Journal of Advances in Computer Research
Quarterly pISSN: 2345-606x eISSN: 2345-6078
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 9, No. 1, February 2018), Pages: 1-16
www.jacr.iausari.ac.ir

An Improved SSPCO Optimization Algorithm for Solve of the Clustering Problem

Rohollah Omidvar✉, Amin Eskandari, Narjes Heydari, Fatemeh Hemmat, Mohammad Feili
Sama Technical and Vocational Training College/ Islamic Azad University, Shiraz Branch, Shiraz, Iran
omidvar@shirazu.info; eskandary.a@gmail.com; narjesheydari977@gmail.com;
hemmatfatemeh15@yahoo.com; arsalan.feili@yahoo.com

Received: 2016/10/30; Accepted: 2017/03/15

Abstract

Swarm Intelligence (SI) is an innovative artificial intelligence technique for solving complex optimization problems. Data clustering is the process of grouping data into a number of clusters. The goal of data clustering is to make the data in the same cluster share a high degree of similarity while being very dissimilar to data from other clusters. Clustering algorithms have been applied to a wide range of problems, such as data mining, data analysis, pattern recognition, and image segmentation. Clustering is a widespread data analysis and data mining technique in many fields of study such as engineering, medicine, biology and the like. The aim of clustering is to collect data points. SSPCO optimization algorithm is a new optimization algorithm that is inspired by the behavior of a type of bird called see-see partridge. One of the things that smart algorithms are applied to solve is the problem of clustering. Clustering is employed as a powerful tool in many data mining applications, data analysis, and data compression in order to group data on the number of clusters (groups). In the present article, an improved chaotic SSPCO algorithm is utilized for clustering data on different benchmarks and datasets; moreover, clustering with artificial bee colony algorithm and particle mass 9 clustering technique is compared. Clustering tests on 13 datasets from UCI machine learning repository have been done. The results show that clustering SSPCO algorithm is a clustering technique which is very efficient in clustering multivariate data.

Keywords: *Clustering, Neighbor, SSPCO, Clustering Error, Algorithm*

1. Introduction

In data mining, Clustering is the most popular, powerful and commonly used unsupervised learning technique. It is a way of locating similar data objects into clusters based on some similarity. Clustering algorithms can be categorized into seven groups, namely Hierarchical clustering algorithm, Density-based clustering algorithm, Partitioning clustering algorithm, Graph-based algorithm, Grid-based algorithm, Model-based clustering algorithm and Combinational clustering algorithm. These clustering algorithms give different result according to the conditions. Some clustering techniques are better for large data set and some gives good result for finding cluster with arbitrary shapes [1]. Knowledge discovery process is introduced as the clustering and is one of the data mining techniques [2], [3]. Furthermore, clustering and classification are two basic tasks of data mining [4]. Analyzing the data to understand various phenomena

plays an essential role. Cluster analysis with no or little previous knowledge comprises the developed research in a range of communities [5]. The purpose of clustering a data set without label is the separation in a discrete finite set of natural [6], [7]. Unanticipated clustering is a mental process in nature that prevents absolute judgment as a relative effect on all clustering techniques [8]. Several nature-inspired algorithms to solve optimization problems came to the aid of human science, algorithms such as birds [9], ant colony algorithm [10], Firefly algorithm [11,12], Artificial bee colony algorithm [13], bees algorithm [14]which are many complex issues in various fields. SSPCO optimization algorithm [15] is also one of the newest algorithms based on the behavior of chicks a type of bird called see-see partridge.

We propose a data clustering algorithm based on improved SSPCO optimization algorithm. The main question in this research is that improved algorithm by our can be a suitable method for raw data clustering. For this purpose, we propose an efficient algorithm which is an improved version of SSPCO optimization algorithm. In addition, in this paper, SSPCO algorithm was improved. Use chaotic population help a lot in creating an initial scenario is suitable for clustering. The idea neighboring particles is effective in faster finding optimal cluster.

Clustering tests on 13 datasets from UCI machine learning repositories have been done. Proposed method compared with 12 clustering technique that the proposed method is better than other methods in most of tests. In this paper, Section 2 is description of clustering, section 3 will be dedicated to introducing the related works. In Section 4, we present the SSPCO algorithm and the experimental study will be discussed in Section 5. Finally, Section 6 is conclusion.

2. The Clustering Problem

Clustering is the process of recognizing natural groupings or clusters in multidimensional data based on some similarity measures [16]. Distance measurement is generally used for evaluating similarities between patterns. In particular the problem is stated as follows: given N objects, allocate each object to one of K clusters and minimize the sum of squared Euclidean distances between each object and the center of the cluster belonging to every such allocated object. The clustering problem minimizing equation 1 is described as in [17]:

$$j(w,z)=\sum_{i=1}^{N}\sum_{j=1}^{K}w_{ij}x_i-z_j{}^2 \tag{1}$$

Where K is the number of clusters, N the number of patterns, $x_i(i=1,\ldots.,N)$is the ith the location of the i th pattern and $z_j(j=1,\ldots.,K)$ is the center of the jth cluster, to be found by equation 2:

$$z_j=\frac{1}{N_j}\sum_{i=1}^{N}w_{ij}x \tag{2}$$

Where N_j is the number of patterns in the jth cluster, w_{ij} the association weight of pattern x_i with cluster j, which will be either 1 or 0 (if pattern i is allocated to cluster j; w_{ij} is 1, otherwise 0). The cost function for the pattern i is given by equation 3 as in:

$$f_i = \frac{1}{D_{Train}} \sum_{j=1}^{D_{Train}} d\left(x_j, P_i^{CL_{know}(x_j)}\right) \quad (3)$$

In which D_{Train} is the number of training patterns, and $P_i^{CL_{know}(x_j)}$ is the class definition.

Where D_{Train} is the number of training patterns which is used to normalize the sum that will range any distance within [0.0, 1.0] and ($P_i^{CL_{know}(x_j)}$) defines the class that instance belongs to according to database.

