# Data Clustering Using by Chaotic SSPCO Algorithm

Rohollah Omidvar[1*], Amin Eskandari[1], Narjes Heydari[1], Fatemeh Hemmat[1], Sara Esmaeili[1]
1-Sama Technical and Vocational Training College, Islamic Azad University, Shiraz Branch, Shiraz, Iran
Email: r.omidvar.uni@gmail.com (Corresponding author), eskandary.a@gmail.com, narjesheydari977@gmail.com
hemmat.fatemeh@yahoo.com, saraesmaeili_uni@yahoo.com

Received: July 2016 Revised: April 2017 Accepted: April 2017

**ABSTRACT**
Data clustering is a popular analysis tool for data statistics in several fields including pattern recognition, data mining, machine learning, image analysis and bioinformatics, in which the information to be analyzed can be of any distribution in size and shape. Clustering is effective as a technique for discerning the structure and unraveling the complex relationship between massive amounts of data. See-See partridge chick's optimization (SSPCO) algorithm is a new optimization algorithm that is inspired by the behavior of a type of bird called see-see partridge. We propose chaotic map SSPCO optimization method for clustering, which uses a chaotic map to adopt a random sequence with a random starting point as a parameter; the method relies on this parameter to update the positions and velocities of the chicks. In this study, twelve different clustering algorithms were compared on thirteen data sets. The results indicate that the performance of the Chaotic SSPCO method is significantly better than the performance of the other algorithms for data clustering problems.

**KEYWORDS:** SSPCO Algorithm, Chaotic, Clustering, Clustering Error, Dataset.

## 1. INTRODUCTION

Chaotic theory is a universal complex dynamical phenomenon, lurking in very nonlinear systems, such as communication systems and meteorological systems. The control and synchronization of chaos has been widely studied [1]. Clustering algorithms have been applied to a wide range of problems, including exploratory data analysis, data mining [2], image segmentation [3], and mathematical programming [4-9], Clustering techniques have been used successfully to address the scalability problem of machine learning and data mining algorithms, where prior to, and during training, training data is clustered, and samples from these clusters are selected for training, thereby reducing the computational complexity of the training process, and even improving generalization performance [9-11]. Evaluation means measuring the similarity between clusters, measuring the compactness, and separation between clusters [12]. Evaluation measurement is also proposed as a key feature in internal and external cluster validation indexes [13]. Unanticipated clustering is a mental process in nature that prevents absolute judgment as a relative effect on all clustering techniques [14]. SSPCO optimization algorithm is also one of the newest algorithms based on the behavior of chickens and a type of bird called see-see partridge. We propose a data clustering algorithm based on chaotic SSPCO optimization algorithm. Using chaotic theory in movement equation helps a lot in creating a scenario and is suitable for clustering. In this paper, Section 1 is introduction and description of clustering, section 2 will be dedicated to introducing the related works. In Section 3, we present the SSPCO algorithm and the experimental study will be discussed in Section 4. Finally, Section 5 is conclusion. Clustering divides data into subsets of similar data so that the same data can be grouped together, while the different species remain in different groups [15]. In general, the evaluation criterion is the distance between the patterns that are alike and when the *N* object is present and objects are assigned to the *K* clusters, clusters to minimize the Euclidean distance of the object from the center of the object are the main issue. Equation 1 is as follows [16]:

$$j\left(w,z\right)=\sum_{i=1}^{N}\sum_{j=1}^{K}w_{ij}x_{i}-z_{j}^{2} \quad (1)$$

Where, K is the number of clusters, N number of models, $x_i(i=1,\ldots,N)$ is the i$^{th}$ pattern location, and $z_j(j=1,\ldots,K)$ that is the j$^{th}$ cluster center and based on Equation (2) shall be regulated as:

$$z_{j}=\frac{1}{N_{j}}\sum_{i=1}^{N}w_{ij}x \quad (2)$$

$N_j$ in this equation is the number of patterns in the j$^{th}$ cluster, and $w_{ij}$ weight $x_i$ is the pattern with the j$^{th}$

cluster which will be one or zero. 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.

## 2. RELATED WORK

In addition to the PSO algorithm, ABC of the 9 Clustering techniques have been used for comparison: Bayes Net [20] of Bayesian, the target of the classification is to attribute an instance based on the values of variables of different attributes to a class. Many class methods of the work-related set of attribute values create a label class. Learning Bayesian estimates the probability of distribution affiliate which was created after such time, and the classified amounts and class are more likely not to be specified. Multilayer Perceptron neural networks or MLP [21] show the non-linear relationships between input and output vectors. They operate through neurons connecting each node to the next one, and the previous layer is created. The output of each neuron is multiplied by weighting coefficients, and non-linear excitation function is given as input. The training is given to the perception of information occurs, and then the weights are adjusted so that the error between the output current and target reaches their least amount of training or the number of preset. Afterward, to evaluate the accuracy of the training process, a series of analytical inputs is applied to the network. The inputs are chosen from the inputs used in the training process of the network. Generally, complex neural network training and an optimization problem have many variables [29].Artificial Neural Networks Radial Basis Function, RBF, [22], unlike the MLP network with several successive layers, comprises three fixed layers: an input layer, a hidden layer, and an output layer. Dissimilar to MLP, the neurons of the hidden layer in RBF networks have a non-linear Gaussian function. The relationship between the neurons of the input layer and the hidden layer are not as simple as MLP network. Hidden layer neurons and the neurons, which are multi-dimensional units equal to the number of entries in both supervised and unsupervised RBF network, have been done. This is a learning process for the first time with a clustering method; the parameters of Gaussian function hidden layer, hidden layer, and output are set, and then the relationship between weight, using a supervised learning algorithm such as the slope of the standard error propagation and conjugate gradient method or procedure, and Marquardt, is regulated [30]. Engineering K-STAR [23], that is the nearest neighbor method, is based on a conversion of the public by the distance from the general function. Engineering Bagging [24] is used to incorporate the anticipated classification of several models. Suppose that you are going to do the prediction model for rating and make the desired data set small. You can select examples from the collection of data and samples obtained for trees category use (for example, C&RT and CHAID). Generally, for example, several different trees will arrive. Then to predict using different trees obtained from the samples, do a simple voting. Final classification, that is a category, predicts that there will be different tree limitations. MultiBoostAB technique [25], the concept for the production of multiple models (for prediction or classification), is therefore used. In boosting the RT&C or CHAID methods, the sequence of the classifier will be produced. NBTree technique [26], a technique that creates such an atmosphere and a decision tree classifier of Bayes N, can be combined. Ridor technique [27] is a technique in which the basic rules are first generated and then for any exception the best exception is generated. The exception to this rule will be based on the lowest classification error and then expand the tree as an exception, which leaves only the default rule to no exception. VFI clustering technique introduced by [28] is another clustering technique in this study which is used to compare performance. In 2007, clustering algorithms were used in studying particle mass, and velocity equations of motion of the particle mass were used for data clustering [19].In the present study, we compare the performance of an algorithm clustering. In addition, we use this clustering algorithm which is based on the artificial bee colony algorithm presented by Karabuga et al. The clustering of the three phases of the optimization algorithm for clustering artificial bee has been used [17].Much work has been carried out on the parameters modification [41], CPSO algorithm is used to improve logistic map [42]. The water discharge and death penalty function are described as the decision variables. In [43], Gaing and Lin propose CPSO to solve short-term unit commitment problems with security constraints. The objective of security-constrained unit commitment is to minimize the total generation cost, which is the total of both transition cost and production cost of the scheduled units. These researches adopt chaotic sequence instead of random sequence in PSO to improve the efficiency of the algorithm. It has successfully been used in the field of image processing as investigated by Chatterjee and siarry [44], Lin et. al. [45], iancastellietal. [46], Wang [47] and wang et. al. [48].

