Journal of Advances in Computer Research
Quarterly ISSN: 2008-6148
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 11, No. 4, November 2020), Pages: 1-22
www.jacr.iausari.ac.ir

An improved opposition-based Crow Search Algorithm for Data Clustering

Rogayyeh Jafari Jabal Kandi, Farhad Soleimanian Gharehchopogh✉
Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran
engineer_gafary@yahoo.com; bonab.farhad@gmail.com

Received: 2020/07/13; Accepted: 2020/09/21

Abstract

Data clustering is an ideal way of working with a huge amount of data and looking for a structure in the dataset. In other words, clustering is the classification of the same data; the similarity among the data in a cluster is maximum and the similarity among the data in the different clusters is minimal. The innovation of this paper is a clustering method based on the Crow Search Algorithm (CSA) and Opposition-based Learning (OBL). The CSA is one of the meat-heuristic algorithms that is difficult at the exploration and exploitation stage, and thus, the clustering problem is susceptible to initialization for centrality of the clusters. In the proposed model, the crows change their position based on the OBL method. The position of the crows is updated using OBL to find the best position for the cluster. To evaluate the performance of the proposed model, the experiments were performed on 8 datasets from the UCI repository and compared with seven different clustering algorithms. The results show that the proposed model is more accurate, more efficient, and more robust than other clustering algorithms. Also, the convergence of the proposed model is better than other algorithms.

Keywords: *Data Clustering; Crow Search Algorithm; Opposition-based Learning; Centrality*

1. Introduction

Clustering is one of the most important methods in data mining that deals with the analysis of unlabeled data [1-3]. Clustering involves the process of clustering the unlabeled datasets into similar clusters. Each cluster includes objects that are similar in clusters and to objects of other clusters are dissimilar [4, 5]. Clustering has been used in many applications such as web, text mining, image processing, stock prediction, signal processing, biology and other fields of science and engineering[6]. Most clustering algorithms face challenges such as the problem of determining the center of clusters, the low accuracy of clustering, inappropriate clustering efficiency of the various datasets and dependence on tangible parameters.

Clustering is including a large-scale search space. Therefore, due to the enormous computing time for data segmentation, we need to apply precise methods. Meta-heuristic algorithms are widely used as an alternative approach to solving high complexity optimization problems[7-10]. The main problem with these algorithms is the trapped of local optimization due to weak searches and loss of good responses due to poor exploitation. Therefore, in order to achieve quality answers at an acceptable time, both processes need to be optimized at a time. An effective strategy is to change the control functions of these algorithms [11-14].

The aim of this paper is to develop a new model for clustering data-based CSA and OBL. The CSA is one of the newest methods for solving optimization problems simulated based on the behavior of the crow in nature. The CSA is presented by Askarzadeh in 2016 [15]. The CSA is a population-based approach that based on this idea works that crows store their extra food in secret places and retrieve the food when needed. The CSA has been optimized to optimize six issues, which indicate that the algorithm is a powerful for solving the optimization problem. The CSA have been widely used in various engineering areas due to advantages such as relatively simple structure, low control parameters, optimized search space and high ability to avoid local optimization. But this algorithm suffers from disadvantages such as low convergence rate and poor exploitation [16]. These two problems often depend on the diversity of the population. A high diversity can guarantee the optimal solution, but also slow convergence. On the other hand, low diversity leads to rapid convergence, but sacrifices against global optimal assurance. Therefore, a good balance between convergence and accuracy should be created.

In this paper, a relatively new meta-heuristic algorithm called the CSA [15] is optimized with the use of OBL [17]. One of the ways to overcome the weaknesses of the CSA is to use the combination techniques. Initial population plays a very important role in the performance of meta-heuristic algorithms. The main contributions of this paper are as follows:

- An Improved Crow Search Algorithm is proposed for Data Clustering
- A method of generating primary population based on learning opposition is presented.
- Creating a balance between exploration and exploitation.
- OBL has been used to accelerate convergence.
- Opposition Based Learning was used with to improve its population diversity.
- This work aims to increase the diversity of solutions and local exploitation of search space in CSA.

CSA has the number of properties such as simplicity and flexibility. However, CSA as other optimization algorithms suffer from some problems such population diversity and local optima. These mentioned reasons and characteristics motivated our study to improve CSA. In proposed model, OBL is used at initialization phase of CSA to improve its population diversity in the search space. In order to evaluate the proposed model, the data presented in the UCI Machine Learning Dataset is used. Comparison with other models proves the efficacy and reliability of the proposed model.

The rest of the study is organized as follows: Section 2 provides the review of the literature. Section 3, introduces the proposed model. In Section 4, the proposed model is analyzed and compared to other models. In Section 5, conclusion is provided.

2. Related Works

Nowadays, finding useful patterns in a large dataset is great interest, and one of important issues is to identify areas of a densely population in a multidimensional dataset called clustering. In this regard, clustering is one of the methods and algorithms that researchers have taken into account in analyzing data and discovering patterns. So far, many studies have been proposed focusing on the application of hybrid clustering based on meta-heuristic algorithms, each of which leads to increasing the strength, stability, accuracy and efficiency of the clustered process.

Chuang et al. [18] have proposed a method for clustering based on the Particle Swarm Optimization (PSO) algorithm. An algorithm based on PSO Gaussian chaos and k-means. In this algorithm, the Gaussian function for the initial population and the update of the velocity and particle position are used. The fitting value of each particle is calculated according to Eq. (1). The fit value is the sum of the total intra-clustering distance of all clusters. The total distance has a profound effect on the error rate. In Eq. (1), k and n are the number of clusters and the number of datasets, respectively. The parameter Z_i is the center of cluster i and Xj of the j data points. The objective of Eq. (1) is to minimize the sum of squared error (SSE) for each particle. In each step, according to Eq. (1), the value of fitness of each particle, which is equal to the SSE relate to k to the center of the cluster. If the value of Eq. (1) is lower, indicative the closeness of the data to the cluster centers and that particle has more fit.

$$fitness = \sum \left| X_j - Z_i \right|, \ i = 1, \dots, k \ , j = 1, \dots, n \tag{1}$$

The evaluation results have done on vowel, Iris, Crude Oil, CMC, Cancer and Wine datasets. The results show that the PSO algorithm based on the Gaussian chaos pattern compared to the k-means, the Genetic Algorithm (GA), the PSO algorithm has a lower error rate.

Wan et al. [19] have proposed a model for clustering data based on chaotic ant swarm (CAS). The proposed algorithm works to achieve optimal clustering based on minimizing the target function. This algorithm has achieved three results. Finding an optimal global solution, the lack of sensitivity of the clusters is appropriate to the size of the dataset and the volume of data and for the multidimensional datasets. One of the most commonly used evaluation functions used for clustering is SSE. The SSE is the total sum of the distances of all sample datasets with the cluster. The formula for calculating the SSE evaluation function is according to Eq. (2). In Eq. (2) k is the number of clusters and kj is the center of the jth cluster.

$$SSE = \sum_{i=1}^{k} \sum_{x_i \in k_j} \left\| x_i - k_j \right\|^2 \tag{2}$$

The evaluation was performed on the 2D-4C dataset (1572 samples), 10D-4C (1289 samples), Iris (150 samples), Wine (178 samples), and Glass (214 samples). The results show that the SSE in CAS is lower than the k-means and the PSO.

