm by applying Equations 8-16. As shown by the results, the values in bold identifying the minimum reconstruction errors with different cluster numbers for each measure. 6 out of 8 cases show that $c = 2$ is the correct number of clusters. This indicates that the proposed FPSO can find the best number of clusters automatically.

Due to the stochastic nature of our proposed algorithm, we tested the proposed algorithm on 30 runs and calculated the average number of clusters as listed in Table 5. The standard deviation values are given as well. The correct numbers of clusters using different validity measures are tabulated, respectively. In all the cases, the number of clusters predicted by FPSO is close to the correct number of clusters. DI, SE, XB and PCAES do not find the correct number of clusters. Wint can identify the correct number of clusters but only for low-dimensional datasets. MPC returns the correct number of clusters, but with larger standard deviation values. PC and PE find the correct number of clusters consistently, however, as the number of dimension increases, the accuracy decreases.

5.4 Visualization of Clustering Results

Since the validity measures reduce the overall evaluation to a certain number, therefore there is some loss of information. In order to better analyze the results, a low-dimensional graphical representation of the clusters is adopted. A toolbox implemented by [45] is used to visualize the cluster results using the proposed algorithm. Principal

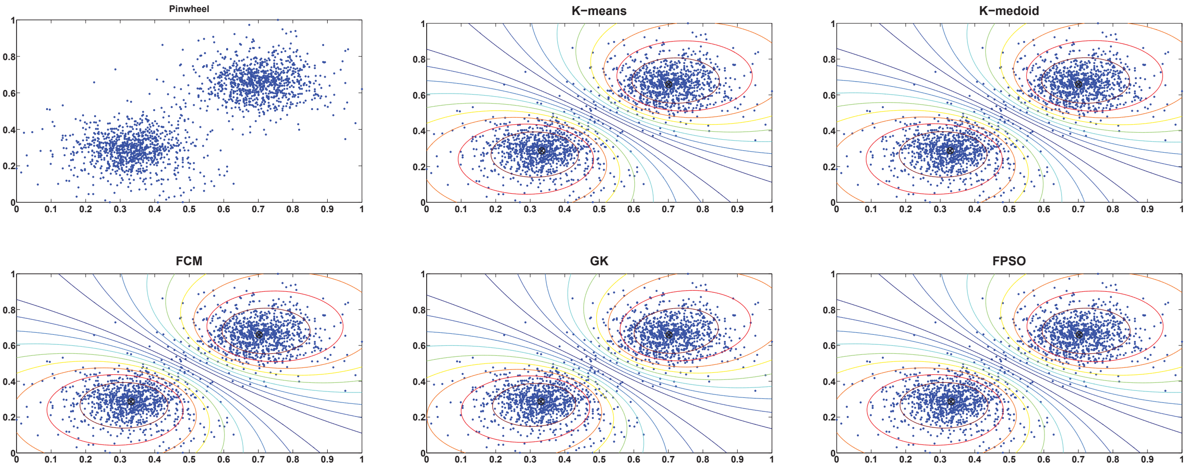

Figure 1: Synthetic data set

Table 3: Index values with varying *c* using pinwheel data set.

Indices	Kmeans	K-medoid	FCM	GK	FPSO
DI	5	5	7	10	7
Wint	2	2	2	2	2
SE	10	10	10	10	10
PC	-	-	2	2	2
PE	-	-	2	2	2
MPC	-	-	2	2	2
FS	-	-	10	10	10
XB	-	-	10	7	10
PCAES	-	-	4	10	8