## 3. CHAOTIC SSPCO 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 [33]. The chicks of this type of bird are located in a regular queue at the time of danger

to reach a safe place and 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 takes us to the best optimal point, and this does not mean that it is minimizing the search space, but it is converging particles after some searches in a regular queue to the best point answers (bird mother). In the algorithm, consider a variable for each particle entitled as priority variable. For particle i, priority variable is defined according to Equation 4:

$$X_i.priority \quad (4)$$

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:

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

$X_i.cost$ is 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 each 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:

$$if \quad P_{best} > G_{best} \rightarrow G_{best} = P_{best} \ and \ X_i.priority = X_i.priority + 1 \quad (6)$$

$G_{best}$ is the global optimum. The motion equation of each particle is set almost similar to the particle swarm algorithm in the form of equation 7:

$$X_i.position = X_i.position + X_i.velocity \quad (7)$$

$X_i.velocity$ is the velocity of each particle or chick. Now the particle velocity equation is calculated based on Equation 8:

$$X_i.velocity = w * X_i.velocity + c * rand() * [position(X_{i+1}.priority)] - X_i.position \quad (8)$$

where, $X_i.velocity$ 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 1 to create a random movement for particles, $[position(X_i.priority + 1)]$ 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, and $X_i.position$ is the current location of the particle. It can be seen that, according to Eq. (7), each particle adjusts its movement based on a particle with one level higher priority. In this way, it does not matter the local and global optimums and at any point in time it only moves to find a particle which is a unit ahead of that particle; and for this reason, the calculation number and time in this algorithm has a great benefit compared to the previous optimization algorithm. According to this equation, particles move until they conduct a particle which is the mother particle to the optimum solution, and remaining particles move behind the particle to the optimum solution

### 3.1. Chaotic Theory

Simulation dynamic behavior of nonlinear systems is called chaos. It has raised enormous interest in different fields of sciences such as synchronization, chaos control, optimization theory, pattern recognition and so on [37]. 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 [38]. 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 [39]. 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:

$$Cr_{(t+1)} = k \times Cr_{(t)} \times \left(1 - Cr_{(t)}\right) \quad (9)$$

In Eq. (5), k =4 and for each independent run, Cr(0) is generated randomly, which Cr(0) is not equal to {0, 0.25, 0.5, 0.75, 1}. Behavior of Cr(t) is controlled by the driving parameter k of the logistic map (as *t* goes to infinity) [40]. Considering to the above descriptions the velocity update equation for chaotic chicks SSPCO optimization can be formulated as:

$$V_{id}^{k} = w * X_i.velocity + c^{Cr} * rand() * [position(X^{k}{}_{id+1}.priority)] - X_{id}{}^{k}.position \quad (10)$$

Where Cr is a function based on the results of the logistic map with values between 0.0 and 1.0.

*1.//initialize all chicken*
*2.Initialize*
*3.Repeat*
*4. For each chicken* $\boldsymbol{i}$
*5. //update the chicken's best position and priority*
*6. If* $\boldsymbol{f(x_i) > f(pb_i)}$ *then*
*7.* $\boldsymbol{pb_i = x_i}$
*8.* $\boldsymbol{prioirity_i = prioirity_i + 1}$
*9. End if*
*10. //update the global best position and priority*
*11. If* $\boldsymbol{f(pb_i) > f(gb)}$ *then*
*12.* $\boldsymbol{gb = pb_i}$
*13.* $\boldsymbol{prioirity_i = prioirity_i + 1}$
*14. End if*
*15. End for*
*16. //update chicken's velocity and position*
*17. For each chicken* $\boldsymbol{i}$
*18. For each dimension* $\boldsymbol{d}$
*19.* $\boldsymbol{V_{id}^k = w * X_i.velocity + c^{Cr} * rand() * [position(X^k{}_{id+1}.priority)] - X_{id}{}^k.position}$
*20.* $\boldsymbol{x_{i,d} = x_{i,d} + v_{i,d}}$
*21. End for*
*22. End for*
*23. //advance iteration*
*24.* $\boldsymbol{it = it + 1}$

**Fig. 1.** The Pseudo code of SSPCO algorithm

### 3.2. 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 the equation No.11:

$$(p_i^{\rightarrow 1}, \dots, p_i^{\rightarrow C}, v_i^{\rightarrow 1}, \dots, v_i^{\rightarrow C}) \quad (11)$$