Saida et al. [20] have proposed a meta-heuristic method for data clustering based on Cuckoo Search Optimization (CSO) algorithm to avoid k-means incompatibility. The main features of CSO is that it is easy to implement and has good computing performance. The experiment was conducted on four datasets (Iris, Wine, Cancer and Vowel) from the UCI Machine Learning dataset. The results show that the CSO is more versatile than k-means, PSO, Gravitational Search Algorithm (GSA), and Black Hole (BH) algorithm.

In [21], researchers used Social Spider Optimization (SSO) algorithm to overcome the problems of k-means. The k-means algorithm is one of the most popular and most important clustering algorithms for its simplicity and ease in execution. However, its function is strongly dependent on the cluster's initial centers and can be close to the local minimum and was converged. To overcome these problems, many researchers have tried to solve the clustering problem using the SSO. So that when the dimensions of a search space and the available data increase, the problem of local optimization and poor convergence rates persist, the effectiveness of these algorithms seem unacceptable. This paper presents an easy-to-use SSO algorithm to overcome the aforementioned drawbacks. The simple method is a possible different strategy, increasing the diversity of the population, while the local search capabilities of the algorithm are also increasing. Using the proposed model in a data clustering problem using 11 datasets from the UCI database confirms the potential power and efficiency of the proposed model. The experimental results showed that the proposed model performs better in terms of accuracy, power and convergence speed than other algorithms.

In [22], a new PSO model is proposed based on the density for data clustering, the proposed model is described to cover the weakness of early convergence and to fine-tune the parameters of the PSO in the clustering problem. The proposed model is compared with other methods based on 11 datasets from the UCI machine learning dataset, which results shows the proposed model superiority to other methods.

One of the most popular and most used clustering algorithms is k-means. Unfortunately, this algorithm is dependent on the initial values of the cluster centers, and therefore does not always perform clustering correctly. One of the best and most widely used methods for eliminating the defects of this algorithm is meta-heuristic and evolutionary algorithms. In [23], a hybrid algorithm based on PSO and Teaching-Learning-Based Optimization (TLBO) algorithm for clustering is proposed. By applying the combined algorithm on the various datasets, optimal solutions are obtained from other methods. So that in terms of speed and achievement, the combined model has a higher superiority than other algorithms.

A new approach to data clustering is proposed using the combination of Harmony Search Algorithm (HAS) and the Simulated Annealing (SA) [24]. The hybrid model steps are as follows: 1) Harmonic memory is filled with random answers and the value of the fitness function for each row is calculated. 2) Create a new harmony for each harmonic memory row and calculate the value of the function. 3) Replacing the new solution with the worst harmonics in the harmonic memory. 4) Send the best current generation harmony for the SA algorithm. 5) If the best solution returned from the SA algorithm is better than the best current generation harmonic, then it will replace the best harmonic. If the condition is not satisfied, the algorithm's steps are repeated. After completing the implementation of the algorithm, the row of the harmonic memory, which has the best value in the function, is returned as the best available solution in memory. To evaluate the efficiency of the UCI dataset has used. The results show that the proposed model is better than HSA, PSO and GA, and it has lower distance compared with other methods.

A new hybrid method has been proposed for data clustering using Ant Lion Optimization (ALO) algorithm and k-means algorithm [25]. In the proposed method, the optimal cluster centers are firstly identified using the ALO algorithm and then the initialization of the k-means algorithm is determined with the improved center. The results on the various datasets indicate that the hybrid model is less distant compared to the k-means algorithm, the PSO, and other models. Table 1 shows a comparison of the proposed models for data clustering.

Table 1. Comparison of the Proposed Models for Data Clustering

Refs	Models	Improved	Datasets	Advantages	Disadvantages
[8]	PSO	Gaussian chaos and k-means	vowel, Iris, Crude Oil, CMC, Cancer and Wine	*Improve Exploration *Improve the searching capability of PSO using k-means	$Gbest$ gets struck with the local minima and optimization is not considered.
[9]	Chaotic Ant Swarm (CAS)	Chaotic	2D-4C, 10D-4C, Iris, Wine, and Glass	*Balance between exploration and exploitation	*Performance of the algorithm is not compared with other meta-heuristic or state-of-art algorithm
[10]	Cuckoo Search Optimization (CSO)	k-means	Iris, Wine, Cancer and Vowel	*good performance *Speed up the execution of applications	*Increasing time complexity
[11]	Social Spider Optimization (SSO)	-	Iris, Wine, Cancer and Vowel	*Increasing accuracy *Increasing power *convergence speed	Algorithm doesn't discuss trade-off solution between time and energy cost
[12]	PSO	-	Iris, Wine, Cancer and Vowel	*Reduce the execution time and improve the utilization ratio.	*Premature convergence
[13]	PSO and Teaching-Learning-Based Optimization (TLBO)	k-means	vowel, Iris, CMC, Cancer and Wine	*Increase efficiency *High convergence speed *Balance between exploration and exploitation	*Complexity of the algorithm is high
[14]	Harmony Search Algorithm (HAS) and the Simulated Annealing (SA)	-	vowel, Iris, CMC, Cancer and Wine	*Balance between exploration and exploitation	*Premature convergence
[15]	Ant Lion Optimization (ALO)	k-means	various datasets	*High convergence speed *Balance between exploration and exploitation	*Increasing time complexity

3. Proposed Model

The proposed model is a combination of CSA and OBL. Clustering is a multivariate data analysis model. Selection of data points as cluster centers plays an important role in the clustering process. In this paper, OBL is used to improve solution vectors in the CSA. In this way, the CSA first selects the optimal points as cluster centers, and then, in updating new situations, an OBL method is used to replace the new centers. In Figure 1, the flowchart of proposed model is shown.

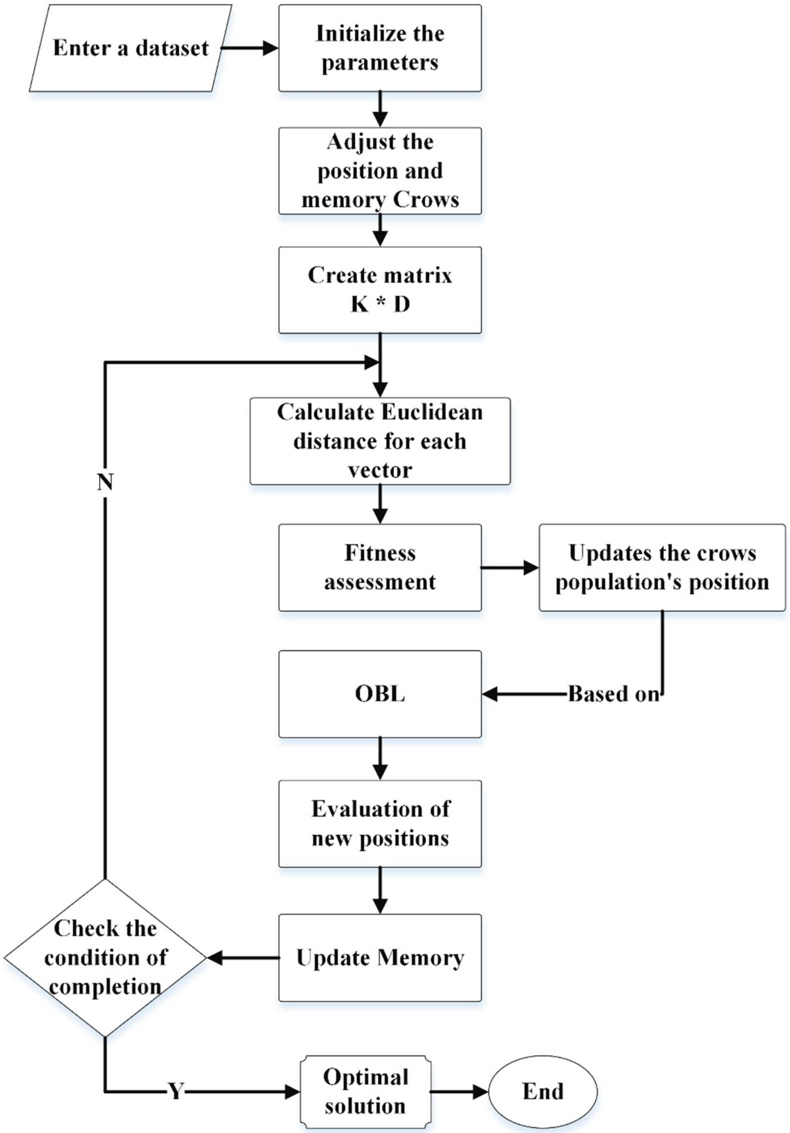
Enter a dataset
Initialize the parameters
Adjust the position and memory Crows
Create matrix K * D
Calculate Euclidean distance for each vector
N
Fitness assessment
Updates the crows population's position
OBL
Based on
Evaluation of new positions
Check the condition of completion
Update Memory
Y
Optimal solution
End