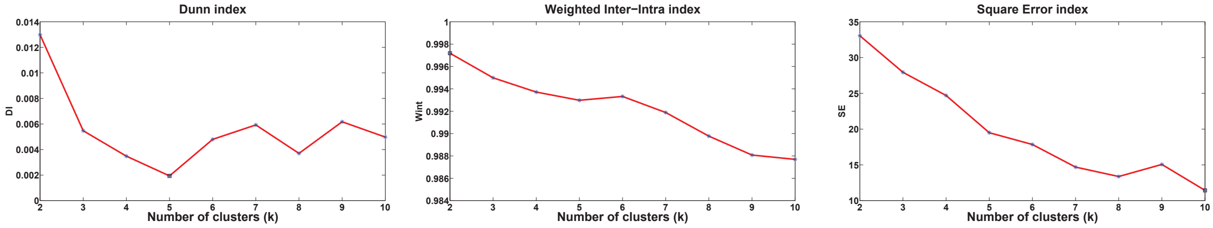

Figure 2: Kmeans using 3 different validity indices

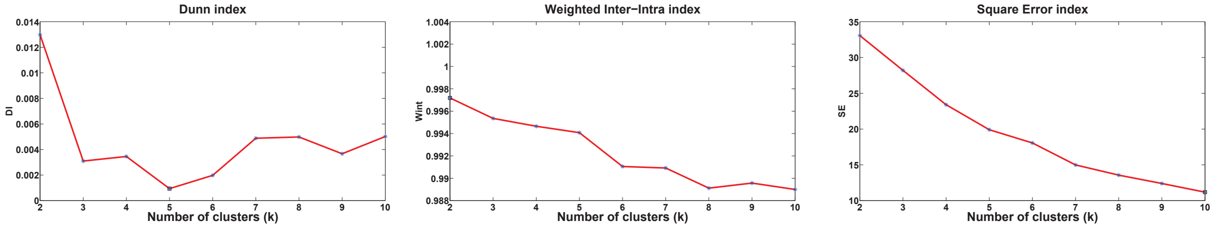

Figure 2: Kmeans using 3 different validity indices

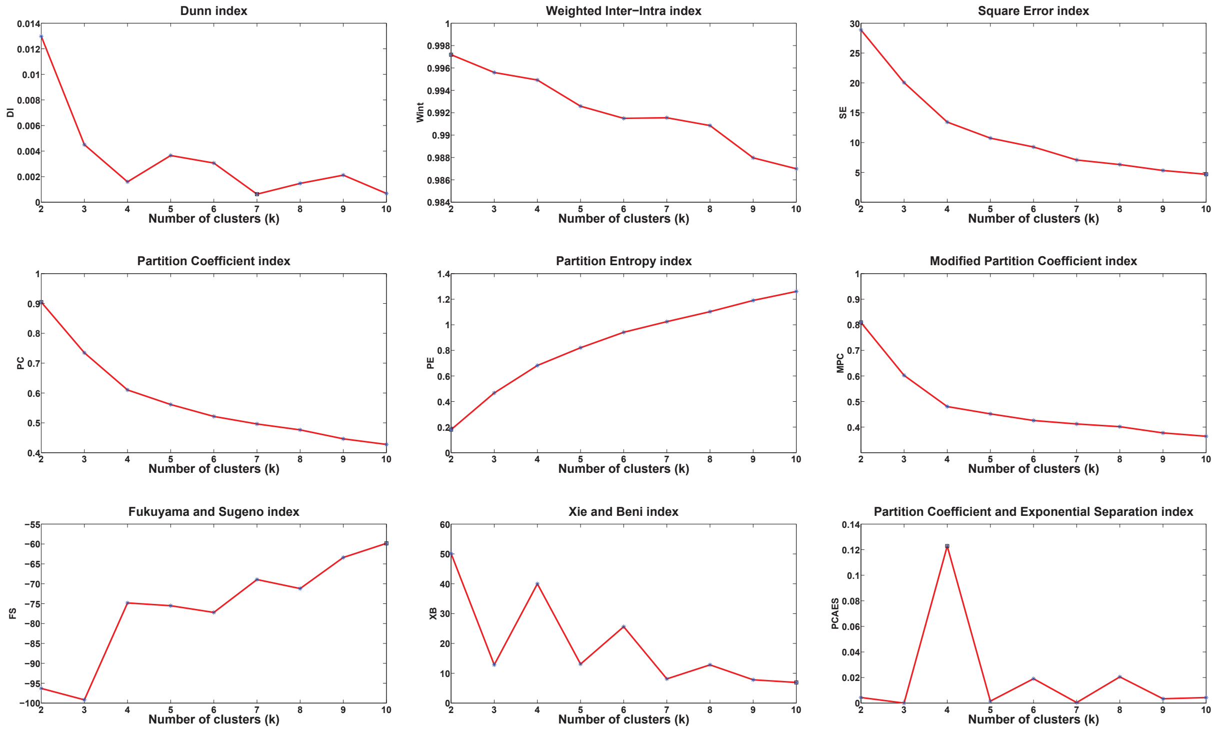

Figure 4: FCM using 9 different validity indices

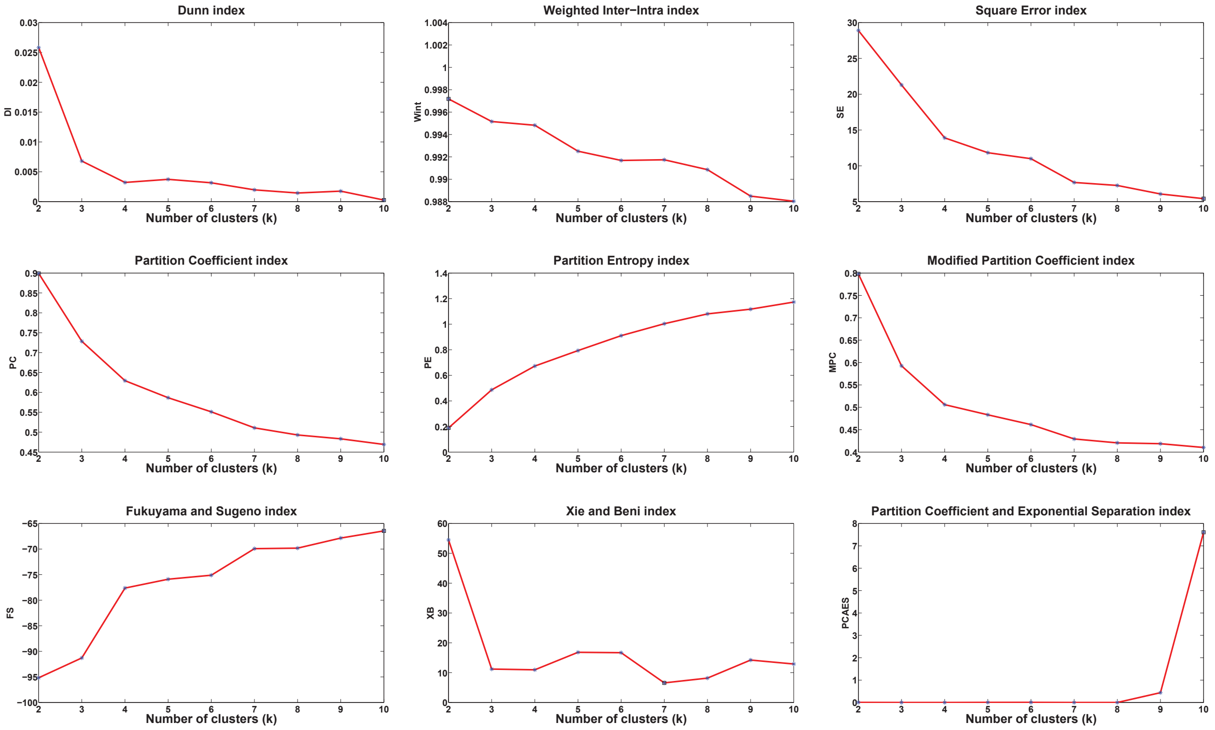

Figure 5: GK using 9 different validity indices