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

$$p_i^{\rightarrow j} = \{p_{1,i}^j, \dots, p_{N,1}^j\} \quad (12)$$

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:

$$v_i^{\rightarrow j} = \{v_{1,i}^j, \dots, v_{N,1}^j\} \quad (13)$$

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

### 3.3. Fitness Function

In symbols, i-th individual fitness is given by equation 14:

$$(i) = \frac{1}{D_{Train}} \sum_{j=1}^{D_{Train}} d(x_j^{\rightarrow}, p_i^{\rightarrow CL_{known}(x_j^{\rightarrow})}) \quad (14)$$

The fitness function is computed in one step as the sum on all training set instances of Euclidean distance in N-dimensional space between generic instance $x_j^{\rightarrow}$and the centroid of the class, it belongs to database($p_i^{\rightarrow CL_{known}(x_j^{\rightarrow})}$). This sum is normalized with respect to $D_{Train}$.

When computing distance, any of its components in the N-dimensional space is normalized with respect to the maximal range in the dimension, and the sum of distance components is divided by N. With this choice, any distance can range within [0.0, 1.0].

## 4. EXPERIMENTAL STUDY

In this article, we compare the clustering algorithm with a two-clustering algorithm introduced earlier in this context. PSO clustering algorithm, in which the collective behavior of birds when flying was inspired by these parameters, has solved the problem of clustering [19]: 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 [17]: the size of the colony is 20, the maximum ring is 1000, and a total of 20,000 is assessed. SSPCO algorithm has been exactly set according to PSO algorithm parameters. In this study, 13 datasets of known database UCI are tested for clustering problem [18]. Clustering of the 13 benchmark criteria is similar to and consistent with all algorithms, and the techniques are compared with SSPCO algorithm. 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 [17]:

**Table 1.** The properties of the problems

| | 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 |

### 4.1. Results and Discussions

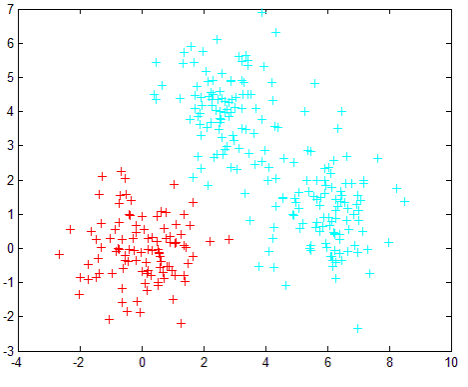
Benchmark comparison clustering techniques are based on the percentage error, and the percentage of models is sorted incorrectly. Each pattern should be part of the cluster closest to Euclidean distance with the cluster's center.-Margins of error classification criteria are compared in this paper based on Equation 15 and set to be [17]:

$$CEP(Classification\ Error\ Percentage) = 100 \times \frac{misclassification\ examples}{size\ of\ test\ data\ set} \quad (15)$$

Results of PSO, ABC 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 |
|---|---|---|---|
| Balance | 15.27 | 15.38 | 25.74 |
| Cancer | 4.00 | 2.81 | 5.81 |
| Cancer-Int | 4.17 | 0 | 2.87 |
| Credit | 15.12 | 13.37 | 22.96 |
| Dermatology | 7.02 | 5.43 | 5.76 |
| Diabetes | 15.23 | 22.39 | 22.50 |
| E. coli | 13.55 | 13.41 | 14.63 |
| Glass | 13.45 | 41.50 | 39.05 |
| Heart | 13.44 | 14.47 | 17.46 |
| Horse | 12.10 | 38.26 | 40.98 |
| Iris | 4..03 | 0 | 2.63 |
| Thyroid | 3.16 | 3.77 | 5.55 |
| Wine | 4.08 | 0 | 2.22 |



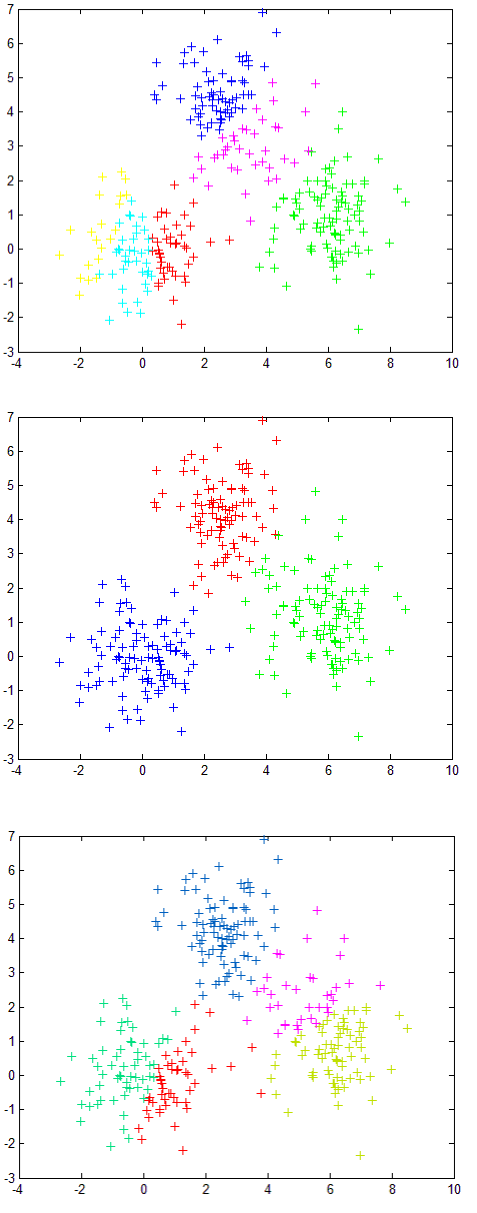


**Fig. 2.** Sample of Clustering Forms in SSPCO Algorithm

It can be seen that the clustering algorithm PSO in 6 data sets from ABC and PSO margins of error has fewer statistically significant errors in the data set compared to the other two algorithms, and the other data collection is ranked second on the error in the 4 clusters and only 3 of the data collection errors are higher than the other two algorithms. The average margin of error for all 13 data sets shows that the clustering algorithm is SSPCO that has the lowest percentage of error. The average margin of error on the full data set for clustering algorithm is with 10.04%, while the percentage errors of ABC and PSO are 13.13% and 15.99%, respectively.