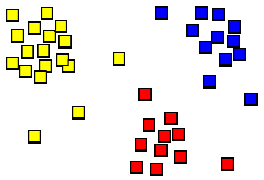

Figure 1. cluster analysis based on the coloring of the squares into three clusters

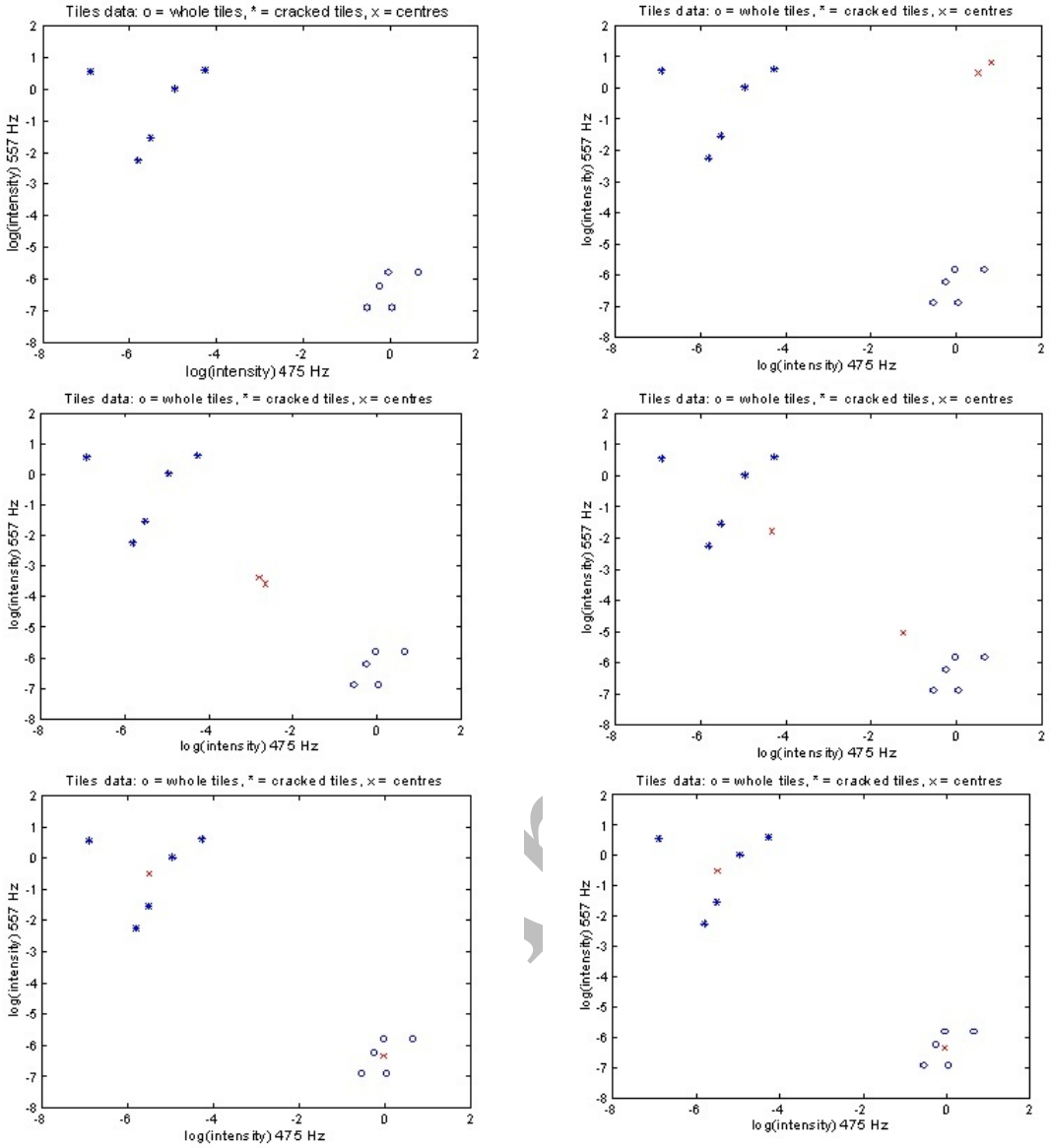

Figure 2. An example of clustering process.

3. Related Work

There are a large number of researchers who have applied different optimization techniques to eliminate clustering problem. For example, Genetic Algorithm based method to solve the clustering problem was proposed by Cowgill et al.1999 [18].

Multilayer Perceptron neural networks or MLP [19], complex neural network training and an optimization problem have many variables [20]. Artificial Neural Networks Radial Basis Function, RBF, [21], in a other research, also unlike the MLP network with several successive layers, comprises three fixed layers [22], Engineering K-STAR [23], Engineering Bagging [24], Multi Boost AB technique [25], NBTree technique [26], Ridor technique [27], VFI clustering technique introduced by [28], PSO-Clustering [29] and ABC-Clustering [30] are another clustering techniques that in this study which is used to compare performance.

Many approaches have been proposed for data clustering. This section reviews some of the recent research works on data clustering based on KM, FKM, and hybrid algorithms. For more details, the interested reader can also refer to pre-vious surveys in the area, among which we point out Nayak et al. [31] that have been presented a comprehensive review on FKM and its applications from 2000 to 2014. Nayak et al. [32] have developed an improved firefly-based fuzzy c-means algorithm (FAFCM) for data clustering. They show the effectiveness and reliability of the proposed method by testing the algorithm with various real-world datasets. Wu et al. [33] have been developed a hybrid fuzzy K-harmonic means (HFKHM) clustering algorithm based on improved possibility c-means clustering (IPCM) and K-harmonic means (KHM).

HFKHM solves the noise sensitivity problem of KHM and improves the memberships of IPCM by combining the merits of KHM and IPCM. The performance of HFKHM is compared with those of KHM and IPCM on several datasets. Experimental results indicate the superiority of HFKHM. Shamshirband et al. [34] proposed a density-based fuzzy imperialist competitive clustering algorithm (D-FICCA) for detecting the malicious behavior in wireless sensor network.

In one of research, PSO based k-means clustering is applied to generate clusters [35]. An improved technique which combines chaotic map particle swarm optimization with an acceleration strategy is proposed, since results of one of the most used clustering algorithm, K-means can be jeopardized by improper choices made in the initializing stage. Accelerated chaotic particle swarm optimization (ACPSO) searches through arbitrary data sets for appropriate cluster centers and can effectively and efficiently find better solutions [36]. In one of research suggest another way of combining K-Means and PSO, using the strength of both algorithms. Most of the methods introduced in the context of clustering, that hybridized K-Means and PSO [37]. In a research a population-based clustering technique, which attempts to integrate different particle swarm optimizers (PSOs) with the famous *k*-means algorithm, is proposed. More specifically, six existing extensively studied PSOs, which have shown promising performance for continuous optimization, are hybridized separately with Lloyd's *k*-means algorithm, leading to six PSO-based clustering methods. These PSO-based approaches use different social communications among neighbors to make some particles escape from local optima to enhance exploration, while *k*-means is utilized to refine the partitioning results for accelerating convergence [38].

