International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Clustering Method Based on Artificial Algae Algorithm

Khaleel İbrahim Anwer 1[1], Sema Servi*[2]

Submitted: 11/01/2021 **Accepted : 10/04/2021**

Abstract: For decades, the researchers have developed many ways as optimization procedures with the aim of find the best solution in short time for many problems under certain conditions in the field of engineering, medicine and banking. These ways also used for parameter updating of algorithms. The most popular Optimization algorithms methods known are mining classification and clustering. In this article, the clustering used to identify the most important point in the best cluster centers of set data. Artificial Algae Algorithm (AAA) optimization algorithm used in the clustering process and implemented on UCI datasets. Balance, Breast Cancer Wisconsin Diagnostic, Breast Cancer Wisconsin original, Pima Diabetes, Glass, Iris, Wine, Urban Land Cover and Hill Valley UCI datasets used to assess the performing of the Algae Algorithm-based clustering algorithm. Euclides method used to calculate the distance between the data. The performance of the AAA based clustering algorithm, Total square distance values in different iteration numbers calculated for each data set. The total square error rate value calculated for each iteration and as the number of iterations progresses, the total square error rate value decreases smoothly. The obtained results compared with k-means, Differential Evolution (DE), Genetic Algorithm (GA), Artificial Bee Colony (ABC), Particle Swarm Optimization (PSO), Whale Optimization Algorithm (WOA) clustering algorithms. According to the experimental results in this study, the proposed AAA-based clustering algorithm achieved better results in iris and wine data sets compared to other clustering algorithms, while it obtained close to good results in other data sets. As a result, the Artificial Algae Algorithm-based clustering algorithm showed that the method showed a stable appearance and the performance of the clusters also increased, which shows that this study successfully achieved its purpose.

Keywords: *Artificial Algae Algorithm, Clustering, Optimization*

1. Introduction

The method of extracting valuable information from a database described as data mining, as of finding hidden relationships between these data. This information is particularly useful for many organizations to grow their businesses and plays great role for helping them as making important decisions [1]. Data mining technology as first level came out as a single machine and a single algorithm for vector data. In the second level, integrated with a database with multiple algorithms. The third level is here support for grid computing. The fourth level is the emergence of data mining algorithms. The fifth level is the location for massive data and parallel data mining algorithms for cloud services [2]. Clustering is one of the data mining techniques. The method of placing related data in a group or cluster, also known as cluster analysis called clustering [3]. Data clustering is the process of grouping or clustering based on similarity criteria of multidimensional data. Clustering is also an essential process, such as pattern recognition and machine learning. In addition, Data clustering is the key method of Artificial Intelligence [1]. Clustering carried out with clustering algorithms. Most clustering algorithms based on two common techniques known as hierarchical clustering and partitioning. Clustering algorithms be used in applications such as image segment vector and colour image inspection, data mining, compression, and machine learning. In the clustering process, a cluster defined by a cluster centre (or centroid). Data clustering is a complicated problem in uncontrolled classification acceptance as it has different shapes and dimensions. Clustering can be formally as summarized as follows: Let X be our dataset, this dataset x= {x1, X2….Xn)} and it reflects Number of experiments used in the dataset. The samples in this dataset are divided into k clusters C= {C1, C2…Ck} and k represents the number of clusters. In clustering, each sample assigned to a cluster, and each cluster must have at least one sample [5]. Distance equations used to determine the process of assigning samples to clusters. After measuring the distance of each sample to all centres, the cluster closest distance assigned to its centres. In the clustering process, cluster centres assigned randomly. Later, these cluster centres updated at each step and each clustering algorithm performs this centre update process differently. The performance and clustering success of clustering algorithms depend on the central update process, which means that the better cluster centres found are the better the clustering process is performed. Clustering algorithms known to be one of the research subjects in data mining. Due to their use in many fields. Clustering algorithms are successfully included in many fields of projects on data such as image segmentation, document clustering and market [6]. Chuang et al. evaluated the k-means clustering algorithm and used it in segmentation of images taken from MRI devices. According to the experimental results, they stated that the fuzzy k-means clustering algorithm showed very good segmentation of the noisy pictures [7]. Also recognized that the efficiency of clustering

[1] *Computer Eng / Master degree, Selcuk University, Konya, TURKEY*
ORCID ID : 0000-0002-5227-1078
[2] *Departmen of Computer Eng, Selcuk University, Konya, TURKEY*
ORCID ID : 0000-0003-2069-9085
** Corresponding Author Email: Halil.kara882@gmail.com*

algorithms focuses on updating cluster centres and obtaining better clustering centres among non-clustered data samples, meaning that the clustering result obtained well. Recently, multiple optimization algorithms used in cluster update process of clustering algorithms. PSO [8], ABC [9], Frog jump algorithm [10], the bat algorithm (BA) [11], the ant algorithm [12] and GA [13] have been used by many researchers and as a result they achieved great success in their work.

Nasiri and Khayabani used a whale optimization algorithm that emerged from foraging behaviours to overcome clustering problem in their study. Nasiri and Khayabani made comparison of their whale-based clustering algorithm with Differential development, GA, PSO and ABC. In their comparison, they used seven data sets Iris, wine, contraceptive method choice (CMC), Breast Cancer Balance, Thyroid and Glass that frequently used from the UCI data sets. Based on their experimental results, the whale optimization algorithm based clustering algorithm gave better results in many data sets than other optimization based clustering algorithm [14].

Maulik and Bandyopadhyay in their studies carried out clustering with genetic algorithm. Genetic algorithm based clustering method has been applied as artificial and real on two different data sets. Genetic algorithm based clustering method was applied artificially and real on two different data sets. It has been determined that the clustering method based on the genetic algorithm performs better than the k-means algorithm in data sets [15].

Karaboga and Ozturk designed a new clustering way of algorithm with the ABC algorithm. They designed the ABC clustering algorithm that tested on 13 UCI data sets and after these tests, they compared their results with nine clustering algorithms. According to their experimental results, they realize better results in many data sets shown by the ABC clustering algorithm compared to other clustering algorithms [16].

In the studies of Shelokar et al., A clustering algorithm focused on this, ACO studied for clustering problems. Testing of the ACO based clustering algorithm carried out on many types of data sets. The efficiency of the proposed clustering algorithm was very promising according to the experimental results obtained by comparing the studies of the ACO based clustering algorithm on the GA with popular optimization algorithms for egg, simulated retrieval and taboo search. The performance of the proposed clustering algorithm was very promising, according to the experimental results obtained by comparing the studies of the ant colony optimization based clustering algorithm with popular optimization algorithms such as "simulated annealing" and "taboo" search on the genetic algorithm [17].

Zhang et al. colony optimization algorithm developed from bees' behavior for clustering problems. They compared their suggested bee colony based clustering algorithm with the simulated annealing, genetic algorithm, and taboo search way and particle swarm optimization algorithms. They used three data sets after comparing "Iris", "wine" and "Thyroid" in their UCI data sets. Based on the experimental findings, it has explained that the bee colony optimization based clustering algorithm gave better results in all data sets than other optimization based clustering algorithms [18].

Balavand et al., in their study, presented a new clustering method using Cluster Validity Indices (CVI) and Data Envelopment Analysis (DEA) in combination. They named their presented clustering method as auto Clustering Based on Data Envelopment Processing (ACDEA)). ACDEA clustering method determines the cluster numbers and cluster centres of data sets with Crow Search Algorithm. It was determined that the ACDEA clustering method was well known in this study and used in many data sets. As a result, it has stated that the ACDEA clustering method was very successful in this study and it can used in many data sets [19].

Karakoyun presented a new clustering method using the frog leap algorithm in his study. The frog leap algorithm based clustering method on UCI datasets compared with many clustering ways in the literature. According to the results, it been emphasized that the frog jump algorithm based clustering algorithm performed better in many UCI data sets [20].

Omran et al. declared that the k-means clustering algorithm applied with the PSO algorithm in which the cluster centres updated. PSO based k-means clustering algorithm that they proposed tested in the image segmentation of data sets. They compared their proposed PSO based k-means clustering algorithm with PSO and k-means clustering algorithm. According as results, the clustering algorithm that they proposed has shown to be very suitable for picture segmentation data sets [21].

In this study, an Artificial Algae Optimization Algorithm AAA based clustering algorithm used to find solutions to clustering problems. In the method in which AAA based clustering algorithm is used, the data in the clustered data sets clustered according to the cluster centres obtained by global search. Global search tests all probabilities in the search space to minimize error between sets of data samples. The UCI data sets were applied in the AAA-based clustering method are algorithm BCWD, BCWO, Pima Diabetes, Glass, Iris and Wine. All of the experimental results obtained from this article study detailed in the conclusion section.

2. CLUSTERING ALGORITHMS

Although all data classification is similar in terms of data entry, the goal is to continue learning it without classes. This process is to classify data whose classes not known and groups according on object similarities [22]. When we examine Figure 1, we can see and understand that there are many types of algorithms for data clustering in the literature [23]. However, the most popular and one most used is the k-means algorithm. The K-means algorithm is very simple in use and suitable for solving fast and compact clusters [24].

Although the k-means clustering algorithm is a popular technique, it does not show the exact number of k for processing before performing the k-means clustering. Therefore, determining the best k number for k-means clustering algorithm poses difficulties [25].

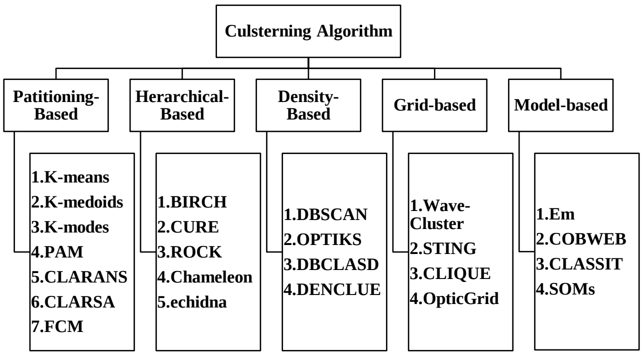


Figure 1. Clustering Algorithms [23]

2.1. *K*-means algorithm

This algorithm is predefining cluster as K in the dataset. This iterative algorithm has each data point belongs only to one cluster and attempts to divide those data points into non-overlapping subsets. While keeping cross-cluster data as different from each other as possible, it also attempts to gather the data points within it as similar as possible. Therefore, these data points often allocated to a cluster in such a sense that the sum of the square space in the data set between data points and the cluster centres is minimal. It has understood that the smaller the variations between clusters, the more homogeneous the data points would be in that cluster. In other words, data assigned to clusters according to the range between the data of the data set and the cluster centres [27].

The number of K sets is determined, Centres are determined by mixing the data set and after we do determine random K data points for the cluster centres without changing them, Iterations will until there is no alteration in cluster centres, which means that the distribution of data points to clusters will not change.