**Table 3**. Classification error percentages of the techniques

| | SSPCO | ABC | PSO | BayesNet | MlpAnn | RBF | KStar | Bagging | MultiBoost | NBTree | Ridor | VFI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Balance | 15.27 | 15.38 | 25.74 | 19.74 | 9.29 | 33.61 | 10.25 | 14.77 | 24.20 | 19.74 | 20.63 | 38.85 |
| Cancer | 4.00 | 2.81 | 5.81 | 4.19 | 2.93 | 20.27 | 2.44 | 4.47 | 5.59 | 7.69 | 6.63 | 7.34 |
| Cancer-Int | 4.17 | 0.00 | 2.87 | 3.42 | 5.25 | 8.17 | 4.57 | 3.93 | 5.14 | 5.71 | 5.48 | 5.71 |
| Credit | 15.12 | 13.37 | 22.96 | 12.13 | 13.81 | 43.29 | 19.18 | 10.68 | 12.71 | 16.18 | 12.65 | 16.47 |
| Dermatology | 7.02 | 5.43 | 5.76 | 1.08 | 3.26 | 34.66 | 4.66 | 3.47 | 53.26 | 1.08 | 7.92 | 7.60 |
| Diabetes | 15.23 | 22.39 | 22.50 | 25.52 | 29.16 | 39.16 | 34.05 | 26.87 | 27.08 | 25.52 | 29.31 | 34.37 |
| E. coli | 13.55 | 13.41 | 14.63 | 17.07 | 13.53 | 24.38 | 18.29 | 15.36 | 31.70 | 20.73 | 17.07 | 17.07 |
| Glass | 13.45 | 41.50 | 39.05 | 29.62 | 28.51 | 44.44 | 17.58 | 25.36 | 53.70 | 24.07 | 31.66 | 41.11 |
| Heart | 13.44 | 14.47 | 17.46 | 18.42 | 19.46 | 45.25 | 26.70 | 20.25 | 18.42 | 22.36 | 22.89 | 18.42 |
| Horse | 12.10 | 38.26 | 40.98 | 30.76 | 32.19 | 38.46 | 35.71 | 30.32 | 38.46 | 31.86 | 31.86 | 41.75 |
| Iris | 4..03 | 0 | 2.63 | 2.63 | 0.00 | 9.99 | 0.52 | 0.26 | 2.63 | 2.63 | 0.52 | 0.00 |
| Thyroid | 3.16 | 3.77 | 5.55 | 6.66 | 1.85 | 5.55 | 13.32 | 14.62 | 7.40 | 11.11 | 8.51 | 11.11 |
| Wine | 4.08 | 0 | 2.22 | 0.00 | 1.33 | 2.88 | 3.99 | 2.66 | 17.77 | 2.22 | 5.10 | 5.77 |

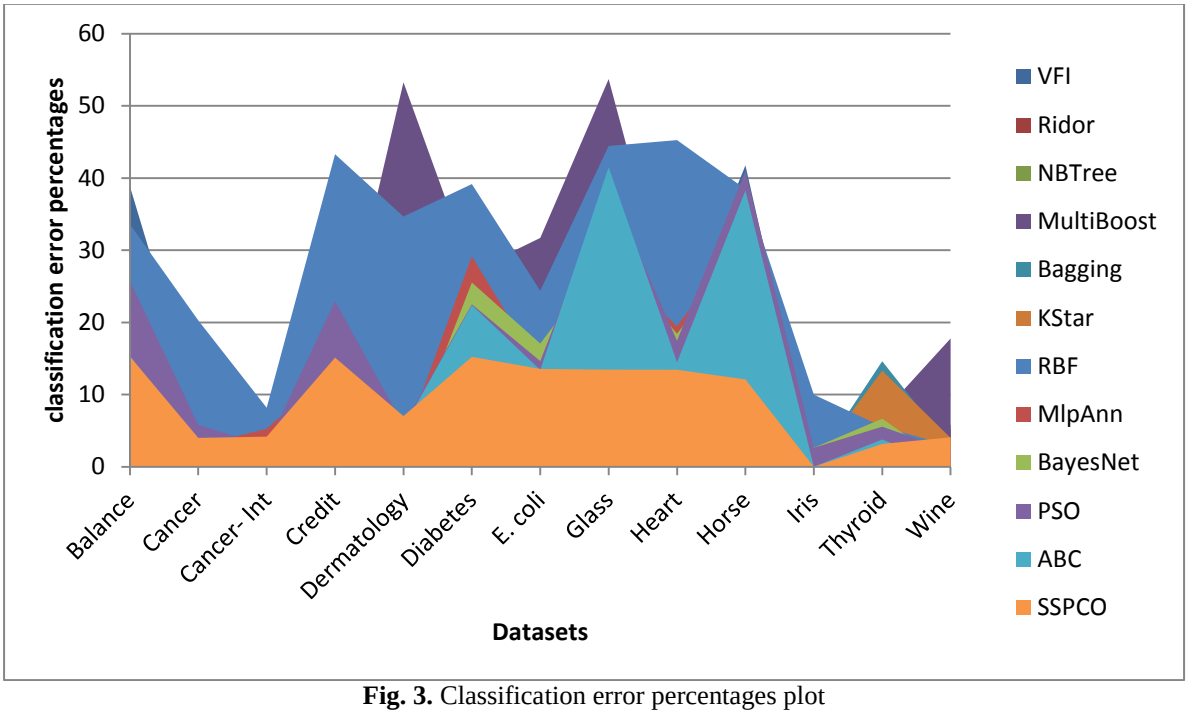



**Fig. 3.** Classification error percentages plot

**Table 4.** The average classification error percentages and ranking of the techniques

| | SSPCO | ABC | PSO | BayesNet | MlpAnn | RBF | KStar | Bagging | MultiBoost | NBTree | Ridor | VFI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Average | 10.04 | 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** |