4. SSPCO Optimization Algorithm

The basic idea of this optimization algorithm is taken from the behavior of the chicks of a type of bird called See-see partridge. The chicks of this type of bird are located in a regular queue at the time of danger to reach a safe place and they start to move behind their mother to reach a safe point. To simulate the behavior of the chicks of this bird in the form of an optimization algorithm, each chick is considered as a particle of the suboptimal problem. The state of each particle should be according to the behavior of this type of chicks in a regular queue that we know this queue takes us to the best optimal point and this does not mean that minimizing the search space, but also, it is converging particles after some searches in a regular queue to the best point answers (bird mother). According to Figure 1, each chick in the search space seeks to find a chick with the priority of a unit higher than itself and it tries to adjust its motion equation based on this chick.

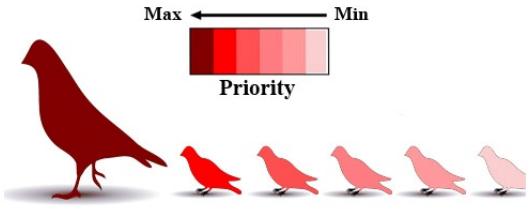

Figure 3. Chicks motion in SSPCO algorithm

In the algorithm, consider a variable for each particle entitled as priority variable. For particle i a priority variable defined. In every assessment, when a particle was better than the best personal experience or local optimum; a unit is added to the priority variable of that particle:

$$if \quad X_i.cost > P_{best} \rightarrow P_{best} = X_i.position \ and \quad X_i.priority = X_i.priority + 1 \qquad (4)$$

$X_i.cost$ The cost of each particle in the benchmark, P_{best} is the best personal experience of each particle, and $X_i.position$ is the location of each particle. In every time of assessment, if the local optimum is better than the global optimum and vice versa, the particle's priority variable goes higher and a unit is added to it:

$$if \quad P_{best} > G_{best} \rightarrow G_{best} = P_{best} \ and \quad X_i.priority = X_i.priority + 1 \qquad (5)$$

G_{best} is the global optimum. The motion equation of each particle is set almost similar to the particle swarm algorithm in the form of equation 4:

$$X_i.position = X_i.position + X_i.velocity \qquad (6)$$

$X_i.velocity$ is the velocity of each particle or chick. Then, Chickens sorted in array based on priority variable.

$$X_i.velocity = w * X_i.velocity + c * rand() * [position(X_{i+1}.priority)] - X_i.position \qquad (7)$$

$X_i.velocity$ is the velocity of the particle, w is the coefficient impact of the previous velocity in the current velocity equation of particle, c is the coefficient impact of position of particle with upper priority in the current velocity equation of particle, $rand()$is a random number between 0 and one to create a random movement for particles, $[position(X_{i+1}.priority)]$ is the location of the particle with one level higher priority than the current particle that the current particle tries to adjust its velocity according to the particle, $X_i.position$ is the current location of the particle. It can be seen that, according to Equation 6, each particle adjusts its movement based on a particle with one level higher priority [8].

4.1 Chaotic Theory

Simulation dynamic behavior of nonlinear systems called chaos. It has raised enormous interest in different fields of sciences such as synchronization, chaos control, optimization theory, pattern recognition and so on [39]. In optimization algorithms based on the chaos theory, the methods using chaotic variables instead of random variables are called chaotic optimization algorithm (COA). COA is a stochastic search methodology that differs from any of the existing swarm intelligence methods and evolutionary computation. COA can carry out overall searches at higher speeds than stochastic searches that depend on probabilities [40]. There are several different chaotic

sequences which the most commonly used such chaotic sequences are logistic maps that are considered in this paper. Logistic maps are frequently used chaotic behavior maps and chaotic sequences can be quickly generated and easily stored. For this reason, there is no need for storage of long sequences [41]. In this study, we substitute the random parameters in PSO with sequences generated by the logistic map. The parameters random are modified by the logistic map based on the following equation:

$$Cr_{(t+1)} = k \times Cr_{(t)} \times \left(1 - Cr_{(t)}\right) \tag{8}$$

In Eq. (5), k =4 and for each independent run, Cr(0) is generated randomly, which Cr(0) not being equal to {0, 0.25, 0.5, 0.75, 1}..

1.//initialize all chicken by $k \times Cr_{(t)} \times (1 - Cr_{(t)})$
2.Initialize by $k \times Cr_{(t)} \times (1 - Cr_{(t)})$
3.Repeat
4. For each chicken i
5. //update the chicken's best position and priority
6. If $f(x_i) > f(pbest_i)$ **then**
7. $pbest_i = x_i$
8. $prioirity_i = prioirity_i + 1$
9. End if
10. //update the global best position and priority
11. If $f(pbest_i) > f(gbest)$ **then**
12. $gbest = pbest_i$
13. $prioirity_i = prioirity_i + 1$
14. End if
15. End for
16. //update chicken's velocity and position
17. For each chicken i
18. For each dimension d
19.
$X_i.velocity =$
$w * X_i.velocity + c * rand() * [position(X_{i+1}.priority)] -$
$X_i.position$
20. $x_{i,d} = x_{i,d} + v_{i,d}$
21. End for
22. End for
23. //advance iteration
24. $itetation = itetation + 1$
25.Until $it > MaxIterations$

Algorithm. 1. Pseudo code of Chaotic SSPCO algorithm

4.2 Improved SSPCO Algorithm

In the original SSPCO Algorithm, since each particle moves in the search space guided only by its historical best solution the location of the particle with one level higher priority than the current particle, it may get trapped in a local optimal solution when current global best solution in a local optimum and not easy for the particle escapes from it. To solve such problem, in this section, one improvement strategies are proposed. From (6), we observe that only the information of the historical best position the location of the particle with one level higher priority than the current particle of all particles is utilized. As a matter of fact, the information of the best neighbor $P^{best}_{N_{(priority)i+1}}$ of the location of the particle with one level higher priority than the current particle i may provide a better guidance than the particle with one level higher priority than the current particle. We explain how to define the neighbors and to determine the best neighbor $P^{best}_{N_{(priority)i+1}}$ of the location of the particle with one level higher priority than the current particle$i$. In order to define appropriate neighbors,

different approaches could be used. In this paper, the neighbors of x_i are defined by using the mean Euclidean distance between x_i and the rest of solutions. Let $d(i,j)$ be the Euclidean distance between x_i and x_j and let md_i be the mean Euclidean distance forx_i . Then md_i can be computed as follows:

$$md_i = \frac{\sum_{j=1}^{M} d(i,j)}{M-1} \tag{9}$$

By (10) if $(i,j) < md_i$, then x_j could be accepted as a neighbor of x_i . In addition, we can also use a more general and flexible definition to determine a neighbor of x_i .

$$if\ d(i,j) \le r \tag{10}$$

$\times\ md_i$, then x_j is a neighbor of x_i .