In the update process of cluster centers, the total square distance calculated between data points and all cluster centers. Every data point assigned to the centers of the nearest cluster. Each data point allocated to the centers of the nearest cluster. The sum of all the data points belonging to each cluster calculated for new cluster centers. According to the given equation, i' of the X data point in the data set is assigned to the closest j of the C cluster centers. In equation 1, m represents the number of data points in the data collection and k represents the number of clusters [28].

$$J = \sum_{i=1}^{m}\sum_{j=1}^{k}\left\|X_i - K_j\right\|^2 \qquad (1)$$

The flow diagram of the K-means clustering algorithm given in Figure 2 [28].

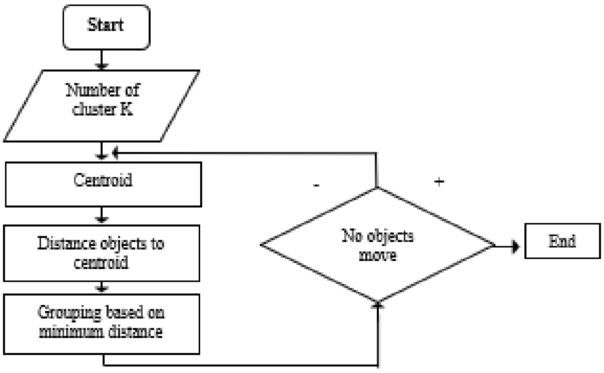

Start
Number of cluster K
Centroid
Distance objects to centroid
Grouping based on minimum distance
No objects move
-
+
End

Figure 2. Flow chart diagram of K- means algorithm [28]

3. Artificial Algae Algorithm (AAA) optimization algorithm

Artificial algae known for characterizing algae characteristics and showing that you can respond to any solution in the problem area. With comparing it to real one, artificial algae shows that once they adapt to the environment by moving towards the light source to photosynthesize by spiral swimming, they can switch superior species and eventually multiply through mitotic division. Thus, this algorithm consists of three basic parts called "Evolutionary Process", "Adaptation" and "Helical Motion". In the AAA, algae will be are main genera population which the latest consists of algae colonies. The name algae colonies given to a group of algae cells living together. Algae colony and algae population given in Equation 2 and 3 respectively.

$$\text{Population} = \begin{bmatrix} x_1^1 & \cdots & x_1^D \\ \vdots & \ddots & \vdots \\ x_N^1 & \cdots & x_N^D \end{bmatrix} \qquad (2)$$

$$\text{Alg colony} = \left[x_i^1, x_i^2 \cdots x_i^D\right] \qquad (3)$$

The algae colony functions as a single cell and passes together and the cells in the colony can die under unfavourable living conditions. The colony present at the optimal point known the optimum colony and consists of optimum algae cells [30].

3.1. Evolutionary process

The growth kinetics of the algae colony calculated with the Monode model given in Equation 4.

$$\mu = \frac{\mu_{max} S}{K_s + S} \qquad (4)$$

Here, μ particular growth rate, μmax is the highest specific growth rate, S is the nutrient concentration and K partitioned into the algal colony. In equation 7, μmax value is set as one (according to the theory of mass conservation, the overall volume converted into biomass must be equal to the amount of substrate consumed per unit time.).

The size of the ith algae colony at t+1 time in the Monod equation is seen in the equation 5:

$$G_i^{t+1} = \mu_i^t G_i^t \qquad i = 1,2,\cdots N \qquad (5)$$

Here G_i^t is the size of the i'th algae colony at time t, N is the number of algae colonies in the system. The algae colony that provides good solutions grows more because of the high number of nutrients they obtain. In each algae cell of the smallest algae colony that passed away in the process of evolution, the algae cell of the largest algae colony reproduced. This duplication done with Equations 6, 7 and 8:

$$biggest^t = \max G_i^t \;\; i = 1,2,\cdots N \qquad (6)$$

$$smallest^t = \min G_i^t \;\; i = 1,2,\cdots N \qquad (7)$$

$$smallest_m^t = \;\; biggest_m^t \;\; m = 1,2,\cdots D \qquad (8)$$

Here D implies that there is a problem factor, which the largest represents the largest colony of algae, and the smallest represents the smallest colony of algae.

3.2. Adaptation

This process ends with a shift in the hunger level of the algorithm, and the preliminary hunger value is zero to every artificial algae.

$$starving^t = \max A_i^t \;\; i = 1,2,\cdots N \qquad (9)$$

$$starving^{t+1} = starving^t + (biggest^t - starving^t)\; rand \quad (10)$$

In Equation 9, A value represents the hunger value of the ith algae colony at time t, and starving^t represents the algae colony with the largest angle value at time t. In Equation 10, the modification Parameter specifies the adaptation process and be utilized at time t. Generally, the modification Parameter value is in the range [0, 1].

3.3. Helical Motion

The motion of the algae cell is helical in AAA, gravity constraining motion displayed as zero, viscous drag showcased as zero, and the viscous drag seen as a shear force commensurate to algae cell size. Algae colonies are spherical in shape and the friction surface is the surface area, which shown by Equations 11 and 12. Here $\tau(x_i)$ is the friction surface.

$$\tau(x_i) = 2\pi r^2 \tag{11}$$

$$\tau(x_i) = 2\pi\left(\sqrt[3]{\frac{3G_i}{4\pi}}\right)^2 \tag{12}$$

Three dimensions are randomly determined by the helical motion of the algae cell, one of them is linear motion as expressed in Equation 13, and the other two dimensions have angular motion in Equations 14 and 15 For one dimensional problems equation 13 is used and the algae cell or the colony is heading in one direction. In two-dimensional problems, algae motions are sinusoidal and thus equations 13 and 15 are used. In the case of three or more dimensions, algae motions helical so equations 13, 14 and 15 are used. The friction surface and the distance to the light source have to decide the step size of the motion [30].

$$x_{im}^{t+1} = x_{im}^{t} + \left(x_{jm}^{t} - x_{im}^{t}\right)\left(\Delta - \tau^{t}(x_i)\right)p \tag{13}$$

$$x_{ik}^{t+1} = x_{ik}^{t} + \left(x_{jk}^{t} - x_{ik}^{t}\right)\left(\Delta - \tau^{t}(x_i)\right)\cos\alpha \tag{14}$$

$$x_{il}^{t+1} = x_{il}^{t} + \left(x_{jl}^{t} - x_{il}^{t}\right)\left(\Delta - \tau^{t}(x_i)\right)\sin\beta \tag{15}$$

Three dimensions for the helical rotation of the algae cell, (x_{im}^t, x_{ik}^t and x_{il}^t) are stated to be the coordinates of the algae cell (x, y and z) at the time t. α and $\beta \in [0, 2]$; $p \in [-1, 1]$; Δ the force to form; t (xi) is the friction surface area of the i th algae cell. The so-called block diagram of the AAA given in Figure 3 [30].

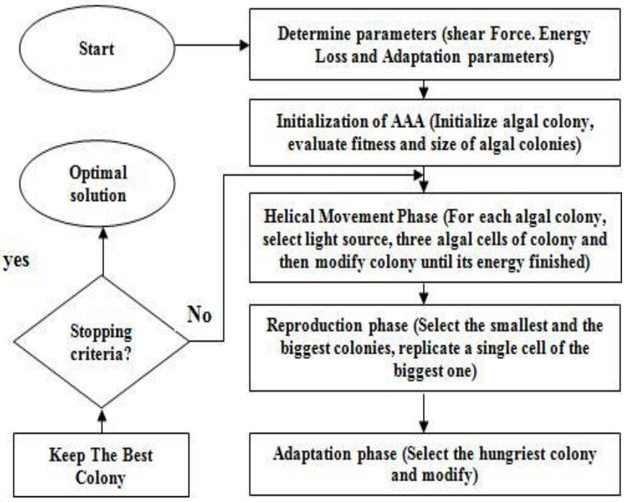

Start
Determine parameters (shear Force. Energy Loss and Adaptation parameters)
Initialization of AAA (Initialize algal colony, evaluate fitness and size of algal colonies)
Optimal solution
Helical Movement Phase (For each algal colony, select light source, three algal cells of colony and then modify colony until its energy finished)
yes
No
Stopping criteria?
Reproduction phase (Select the smallest and the biggest colonies, replicate a single cell of the biggest one)
Keep The Best Colony
Adaptation phase (Select the hungriest colony and modify)

Figure 3. Block diagram of Artificial Algae Algorithm [30]

4. Artificial Algae Algorithm (AAA) based clustering algorithm

In this study, data clustering performed as an AAA-based clustering algorithm to solve problems. The AAA presented in this paper as an optimization algorithm for this task. The most important process in clustering problems is that the cluster centres updated and the better the cluster centres updated for this, the better the data clustered. The proposed artificial algae based clustering algorithm used in this study to update cluster centres in clustering. The flow diagram of the AAA based clustering algorithm used shown in Figure 4. When we look at Figure 4, clustered data are required from the user for clustering, so the number of k is determined according to the class information of this clustered data. At the same time, the number of repetitions and the number of algae colonies in the population of the AAA, the number of iterations, shear force, energy loss and adaptation parameters are required for the update process.

Then, a random cluster centre generated for the data set between certain intervals, thus the data set clustered and finally the total square length error is calculated. After updating the cluster centres for a specific of iteration numbers with the AAA, the calculation of the total square length error performed after each update process. Rand index calculations made for the clustered data set depending on the realization of the update as the number of iterations entered by the user.

The performance of the clustering algorithm based on the proposed AAA applied to the Balance, BCWD, BCWO Pima Diabetes, Glass, Iris, Wine, Urban Land Cover and Hill Valley datasets. Meanwhile, these obtained results discussed in details at the section of experimental results.

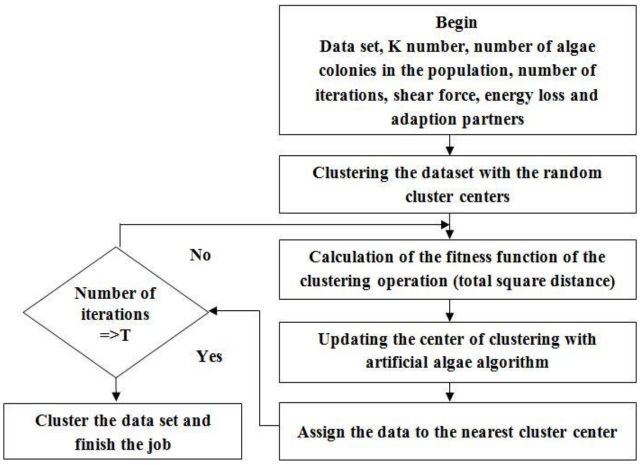

Begin
Data set, K number, number of algae colonies in the population, number of iterations, shear force, energy loss and adaption partners
Clustering the dataset with the random cluster centers
No
Calculation of the fitness function of the clustering operation (total square distance)
Number of iterations =>T
Yes
Updating the center of clustering with artificial algae algorithm
Cluster the data set and finish the job
Assign the data to the nearest cluster center

Figure 4. Flowchart of the AAA-based clustering algorithm

The AAA-based clustering algorithm used in this study applied to any non-clustered data set (Figure 4). These stages consist of several steps. All of these steps presented in a very descriptive and detailed manner.