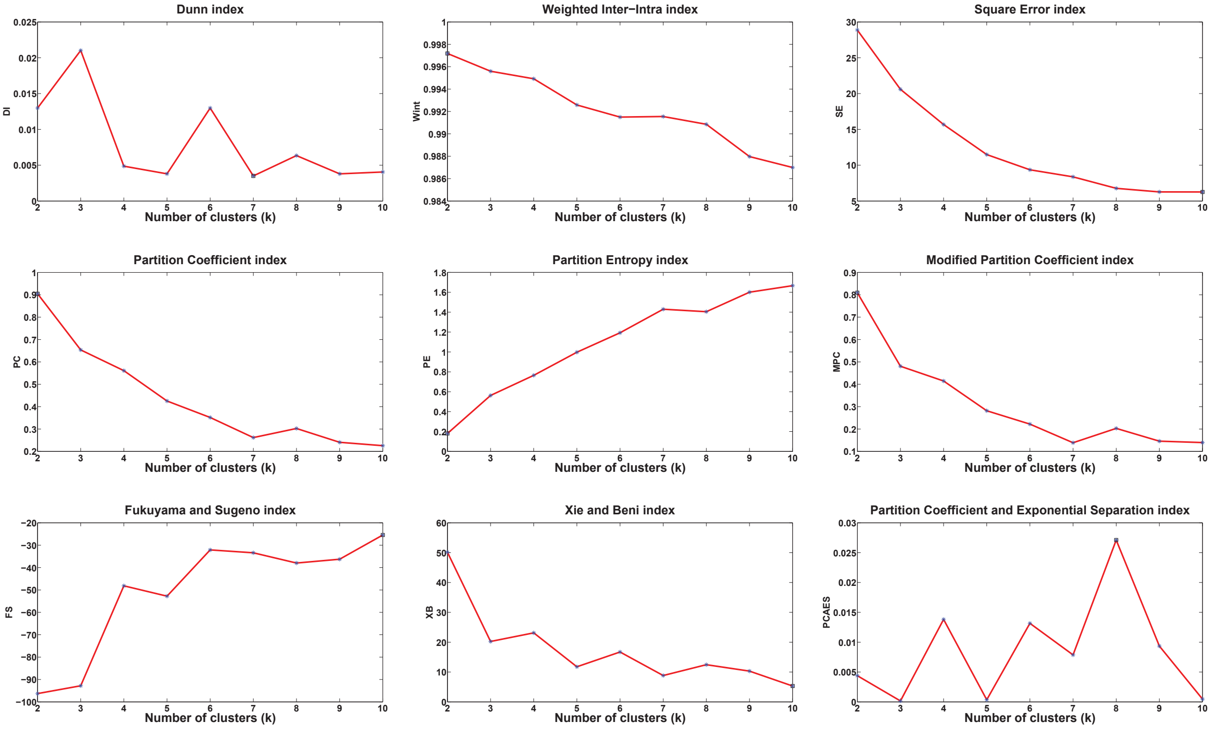

Figure 6: FPSO using 9 different validity indices

Table 4: Reconstruction error with varying *c* using transfusion data set.

	c=2	3	4	5	6	7	8	9
DI	31.2	30.4	**24.5**	61.3	31.3	**24.5**	59.8	25.0
Wint	**23.6**	29.3	25.4	47.4	51.5	68.8	42.5	36.5
SE	**23.6**	25.6	26.5	23.6	28.7	33.7	24.3	25.9
PC	**16.3**	29.6	34.2	81.8	22.7	75.1	31.1	21.4
PE	26.3	85.0	26.0	57.0	24.9	39.3	**16.9**	30.3
MPC	17.0	44.8	24.6	60.4	93.3	**12.2**	166.7	12.5
FS	**23.6**	**23.6**	**23.6**	25.6	29.5	24.9	33.6	23.7
XB	**23.6**	43.0	25.1	36.1	27.1	61.3	68.8	63.1
PCAES	**38.9**	81.7	82.5	48.1	53.8	57.5	97.2	93.4

Table 5: Nine different indices using the proposed algorithm.