If (11) is used, then a new parameter r ($r \geq 0$), which refers to the "neighborhood radius," will be added to the parameters of SSPCO. If $r = 0$, it turns to the standard PSO. With the value of r increasing, the neighborhood of x_i enlarges or its neighborhood shrinks as the value of r decreases. Once the neighbors are determined, we select the best position among the neighbors of x_i as the best neighbor$P^{best}_{N_{(priority)i+1}}$.

After determining the best neighbor$P^{best}_{N_{(priority)i+1}}$, we give a new way moving for each particle. If $f(P^{best}_{N_{(priority)i+1}}) < f(position(X_{i+1}.priority))$, we set:

$$X_i.velocity = w * X_i.velocity + c * rand() * P^{best}_{N_{(priority)i+1}} - X_i.position \tag{11}$$

If $f(P^{best}_{N_{(priority)i+1}}) > f(position(X_{i+1}.priority))$, then set:

$$X_i.velocity = w * v_{i,d} + C * Rand(0,1) * [position(X_{i+1}.priority] - x_{i,d} \tag{12}$$

Also moves to a new position by the following equation.

$$X_i.position = X_i.position + X_i.velocity \tag{13}$$

4.3 SSPCO Applied to Clustering

Given a database with C classes and N parameters, the classification problem can be seen as that of finding the optimal positions of C center in an N-dimensional space i.e. That of determining for any center its N coordinates, each of which can take on, in general, real values. With these premises, the i-th individual of the population is Encoded as it equation 9.

$$(p_i^{\to 1}, \ldots, p_i^{\to C}, v_i^{\to 1}, \ldots, v_i^{\to C}) \tag{14}$$

Where $p_i$the position of the j-th center is constituted by N real numbers representing, its N coordinates in the problem space.

$$p_i^{(?j)} = ?\ p\ ?^j_{(1,i)}, \mathrm{n}, p^j_{(N,1)} \tag{15}$$

And similarly the velocity of the j-th center is made up of N real numbers representing, its N velocity components in the problem space.

$$v_i^{\to j} = \{v^j_{1,i}, \ldots, v^j_{N,1}\} \tag{16}$$

Then, each individual in the population is composed of 2*C*N components, each represented by a real value.

1.//initialize all chicken by $k \times Cr_{(t)} \times (1 - Cr_{(t)})$ *(input is a clustering form according chaotic theory)*
2.Initialize by $k \times Cr_{(t)} \times (1 - Cr_{(t)})$
3.Repeat
4. For each chicken i
5. //update the chicken's best position and priority
6. $(\vec{p_i^1}, \dots, \vec{p_i^C}, \vec{v_i^1}, \dots, \vec{v_i^C})$
$\vec{p_i^j} = \{p_{1,i}^j, \dots, p_{N,1}^j\}$

7. If $f(x_i) > f(pbest_i)$ ***then***
8. $pbest_i = x_i$
9. $prioirity_i = prioirity_i + 1$
10. End if
11. //update the global best position and priority
12. If $f(pbest_i) > f(gbest)$ ***then***
13. $gbest = pbest_i$
14. $prioirity_i = prioirity_i + 1$
15. End if
16. End for
17. //update chicken's velocity and position
18. For each chicken i
19. If $f(P^{best}_{N_{(priority)i+1}}) < f(position(X_{i+1}.priority))$
20. $v_{i,d} = v_{i,d} + C * Rand(0,1) * [P^{best}_{N_{(priority)i+1}}] - x_{i,d}$
21. *If* $f(P^{best}_{N_{(priority)i+1}}) > f(position(X_{i+1}.priority))$
22. $v_{i,d} = v_{i,d} + C * Rand(0,1) * [position(X_{i+1}.priority] - x_{i,d}$
23. $x_{i,d} = x_{i,d} + v_{i,d}$
24. End for
25. End for
26. //advance iteration
27. $itetation = itetation + 1$
28.Until $it > MaxIteration$***s***
29. clustering form=gbest(index of cluster-heads and members)

Algorithm.2. Pseudo code of Proposed Method

5. Simulation and Result

In this paper, proposed algorithm was compared with PSO-Clustering and ABC-Clustering. PSO clustering algorithm by these parameters has solved the problem of clustering. n = 50, $T_{\max}$=1000, $v_{\max}$=. 05,$v_{\min}$ = -. 05, C_1= 2, C_2= 2, $w_{\max}$= .09, $w_{\min}$= .04. Artificial bee colony clustering algorithm has the following parameters [30]. The size of the colony is 20, the maximum ring is 1000, and a total of 20,000 are assessed. In this study, 13 datasets of UCI database are tested for clustering problem [42]. 75% of the data for each data set is dedicated to education and 25% to testing. First, to briefly discuss data collections in this study, all the attributes are expressed and presented in Table 1 [22].