Figure 1. Flowchart of proposed model

3.1 Crow Search Algorithm

The process of implementing the CSA is as follows:

1) Initially, decision variables and constraints are defined. Number of crows (N), flight length (fl), maximum repeat and awareness probability (AP).

2) In the second step, the memory and population of the algorithm are determined. Each crow is a logical answer, and d is the number of decision variables. N crows accidentally fall into a d-dimensional search space as members of the population. Each crow represents a solvable solution of the problem, and d is the number of decision variables. The memory of each crow is initialized. Because in the early position, the crows do not have experience, it is assumed that they hid their foods in their early situations [15].

$$Crows = \begin{bmatrix} x_1^1 & x_2^1 & \ldots & x_d^1 \\ x_1^2 & x_2^2 & \ldots & x_d^2 \\ \vdots & \vdots & \vdots & \vdots \\ x_1^N & x_2^N & \cdots & x_d^N \end{bmatrix} \tag{3}$$

$$memory = \begin{bmatrix} m_1^1 & m_2^1 & \ldots & m_d^1 \\ m_1^2 & m_2^2 & \ldots & m_d^2 \\ \vdots & \vdots & \vdots & \vdots \\ m_1^N & m_2^N & \cdots & m_d^N \end{bmatrix} \tag{4}$$

3) Production of a new position: The new position of the crows in the search space is as follows: it is assumed ith crow create a new position. For this purpose, this crow randomly selects one crow population (for example, the crow j) and by following it then discovers the position of the food hid by the crow (m_j). The new position crow i is calculated by Eq. (5). This process is repeated for all crows. In Eq. (5) r_j is a random number with uniform distribution between 0 and 1, and $AP^{i,iter}$ is the probability of awareness of the jth crow at each step of the repetition. In the algorithm, diversity and resonance are controlled by the *AP* parameter. As small amounts of *AP* result in increased resonance and large amounts of *AP*, diversity increases [15].

$$x^{i,iter+1} = \begin{cases} x^{i,iter} + r_i \times fl^{i,iter} \times (m^{j,iter} - x^{i,iter}) & r_j \geq AP^{j,iter} \\ a\ random\ position & otherwise \end{cases} \tag{5}$$

4) In the fourth step, the objective function is evaluated (Euclidean distance) and the values of the objective function for each crow are calculated.

5) The memory of the crows is updated according to Eq. (6) [15].

$$m^{i,iter+1} = \begin{cases} x^{i,iter+1} \leftarrow f\left(x^{i,iter}\right) is (better) than\ (f\left(m^{i,iter}\right)) \\ m^{i,iter} \leftarrow otherwise \end{cases} \tag{6}$$

6) At this stage, the condition of convergence is controlled, and if it is satisfactory, the algorithm will be completed.

3.2 Opposition-based Learning

The OBL is a new method in machine intelligence that is widely used in optimization, neural networks and reinforcement learning. In dealing with optimization issues, the OBL method uses opposite numbers to search for the optimal point. Assume that $X = (x_1, x_2, \ldots, x_n)$ is a search point in the n-dimensional space and $x_i = [a_i, b_i]$

such that $i = 1,2, \dots, n$. The opposite number $X^* = (x_1^*, x_2^*, \dots, x_n^*)$ is calculated as the Eq. (7).

$$x_i^* = a_i + b_i - x_i \quad i = 1,2, \dots, n \tag{7}$$

The principle of OBL method in optimization based on X and X* to search and find optimal answers in such a way that in each iteration X* is calculated from X and then f (X) and f (X*) as the value of fitness X and X* are calculated respectively. In iterations where f (X) ≥f (X*), X is considered as the response vector.

So far, various algorithms have been able to report better than their original version by using OBL, including Grasshopper Optimization Algorithm [26], Monarch Butterfly Optimization [27], Gray Wolf Optimization Algorithm [28], Sine and Cosine Optimization Algorithm [29], Differential Evolution Algorithm [30], Shuffled Frog Leaping Algorithm [31], Krill Herd Algorithm [32], Harmony Search Algorithm [33], and Scatter Search Algorithm [34].

3.3 Innovation

3.3.1 Preprocessing

The preprocessing step is used to unify the values of the data set properties. If the dataset is not specified in a range, it reduces accuracy. In the proposed model, based on Eq. (8), we perform preprocessing and convert the properties values to 0 and 1. In Eq. (8), x is the value of the selected features, and min and max are respectively the least and most values of the features.

$$x_i = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{8}$$

3.3.2 Generating Opposition Solutions for Population

Due to the lack of discover of new positions, the optimal responses found in the search cycle may not be optimal, and if the position of the crows is not updated, it will mislead the population and move the crows to undesirable responses. Therefore, in the proposed model, instead of using the predetermined interval, the minimum and maximum values of each decision variable (cluster center) are used in the current population. After updating the crows, the best crow in each vector is considered as an opposite solution, and initializing each variable of the decision is made to improve performance by this model. Each vector has the lowest and highest values, and based on them, the probability of the best center for the cluster j^{th} is discovered.

$$\tilde{x}_{ij} \leftarrow x_{minj}^{z} + x_{maxj}^{z} - x_j \tag{9}$$

In Eq. (9) $maxj$ and $minj$ are respectively the maximum and minimum values of the decision variable (dimensions) j^{th} in the current population *Z*. After the end of this phase, the fitness of the generated answers is calculated and with the greedy selection method, better solutions are selected. In Eq. (9), diversity is determined by the x position. Then, based on $maxj$ and $minj$, and position *x*, estimated the value of the contradiction x. In the OBL approach, there is a type of learning and an estimate is obtained based on available values. In contrast, in a random system, the current state of the system is independent of its previous state. In the randomized model, it is not possible to estimate the next situation. Usually in a mathematical model, the set of variables is represented by the vector $\mathrm{x} = [\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_n]$, where n is the number of variables. The main goal is to achieve the best possible set of values in relation to clustering. This better set is called the optimal variable, usually $\mathrm{x} = [x_1^*, x_2^*, \dots, x_n^*]$.