Step 1: Entries of the algorithm it is known that at this step, the AAA algorithm needs more than one parameter values to perform the clustering process. The user must give these parameters. The values of the parameters that belong to the AAA algorithm are the values used by their authors in the literature, also at the same time all these values known for trial purposes. Those parameters are the non-clustered data set, iteration number, cluster number, shear force, algae number and energy loss.

Step 2: Random clustering process the non-clustered dataset in this step clustered by random cluster centres generated by AAA. The cluster centres formed equal to the number of algae. After this formation of clustering according to the cluster centres, the error rate between clusters calculated for each alga known as the fitness function. In addition, at this stage, the Total square distance function used to calculate the error between clusters.

Step 3: Updating cluster centres with AAA in this step, cluster centres updated with the AAA algorithm. This motivates the fitness functions of other algae with respect to the alga, which has the best fitness function for the AAA algorithm. This update process performed according to a certain number of iterations, and as the number of iterations increases, the value of the fit function decreases, and the error rate between sets decreases. After each update, the fitness function recalculated for each algae.

Step 4: Clustering process in this last step, the non-clustered data set clustered according to the last cluster centers accessed in the preceding step. Finally, the AAA-based clustering algorithm finishes the process.

5. Experimental

In this study, it realized that different parameters used when data sets applied to the proposed AAA based clustering algorithm. In this section, the total square distance values of Balance, BCWD, BCWO, Pima Diabetes, Glass, Iris, Wine, Urban Land Cover and Hill Valley UCI data sets given in the AAA-based clustering algorithm we use for clustering problems calculated at table 1. When the parameters of the proposed AAA-based clustering algorithm examined, the number of repetitions and the number of populations have more than one different values. The proposed AAA-based clustering algorithm run with three different iteration numbers of 1000, 3000, 5000 and separate results obtained for each iteration result. On the other hand, in the same way, the results obtained by applying the population number in 2 different ways, 30 and 50. Forming force 2, Energy loss 0.3 and Adaptation probability 0.5 parameters are the best parameters applied by the author on many problems, the AAA-based clustering algorithm coded with Matlab programming language all tests performed on computer with Intel 2.3 GHz processor 8 GB ram and Windows 10 feature.

5.1. Data Sets Used in Clustering

The performance of the clustering method based on AAA taken for evaluation in UCI data sets. The characteristics of the Balance, BCWD, BCWO, Glass, Iris, Wine, Urban Land Cover and Hill Valley data sets selected from the UCI datasets [26].

Collection of balance data; This data set has 4 features, 3 classes and 625 data samples which is used to model psychological states. 625 Of these data samples include 49 are determined as balanced class, 288 as scale type to the left class and 288 as scale type to the right.

Breast Cancer Wisconsin Diagnostic (BCWD) data collections; this data has 32 features and 2 classes and 569 data samples which is token from chest area. 32 features used to determine whether tumours in the chest area are benign or malignant. 357 of these 569 data samples were determined as benign class and 212 as malignant class.

Breast Cancer Wisconsin Original (BCWO) dataset; has 9 features, 699 samples and 2 classes that is similar to the BCWD dataset with the difference of that BCWO dataset identifies tumours with fewer features. Thirty features used to determine whether the tumours in the chest area are benign or malignant. Of these 699 data samples, 458 classified as Benign class and 241 as Malignant class.

Glass data set; has 9 features and 6 classes consists of 214 data samples in total. In this data set specially, sodium, magnesium, aluminium, silicon, potassium, barium, calcium and iron values are used. Building Windows float, building windows non-float, vehicle windows float, containers, tableware and headlamps glass classes are determined according to these property values.

Iris data set; contains 4 features, 3 class and 150 samples in total. The values of the properties taken from the width and length of the leaves of the iris flower and 50 samples of the first class are Iris Setosa, 50 samples in the second class are Iris Versicolour and finally 50 samples are Iris Virginica.

Wine data set; there are 13 features, 3 class information and there are 178 data samples in total. In the Wine data set, 59 of 178 data samples were determined as first class, 71 as second class and 48 as positive class.

Urban Land Cover data set; has 148 features and 9 classes, which consists of 168 samples. This data set consists of urban land cover data.

Hill Valley data set; has 101 features, 2 classes and consists of 606 samples. These data sets consist of the hill valley characteristics data.

5.2. Evaluation Criteria of Used Clustering Algorithm

It is very important for us to evaluate the clustering algorithm used in this study after the data set clustered with clustering algorithms. The way to evaluate clustering algorithms performed by evaluation methods such as rand index and total error rate. This evaluation process reveals which clustering algorithm performs the clustering process better than others do.

5.2.1. Distance calculation

The space between the elements of the data set calculated by distance methods such as Euclidean (Euclidean) space, Manhattan distance, Minkowski distance and Pearson distance. In this study, the space between the data made using the Euclidean (Euclidean) distance method. The Euclidean (Euclidean) distance method given as in Equation 16:

$$d(i,j) = \sqrt{\sum_{m=1}^{n} \left(x_{im} - x_{jm}\right)^2} \qquad (16)$$

$\boldsymbol{d(i,j)}$: $\boldsymbol{i}$. and **j**. represents distance of sample

x_{im}: i. for sample m. value of attribute

x_{jm} : j. for sample m. value of attribute

i= 1, 2… N

j= 1, 2… N

N = Total number of samples in the data set.

5.2.2. Rand Index Measurement

The Rand Index (RI) measurement used in many studies to test the output of cluster algorithms in this study. The Rand index equation given in Equation 3 [29].

When looking at the Rand index equation, we can see how many data samples correctly assigned to the correct cluster within the clustered data set by the clustering algorithm. Therefore, the value obtained from the Rand index is between 0 and 1, the closer the value is to 17, the better the clustering algorithm performs.

$$\text{Rand index } = \frac{\text{correct number}}{\text{correct number} + \text{uncorrect number}} \qquad (17)$$

5.2.3. Total Squared Distance Calculation

Total Quadratic Distance (TKU) measures the error between clusters in the clustered dataset through the clustering algorithm. Total Squared Distance Calculation equation given in Equation 18 [30].

$$\text{Total Square Distance} = \sum_{i=1}^{k} \sum_{x_j \in C_i} \left\| x_j - c_i \right\|^2 \qquad (18)$$

k is the number of clusters and $\left\| x_j - c_i \right\|$ is the distance from the centres of c_i to x_j. The smaller value amount of the total quadratic distance means the better clustering of data samples. Thus, total square distance method known as one of the clustering evaluation methods frequently used by many researchers.

6. Results

By giving the parameters of the proposed AAA -based clustering algorithm, the results obtained for each data set examined. All the results obtained from the data sets given as in tables.

After each non-clustered data set given to the AAA-based clustering algorithm, the AAA algorithm in this case determines the most suitable cluster centres for the data samples of this data set. Then, Data samples clustered according to the cluster centres obtained because of each iteration, and a total square distance and rand index values calculated for this data set. Each data set run by the program for five cuts and the average with standard deviation values of the best, worst, best and worst values for total square distance and rand index criteria given in Tables 1 and 2. With the examining the results of Table 1 and 2, the proposed AAA-based clustering algorithm has been obtained very consistently with the standard deviation values of either the total quadratic distance or the rand index values.

Graphical representation of the total square distance values of 1000, 3000, 5000 iteration numbers and 30, 50 populations for all data sets is given in Figure 5,6,7,8,9,10,11,12 and 13. When we look at the graphics the proposed AAA-based clustering algorithm as seen in all data sets as the number of iterations increases, it is seen that the total square distance error value decreases smoothly.

Very close and similar results gets every time with the running of these statistics results of proposed YAA based clustering algorithm. The evaluation of the proposed clustering algorithm according to the fitness function (total square distance) has done separately for each data set. This total square distance value starts with a large value in the first iteration and this error value decreases as the number of iterations progresses, it indicates that the algorithm is working correctly.

The proposed AAA-based clustering algorithm in the Table 3 has obtained the values of 52246587.3 and 755926.14, respectively, in Urban Land Cover and Hill Valley data sets. This shows the superiority of the proposed AAA based clustering algorithm with the total square distance error values.

In addition to the tests have been resulted, the projected AAA-based clustering algorithm has compared with the study of Nasiri and Khiyabani [14] as given in Table 4.

As when we look at Table 4, we can see the outcome of the proposed AAA-based cluster algorithm compared with the k-means, DE, GA, ABC, PSO, WOA clustering algorithms. In this comparison, iris, wine, CMC, balance, cancer, glass and thyroid data sets used and the total square distance value obtained from 1000 iterations for each data set.

The average value of the total square distance values obtained by running the order 20 times according to the standard deviation and the total square distance value. According to Table 4, it was shown that the suggested AAA based clustering algorithm provides superior result than other clustering algorithms with total square distance values of 96,6555 and 16292,31 in iris and wine data sets. In CMC, Balance, Glass and Thyroid datasets, the proposed AAA based clustering algorithm is in the second place. While the proposed clustering algorithm based on AAA shows slightly poor performance in the Cancer dataset as taking fourth in place.

As when we check Table 5, the proposed AAA-based clustering algorithm according to the Friedman statistical test takes the second place with a value of 2.00 while bypassing the DE, GA, ABC, PSO clustering algorithms. On the other hand, the WOA based clustering algorithm known to be in the first place with a value of 1.71. When looking at the test results in general, the proposed AAA clustering method seems appropriate for clustering.

7. Conclusion

In this study we have learned, it known that clustering methods are one of the frequently used methods in data mining and this uncontrolled method "clustering" is common as in many statistical data analyses such as engineering and medicine, market, segmentation and banking. More than one parameter affects the clustering process; the most important of these parameters is to find the most suitable cluster centres for the clustered data. Finding the most suitable cluster centres known as an optimization problem. Optimization algorithms are the most common used to find the optimum values in optimization problems, and in this study, AAA utilized to obtain appropriate cluster centres and to refine cluster success. The performance of the AAA based clustering algorithm has been evaluated with frequently used data sets of Balance, BCWD, BCWO, Pima Diabetes, Glass, Iris, Wine, Urban Land Cover and Hill Valley UCI. The performance of the proposed Artificial Algae Algorithm based clustering algorithm evaluated by using the total quadratic distance and rand index evaluation criteria. The results obtained proved that the proposed clustering method clustered the non-clustered data sets very well. The proposed clustering method were compared with the k-means, DE, GA, ABC, PSO and WOA based clustering methods which proposed in the literature. The proposed clustering method with comparison resulted as the first in two data sets, the second in four data sets, and the fourth in only one data set in all six data sets. The Artificial Algae-based clustering algorithm proposed for the clustering problem in this study can achieved successful with using these data sets. The error rate value been calculated in the total square iteration of the data set. Thus, as the number of iterations progressed, it seen that the total square error rate decreased smoothly and as a result, the AAA-based clustering algorithm used in this study shows quite stable in these data sets. Based on these great results with high performance, the proposed Artificial Algae Algorithm-based clustering algorithm can applied to real-world data sets.