Table 1. Properties of the problems. [17]

	Data	Train	Test	Input	Class
Balance	625	469	156	4	3
Cancer	569	427	142	30	2
Cancer-Int	699	524	175	9	2
Credit	690	518	172	51	2
Dermatology	366	274	92	34	6
Diabetes	768	576	192	8	2
E. coli	327	245	82	7	5
Glass	214	161	53	9	6
Heart	303	227	76	35	2
Horse	364	273	91	58	3
Iris	150	112	38	4	3
Thyroid	215	162	53	5	3
Wine	178	133	45	13	3

5.1 Results and Discussions

Each pattern should be part of the cluster closest to Euclidean distance with the cluster's center. The data is divided into two pieces, 75% of the training data and 25% of the final test data. Classification Error Percentage According to equation 17 is calculated [30]:

$$CEP\left(Classification\ Error\ Percentage\right)=100\times\frac{misclassification\ examples}{size\ of\ \ test\ data\ set} \tag{17}$$

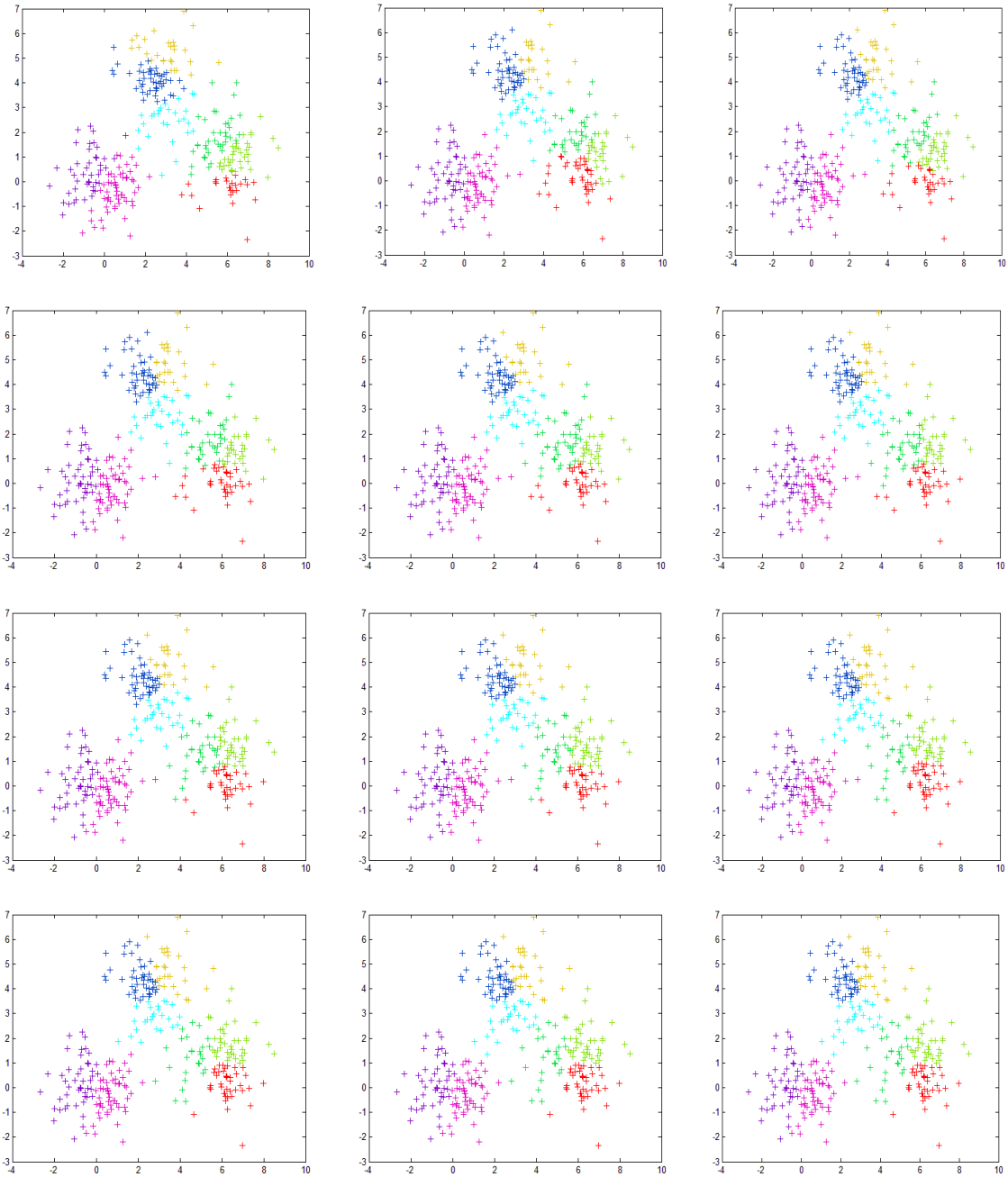

Figure 4. Clustering Form in Several Iterations

Results of PSO, ABC, GA, DE and SSPCO algorithms on the basis of the classification error on 13 issues of clustering are given in Table 2. All of the results were obtained from an average of 20 runs.

Table 2. Classification error percentages of the meta-heuristic algorithm

	SSPCO	ABC	PSO	GA	DE
Balance	15.36	15.38	25.74	29.65	28.53
Cancer	4.15	2.81	5.81	8.54	3.16
Cancer-Int	4.49	0	2.87	3.98	1.50
Credit	15.92	13.37	22.96	25.10	16.35
Dermatology	16.11	5.43	5.76	8.06	4.20
Diabetes	16.66	22.39	22.50	26.25	19.42
E. coli	13.89	13.41	14.63	15.02	13.85
Glass	13.56	41.50	39.05	40.16	42.15
Heart	14.03	14.47	17.46	18.59	15.12
Horse	12.22	38.26	40.98	42.40	40.21
Iris	4.48	0	2.63	3.60	1.92
Thyroid	3.31	3.77	5.55	6.59	4.17
Wine	5.12	0	2.22	3.43	3.61

The average Classification error percentage for all 13 data sets showed that the proposed algorithm has the lowest percentage of error. The average Classification error percentage on the full data set for proposed algorithm is 10.71%, while the percentage errors of ABC and PSO are 13.13% and 15.99%.