In the first step, firstly, the crows are randomly initialized. The structure of the crows is according to Eq. (10). Eq. (10) explanatory the creation of the first crows consisting of the k center of the cluster. In the proposed model, the number of clusters is not initially determined. Therefore, the calculation of vectors must be such that the number of clusters is calculated during the clustering of the data. Therefore, assuming that the maximum number of clusters is equal to k and the number of dimensions of the dataset is equal with *d*, then each possible vector is a matrix equal with k × (d + 1) because, in order to obtain the optimal number of clusters, one the decision variable is added to the matrix vector for clustering. In other words, each center of the cluster, which has d-dimension, contains a variable called the decision variable in order to determine the optimality of the cluster. Therefore, each crow has k × (d + 1) dimension. Valid values for the decision variable are real numbers in the interval (0, 1). The center of a cluster is chosen if the value of the decision variable is larger than 0.5. Otherwise the center of the cluster will not be selected. In fact, 0.5 is the threshold for choosing a cluster. The objective function to be minimized is Euclidean distance between points. Then for each of the crows, the direction to move is randomly assigned. This step is repeated until all the initial population is created. Each of the crows is considered as a possible answer to the problem.

$$(Z_{1,1}, Z_{1,2}, \dots, Z_{1,D}, Z_{2,1}, Z_{2,2}, \dots, Z_{2,D}, \dots Z_{k,1}, Z_{k,2}, \dots, Z_{k,D}) \qquad (10)$$

In the CSA, the problem space is modeled as data vectors in multi-dimensional spaces. A crow in the group shows a possible solution for clustering the dataset. Let's consider a set of M crows $X = (X_1, X_2, \dots, X_M)$, so that each M crow is a d-dimensional vector. Each crow is represented by a matrix $X_i = (Z_1, Z_2, \dots., Z_j, \dots, Z_k)$, so that Zj is a vector of the center of cluster j^{th}, and the number of clusters is determined by *k*. The crows update the positions of the center of cluster in each repetition in accordance with their own knowledge and experience and neighboring crows. The amount of fitness is the sum of the intra-clustering distance of all clusters. The sum of intervals is small when high cluster results are observed. These total distances have a profound effect on the amount of error.

To measure the overall quality of the clusters must be used the evaluation function. One of the most commonly evaluation function is SSE function that for clustering used in this paper. The SSE is the total sum of the distances of all sample datasets with the cluster center in which they are located. The SSE function is defined by Eq. (11). In Eq. (11) k is the number of clusters and zj is the center of the j^{th} cluster [35].

$$SSE = \sum_{i=1}^{k} \sum_{x_i \in z_j} \left\| x_i - z_j \right\|^2 \qquad (11)$$

Eq. (12) is used to calculate standard deviations for clusters. The parameter c represents the number of clusters and Vi is equal to the center of i^{th} cluster. If the standard deviation of a set of data to be close to 0, it indicates that the data are close to each other and have little dispersion, while the large deviation represents a significant dispersion of the data.

$$std = \frac{1}{c}\sqrt{\sum_{i=1}^{n_c} \|\sigma(V_i)\|} \qquad (12)$$

The error rate is equal to the number of false examples, divided by the total number of samples. The error rate is calculated according to Eq. (13) [36].

$$\text{ER} = \left(\sum_{i=1}^{n}(if\,(A_i = B_i)\ then\ 0\ else\ 1) \div n\right) \times 100 \qquad (13)$$

In Eq. (13), n represents the total number of samples, and Ai and Bi represent the data sets with i^{th} point of that member before and after clustering.

4. Evaluation and Results

In this section, details of experiments and empirical results obtained from algorithms in the VC#.NET 2017 are presented. According to the sensitivity analysis of the algorithm parameters, the best number of populations for the crows is 30, and the best value of the fl and *AP* parameters is 2 and 0.3 respectively. In this paper, validating and evaluating the proposed model and comparing its performance with other methods has been done. The proposed model on eight UCI [37] standard datasets such as Glass, Vowel, CMC, Iris, Wine, Wisconsin, Seeds and Heart was assessed. Each dataset has a number of different clusters for a number of samples that share a common set of features. Table 2 shows a summary of the properties of this data set.

Table 2. Standard datasets from UCI

No	Datasets	No. Features	No. Class	Size	Missed value
1	Glass	10	6	214	N
2	Vowel	3	6	871	N
3	CMC	10	3	1473	N
4	Iris	4	3	150	N
5	Wine	13	3	178	N
6	Wisconsin breast-cancer	10	2	699(683)	16
7	Seeds	7	3	210	N
8	Heart	13	2	270	N

In Table 3 and 4 the comparison of the proposed model with other models is shown on datasets of different based on the objective function. In reviewing the results, the best cost is considered as the main criterion for evaluation. The proposed model is repeated 100 times on each dataset. Based on the difference between 100 answers obtained from the first repetition of the method and the last iteration, the standard deviation is calculated and shows the speed and accuracy of the convergence of the proposed model. This means that if standard deviation to be lower, this indicates that the proposed model converges to a close and identical answer at each run. It is clear that the solution obtained from the proposed model has the best possible cost among several models and has the least standard deviation.

Table 3. Comparison of the proposed model with GSA, PSO and HBMO on different datasets based on quality

datasets	criterion	Proposed Model	GSA [38]	PSO [38, 39]	HBMO [38]
Glass	Best	201.16	220.78	270.57	245.73
	Mean	201.24	225.70	275.71	247.71
	Worst	201.53	229.45	283.52	249.54
	Std.	1.1214	3.4008	4.5571	2.4381
Vowel	Best	148976.0196	-	148976.0152	149201.6320
	Mean	148976.2084	-	148999.8251	161431.0431
	Worst	148976.4580	-	149121.1834	165804.6710
	Std.	8.5309	-	28.8134	2746.0416
CMC	Best	5695.1195	5698.1500	5700.9853	5699.2670
	Mean	5695.1249	5699.8400	5820.9647	5713.9800
	Worst	5695.1542	5702.0900	5923.2490	5725.3500
	Std.	0.2581	1.7240	46.9501	12.6900
Iris	Best	96.6512	96.6980	96.8942	96.7520
	Mean	96.6507	96.7230	97.2328	96.9531
	Worst	96.6542	96.7640	97.8973	97.7576
	Std.	0.0096	0.0123	0.3480	0.5310
Wine	Best	16295.2901	16315.3501	16345.9670	16357.2843
	Mean	16295.3509	16376.6101	16417.4725	16357.2843
	Worst	16295.7216	16425.5801	16562.3180	16357.2843
	Std.	0.0356	31.3401	85.4974	-
cancer	Best	2963.31	2967.96	2973.50	2989.94
	Mean	2963.31	2973.58	3050.04	3112.42
	Worst	2963.31	2990.83	3318.88	3210.78
	Std.	0.0535	8.1731	110.8013	103.4710
Seeds	Best	311.1297	311.7980	312.6837	-
	Mean	311.1297	311.7980	313.8597	-
	Worst	311.1297	311.7980	313.8597	-
	Std.	1.0051	-	33.3095	-
Heart	Best	10622.5681	-	10622.9924	-
	Mean	10622.5681	-	10623.0776	-
	Worst	10622.5681	-	10623.7094	-
	Std.	0.0136	-	0.1711	-