Table 1: Total square distance final values for data sets

Data sets		T = 1000		T = 3000		T = 5000	
		P = 30	**P = 50**	**P = 30**	**P = 50**	**P = 30**	**P = 50**
Balance	Good	1423.82	1423.82	1423.82	1423.82	1423.82	1423.82
	Bad	1423.82	1423.82	1423.82	1425.72	1423.82	1423.82
	Average	1423.82	1423.82	1423.82	1424.20	1423.82	1423.82
	SD	3.01E-05	3.74E-08	1.38E-12	0.8511	7.79E-13	6.53E-13
bcwd	Good	149473.86	149473.86	149473.86	149473.86	149473.86	149473.86
	Bad	149473.87	149473.89	149473.86	149473.86	149473.86	149473.86
	Average	149473.86	149473.87	149473.86	149473.86	149473.86	149473.86
	SD	0.0039983	0.016703	4.11E-10	1.45E-10	1.29E-10	2.91E-11
bcwo	Good	2964.39	2964.39	2964.39	2964.39	2964.39	2964.39
	Bad	2964.39	2964.39	2964.39	2964.39	2964.39	2964.39
	Average.	2964.39	2964.39	2964.39	2964.39	2964.39	2964.39
	SD	4.11E-10	8.64E-10	1.86E-12	1.09E-12	1.58E-12	8.81E-13
Glass	Good	214.67	218.53	210.51	210.48	218.85	210.45
	Bad	242.07	245.38	237.82	214.90	237.90	237.82
	Average.	231.22	234.10	222.59	212.57	234.05	222.91
	SD	11.3779	10.9553	13.1875	2.1742	8.4934	13.6628
İris	Good	96.66	96.66	96.66	96.66	96.66	96.66
	Bad	96.66	96.66	96.66	96.66	96.66	96.66
	Average.	96.66	96.66	96.66	96.66	96.66	96.66
	SD	2.39E-08	5.93E-09	2.66E-14	1.00E-14	2.66E-14	1.74E-14
Pima Indian Diabetes	Good	47561.13	47561.13	47561.13	47561.13	47561.13	47561.13
	Bad	47561.13	47561.13	47561.13	47561.13	47561.13	47561.13
	Average.	47561.13	47561.13	47561.13	47561.13	47561.13	47561.13
	Sta. Sap.	8.15E-11	2.81E-10	3.04E-11	2.91E-11	1.36E-11	2.18E-11
Wine	Good	16292.19	16292.20	16292.18	16292.18	16292.18	16292.18
	Bad	16292.68	16292.33	16292.67	16292.18	16292.67	16292.18
	Average.	16292.30	16292.24	16292.28	16292.18	16292.47	16292.18
	SD	0.21524	0.058177	0.21581	6.14E-10	0.26431	1.10E-11
Urban Land Cover	Good	725562.56	708106.74	613365.76	624308.32	601396.21	613897.85
	Bad	755926.14	747798.39	660899.54	659587.99	642100.52	635490.91
	Average.	733892.77	723066.79	638695.50	643797.48	617492.91	620788.12
	SD	12797.3685	16541.3528	19442.0274	14120.5626	14979.7324	8912.2746
Hill valley	Good	51607777.14	51596571.24	50238591.20	50238944.81	50229397.88	50229333.85
	Bad	52246587.30	52264566.75	50246203.28	50240367.09	50230256.10	50229626.19
	Average	52009785.05	51895171.20	50242103.45	50239681.22	50229872.31	50229530.96
	SD	259355.386	280247.573	3124.6223	603.6838	313.4741	117.0006

Table 2 Rand index results

Data sets		T = 1000		T = 3000		T = 5000	
		P = 30	P = 50	P = 30	P = 50	P = 30	P = 50
Balance	Good	58.969	60.039	58.983	59.509	58.983	59.509
	Bad	57.831	57.831	57.831	58.712	57.831	58.756
	verage	58.449	59.02	58.522	59.031	58.246	59.012
	SD	.0052583	.0083136	.0063109	.0029114	.0057456	.0029495
Bcwd	Good	77.153	77.153	77.153	77.153	77.153	77.153
	Bad	77.153	77.153	77.153	77.153	77.153	77.153
	verage	77.153	77.153	77.153	77.153	77.153	77.153
	SD	0	0	0	0	0	0
Bcwo	Good	93.229	93.229	93.229	93.229	93.229	93.229
	Bad	93.229	93.229	93.229	93.229	93.229	93.229
	verage	93.229	93.229	93.229	93.229	93.229	93.229
	SD	1.24E-16	1.24E-16	1.24E-16	1.24E-16	1.24E-16	.24E-16
Glass	Good	67.02	65.999	67.338	68.763	67.338	67.633
	Bad	59.691	62.504	59.691	66.346	59.691	59.691
	verage	63.956	64.992	64.968	67.611	61.221	64.918
	SD	0.036497	0.014132	0.031989	.0085962	0.034195	.034862
İris	Good	88.742	88.742	88.742	88.742	88.742	88.742
	Bad	88.742	88.742	88.742	88.742	88.742	88.742
	verage	88.742	88.742	88.742	88.742	88.742	88.742
	SD	0	0	0	0	0	0

Data sets		T = 1000		T = 3000		T = 5000	
		P = 30		P = 50		P = 30	
Pima Indian Diabetes	Good	52.398	52.398	52.398	52.398	52.398	52.398
	Bad	52.398	52.398	52.398	52.398	52.398	52.398
	Average	52.398	52.398	52.398	52.398	52.398	52.398
	SD	0	0	0	0	0	0
Wine	Good	73.115	73.115	73.115	73.115	73.115	73.115
	Bad	72.582	73.115	72.582	73.115	72.582	73.115
	Average	73.009	73.115	73.009	73.115	72.795	73.115
	SD	0.0023861	0	0.0023861	0	0.0029224	0
Urban Land Cover	Good	61.264	61.151	70.886	69.611	69.139	71.14
	Bad	58.502	56.65	64.258	59.629	65.089	60.883
	Average	59.384	59.508	66.886	64.455	67.15	66.22
	SD	0.011146	0.018487	0.024526	0.048224	0.015499	0.039717
Hill valley	Good	50.085	50.085	50.085	50.085	50.085	50.085
	Bad	50.085	50.085	50.085	50.085	50.085	50.085
	Average	50.085	50.085	50.085	50.085	50.085	50.085
	SD	0	0	0	0	0	0

Table 3: Comparison results of data sets

Data	Criteria	k-Means	DE	GA	ABC	PSO	WOA	AAA
	Mean	103,58	125,3395	120,0766	97,0986	97,4058	96,7993	**96,6555**
Iris	SD.	12,48	1,13	0,36	0,43	0,26	0,1	**0,037**
	Rank	5	7	6	3	4	2	1
	Mean	18560	16530,11	16493,93	16394,82	16292,96	16295	**16292,31**
Wine	SD	2869	1,32	0,75	1,29	8,05	**0,72**	3,942
	Rank	7	6	5	4	2	3	1
	Mean	5853,57	5794,273	5732,869	5643,849	**5532,096**	5539,72	5532,193
CMC	SD	1,63	30,9	**0,64**	10	14,78	0,79	1,48
	Rank	7	6	5	4	1	3	2
	Mean	1427,52	1427,869	1425,489	1429,945	1424,754	**1423,8**	1423,916
Balance	SD	3,49	1,05	0,94	1,32	**0,35**	0,98	0,42552
	Rank	5	6	4	7	3	1	2
	Mean	3262,35	3237,142	3237,926	3036,955	3037,963	**3036,12**	3064,387
Cancer	SD	0,16	0,19	0,25	**0,048**	6,44	0,2	0,13
	Rank	7	5	6	2	3	1	4
	Mean	255,073	261,0285	252,2135	256,0595	240,8885	**231,2912**	237,6462
Glass	SD	4,72	7,66	**1,56**	2,94	12,06	4,51	7,8821
	Rank	5	7	4	6	3	1	2
	Mean	1995,189	1877,237	1888,209	1897,421	1890,207	**1870,93**	1873,598
Thyroid	SD	10,78	6,05	6,9	6,62	**0**	1,3	11,1468
	Rank	7	3	4	6	5	1	2