Table 3. Classification error percentages of the techniques. [17]

	SSPCO	ABC	PSO	Bayes et	MlpAı	RBF	KSta	Baggir	MultiBoos	NBTre	Ridc	VFI
Balance	15.36	15.38	25.74	19.74	9.29	33.6	10.25	14.77	24.20	19.7	20.6	38.85
Cancer	4.15	2.81	5.81	4.19	2.93	20.2	2.44	4.47	5.59	7.69	6.63	7.34
Cancer- Int	4.49	0.00	2.87	3.42	5.25	8.17	4.57	3.93	5.14	5.71	5.48	5.71
Credit	15.92	13.37	22.96	12.13	13.8	43.2	19.18	10.68	12.71	16.18	12.6	16.47
Dermatology	16.11	5.43	5.76	1.08	3.26	34.6	4.66	3.47	53.26	1.08	7.92	7.60
Diabetes	16.66	22.39	22.50	25.52	29.1	39.1	34.05	26.87	27.08	25.52	29.3	34.37
E. coli	13.89	13.41	14.63	17.07	13.5	24.3	18.29	15.36	31.70	20.73	17.0	17.07
Glass	13.56	41.50	39.05	29.62	28.5	44.4	17.58	25.36	53.70	24.07	31.6	41.11
Heart	14.03	14.47	17.46	18.42	19.4	45.2	26.70	20.25	18.42	22.36	22.8	18.42
Horse	12.22	38.26	40.98	30.76	32.1	38.4	35.71	30.32	38.46	31.86	31.8	41.75
Iris	4.48	0	2.63	2.63	0.00	9.99	0.52	0.26	2.63	2.63	0.52	0.00
Thyroid	3.31	3.77	5.55	6.66	1.85	5.55	13.32	14.62	7.40	11.11	8.51	11.11
Wine	5.12	0	2.22	0.00	1.33	2.88	3.99	2.66	17.77	2.22	5.10	5.77

Result showed in data set of Balance, Diabetes, Glass, Heart, Horse and Thyroid proposed algorithm have minimum Classification error. In data set of Cancer, KStar has best performance, in data set of Cancer-Int ABC algorithm has the best performance, in Credit data set, method of BayesNet has best performance, in Dermatology methods of BayesNet and NBTree has minimum Classification error, in data set of E.coli ABC has best performance, in Iris, ABC and MlpAnn are best algorithms for clustering and in Wine, ABC and BayesNet has best Classification error.

Table 4. Average classification error percentages and ranking of the techniques[17]

	SSPCO	ABC	PSO	BayesNet	MlpAnn	RBF	KStar	Bagging	MultiBoost	NBTree	Ridor	VFI
Average	10.71	13.13	15.99	13.17	12.35	26.93	14.71	13.30	22.92	14.68	15.38	**18.89**
Rank	1	3	9	4	2	12	6	5	11	7	8	**10**

In total, the proposed method has been the best performance. proposed algorithm in 6 data set is best rank, ABC in 3 data set is best rank, KStar in 1 data set is best, BayesNet in 3 data set is best, MlpAnn in a data set and NBTree also in an data set is best rank.

Table 5. Standard deviation classification in the techniques[17]

	SSPCO	ABC	PSO	BayesNet	MlpAn	RBF	KStar	Bagging	MultiBoost	NBTree	Ridor	VFI
Balance	6.89	7.23	9.23	8.86	4.25	5.85	3.35	6.30	5.55	7.03	9.33	**3.36**
Cancer	0.53	0.21	1.22	0.13	0.96	0.71	0.77	0.33	0.22	0.88	1.03	**0.11**
Cancer- Int	0.86	0.88	2.30	0.55	0.54	0.68	1.35	0.96	0.64	0.37	2.63	**0.55**
Credit	9.08	7.05	4.58	9.00	8.80	6.63	14.58	8.57	7.66	9.36	6.69	**6.23**
Dermatology	7.83	6.25	5.25	6.68	7.66	7.00	8.03	6.66	5.63	9.35	5.55	**5.22**
Diabetes	6.61	6.10	9.98	5.59	6.69	4.05	2.36	6.22	5.69	6.21	9.54	**5.66**
E. coli	6.52	5.25	6.69	6.01	5.53	5.32	4.38	6.33	8.55	3.23	12.05	**4.30**
Glass	7.30	6.68	9.35	7.00	7.22	7.05	3.36	4.69	7.35	6.35	9.23	**6.28**
Heart	5.74	4.47	7.78	6.52	5.02	4.45	2.20	4.66	6.65	4.33	8.32	**4.41**
Horse	7.45	6.68	10.03	7.04	6.69	7.50	2.68	9.98	4.63	6.61	8.00	**6.60**
Iris	1.03	0.98	0.22	0.98	1.01	0.78	1.58	0.99	0.69	1.08	3.02	**0.88**
Thyroid	0.67	0.55	0.55	0.44	0.65	0.44	0.20	0.65	0.36	1.00	1.50	**0.22**
Wine	1.12	2.30	1.11	1.56	0.65	2.00	0.69	1.10	2.03	1.32	1.01	**3.03**

Table 6. Mean of Standard deviation classification in the techniques[17]

	SSPCO	ABC	PSO	BayesNet	MlpAnn	RBF	KStar	Bagging	MultiBoost	NBTree	Ridor	VFI
Balance	12.59	14.65	11.65	14.29	8.64	9.38	13.30	13.10	10.65	9.14	13.65	**9.99**
Cancer	5.30	3.32	3.35	1.25	4.02	3.33	1.03	6.31	0.88	1.26	4.29	**1.00**
Cancer- Int	2.21	5.52	6.35	2.16	4.26	1.22	3.63	1.30	1.03	0.56	3.31	**2.08**
Credit	13.55	11.03	8.42	11.02	11.26	12.28	25.20	2.36	5.23	13.10	12.06	**7.59**
Dermatology	14.20	12.20	12.27	10.08	9.64	9.02	1.02	15.24	4.36	16.35	11.89	**9.19**
Diabetes	19.12	16.43	16.61	15.50	11.07	16.46	19.00	16.61	12.23	11.23	14.10	**8.05**
E. coli	23.35	20.25	18.25	8.55	6.62	22.29	10.03	14.41	9.46	6.64	10.00	**9.18**
Glass	17.21	17.20	16.44	14.20	9.18	14.59	16.66	16.55	13.64	11.90	12.21	**14.88**
Heart	10.05	10.00	8.29	9.68	6.64	9.94	8.92	8.89	9.80	6.55	9.91	**7.78**
Horse	27.20	25.31	22.29	15.28	9.75	12.03	4.42	11.06	20.06	10.71	15.15	**12.75**
Iris	9.32	2.56	6.54	6.64	6.89	6.30	8.84	6.68	1.06	6.45	6.44	**1.73**
Thyroid	4.30	6.25	2.23	2.24	4.21	2.05	2.22	4.26	0.85	4.22	2.69	**0.88**
Wine	8.54	4.58	4.02	3.65	1.88	7.89	8.48	6.64	1.06	6.44	6.64	**7.17**

The standard deviation of proposed method is showing of diversity of the solutions for find the best clusters. Also in table 7 is showing that run time of proposed algorithm for clustering problem is low.