Table 4. Comparison of the proposed model with GSA, PSO and HBMO on different datasets based on quality

datasets	criterion	Proposed Model	ACO [38]	SA [38]	GA [38, 40]	k-means [38]
Glass	Best	201.16	269.72	275.16	278.37	215.74
	Mean	201.24	273.46	282.19	282.32	235.50
	Worst	201.53	280.08	287.18	286.77	255.38
	Std.	1.1214	3.5848	4.2384	4.1387	12.4710
Vowel	Best	148976.0196	149395.6020	149370.4700	149513.7350	149422.2601
	Mean	148976.2084	159458.1438	161566.2810	159153.4980	159242.8901
	Worst	148976.4580	165939.8260	165986.4200	165991.6520	161236.8101
	Std.	8.5309	3485.3816	2847.8594	3105.5445	916.0000
CMC	Best	5695.1195	5701.9230	5849.0380	5705.6301	5842.2001
	Mean	5695.1249	5819.1347	5893.4823	5756.5984	5893.6001
	Worst	5695.1542	5912.4300	5966.9470	5812.6480	5934.4301
	Std.	0.2581	45.6340	50.8670	50.3690	47.1601
Iris	Best	96.6512	97.1007	97.4573	113.9865	97.3330
	Mean	96.6507	97.1715	99.9570	125.1970	106.0500
	Worst	96.6542	97.8084	102.0101	139.7782	120.4500
	Std.	0.0096	0.367	2.018	14.563	14.6311
Wine	Best	16295.2901	16530.5338	16473.4825	16530.5338	16555.68
	Mean	16295.3509	16530.5338	17521.0940	16530.5338	18061.0001
	Worst	16295.7216	16530.5338	18083.2510	16530.5338	18563.1201
	Std.	0.0356	-	753.0840	-	793.2101
cancer	Best	2963.31	2970.49	2993.45	2999.32	2999.19
	Mean	2963.31	3046.06	3239.17	3249.46	3251.21
	Worst	2963.31	3242.01	3421.95	3427.43	3521.59
	Std.	0.0535	90.5002	230.1920	229.7340	251.1401
Seeds	Best	311.1297	-	-	312.9476	587.3195
	Mean	311.1297	-	-	327.6251	588.1048
	Worst	311.1297	-	-	382.5527	589.0491
	Std.	1.0051	-	-	18.7126	-
Heart	Best	10622.5681	10654.3219	-	-	10681.4447
	Mean	10622.5681	10751.2534	-	-	10688.6493
	Worst	10622.5681	10963.9506	-	-	10700.8385
	Std.	0.0136	7.5107	-	-	8.3298

The minimum value of the Glass dataset is 201.16, which belongs to the proposed model. Also, the total distance in the GSA and k-means is greater than the other models. The best value for the Vowel dataset is 148976.0152, which belongs to the PSO algorithm. Also, the total distance in the proposed model is lower compared to other models. The minimum amount of the CMC dataset is 5695.1195, which belongs to the proposed model. Also, the total distance in the Honey Bees Mating Optimization (HBMO) algorithm and PSO algorithm are lower than other models. The minimum value of the Iris dataset is 96.6512, which belongs to the proposed model. Also, the total distance in the HBMO algorithm, the GSA and the PSO algorithm are lower than the other models. The minimum value of the Wine dataset is 16295.2901, which belongs to the proposed model. Also, the total distance in the GSA and the PSO algorithm is lower than that of other models. The smallest amount of data from the Cancer is 2963.31, which belongs to the proposed model. Also, the total distance in the GSA is lower than the other models. The minimum value of the Seed is 311.1297, which belongs to the proposed model. Also, the total distance in the GSA and the PSO algorithm is lower than that of other models. The minimum value of the Heart is 10622.5681, which belongs to the proposed model. Also, the total distance in the PSO algorithm is lower than the other models.

Based on the results of simulation in Table 3 and 4 for each dataset, the proposed model has the highest quality in providing solutions, including the best, worst, and average intra-cluster distance for dataset samples, compared to other models. The results for each dataset are better and more desirable than all other models in all cases. The low amount for the standard deviation includes the proposed model in the process of data clustering for each dataset, which can be used to find an optimally close solution in many independent implementations discovers and has a high power and capability is efficient in convergence to the optimal solution.

Figures 2 to 5 show the process of convergence of algorithms based on the calculation of the cost function and the number of repetitions. In order to have a general view on the convergence of algorithms, changes the best solutions algorithms of GA, SA, Ant Colony Optimization (ACO), HBMO, PSO algorithm, GSA and CSA in 100 iterations for different dataset have been shown. In Figures 2 to 5, the convergence of algorithms is indicated for the purpose of efficiency and search precision, so that the best answers can be compared. The CSA, from the persistent iterate point of view, gradually reduces its search range to a region, thereby increasing the speed of convergence. The proposed model is sufficiently capable of preventing trapping in the local minimum.

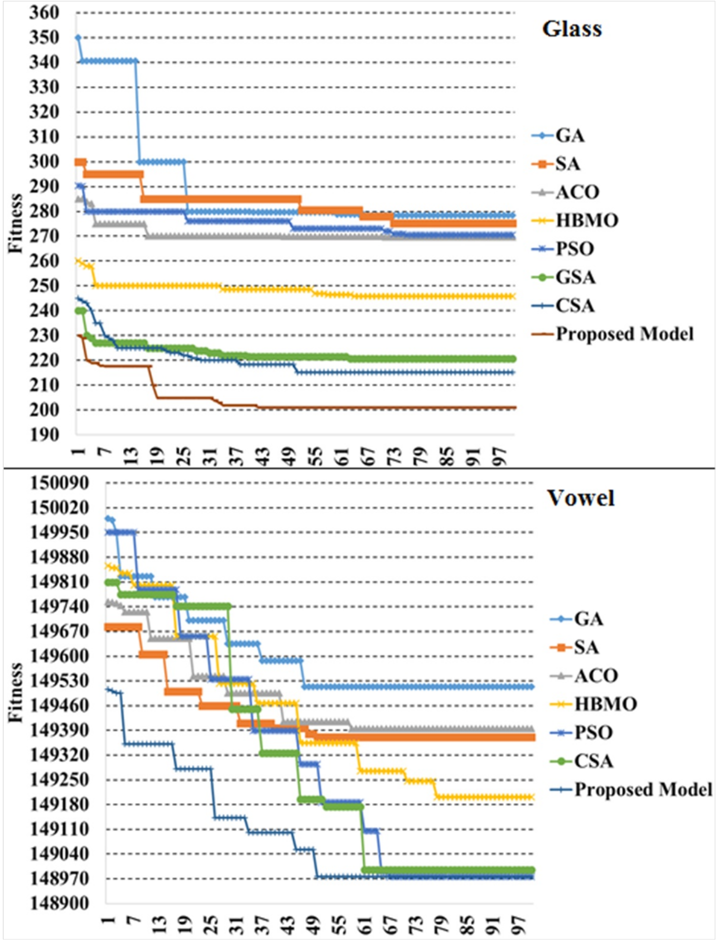

Figure 2. Display the convergence of models on the Glass and Vowel datasets

According to Figure 3, the proposed model converges in 45 iterations on CMC, but the ACO algorithm in 69 iterations, the GSA in 64 iterations and the PSO in 71 iterations; therefore, the proposed model converges in less iterations and this means that the convergence rate of the proposed model is much more acceptable than the speed of convergence of GA, SA, ACO, HBMO, PSO and GSA. In the proposed model, when the current optimal answer does not show any improvement in the value of the cost function in continuous iterations, the algorithm assumes that the necessary convergence is achieved and the execution of the program ends.