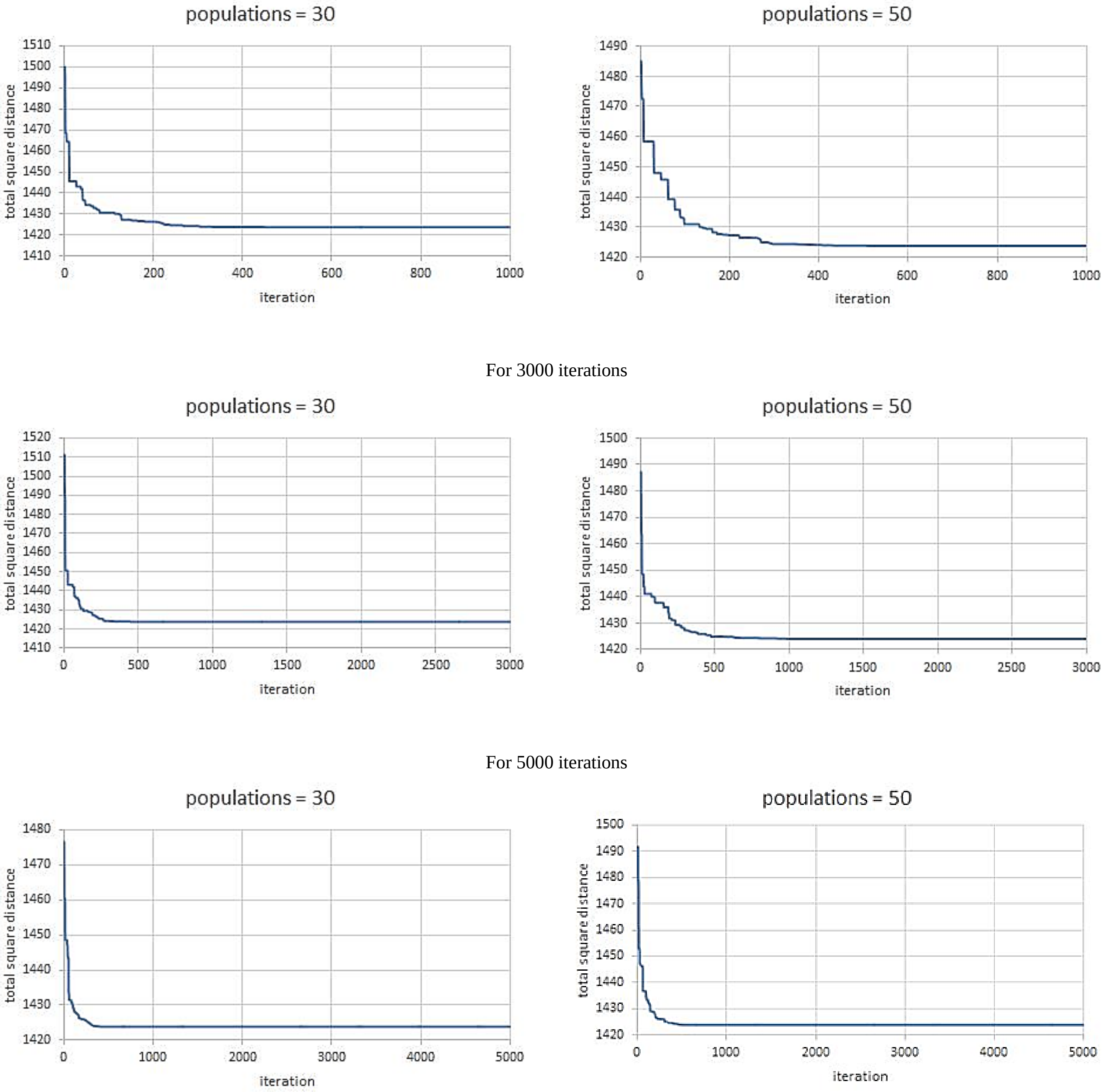
Table 4. Comparison results of Urban Land Cover and Hill Valley data sets

Data sets	K-means	AAA
Urban Land Cover	53686200.5	**52246587.3**
Hill Valley	762291.83	**755926.14**