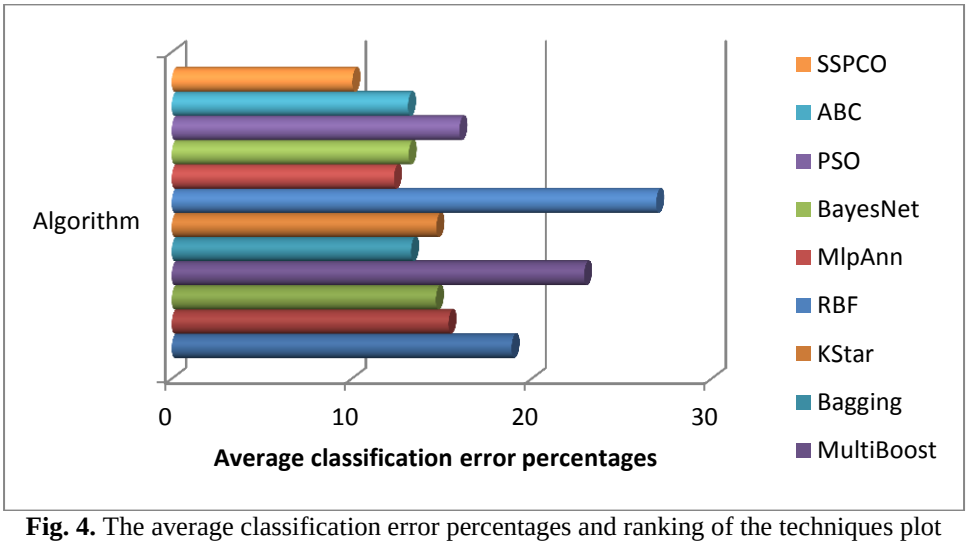



**Fig. 4.** The average classification error percentages and ranking of the techniques plot

In Table 4, 11% error clustering algorithm on 13 data sets [19] was found for SSPCO clustering algorithm. The proposed algorithm in 5 of the 12 data clustering techniques dwarfed, and 5 datasets with good performance and a good level of error have been clustered in three high sets of data error. The proposed algorithm in data collection Horse Glass and error clustering is very good compared to other techniques. In Table 3, the average error clustering techniques on the entire data set are reviewed, and the best error clustering algorithm is SSPCO, the second is MlpAnn, and techniques of clustering algorithm clustering ABC is the third. Totally, SSPCO algorithm is better in good times compared to the mechanisms of the other clustering techniques. Table 5 shows the standard deviation clustering algorithm proposed in this paper on different benchmarks.

**Table 5.** Standard deviation classification in the techniques

| | SSPCO | ABC | PSO | BayesNet | MlpAnn | RBF | KStar | Bagging | Multi-Boost | NBTree | Ridor | VFI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Balance | 6.30 | 7.23 | 9.23 | 8.86 | 4.25 | 5.85 | 3.35 | 6.30 | 5.55 | 7.03 | 9.33 | 3.36 |
| Cancer | 0.22 | 0.21 | 1.22 | 0.13 | 0.96 | 0.71 | 0.77 | 0.33 | 0.22 | 0.88 | 1.03 | 0.11 |
| Cancer- Int | 0.45 | 0.88 | 2.30 | 0.55 | 0.54 | 0.68 | 1.35 | 0.96 | 0.64 | 0.37 | 2.63 | 0.55 |
| Credit | 9.00 | 7.05 | 4.58 | 9.00 | 8.80 | 6.63 | 14.58 | 8.57 | 7.66 | 9.36 | 6.69 | 6.23 |
| Dermatology | 7.55 | 6.25 | 5.25 | 6.68 | 7.66 | 7.00 | 8.03 | 6.66 | 5.63 | 9.35 | 5.55 | 5.22 |
| Diabetes | 6.31 | 6.10 | 9.98 | 5.59 | 6.69 | 4.05 | 2.36 | 6.22 | 5.69 | 6.21 | 9.54 | 5.66 |
| E. coli | 5.43 | 5.25 | 6.69 | 6.01 | 5.53 | 5.32 | 4.38 | 6.33 | 8.55 | 3.23 | 12.05 | 4.30 |
| Glass | 6.38 | 6.68 | 9.35 | 7.00 | 7.22 | 7.05 | 3.36 | 4.69 | 7.35 | 6.35 | 9.23 | 6.28 |
| Heart | 7.14 | 4.47 | 7.78 | 6.52 | 5.02 | 4.45 | 2.20 | 4.66 | 6.65 | 4.33 | 8.32 | 4.41 |
| Horse | 6.35 | 6.68 | 10.03 | 7.04 | 6.69 | 7.50 | 2.68 | 9.98 | 4.63 | 6.61 | 8.00 | 6.60 |
| Iris | 0.77 | 0.98 | 0.22 | 0.98 | 1.01 | 0.78 | 1.58 | 0.99 | 0.69 | 1.08 | 3.02 | 0.88 |
| Thyroid | 0.26 | 0.55 | 0.55 | 0.44 | 0.65 | 0.44 | 0.20 | 0.65 | 0.36 | 1.00 | 1.50 | 0.22 |
| Wine | 1.02 | 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