Table 7. Run Time of classification in the techniques

	SSPCO	ABC	PSO	BayesNe	MlpAn	RBF	KStar	Bagging	MultiBoost	NBTre	VFI
Balance	29s	57s	27s	25s	105s	65s	33s	6.30	5.55	7.03	**3.36**
Cancer	32s	33s	30s	28s	84s	54s	36s	0.33	0.22	0.88	**0.11**
Cancer- Int	10s	53s	17s	89s	66s	16s	9s	0.96	0.64	0.37	**0.55**
Credit	35s	102s	21s	44s	94s	40s	33s	8.57	7.66	9.36	**6.23**
Dermatology	30s	85s	35s	25s	37s	28s	26s	6.66	5.63	9.35	**5.22**
Diabetes	32s	92s	39s	26s	115s	39s	58s	6.22	5.69	6.21	**5.66**
E. coli	27s	51s	25s	18s	64s	28s	21s	6.33	8.55	3.23	**4.30**
Glass	19s	41s	19s	14s	71s	25s	17s	4.69	7.35	6.35	**6.28**
Heart	12s	27s	15s	22s	95s	19s	09s	4.66	6.65	4.33	**4.41**
Horse	11s	18s	14s	66s	73s	52s	10s	9.98	4.63	6.61	**6.60**
Iris	8s	29s	13s	32s	112s	12s	19s	0.99	0.69	1.08	**0.88**
Thyroid	10s	17s	18s	17s	19s	16s	21s	0.65	0.36	1.00	**0.22**
Wine	10s	40s	9s	19s	71s	14s	22s	1.10	2.03	1.32	**3.03**

Table 8. The total results of the proposed algorithm

	least error	Average error	maximum error	average standard deviation	standard deviation
Balance	15.36	15.91	16.01	12.59	6.89
Cancer	4.15	5.1	5.68	5.30	0.53
Cancer- Int	4.49	4.85	5.30	2.21	0.86
Credit	15.92	16.20	19.01	13.55	9.08
Dermatology	16.11	17.06	19.05	14.20	7.83
Diabetes	16.66	16.89	17.20	19.12	6.61
E. coli	13.89	14.99	15.24	23.35	6.52
Glass	13.56	13.88	13.95	17.21	7.30
Heart	14.03	14.44	15.40	10.05	5.74
Horse	12.22	13.30	14.85	27.20	7.45
Iris	4.48	4.68	4.82	9.32	1.03
Thyroid	3.31	3.55	3.64	4.30	0.67
Wine	5.12	5.30	6.11	8.54	1.12

In this section, proposed algorithm compared to the performance of 3 chaotic particle swarm optimization. ACPSO [36], a clustering algorithm based on integration of K-Means and PSO [37], a population-based clustering technique using particle swarm optimization and k-means: PCPSO-Kmeans [38].

Table 9. Classification error percentages of the new techniques

	ACPSO	K-Means-PSO	PCPSO-KMeans	SSPCO
Balance	16.88 (4)	16.45 (3)	16.09 (2)	15.36 (1)
Cancer	4.19 (2)	4.52 (4)	5.06 (3)	4.15 (1)
Cancer- Int	4.79 (4)	0.23 (1)	0.80 (2)	4.49 (3)
Credit	16.89 (4)	16.10 (2)	17.80 (3)	15.92 (1)
Dermatology	17.46 (3)	17.89 (4)	16.38 (2)	16.11 (1)
Diabetes	17.89 (4)	17.55 (3)	16.08 (1)	16.66 (2)
E. coli	14.12 (3)	14.60 (4)	14.00 (2)	13.89 (1)
Glass	12.55 (2)	12.90 (3)	11.50 (1)	13.56(4)
Heart	15.46 (4)	15.36 (3)	15.23 (2)	14.03 (1)
Horse	12.49 (3)	12.83 (4)	12.29 (2)	12.22 (1)
Iris	4.87 (4)	0.88 (1)	1.29 (2)	4.48 (3)
Thyroid	4.52 (3)	4.61 (4)	4.12 (2)	3.31 (1)
Wine	0.00 (1)	0.00 (1)	0.12 (2)	5.12(3)

6. Conclusion

Clustering analysis, used in many disciplines and applications, is an important tool and a descriptive task seeking to identify homogeneous groups of objects based on the values of their attributes. In this paper by using chaotic improved SSPCO optimization algorithm, a clustering technique was presented at 13 benchmark tests which were compared with 11 other clustering algorithms on the benchmarks. The technique measures the performance of similar clustering patterns, which are classified in a cluster with other clusters, as well as the diversity and specific clustering of error, as compared to the techniques of clustering index, defining that the proposed algorithm in 5 sets with the lowest error clustering in clustering techniques were compared between 12 techniques and 5 other data collections have been good, and a total of 13 benchmarks have had the lowest average error. Results show that proposed clustering algorithm has much potential.

References

[1] H. Rehioui, A. Idrissi, M. Abourezq, F. Zegeari, DENCLUE-IM: A New Approach for Big Data Clustering, Elseivier, doi: 10.1016/j.procs.2016.04.265, 2016.

[2] J. Han, M. Kamber. *Concept and Techniques,* Morgan Kaufmann, 2001.

[3] D J. Hand, H. Mannila, P. Smyte, *Principles of Data Mining,* The MIT Press, 2001.

[4] M P. Veyssieres, R E. Plant , *Identification of vegetation state and transition domains in California's hardwood rangelands,* University of California, 1998.

[5] R. Xu, D. Wunsch, *Survey of Clustering Algorithms,* IEEE TRANSACTIONS ON NEURAL NETWORKS, Vol. 16. 3, 2005.

[6] A. Barladi, Alpaydin, *Constructive feedforward ART clustering networks—Part I and II.* IEEE Trans. Neural Netw, Vol. 13. 3, 2002.

[7] V. Cherkassky, F. Mulier, *Learning From Data: Concepts, Theory, and Methods.* New York : Wiley, 1998.

[8] Jain, M. Murty, P. Flynn, *Data clustering: A review,* ACM Comput. Surv, Vol. 31. 3, 1999.

[9] J. Kennedy, R. Eberhart, *Particle Swarm Optimization*, Proceedings of IEEE International Conference on Neural Networks, 1995.

[10] M. Dorigo, M. Birattari, T. Stutzle, *Ant Colony Optimization: Artificial Ants as a Computational Intelligence Technique.* IEEE Computational Intelligence Magazine, 2006.