Table 5 Friedman test results

Methods	Results
K-means	6.14
DE	5.71
GA	4.86
ABC	4.57
PSO	3.00
WOA	1.71
AAA	2.00

For 1000 iterations

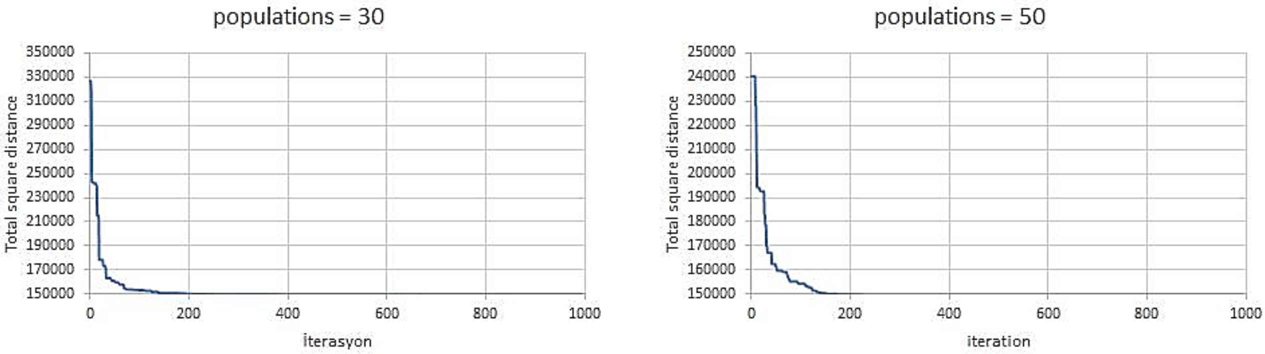


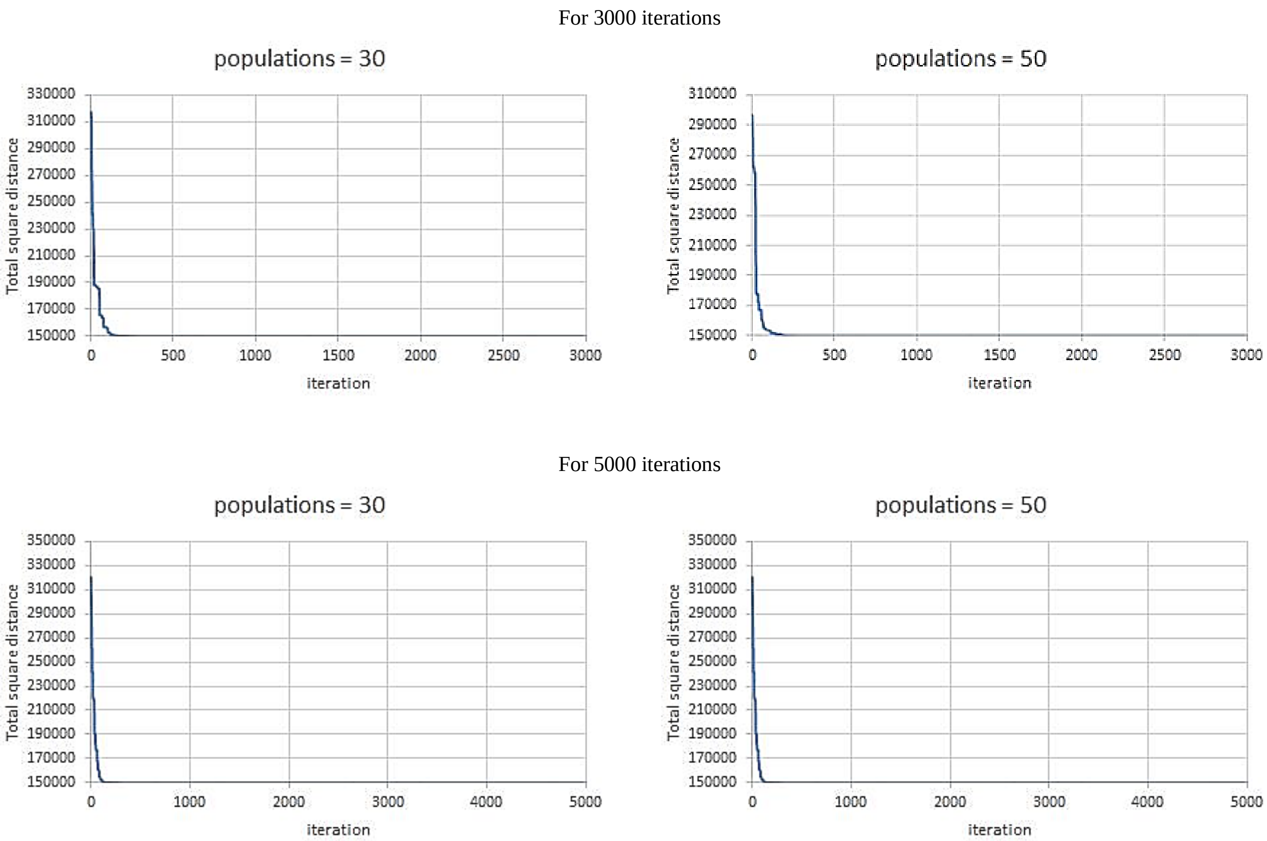
For 3000 iterations

For 5000 iterations

Figure 5: Convergence of the total square distance for Balance data set

For 1000 iterations

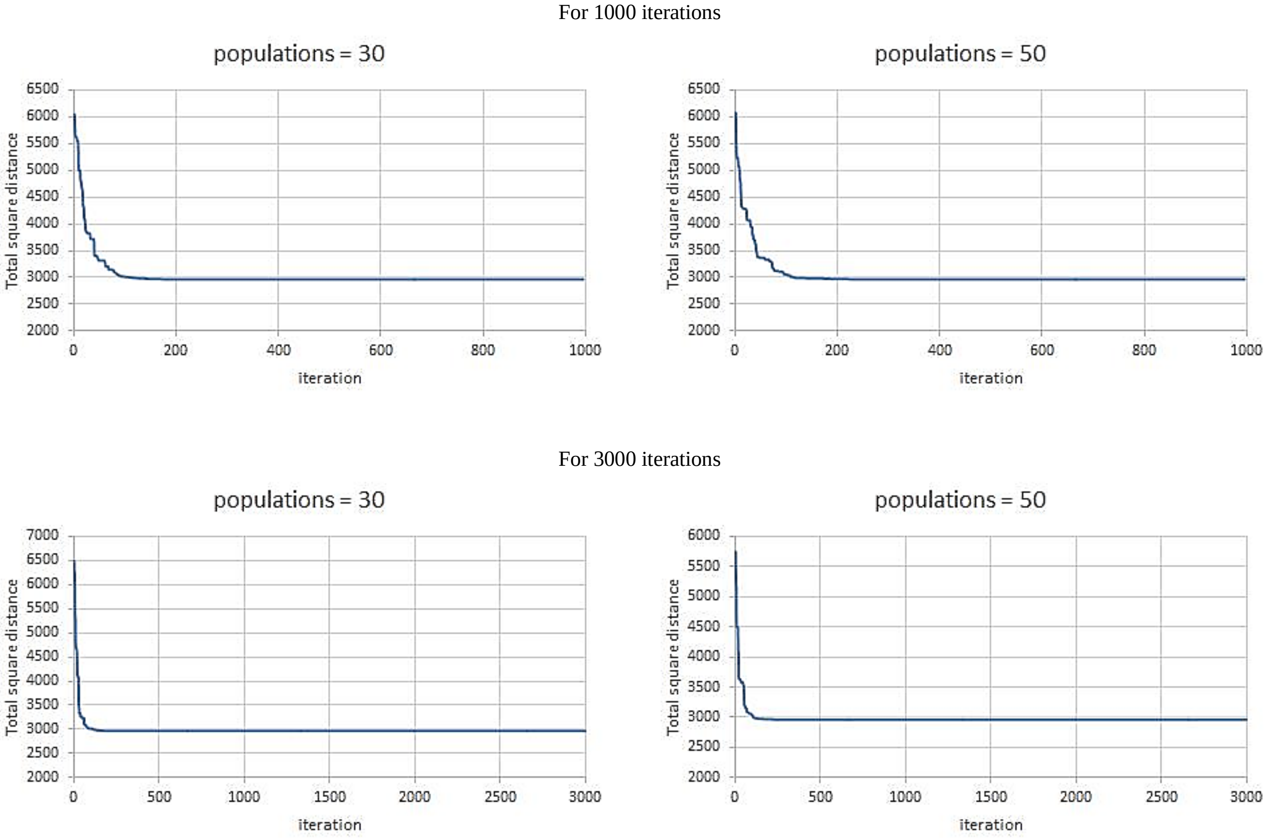




For 3000 iterations

For 5000 iterations

Figure 6: Convergence of the total square distance for bcwd data set



For 1000 iterations

For 3000 iterations

For 5000 iterations

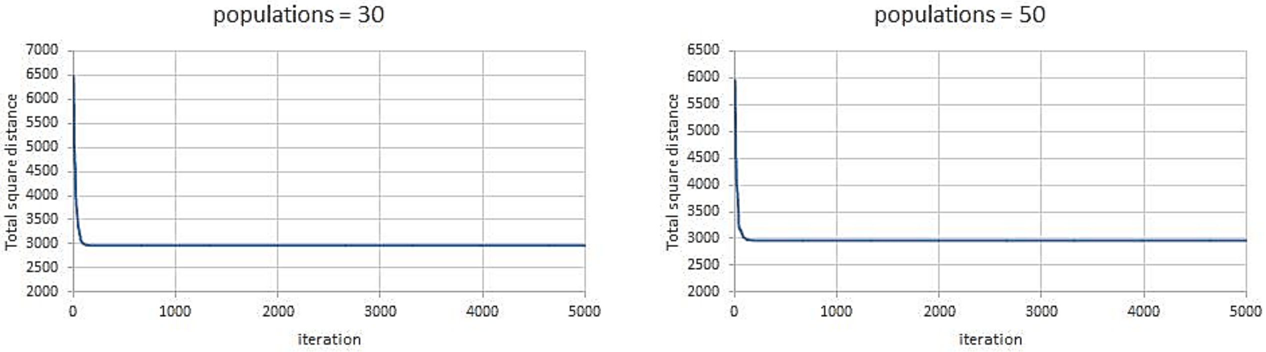
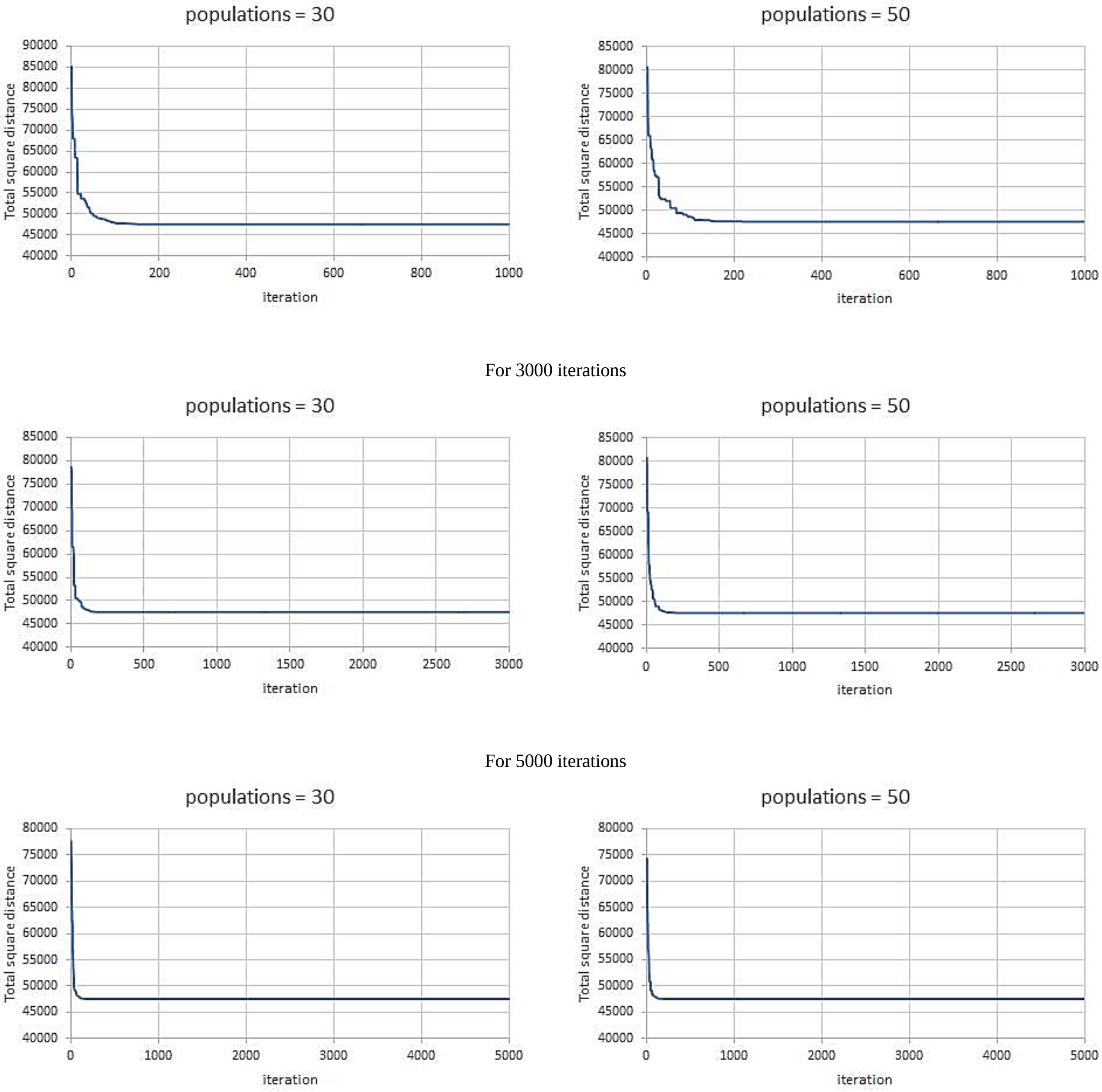