| | SSPCO | ABC | PSO | BayesNet | MlpAnn | RBF | KStar | Bagging | MultiBoost | NBTree | Ridor | VFI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Balance | 12.33 | 14.65 | 11.65 | 14.29 | 8.64 | 9.38 | 13.30 | 13.10 | 10.65 | 9.14 | 13.65 | 9.99 |
| Cancer | 4.33 | 3.32 | 3.35 | 1.25 | 4.02 | 3.33 | 1.03 | 6.31 | 0.88 | 1.26 | 4.29 | 1.00 |
| Cancer- Int | 1.66 | 5.52 | 6.35 | 2.16 | 4.26 | 1.22 | 3.63 | 1.30 | 1.03 | 0.56 | 3.31 | 2.08 |
| Credit | 12.69 | 11.03 | 8.42 | 11.02 | 11.26 | 12.28 | 25.20 | 2.36 | 5.23 | 13.10 | 12.06 | 7.59 |
| Dermatology | 13.77 | 12.20 | 12.27 | 10.08 | 9.64 | 9.02 | 1.02 | 15.24 | 4.36 | 16.35 | 11.89 | 9.19 |
| Diabetes | 15.68 | 16.43 | 16.61 | 15.50 | 11.07 | 16.46 | 19.00 | 16.61 | 12.23 | 11.23 | 14.10 | 8.05 |
| E. coli | 18.64 | 20.25 | 18.25 | 8.55 | 6.62 | 22.29 | 10.03 | 14.41 | 9.46 | 6.64 | 10.00 | 9.18 |
| Glass | 12.38 | 17.20 | 16.44 | 14.20 | 9.18 | 14.59 | 16.66 | 16.55 | 13.64 | 11.90 | 12.21 | 14.88 |
| Heart | 9.35 | 10.00 | 8.29 | 9.68 | 6.64 | 9.94 | 8.92 | 8.89 | 9.80 | 6.55 | 9.91 | 7.78 |
| Horse | 13.77 | 25.31 | 22.29 | 15.28 | 9.75 | 12.03 | 4.42 | 11.06 | 20.06 | 10.71 | 15.15 | 12.75 |
| Iris | 8.37 | 2.56 | 6.54 | 6.64 | 6.89 | 6.30 | 8.84 | 6.68 | 1.06 | 6.45 | 6.44 | 1.73 |
| Thyroid | 3.65 | 6.25 | 2.23 | 2.24 | 4.21 | 2.05 | 2.22 | 4.26 | 0.85 | 4.22 | 2.69 | 0.88 |
| Wine | 5.66 | 4.58 | 4.02 | 3.65 | 1.88 | 7.89 | 8.48 | 6.64 | 1.06 | 6.44 | 6.64 | 7.17 |

**Table 7.** Run Time of classification in the techniques

| | SSPCO | ABC | PSO | BayesNet | MlpAnn | RBF | KStar | Bagging | Multi-Boost | NBTree | Ridor | VFI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Balance | 14s | 57s | 27s | 25s | 105s | 65s | 33s | 6.30 | 5.55 | 7.03 | 9.33 | 3.36 |
| Cancer | 22s | 33s | 30s | 28s | 84s | 54s | 36s | 0.33 | 0.22 | 0.88 | 1.03 | 0.11 |
| Cancer- Int | 09s | 53s | 17s | 89s | 66s | 16s | 9s | 0.96 | 0.64 | 0.37 | 2.63 | 0.55 |
| Credit | 18s | 102s | 21s | 44s | 94s | 40s | 33s | 8.57 | 7.66 | 9.36 | 6.69 | 6.23 |
| Dermatology | 22s | 85s | 35s | 25s | 37s | 28s | 26s | 6.66 | 5.63 | 9.35 | 5.55 | 5.22 |
| Diabetes | 19s | 92s | 39s | 26s | 115s | 39s | 58s | 6.22 | 5.69 | 6.21 | 9.54 | 5.66 |
| E. coli | 24s | 51s | 25s | 18s | 64s | 28s | 21s | 6.33 | 8.55 | 3.23 | 12.05 | 4.30 |
| Glass | 14s | 41s | 19s | 14s | 71s | 25s | 17s | 4.69 | 7.35 | 6.35 | 9.23 | 6.28 |
| Heart | 08s | 27s | 15s | 22s | 95s | 19s | 09s | 4.66 | 6.65 | 4.33 | 8.32 | 4.41 |
| Horse | 10s | 18s | 14s | 66s | 73s | 52s | 10s | 9.98 | 4.63 | 6.61 | 8.00 | 6.60 |
| Iris | 09s | 29s | 13s | 32s | 112s | 12s | 19s | 0.99 | 0.69 | 1.08 | 3.02 | 0.88 |
| Thyroid | 11s | 17s | 18s | 17s | 19s | 16s | 21s | 0.65 | 0.36 | 1.00 | 1.50 | 0.22 |
| Wine | 08s | 40s | 9s | 19s | 71s | 14s | 22s | 1.10 | 2.03 | 1.32 | 1.01 | 3.03 |

**Table 8.** The total results of the proposed algorithm

| | least error | Average error | maximum error | average standard deviation | standard deviation |
|---|---|---|---|---|---|
| Balance | 15.27 | 15.22 | 15.99 | 12.33 | 6.30 |
| Cancer | 4.00 | 4.58 | 4.90 | 4.33 | 0.22 |
| Cancer- Int | 4.17 | 4.22 | 5.03 | 1.66 | 0.45 |
| Credit | 15.12 | 15.78 | 16.02 | 12.69 | 9.00 |
| Dermatology | 7.02 | 7.55 | 7.98 | 13.77 | 7.55 |
| Diabetes | 15.23 | 15.80 | 17.00 | 15.68 | 6.31 |
| E. coli | 13.55 | 13.97 | 14.07 | 18.64 | 5.43 |
| Glass | 13.45 | 13.64 | 13.93 | 12.38 | 6.38 |
| Heart | 13.44 | 13.77 | 14.06 | 9.35 | 7.14 |
| Horse | 12.10 | 12.80 | 13.15 | 13.77 | 6.35 |
| Iris | 4..03 | 4.66 | 4.97 | 8.37 | 0.77 |
| Thyroid | 3.16 | 3.64 | 6.58 | 3.65 | 0.26 |
| Wine | 4.08 | 4.71 | 4.99 | 5.66 | 1.02 |

### 4.1. T-Test

The statistics t-test allows us to answer this question by using the t-test statistic to determine a p-value that indicates how likely we could have gotten these results by chance, if in fact the null hypothesis were true (i.e. no difference in the population)[34]. By convention, if there is less than 5% chance of getting the observed differences by chance, we reject the null hypothesis and say we found a statistically significant difference between the two groups. See Statistical Data Analysis for more information about hypothesis testing [35], [36].In this study $H_1$ is defined as follow: the obtained results are based on the random nature of the problem. If the value of the significant level for the example is zero, then it indicates that the probability of $H_1$ being incorrect will be zero. Therefore, in this particular example, it is safe to say that the obtained results are independent of the random circumstances of the problem.