	DI	Wint	SE	PC	PE	MPC	FS	XB	PCAES	FPSO	*k*
Transfusion	5.70	2.63	9.43	2.17	2.13	3.67	2.73	9.63	6.87	2.42	2
std.	1.84	0.49	0.86	0.38	0.35	2.15	1.05	0.56	2.29	0.58	
Haberman	5.90	2.93	9.97	2.00	2.00	6.50	2.07	9.83	8.80	2.24	2
std.	0.55	0.25	0.18	0.00	0.00	1.50	0.25	0.75	2.51	0.15	
Breast	6.13	3.00	8.43	3.27	2.17	5.30	2.13	9.77	6.83	2.64	2
std.	1.96	0.00	1.43	1.72	0.38	2.45	0.35	0.63	2.09	0.62	
Jain	7.30	2.97	9.57	2.87	2.10	3.60	3.37	7.90	6.47	2.25	2
std.	1.64	0.56	0.68	0.82	0.31	1.52	2.75	2.20	2.22	0.15	
Thyroid	4.27	2.83	9.90	2.10	2.03	4.27	2.00	9.53	8.10	2.24	2
std.	1.87	0.38	0.31	0.40	0.18	2.21	0.00	0.68	1.99	0.24	
Iris	3.83	2.93	9.00	2.53	2.30	4.13	2.40	9.57	6.30	2.55	2
std.	1.58	0.52	1.62	0.94	0.47	2.06	0.89	0.57	2.58	0.7	
DIM032	6.33	2.73	7.90	4.60	4.00	4.90	3.73	7.40	6.40	5.34	5
std.	1.07	0.64	1.83	1.43	1.91	1.44	1.10	1.30	2.04	2.08	
DIM064	7.17	3.20	7.77	5.47	5.50	5.53	2.00	9.43	7.40	5.94	5
std.	2.63	1.92	1.83	1.33	1.41	1.46	0.00	1.01	1.92	1.50	
DIM128	6.90	2.57	8.27	5.63	5.57	5.77	2.00	8.37	8.03	5.90	5
std.	2.02	0.57	1.55	1.25	1.38	1.04	0.00	1.97	1.35	1.24	
DIM256	8.57	2.57	8.57	6.43	6.13	6.70	2.00	9.13	8.57	6.52	5
std.	1.19	0.50	1.43	1.65	1.72	1.56	0.00	1.07	1.33	1.16	

Component Analysis (PCA) projection, Conventional Sammon Mapping (CSM), and Fuzzy Sammon Mapping (FSM) are used. The dimensions of DIM032, DIM064, DIM128 and DIM256 are 32, 64, 128 and 256, respectively. The correct number of clusters is 5. Figure 7 lists the performance of PCA, CSM and FSM using the four data sets. The black cross represents the identified cluster centers. Obviously, the correct number of clusters can be seen by looking at the figure. As the number of dimension increases, the performance of the PCA, CSM and FSM decrease.

The performance of the mapping are listed in Table 6. The mean square error of the re-calculated membership values (P), two different original and re-calculated validity measures (F and F*), and the Sammon stress coefficient (S) are listed in the parenthesis. As the number of dimension increases, the FSM is better than PCA and CSM in terms of smaller P, F* and S values. The performance of PCA, CSM and FSM are the same in terms of F values.

6 Conclusion

This paper proposed an algorithm to overcome the drawbacks of traditional partition clustering, which is that the number of clusters needs to be predefined. The proposed algorithm uses using PSO and FCM with a threshold vector to control and identify the optimal number of clusters. The algorithm solves the clustering problem via an iterative fuzzy partition process.

For the evaluation of our algorithm we generated a synthetic dataset as well as used 6 datasets from the UCI repository. We compared our algorithm with hard clustering approaches such as Kmeans and K-medoid as well as with fuzzy clustering algorithms such as FCM and GK. Nine different validity indices were used to evaluate the performance. Furthermore, measures such principal component analysis projection, conventional sammon mapping, and fuzzy sammon mapping were used to visualize the clustering results. Overall, the results show that the proposed algorithm can identify the correct number of clusters on all the data set tested. However, due to the slow convergence and the stochastic nature of the PSO algorithm, the prediction results of a single run vary and thus make it