Figure 7: Convergence of the total square distance for bcwo data set

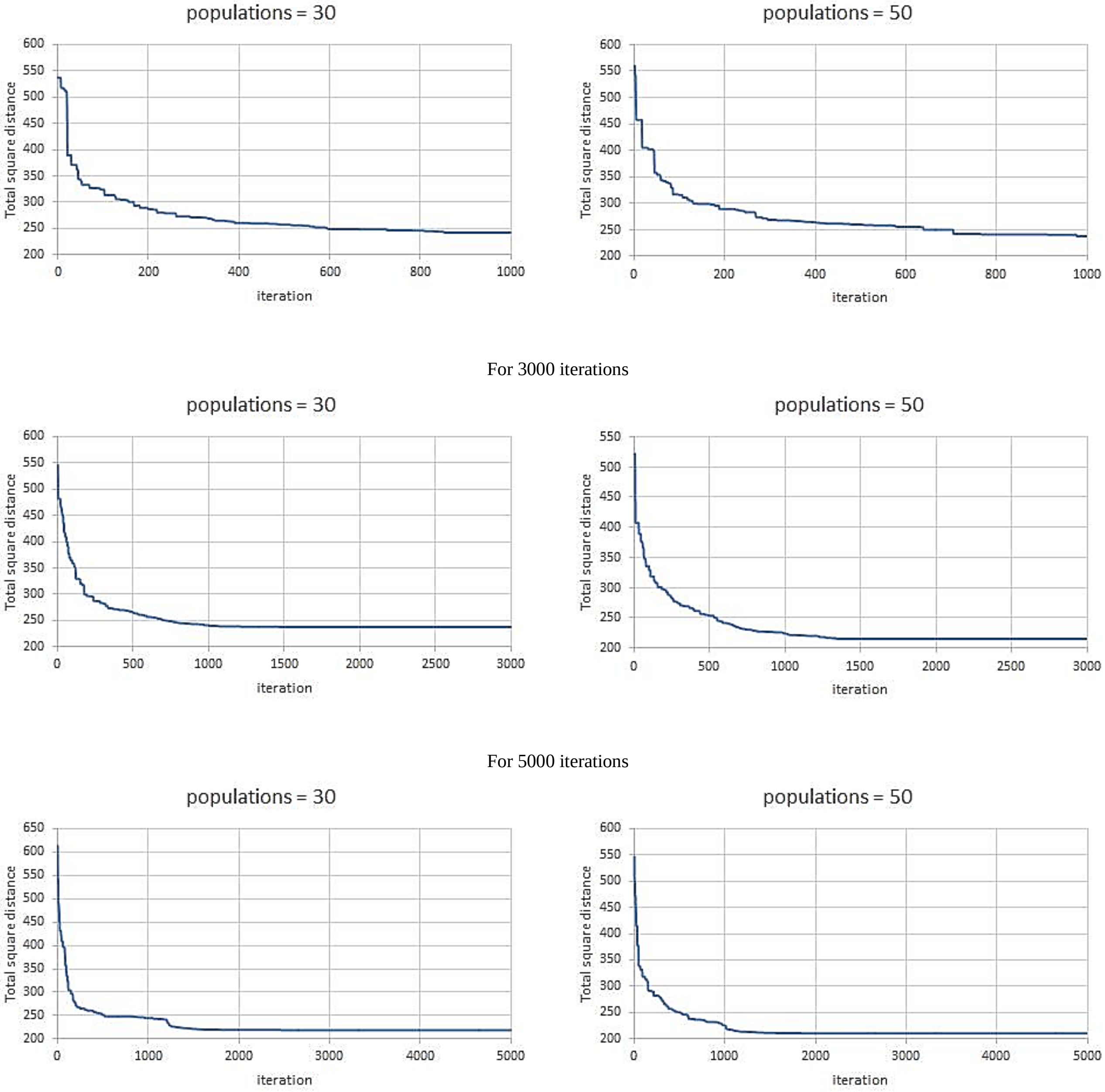
For 1000 iterations



For 3000 iterations

For 5000 iterations

Figure 8: Convergence of the total square distance for Diabets data set

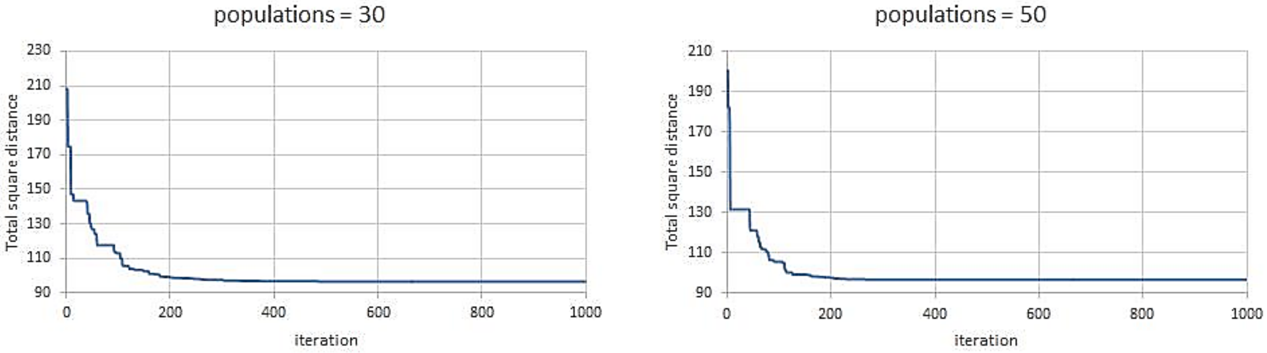

For 1000 iterations

For 3000 iterations

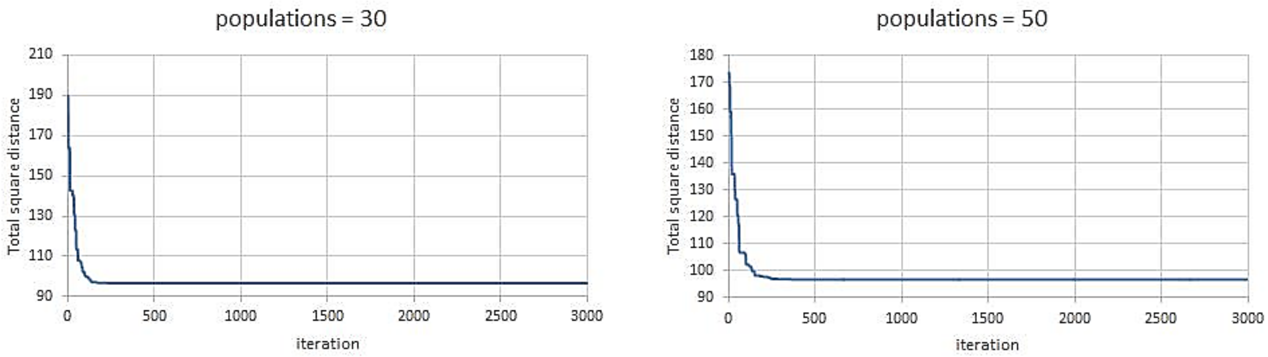
For 5000 iterations

Figure 9: Convergence of the total square distance for Glass data set

For 1000 iterations



For 3000 iterations



For 5000 iterations

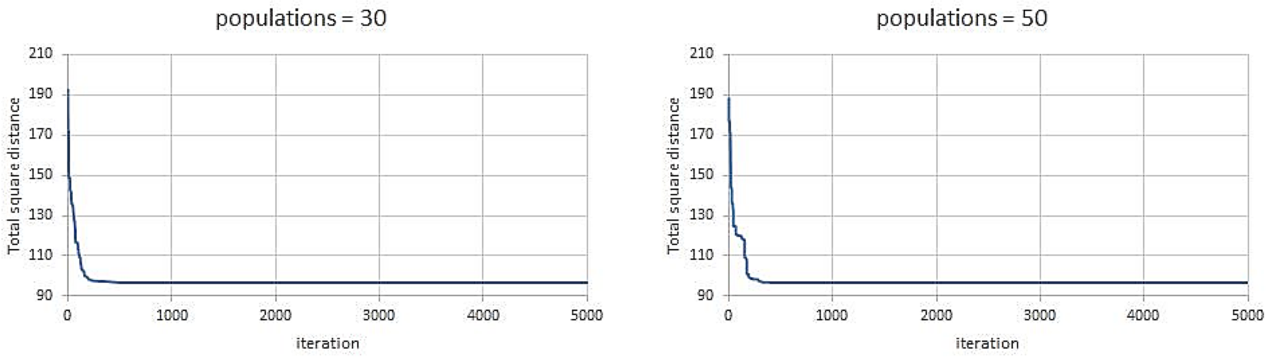
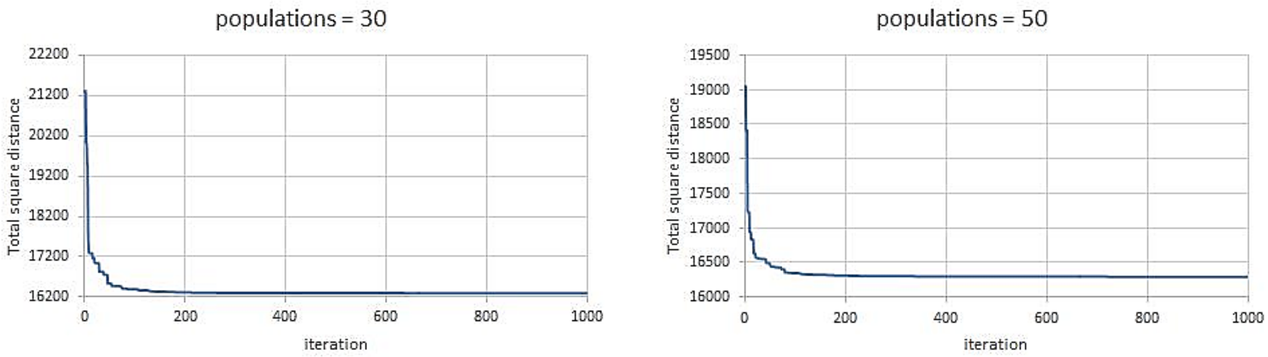
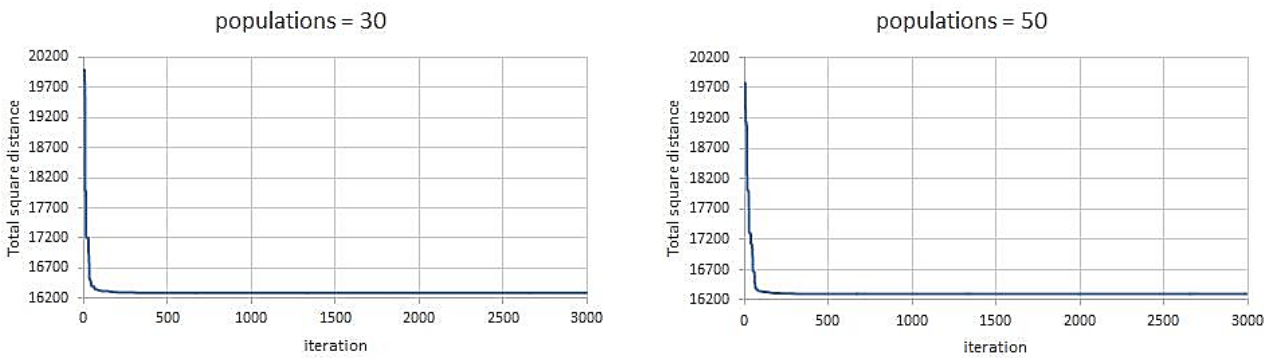


Figure 10: Convergence of the total square distance for İRİS data set

For 1000 iterations



For 3000 iterations



For 5000 iterations

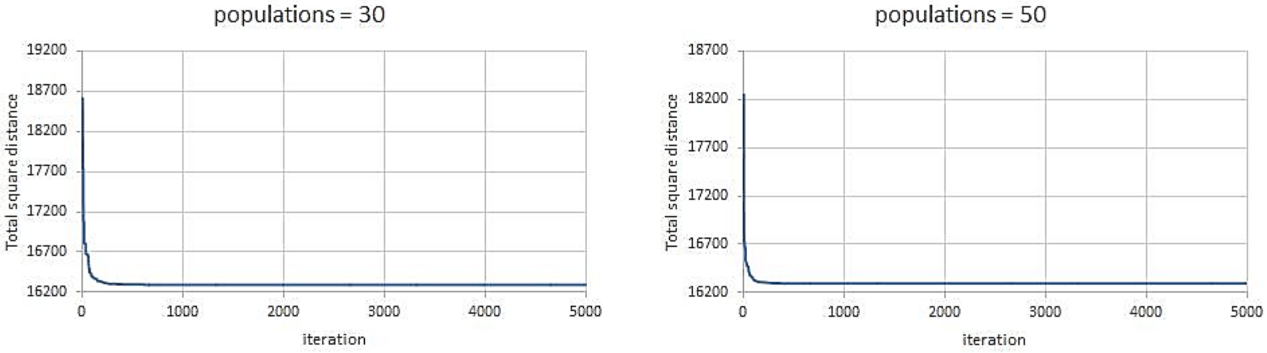
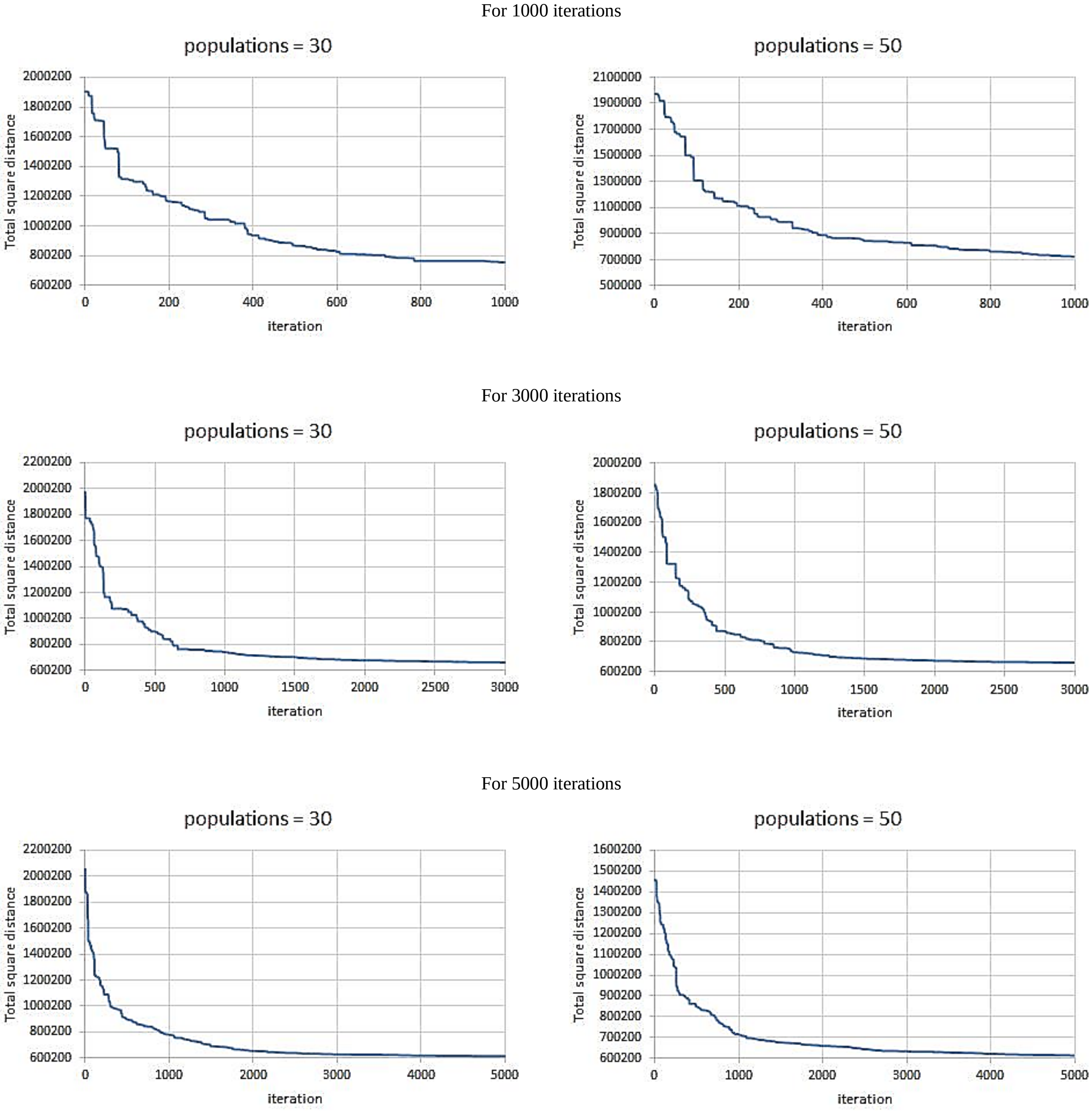