**Table 9.** T-Test results For Classification error of SSPCO Algorithm

| Dataset | Confidence interval 95% | | Test Value | p-value | $H_0/H_1$ |
|---|---|---|---|---|---|
| | Low | upper | | | |
| Balance | 9.29 | 24.95 | 17 | 0 | $H_1$ |
| Cancer | 0.747 | 5.203 | 3 | 0 | $H_1$ |
| Cancer-Int | 2.18 | 7.227 | 4.5 | 0 | $H_1$ |
| Credit | 5.31 | 25.02 | 15.5 | 0 | $H_1$ |
| Dermatology | 7.33 | 22.77 | 15.5 | 0 | $H_1$ |
| Diabetes | 3.49 | 27.36 | 16 | 0 | $H_1$ |
| E. coli | 3.49 | 27.36 | 12.5 | 0 | $H_1$ |
| Glass | 5.10 | 22.22 | 13 | 0 | $H_1$ |
| Heart | 1.61 | 23.92 | 13 | 0 | $H_1$ |
| Horse | 14.00 | 27.00 | 22 | 0 | $H_1$ |
| Iris | 2.404 | 6.146 | 4.5 | 0 | $H_1$ |
| Thyroid | 2.11 | 3.627 | 3 | 0 | $H_1$ |
| Wine | 1.518 | 5.909 | 3 | 0 | $H_1$ |

**Table 10.** T-Test results For Standard deviation of SSPCO Algorithm

| Dataset | Confidence interval 95% | | Test Value | p-value | $H_0/H_1$ |
|---|---|---|---|---|---|
| | Low | upper | | | |
| Balance | 0 | 6.43 | 17 | 0 | $H_1$ |
| Cancer | 0.275 | 0.898 | 3 | 0 | $H_1$ |
| Cancer-Int | 0.398 | 1.982 | 4.5 | 0 | $H_1$ |
| Credit | 0.48 | 7.18 | 15.5 | 0 | $H_1$ |
| Dermatology | 0.55 | 15.78 | 15.5 | 0 | $H_1$ |
| Diabetes | 1.09 | 15.43 | 16 | 0 | $H_1$ |
| E. coli | 0 | 15.08 | 12.5 | 0 | $H_1$ |
| Glass | 4.66 | 19.19 | 13 | 0 | $H_1$ |
| Heart | 1.07 | 7 | 13 | 0 | $H_1$ |
| Horse | 3.03 | 15.84 | 22 | 0 | $H_1$ |
| Iris | 0.026 | 0.97 | 4.5 | 0 | $H_1$ |
| Thyroid | 0 | 1.773 | 3 | 0 | $H_1$ |
| Wine | 0.317 | 1.313 | 3 | 0 | $H_1$ |

### 4.2. Friedman Test

We have to correct the results of the proposed clustering algorithm in order to ensure that a test is carried out on the results. Friedman test, a nonparametric test, is an analysis of variance with repeated measures and is equivalent to that of the comparison between the K variables used in the average rating. The test status variables are assessed in several related cases. More information about Friedman's test is available [31], [32]. We have to consider the validity of the results of the proposed algorithm. We have evaluated the results of the proposed algorithm in 4 different repetitions on each of the 13 benchmark Friedman tests. As it can be seen in Figure 5, the average error of clustering rankings in 4 different iterations of the proposed clustering algorithm that is specified in various iterations is similar to clustering. The main samples taken from Friedman test show this on 13 benchmarks. The final answer of this test is 0.502, because it is more indicative of the value of 0.50; this is the natural course that answers the same level and between different repetitions compliance on each of the 13 benchmarks, and the results are reliable.

**Table 11.** Friedman test for Classification techniques of SSPCO algorithm in 4 times

**Descriptive Statistics**

| | N | Mean | Std. Deviation | Minimum | Maximum | Percentiles | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | 25th | 50th(Median) | 75th |
| Var00001 | 13 | 10.7154 | 5.41148 | 3.31 | 16.66 | 4.4850 | 13.5600 | 15.6400 |
| Var00002 | 13 | 10.7646 | 5.67621 | 3.35 | 19.33 | 4.8200 | 12.0500 | 15.7800 |
| Var00003 | 13 | 11.6962 | 7.56971 | 3.09 | 27.22 | 4.3750 | 9.5500 | 17.5800 |
| Var00004 | 13 | 10.9092 | 5.36789 | 4.00 | 18.51 | 5.5200 | 13.2900 | 15.2100 |

| Rank | | Test Statistics | |
|---|---|---|---|
| Var00001 | 2.08 | N | 13 |
| Var00002 | 2.54 | Chi-Square | 2.354 |
| Var00003 | 2.54 | df | 3 |
| Var00004 | 2.85 | Asymp. Sig. | .502 |

## 5. CONCLUSION

Given the crucial role of clustering in various sciences and the need for progress in this area, in this paper by using a chaotic optimization algorithm, a clustering technique was presented at 13 benchmark tests which were compared with 11 other clustering algorithms on the benchmarks. Chaotic SSPCO clustering algorithm was to simulate the behavior of a type of bird called see-see partridge and was compared with ABC and PSO clustering techniques and other known techniques. 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 the lowest average error. The results of Friedman's test proposed the accuracy and reliability of the clustering algorithm and the results of simulations showed the effectiveness of the algorithm for clustering data.

## REFERENCES


[1] J. Wang, B. Zhou, Sh. Zhou, "**An Improved Cuckoo Search Optimization Algorithm for the Problem of Chaotic Systems Parameter Estimation"**, *Hindawi Publishing Corporation Computational Intelligence and Neuroscience*, 2016, Article ID 2959370, 8 pages, 2016.

[2] IE Evangelou, DG Hadjimitsis, A. A. Lazakidou, C Clayton, "**Data Mining and Knowledge Discovery in Complex Image Data using Artificial Neural Networks**", *Workshop on Complex Reasoning an Geographical Data*, Cyprus, 2001.