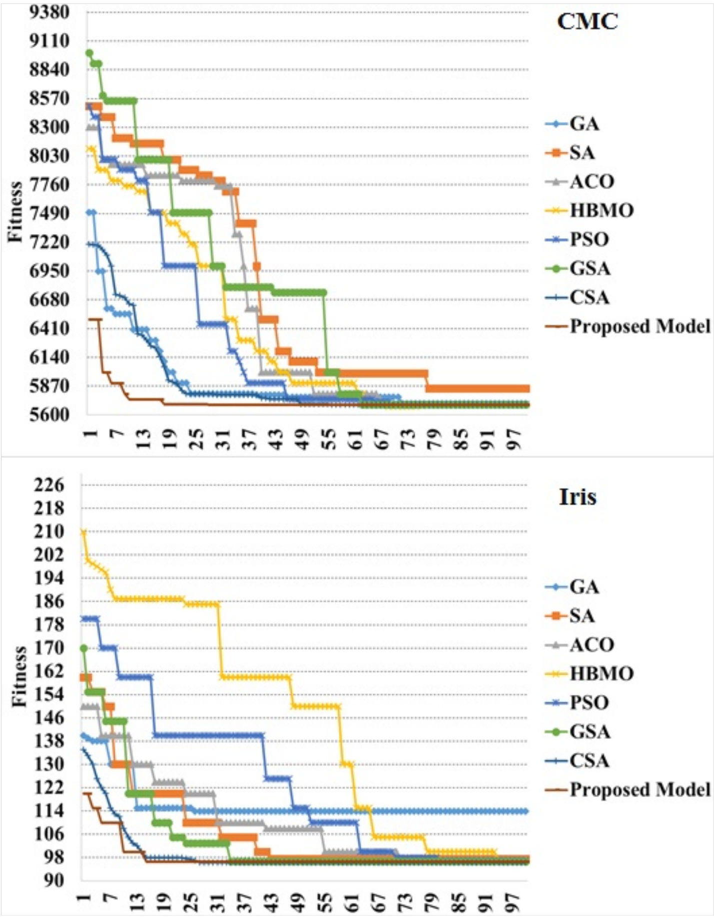

Figure 3. Display the convergence of algorithms on the CMC and Iris dataset

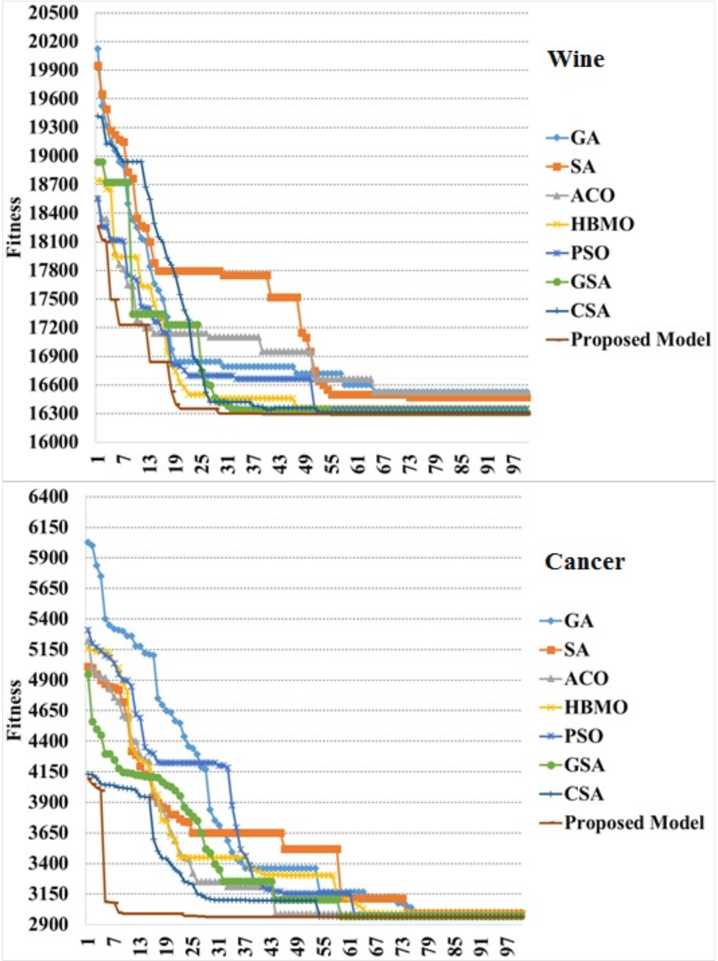

Figure 4. Display the convergence of algorithms on the Wine and Cancer dataset

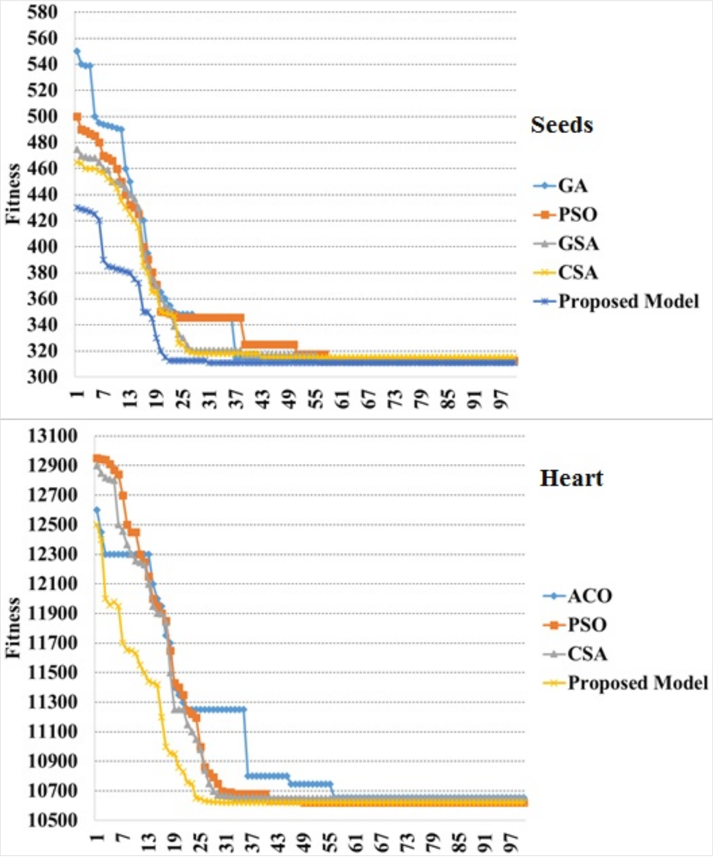

Figure 5. Display the convergence of algorithms on the Seeds and Heart Dataset

As shown in Table 5, the proposed model eventually achieves the lowest error compared to other models. The proposed model searches problem space whole in order to the better performance. So that each crow with different initial values, in a different way, discovers the space of the problem. This makes the algorithm unlocked at the local minimum and, on the other hand, the search process does not increase. The proposed error rate for Glass, Vowel, CMC, Iris, Wine, cancer, seeds and heart is 32.46, 41.58, 53.12, 9.98, 28.49, 3.58, 10.25 and 11.26, respectively. The HBMO, PSO algorithm and GSA have less error rate.

Table 5. Comparison of proposed model with other models on various datasets based on error rate

datasets	Proposed Model	GSA [38]	PSO [38, 39]	HBMO [38]	ACO [38]	SA [38]	GA [38, 40]	k-means [38]
Glass	32.46	45.92	45.59	45.73	46.12	46.00	47.68	37.81
Vowel	41.58	42.25	44.65	42.68	42.38	43.29	45.03	44.26
CMC	53.12	53.87	54.41	54.35	54.63	54.46	54.98	54.49
Iris	9.98	10.35	10.61	10.13	10.16	10.12	11.42	13.67
Wine	28.49	29.25	28.15	29.00	28.31	28.09	28.47	31.12
cancer	3.58	4.02	5.11	3.68	5.00	4.68	4.72	4.08
Seeds	10.25	10.15	10.36	11.05	11.39	12.09	13.47	11.55
Heart	11.26	25.21	16.82	19.58	19.35	20.12	28.36	37.81

Figure 6 shows the comparison diagram of the proposed model with other models on the Glass, Vowel, CMC, and Iris datasets based on the error rate. The error rate graph on the Glass, Vowel, CMC, and Iris dataset indicates that the proposed model has less error rate than other models.