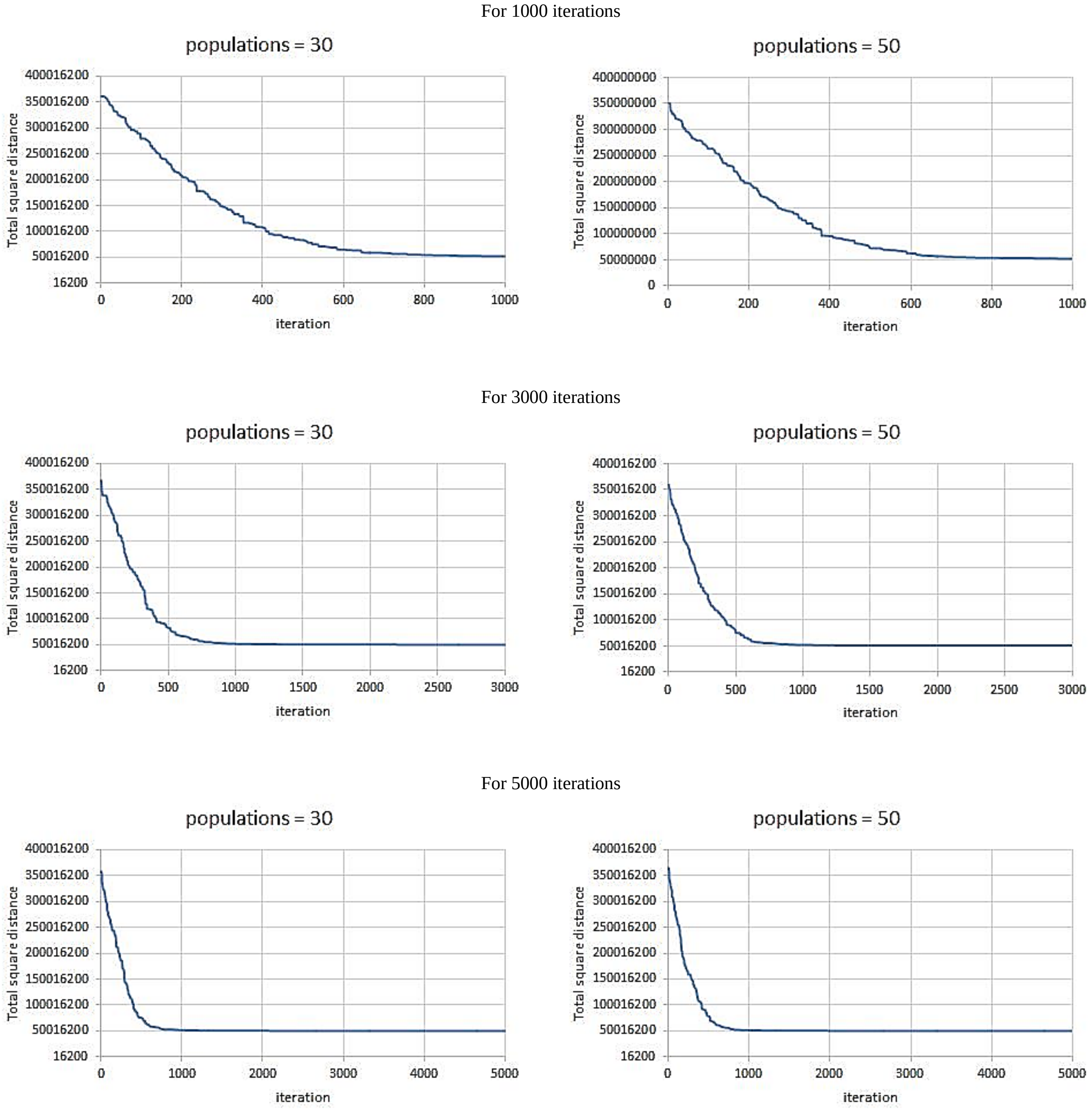
Figure 11: Convergence of the total square distance for Wine data set


For 1000 iterations

For 3000 iterations

For 5000 iterations

Figure 12: Convergence of the total square distance for Urban Land Cover data set


For 1000 iterations

For 3000 iterations

For 5000 iterations

Figure 13: Convergence of the total square distance for Hill Valley data set

8. References

[1] Han, J., Pei, J. ve Kamber, M., 2011, Data mining: concepts and techniques, Elsevier, p.

[2] Hu, T., Chen, H., Huang, L. ve Zhu, X., 2012, A survey of mass data mining based on cloud-computing, *Anti-counterfeiting, Security, and Identification*, 1-4.

[3] Liu, Y., Li, Z., Xiong, H., Gao, X. ve Wu, J., 2010, Understanding of internal clustering validation measures, *2010 IEEE International Conference on Data Mining*, 911-916.

[4] Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d. F. ve Rodrigues, F. A., 2019, Clustering algorithms: A comparative approach, *PloS one*, 14 (1).

[5] Ansari, S., Chetlur, S., Prabhu, S., Kini, G. N., Hegde, G. ve Hyder, Y., 2013, An overview of clustering analysis techniques used in data mining, *International Journal of Emerging Technology and Advanced Engineering*, 3 (12), 284-286.

[6] Haralick, R. M. ve Shapiro, L. G., 1985, Image segmentation techniques, *Computer vision, graphics, and image processing*, 29 (1), 100-132.

[7] Chuang, K.-S., Tzeng, H.-L., Chen, S., Wu, J. ve Chen, T.-J., 2006, Fuzzy c-means clustering with spatial information for image segmentation, *computerized medical imaging and graphics*, 30 (1), 9-15.

[8] Poli, R., Kennedy, J. ve Blackwell, T., 2007, Particle swarm optimization, *Swarm intelligence*, 1 (1), 33-57.

[9] Karaboga, D. ve Basturk, B., 2008, On the performance of artificial bee colony (ABC) algorithm, *Applied soft computing*, 8 (1), 687-697.

[10] Baghmisheh, M. V., Madani, K. ve Navarbaf, A., 2011, A discrete shuffled frog optimization algorithm, *Artificial Intelligence Review*, 36 (4), 267.

[11] Yang, X. S. ve Gandomi, A. H., 2012, Bat algorithm: a novel

approach for global engineering optimization, *Engineering computations*.

[12] Mou, C., Qing-xian, W. ve Chang-sheng, J., 2008, A modified ant optimization algorithm for path planning of UCAV, *Applied soft computing*, 8 (4), 1712-1718.

[13] Whitley, D., 1994, A genetic algorithm tutorial, *Statistics and computing*, 4 (2), 65-85.

[14] Nasiri, J. ve Khiyabani, F. M., 2018, A whale optimization algorithm (WOA) approach for clustering, *Cogent Mathematics & Statistics*, 5 (1), 1483565.

[15] Van der Merwe, D. ve Engelbrecht, A. P., 2003, Data clustering using particle swarm optimization, *The 2003 Congress on Evolutionary Computation, 2003. CEC'03.*, 215-220.

[16] Karaboga, D. ve Ozturk, C., 2011, A novel clustering approach: Artificial Bee Colony (ABC) algorithm, *Applied soft computing*, 11 (1), 652-657.

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

[18] Zhang, C., Ouyang, D. ve Ning, J., 2010, An artificial bee colony approach for clustering, *Expert Systems with Applications*, 37 (7), 4761-4767.

[19] Balavand, A., Kashan, A. H. ve Saghaei, A., 2018, Automatic clustering based on Crow Search Algorithm-Kmeans (CSA-Kmeans) and Data Envelopment Analysis (DEA), *International Journal of Computational Intelligence Systems*, 11 (1), 1322-1337.

[20] Karakoyun, M., 2015, Kurbağa sıçrama algoritmasının kümeleme problemlerine uygulanması, *Selçuk Üniversitesi Fen Bilimleri Enstitüsü*.

[21] Omran, M. G., Salman, A. ve Engelbrecht, A. P., 2006, Dynamic clustering using particle swarm optimization with application in image segmentation, *Pattern Analysis and Applications*, 8 (4), 332.

[22] Mining, W. I. D., 2006, Data mining: Concepts and techniques, *Morgan Kaufinann*, 10, 559-569.

[23] Gulati, H. ve Singh, P., 2015, Clustering techniques in data mining: A comparison, *2015 2nd international conference on computing for sustainable global development (INDIACom)*, 410-415.

[24] Shahbaba, M. ve Beheshti, S., 2014, MACE-means clustering, *Signal processing*, 105, 216-225.

[25] Aliguliyev, R. M., 2009, Clustering of document collection–a weighting approach, *Expert Systems with Applications*, 36 (4), 7904-7916.

[26] Asuncion, A. ve Newman, D., 2007, UCI machine learning repository.

[27] Dhanachandra, N., Manglem, K. ve Chanu, Y. J., 2015, Image segmentation using K-means clustering algorithm and subtractive clustering algorithm, *Procedia Computer Science*, 54, 764-771.

[28] Paz, A. M., Gerardo, B. D. ve Tanguilig III, B. T., 2014, Development of college completion model based on K-Means clustering algorithm, *International Journal of Computer and Communication Engineering*, 3 (3), 172.

[29] Servi, T., 2009, Çok Değişkenli Karma Dağılım Modeline Dayalı Kümeleme Analizi, *Yayımlanmamış Doktora Tezi, Çukurova Üniversitesi, Adana*.

[30] Uymaz, S. A., Tezel, G. ve Yel, E., 2015, Artificial algae algorithm (AAA) for nonlinear global optimization, Applied soft computing, 31, 153-171.