[3] T. Lillesand, R. Keifer, "**Remote Sensing and Image Interpretation**", *John Wiley & Sons*, 1994.

[4] H.C. Andrews, "**Introduction to Mathematical Techniques in Pattern Recognition**", *John Wiley & Sons*, New York, 1972.

[5] G. Ball, D. Hall, "**A Clustering Technique for Summarizing Multivariate Data**", *Behavioral Science*, Vol. 12, pp 153-155, 1967.

[6] A.P. Engelbrecht. "**Sensitivity Analysis of Multilayer Neural Networks**", *PhD Thesis, Department of Computer Science, University of Stellenbosch, Stellenbosch, South Africa*, 1999.

[7] I.E. Evangelou, D.G. Hadjimitsis, A.A. Lazakidou, C Clayton, "**Data Mining and Knowledge Discovery in Complex Image Data using Artificial Neural Networks**", *Workshop on Complex Reasoning an Geographical Data*, Cyprus, 2001.

[8] L.V. Fausett, "**Fundamentals of Neural Networks**", *prentice Hall*, 1994.

[9] D. Fisher, "**Knowledge Acquisition via Incremental Conceptual Clustering**", *Machine Learning*, Vol. 2, pp. 139-172, 1987.

[10] J.R. Quinlan, "**C4.5: Programs for Machine Learning**", *Morgan Kaufmann, San Mateo*, 1993.

[11] M.R. Rao, "**Cluster Analysis and Mathematical Programming**", *Journal of the American Statistical Association*, Vol. 22, pp 622-626, 1971.

[12] L. Jegatha Deborah, R. Baskaran, and A. Kannan, "**A survey on internal validity measure for cluster validation,**" *International Journal of Computer Science & Engineering Survey*, Vol. 1, No. 2, pp. 85–102, 2010.

[13] E. Rend´on, I. M. Abundez, and C. Gutierrez, "**A comparison of internal and external cluster validation indexes,**" *in Proceedings of the 5th WSEAS International Conference on Computer Engineering and Applications*, pp. 158–163, 2011.

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

[15] 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 .

[16] 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.

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

[18] C. L. Blake, C. J. Merz. "**The University of California at Irvine Repository of Machine"**, http://www.ics.uci.edu/ mlearn/MLRepository., 1998.

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

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

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

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

[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] D. Rumelhart, E. Hinton, J. Williams, Learning "**Internal representation by error propagation"**, *Parallel Distribute Processing*, Vol. 1, 318-362, 1986.

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

[31] S. Garcia, D. Molina, M. Lozano, F. Herrera, "**A Study on The Use of Non-parametric Tests for Analyzing The Evolutionary Algorithms' Behaviour: A Case Study on The CEC'2005 Special Session on Real Parameter Optimization"**, *Journal of Heuristics*, Vol. 15(6), pp. 617–644, 2009.

[32] J. G. Villegas, "**Using Nonparametric Test to Compare the Performance of Metaheuristics"**, http://juangvillegas.files.wordpress.com/2011/08/friedman-test-24062011.pdf, 2001.

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

[34] Statistical Consultant for Doctoral Students and Researchers, http://www.statisticallysignificantconsulting.com/Ttest.htm.

[35] J. K. Kruschke, **"Bayesian estimation supersedes the *t* test"**, *Journal of Experimental Psychology: General* Version of May 31, 2012.

[36] J. C. F. De. Winter, "**Using the Student's t-test with extremely small sample sizes"**, *Practical Assessment, Research & Evaluation*, Vol 18, No 10, 2013.

[37] 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*, Vol. 42, pp. 3169-76, 2009.

[38] Coelho L, Mariani V. "**Use of chaotic sequences in a biologically inspired algorithm for engineering design optimization",** *Expert Syst. Appl*. Vol. 34, pp. 1905-13, 2008.

[39] H. Gao, Y. Zhang, S. Liang, D. Li. "**A new chaotic algorithm for image encryption",** *Chaos Solitons Fractals*, Vol. 29, Vol. 393-9, 2006.

[40] Kuo D. "**Chaos and its computing paradigm",** *IEEE Potentials Mag*., Vol. 24, pp. 13-5, 2005.

[41] S. Kianpisheh and N. M. Charkari, "**A grid workflow Qualityof-Service estimation based on resource availability prediction**", *The Journal of Supercomputing*, vol. 67, no. 2, pp. 496–527, 2014.

[42] Y. He, S. Yang, and Q. Xu, **"Short-term cascaded hydroelectric system scheduling based on chaotic particle swarm optimization using improved logistic map,**" *Communications in Nonlinear Science and Numerical Simulation*, Vol. 18, No. 7, pp.1746-1756, 2013.

[43] Z.-L. Gaing and G.-N. Lin, "**Unit commitment with security assessment using chaotic PSO algorithm**," *Journal of Circuits, Systems and Computers*, Vol. 20, No. 7, pp. 1357–1376, 2011.

[44] A. Chatterjee, P. Siarry. "**Computational Intelligence in Image Processing",** *Berlin*, Springer, 2013, pp. 57-65.

[45] T. Lin, P. Wu, F. Gao. "**Study on SVM Temperature Compensation of Liquid Ammonia Volumetric Flowmeter Based on Variable Weight PSO",** *International Journal of Heat and Technology*, Vol. 33, 2015, No 2, pp. 151-156.

[46] L. Piancastelli, L. Frizziero, G. Zanuccoli. "**Design and Heuristic Optimization of Low Temperature Differential Stirling Engine for Water**

**Pumping",** *International Journal of Heat and Technology*, Vol. 31, 2013, No 1, pp. 9-16.

[47] C. A. Wang. **"Rapid Auto-Focus Method in the Telephoto Lens**", *Review of Computer Engineer Studies*, Vol. 2, 2015, No 2, pp. 13-18.

[48] T. Wang, H. Yan, S. Zhong. "**Research of Fire Alarm System Based on Extension Neural Network",** *Review of Computer Engineer Studies*, Vol. 2, 2015, No 1, pp. 9-14.