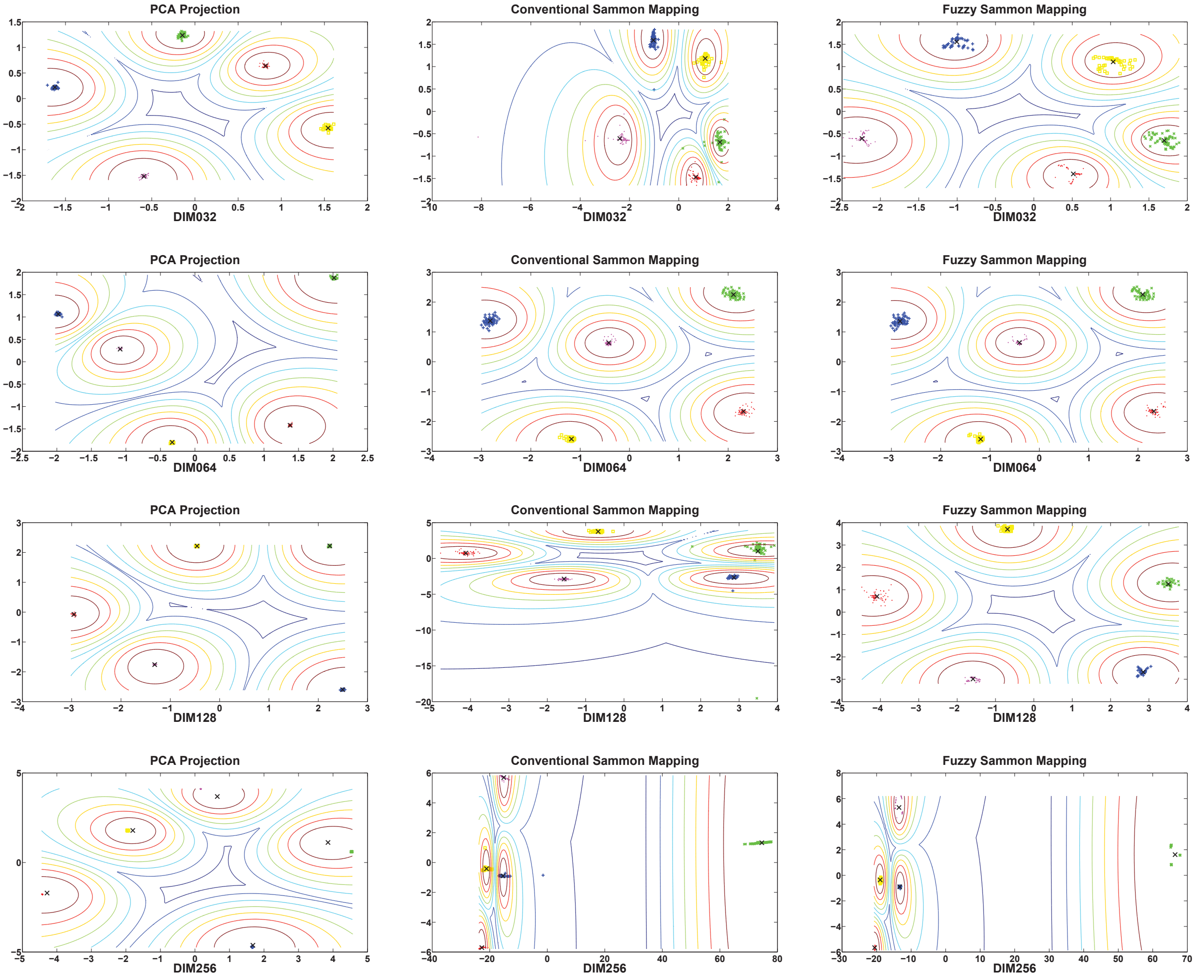

Figure 7: Data sets using Principal Component Analysis (PCA) projection, Sammon mapping and fuzzy Sammon mapping

Table 6: Mapping using the proposed algorithm.

Datasets	PCA	CSM	FSM
DIM032	(0.0030 0.9826 0.9975 0.0821)	(0.0104 0.9826 0.9495 0.2953)	(0.0023 0.9826 0.9742 0.0780)
DIM064	(0.0011 0.9939 0.9995 0.1329)	(0.0005 0.9939 0.9931 0.0553)	(0.0005 0.9939 0.9920 0.0556)
DIM128	(0.0008 0.9956 0.9998 0.1252)	(0.0050 0.9956 0.9765 1.5603)	(0.0004 0.9956 0.9941 0.0498)
DIM256	(0.0046 0.2003 0.2010 0.1138)	(0.0262 0.2003 0.2081 34.2770)	(0.0044 0.2003 0.2001 0.0482)

difficult to prediction the correct number of clusters. Unlike K-means and FCM, the proposed algorithm needs to be executed repeatedly in order to find the correct solution. In addition, the maximum number of clusters has to be predefined, and the iterative process to identify the optimal number of clusters is computationally expensive.