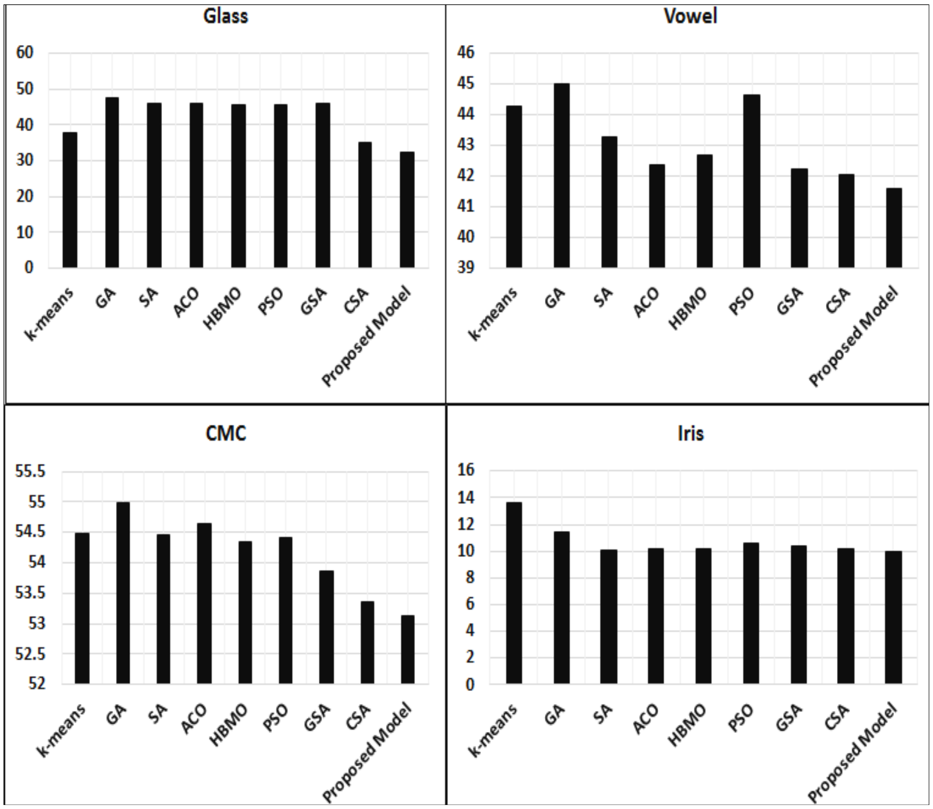

Figure 6. Comparison of the error rate on the Glass, Vowel, CMC, and Iris datasets

Figure 7 shows the comparison chart of the proposed model with other models based on the Wine, Seeds, Heart and Heart datasets based on the error rate.

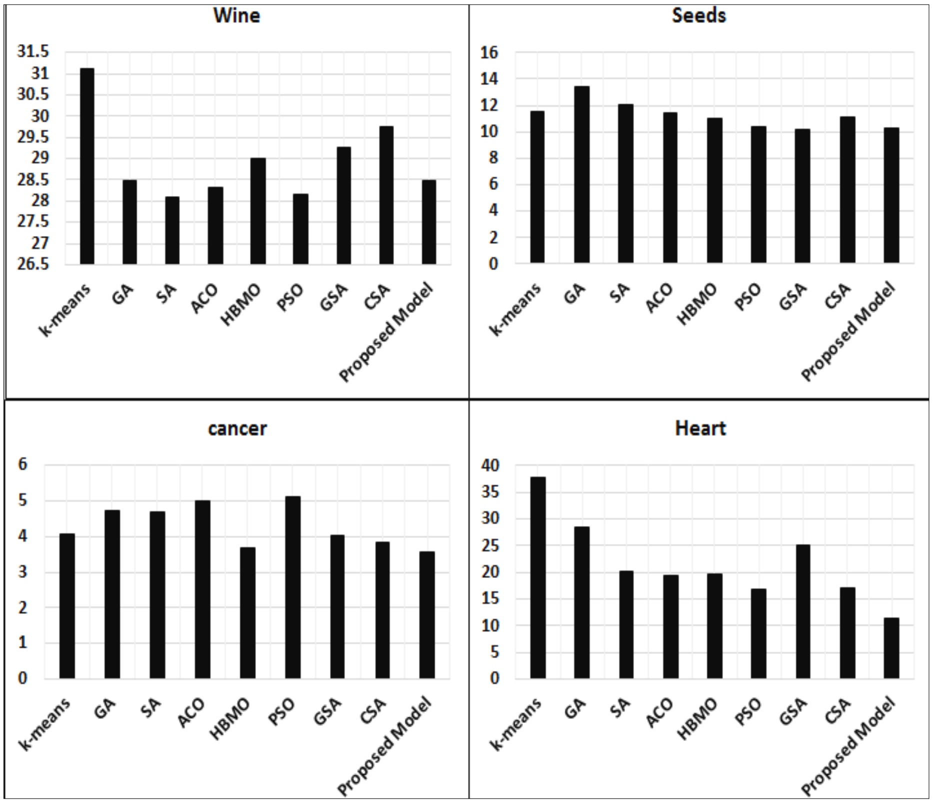

Figure 7. Comparison of the error rate on the Wine, Seeds, Cancer, and Heart datasets

According to the results, it can be concluded that the proposed model is suitable for data clustering and has high ability to find the center of clusters and similar samples. The error rate in the proposed model is lower than the other models, and the SSE is less.

5. Conclusions and Future Works

Data clustering is difficult and challenging due to its specific features. Data clustering is a data grouping process to put similar instances in a cluster. According to the results, the proposed model was successful in solving the clustering problem. The CSA has a good ability to explore, but due to the speed of the agents that search for the solution space, it lacks the proper operation mechanism, which leads to an increase in iterations. The OBL method makes the variation in the algorithm higher and does not just move towards the best, which prevents early convergence. The proposed model utilized the benefits of OBL strategy to improve the diversity of CSA population. In order to balance between exploring and exploiting has been used an OBL method. This strategy enhances the search-efficiency of CSA and help to alleviate the issues of stagnation at the sub-optimal solution and premature convergence. To prevent early convergence at each stage of the update, we changed the central positions of the clusters. The simulation results on the eight UCI datasets showed that the proposed model compared the GA, SA, ACO, HBMO, PSO, and k-means in terms of solution quality and the speed of convergence has worked better.

Two essential components which justify the quality of the solution are exploration and exploitation. Each algorithm has its individual capacity and it may concentrate on either exploration or exploitation. When it was combined more than one algorithm, it will help to improve the quality of the solution by focusing on both exploration and exploitation. The potential of CSA is high because it had performed better than other optimization algorithms. For future works, we can use crossover and mutation operators to strengthen CSA.

References

[1] Zhao, X., J. Liang, and C. Dang, *A stratified sampling based clustering algorithm for large-scale data*. Knowledge-Based Systems, 2019. 163: p. 416-428.

[2] Ramezani, F., *Solving Data Clustering Problems using Chaos Embedded Cat Swarm Optimization*. Journal of Advances in Computer Research, 2019. 10(1): p. 1-10.