[11] S. Arora, S. Singh, *The Firefly Optimization Algorithm: Convergence Analysis and Parameter Selection.* International Journal of Computer Applications, Vol. 69. 3, 2013.

[12] X. Fister, J. She yang, Brest , *A comprehensive review of firefly algorithm*: ELSEVIER, 2013.

[13] D. Karaboga, B. Basturk, *On the performance of artificial bee colony algorithm,* Applied Soft Computing, Vol. 8, 2008.

[14] D. Pham, A. Ghanbarzadeh, A. Koc, S. Otri, S. Rahim, M. Zaidi ,*The bees algorithm,* Technical note, Cardiff university, UK: Manufactoring Engineering center, 2005.

[15] R. Omidvar, H. Parvin, F. Rad, SSPCO Optimization Algorithm (See-See Partridge Chicks Optimization), 14 th-Mexican international conferences on artificial intelligence, IEEE, 2015.

[16] L. Rokach, *A survey of Clustering Algorithms*, Data Mining and Knowledge Discovery Handbook, 2nd ed. Springer Science. 10.1007/978-0-387-09823-4_14, 2010 .

[17] Y. Marinakis, M. Marinaki, M. Doumpos, N. Matsatsinis, C. Zopounidis, *A hybrid stochastic genetic—GRASP algorithm for clustering analysis.*, Oper. Res. Int. J.(ORIJ) , Vol. 8. 1, 2008.

[18] F. Jensen, *An Introduction to Bayesian Networks*, UCL Press/Springer–Verlag, 1996.

[19] D E. Rumelhart, G E. Hinton, R J. Williams, *Learning representation by backpropagation errors,* Nature 323, Vols. 533-536, 1986.

[20] D. Rumelhart, E. Hinton, J. Williams, Learning internal representation by error propagation, *Parallel DistributeProcessing*, vol. 1, 318-362, 1986.

[21] M H. Hassoun, *Fundamentals of Artificial Neural Networks,* The MIT Press, Cambridge, MA, 1995.

[22] M. B. Menhaj, Principles of Neural Networks, Amirkabir University of Technology, second edition, pp.715, 2002.

[23] J C. Cleary, L E. Trigg, *an instance-based learner using an entropic distance measure,* Proceedings of the 12th International Conference on Machine Learning. pp. 108–114, 1995.

[24] L. Breiman, *Bagging predictors,* Mach. Learn, Vol. 24. 2, 1996.

[25] G I. Webb, *Multiboosting: a technique for combining boosting and wagging,* Mach. Learn, Vol. 40. 2, 2000.

[26] R. Kohavi, *Scaling up the accuracy of naive-Bayes classifiers: a decision tree hybrid, in: E. Simoudis, J.W. Han, U. Fayyad (Eds.)* ,Proceedings of the Second International ConferenceonKnowledge Discovery and Data Mining, AAAI Press. pp. 202–207, 1996.

[27] P. Compton, R. Jansen, *Knowledge in context: a strategy for expert system maintenance, in: C.J., Barter, M.J., Brooks (Eds.),* Proceedings of Artificial Intelligence LNAI, Berlin, Springer–Verlag, Adelaide, Australia, Vol. 406. pp. 292–306, 1988.

[28] G. Demiroz, A. Guvenir, Proceedings *Classification by voting feature intervals.* of the Seventh European Conference on Machine Learning, pp. 85–92, 1997.

[29] A. De Falco, E. Della Cioppa, Tarantino, *Facing classification problems with Particle Swarm Optimization.* Appl. Soft Comput, Vol. 7. 3, 2007.

[30] D. Karaboga, C. Ozturk, *A novel clustering approach: Artificial Bee Colony (ABC) algorithm,* Applied Soft Computing. Elsevier, 10.1016/j.asoc.12.025, 2009.

[31] J. Nayak, B. Naik, H.S. Behera, Fuzzy C-Means (FCM) lustering algorithm: a decade review from 2000 to 2014. In: Jain, L.C. et al. (eds.) Comput. Intell. Data Min. vol. 2, Smart Innov. Syst. Technol. 32, vol. 2, pp. 133–149 (2014).

[32] J. Nayak, M. Nanda, K. Nayak, B. Naik, H.S. Behera, An improved firefly fuzzy c-means (FAFCM) algorithm for clustering real world data sets. Smart Innov. Syst. Technol. **27**, 339–348 (2014).

[33] X. Wu,B. Wu, J. Sun, S. Qiu, X. Li , A hybrid fuzzy K-harmonic means clustering algorithm. Appl. Math. Model. **39**(12), 3398–3409 (2015).

[34] S. Shamshirband, A. Amini, N B. Anuar, L M. Kiah, ICCA: a density-based fuzzy imperialist competitive clustering algorithm for intrusion detection in wireless sensor networks. Measurement 55, 212–226 (2014).

[35] K. Shankar Bopche, A. Jain, A Hybrid Clustering Technique Combining A PSO Algorithm with K-Means, International Journal of Computer Applications (0975 – 8887) Volume 137 – No.1, March 2016.

[36] L Y. Chuang, Ch J. Hasio, Ch H. Ho, Chaotic particle swarm optimization for data clustering, Expert Systems with Applications 38(12):14555-14563 · November 2011.

[37] S. Shamshirband, A. Amini, N B. Anuar, L M. Kiah, ICCA: a density-based fuzzy imperialist competitive clustering algorithm for intrusion detection in wireless sensor networks. Measurement 55, 212–226 (2014).

[38] B. Niu, Q. Duan, L. Tan, Y. Liu, A population-based clustering technique using particle swarm optimization and k-means, Springer, Natural Computing journal, 10.1007/s11047-016-9542-9, 2016.

[39] Y. He, J. Zhou, X. Xiang, H. Chen, H. Qin. Comparison of different chaotic maps in particle swarm optimization algorithm for long-term cascaded hydroelectric system scheduling. Chaos Solitons Fractals 2009;42:3169-76.

[40] L. Coelho, V. Mariani, Use of chaotic sequences in a biologically inspired algorithm for engineering design optimization. Expert Syst Appl 2008;34:1905-13.

[41] H. Gao, Y. Zhang, S. Liang, D. Li. A new chaotic algorithm for image encryption. Chaos Solitons Fractals 2006;29:393-9.