As for future work, it would be interesting to improve the proposed algorithm to achieve more stable predictions with fewer runs. Moreover, we are planning to explore the proposed algorithm with big datasets, and therefore parallelization techniques are necessary.

References

[1] H. A. Edelstein, Introduction to data mining and knowledge discovery (3rd ed), Potomac, MD: Two Crows Corp. 1999.

[2] B. Mirkin, Clustering: A Data Recovery Approach, Second Edition (Chapman & Hall/CRC Computer Science & Data Analysis).

[3] R. B. Cattell, The description of personality: Basic traits resolved into clusters, Journal of Abnormal and Social Psychology, 38, 476-506,1943.

[4] S. Theodoridis, and K. Koutroubas, Pattern Recognition, Academic Press, 1999.

[5] T. N. Pappas, An adaptive clustering algorithm for image segmentation, IEEE Trans. Signal process, vol. 40, pp.901-914, 1992

[6] A. Likas, N. Vlassis, and J. Verbeek, The global k-means clustering algorithm (Technical Report), Computer Science Institute, University of Amsterdam, The Netherlands. ISA-UVA-01-02. 2001.

[7] V. P. Guerrero-Bote, et al., Comparison of neural models for document clustering, Int. Journal of Approximate Reasoning, vol. 34, pp.287-305, 2003.

[8] G. L. Carl, A fuzzy clustering and fuzzy merging algorithm, Technical Report, CS-UNR-101, 1999.

[9] L. A. Zadeh, Fuzzy sets, Information and Control, Vol. 8, pp. 338-353, 1965.

[10] R. Babuska, Fuzzy Modelling for Control, Kluwer Academic, USA, 1998.

[11] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, ISBN 0-306-40671-3, 1981.

[12] J. C. Dunn, A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, Journal of Cybernetics 3: 32-57, 1973.

[13] I. Gath and A. B. Geva, Unsupervised optimal fuzzy clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 11(7), pp 773-781, 1989.

[14] J. C. Bezdek, C. Coray, R. Gunderson and J. Watson, Detection and characterization of cluster substructure i. linear structure: Fuzzy c-lines, SIAM Journal on Applied Mathematics, 40(2), 339-357.

[15] J. Yu, and M. S Yang, A generalized fuzzy clustering regularization model with optimality tests and model complexity analysis, Fuzzy Systems, IEEE Transactions on 15.5 (2007): 904-915.

[16] Y. T. Kao, E. Zahara, I. W. Kao, A hybridized approach to data clustering, Expert Systems with Applications 34 (3), 1754-1762, 2008.

[17] D. N. Cao, J. C. Krzysztof, GAKREM: a novel hybrid clustering algorithm, Information Sciences 178, 4205-4227, 2008.

[18] K. R. Zalik, An efficient k-means clustering algorithm, Pattern Recognition Letters 29, 1385-1391, 2008.

[19] K. Krishna, Murty, Genetic k-means algorithm, IEEE Transactions of System Man Cybernetics Part B-Cybernetics 29, 433-439, 1999.

[20] U. Mualik, S. Bandyopadhyay, Genetic algorithm-based clustering technique, Pattern Recognition 33, 1455-1465, 2000.

[21] M. Laszlo, S. Mukherjee, A genetic algorithm that exchanges neighboring centers for k-means clustering, Pattern Recognition Letters 28 (16), 2359-2366, 2007.

[22] P. S. Shelokar, V. K. Jayaraman, B. D. Kulkarni, An ant colony approach for clustering, Analytica Chimica Acta 509 (2), 187-195, 2004.

[23] S. A. Ludwig, Clonal Selection based Fuzzy C-Means Algorithm for Clustering, Proceedings of Genetic and Evolutionary Computation Conference (ACM GECCO), Vancouver, BC, Canada, July 2014.

[24] T. A. Runkler, and C. Katz. Fuzzy clustering by particle swarm optimization, Fuzzy Systems, 2006 IEEE International Conference on. IEEE, 2006.