[3] Rahnema, N. and F.S. Gharehchopogh, *An improved artificial bee colony algorithm based on whale optimization algorithm for data clustering*. Multimedia Tools and Applications, 2020: p. 1-26.

[4] Aliniya, Z. and S.A. Mirroshandel, *A novel combinatorial merge-split approach for automatic clustering using imperialist competitive algorithm*. Expert Systems with Applications, 2019. 117: p. 243-266.

[5] Omidvar, R., et al., *An Improved SSPCO Optimization Algorithm for Solve of the Clustering Problem*. Journal of Advances in Computer Research, 2018. 9(1): p. 1-16.

[6] Allahverdipour, A. and F. Soleimanian Gharehchopogh, *An improved k-nearest neighbor with crow search algorithm for feature selection in text documents classification*. Journal of Advances in Computer Research, 2018. 9(2): p. 37-48.

[7] Jafari, N. and F. Soleimanian Gharehchopogh, *An Improved Bat Algorithm with Grey Wolf Optimizer for Solving Continuous Optimization Problems*. Journal of Advances in Computer Engineering and Technology, 2020.

[8] Gharehchopogh, F.S. and S. Haggi, *An Optimization K-Modes Clustering Algorithm with Elephant Herding Optimization Algorithm for Crime Clustering.*

[9] Abedi, M. and F.S. Gharehchopogh, *An improved opposition based learning firefly algorithm with dragonfly algorithm for solving continuous optimization problems*. Intelligent Data Analysis, 2020. 24(2): p. 309-338.

[10] Amjad, S. and F. Soleimanian Gharehchopogh, *A Novel Hybrid Approach for Email Spam Detection based on Scatter Search Algorithm and K-Nearest Neighbors*. Journal of Advances in Computer Engineering and Technology, 2019. 5(3): p. 181-194.

[11] Rabani, H. and F. Soleimanian Gharehchopogh, *An Optimized Firefly Algorithm based on Cellular Learning Automata for Community Detection in Social Networks*. Journal of Advances in Computer Research, 2019. 10(3): p. 13-30.

[12] Gharehchopogh, F.S. and H. Gholizadeh, *A comprehensive survey: Whale Optimization Algorithm and its applications*. Swarm and Evolutionary Computation, 2019. 48: p. 1-24.

[13] Shayanfar, H. and F.S. Gharehchopogh, *Farmland fertility: A new metaheuristic algorithm for solving continuous optimization problems*. Applied Soft Computing, 2018. 71: p. 728-746.

[14] Gharehchopogh, F.S., H. Shayanfar, and H. Gholizadeh, *A comprehensive survey on symbiotic organisms search algorithms*. Artificial Intelligence Review, 2019: p. 1-48.

[15] Askarzadeh, A., *A novel metaheuristic method for solving constrained engineering optimization problems: crow search algorithm*. Computers & Structures, 2016. 169: p. 1-12.

[16] Rizk-Allah, R.M., A.E. Hassanien, and S. Bhattacharyya, *Chaotic crow search algorithm for fractional optimization problems*. Applied Soft Computing, 2018. 71: p. 1161-1175.

[17] Rahnamayan, S., H.R. Tizhoosh, and M.M. Salama, *Opposition-based differential evolution.* IEEE Transactions on Evolutionary computation, 2008. 12(1): p. 64-79.

[18] Chuang, L.-Y., Y.-D. Lin, and C.-H. Yang. *An improved particle swarm optimization for data clustering.* in *Proceedings of the International MultiConference of Engineers &Computer Scientist 2012 I, IMECS.* 2012.

[19] Wan, M., et al., *Chaotic ant swarm approach for data clustering.* Applied Soft Computing, 2012. 12(8): p. 2387-2393.

[20] Saida, I.B., K. Nadjet, and B. Omar, *A new algorithm for data clustering based on cuckoo search optimization*, in *Genetic and Evolutionary Computing.* 2014, Springer. p. 55-64.

[21] Zhou, Y., et al., *A simplex method-based social spider optimization algorithm for clustering analysis.* Engineering Applications of Artificial Intelligence, 2017. 64: p. 67-82.

[22] Alswaitti, M., M. Albughdadi, and N.A.M. Isa, *Density-based particle swarm optimization algorithm for data clustering.* Expert Systems with Applications, 2018. 91: p. 170-186.

[23] Panigrahi, S.K. and S. Pattnaik, *Empirical study on clustering based on modified teaching learning based optimization.* Procedia Computer Science, 2016. 92: p. 442-449.

[24] Abualigah, L.M., et al., *A novel hybridization strategy for krill herd algorithm applied to clustering techniques.* Applied Soft Computing, 2017. 60: p. 423-435.

[25] Majhi, S.K. and S. Biswal, *Optimal cluster analysis using hybrid K-Means and Ant Lion Optimizer.* Karbala International Journal of Modern Science, 2018. 4(4): p. 347-360.

[26] Ewees, A.A., M.A. Elaziz, and E.H. Houssein, *Improved grasshopper optimization algorithm using opposition-based learning.* Expert Systems with Applications, 2018. 112: p. 156-172.

[27] Feng, Y., et al., *Opposition-based learning monarch butterfly optimization with Gaussian perturbation for large-scale 0-1 knapsack problem.* Computers & Electrical Engineering, 2018. 67: p. 454-468.

[28] Ibrahim, R.A., M.A. Elaziz, and S. Lu, *Chaotic opposition-based grey-wolf optimization algorithm based on differential evolution and disruption operator for global optimization.* Expert Systems with Applications, 2018. 108: p. 1-27.

[29] Elaziz, M.A., D. Oliva, and S. Xiong, *An improved opposition-based sine cosine algorithm for global optimization.* Expert Systems with Applications, 2017. 90: p. 484-500.

[30] Ahandani, M.A., *Opposition-based learning in the shuffled bidirectional differential evolution algorithm.* Swarm and Evolutionary Computation, 2016. 26: p. 64-85.

[31] Ahandani, M.A. and H. Alavi-Rad, *Opposition-based learning in shuffled frog leaping: An application for parameter identification.* Information Sciences, 2015. 291: p. 19-42.

[32] Bulbul, S.M.A., et al., *Opposition-based krill herd algorithm applied to economic load dispatch problem.* Ain Shams Engineering Journal, 2018. 9(3): p. 423-440.

[33] Sarkhel, R., et al., *An improved harmony search algorithm embedded with a novel piecewise opposition based learning algorithm.* Engineering Applications of Artificial Intelligence, 2018. 67: p. 317-330.

[34] Remli, M.A., et al., *An enhanced scatter search with combined opposition-based learning for parameter estimation in large-scale kinetic models of biochemical systems.* Engineering Applications of Artificial Intelligence, 2017. 62: p. 164-180.

[35] Niknam, T., J. Olamaie, and B. Amiri, *A hybrid evolutionary algorithm based on ACO and SA for cluster analysis.* Journal of Applied Science, 2008. 8(15): p. 2695-2702.

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

[37] dataset, https://archive.ics.uci.edu/ml/index.php.

[38] Niknam, T. and B. Amiri, *An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis.* Applied soft computing, 2010. 10(1): p. 183-197.

[39] Zhou, Y., et al., *Automatic data clustering using nature-inspired symbiotic organism search algorithm.* Knowledge-Based Systems, 2019. 163: p. 546-557.

[40] Boushaki, S.I., N. Kamel, and O. Bendjeghaba, *A new quantum chaotic cuckoo search algorithm for data clustering.* Expert Systems with Applications, 2018. 96: p. 358-372.