[25] H. C. Liu, J. M. Yih, D. B Wu, S. W. Liu, Fuzzy C-mean clustering algorithms based on Picard iteration and particle swarm optimization, Education Technology and Training, 2008. and 2008 International Workshop on Geoscience and Remote Sensing. ETT and GRS 2008. International Workshop on. Vol. 2. IEEE, 2008.

[26] B. J. Zhao, An ant colony clustering algorithm, Machine Learning and Cybernetics, 2007 International Conference on. Vol. 7. IEEE, 2007.

[27] G. Gan, J. Wu, and Z. Yang, A genetic fuzzy k-Modes algorithm for clustering categorical data, Expert Systems with Applications 36.2 (2009): 1615-1620.

[28] F. Yang, T. Sun, and C. Zhang, An efficient hybrid data clustering method based on K-harmonic means and Particle Swarm Optimization, Expert Systems with Applications 36.6 (2009): 9847-9852.

[29] T. Niknam, B. Amiri, An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis, Applied Soft Computing, 10(1), 183-197.

[30] P. Melin, F. Olivas, O. Castillo, F. Valdez, J. Soria, and M. Valdez, Optimal design of fuzzy classification systems using PSO with dynamic parameter adaptation through fuzzy logic, Expert Systems with Applications 40, no. 8 (2013): 3196-3206.

[31] L. Y. Chuang, C. J. Hsiao, and C. H. Yang, Chaotic particle swarm optimization for data clustering, Expert systems with Applications 38.12 (2011): 14555-14563.

[32] H. Izakian, and A. Abraham, Fuzzy C-means and fuzzy swarm for fuzzy clustering problem, Expert Systems with Applications 38.3 (2011): 1835-1838.

[33] S. Das, A. Abraham, and A. Konar, Automatic kernel clustering with a multi-elitist particle swarm optimization algorithm, Pattern Recognition Letters, 29(5), 688-699, 2008.

[34] M. Chen and S. A. Ludwig, Fuzzy Clustering Using Automatic Particle Swarm Optimization, Proceedings of 2014 IEEE International Conference on Fuzzy Systems, Beijing, China, July 2014.

[35] R. C. Eberhart and J. Kennedy, A New Optimizer using Particle Swarm Theory, In Proc. 6th Symp. Micro Machine and Human Science, Nagoya, Japan 1995, 29-43.

[36] J. C. Bezdek, Cluster validity with fuzzy sets, (1973): 58-73.

[37] R. N. Dave, Validating fuzzy partitions obtained through c-shells clustering, Pattern Recognition Letters, 17(6), 613-623, 1996.

[38] Y. Fukuyama and M. Sugeno, A new method of choosing the number of clusters for the fuzzy c-means method, Proceeding of fifth fuzzy Syst. Sympo., pp.247-250, 1989.

[39] X. L. Xie, and G. Beni, A validity measure for fuzzy clustering. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 13(8), 841-847, 1991.

[40] N. R. Pal, and J. C. Bezdek, On cluster validity for the fuzzy c-means model, Fuzzy Systems, IEEE Transactions on, 3(3), 370-379, 1995.

[41] K. L. Wu, and M. S. Yang, A cluster validity index for fuzzy clustering, Pattern Recognition Letters, 26(9), 1275-1291, 2005.

[42] A. Strehl, Relationship-based clustering and cluster ensembles for high-dimensional data mining, 2002.

[43] W. Pedrycz and J. V. de Oliveira, A development of fuzzy encoding and decoding through fuzzy clustering, IEEE Trans. Instrum. Meas., vol. 57, no. 4, pp. 829?837, Apr. 2008.

[44] A. Frank & A. Asuncion, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2010.

[45] B. Balasko, J. Abonyi, and B. Feil. Fuzzy clustering and data analysis toolbox, Department of Process Engineering, University of Veszprem, Veszprem, 2005.

[46] J. MacQueen, Some methods for classification and analysis of multivariate observations, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281–297, University of California Press, Berkeley, Calif., 1967.

[47] L. Kaufman and P.J. Rousseeuw, Clustering by means of Medoids, in Statistical Data Analysis Based on the L_1 Norm and Related Methods, edited by Y. Dodge, North-Holland, 405–416, 1987.