E-ISSN 2147-9429

Journal of Polytechnic
POLİTEKNİK
DERGİSİ

GAZİ ÜNİVERSİTESİ

POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

WBBA-KM: A hybrid weight-based bat algorithm with the k-means algorithm for cluster analysis

WBBA-KM: Küme analizi için K-ortalama algoritmalı ve ağırlık tabanlı yarasa algoritmalı hibrit algoritma

***Yazar(lar) (Author(s))**: Mohammed H. IBRAHIM[1]*

ORCID[1]: 0000-0002-6093-6105

Bu makaleye şu şekilde atıfta bulunabilirsiniz(To cite to this article): IBRAHIM M. H., "WBBA-KM: A hybrid weight-based bat algorithm with the k-means algorithm for cluster analysis", ***Politeknik Dergisi***, 25(1): 65-73, (2022).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: 10.2339/politeknik.689384

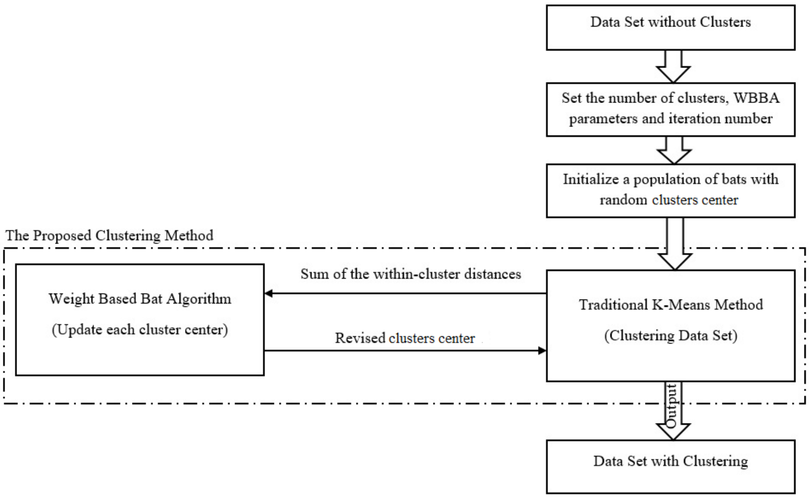
WBBA-KM: A hybrid weight-based bat algorithm with the k-means algorithm for cluster analysis

Highlights

- *The performance of the bat algorithm has been improved by updating the velocity equation.*
- *The proposed weight-based bat algorithm (WBBA) hybridized with the k-mean (KM) clustering algorithm.*
- *A clustering method called WBBA-KM has been proposed to increase the efficiency of cluster analysis.*
- *The WBBA-KM applied to UCI datasets has been compared with clustering methods in the literature.*
- *The WBBA-KM has shown very promising results and can be used to cluster real-world datasets.*

Graphical Abstract

In this study, a variant bat algorithm named weight-based bat algorithm (WBBA) has been integrated with the K-mean (KM) clustering method. The proposed WBBA-KM clustering method has been used to cluster analysis.



Data Set without Clusters
Set the number of clusters, WBBA parameters and iteration number
Initialize a population of bats with random clusters center
The Proposed Clustering Method
Weight Based Bat Algorithm
(Update each cluster center)
Sum of the within-cluster distances
Revised clusters center
Traditional K-Means Method
(Clustering Data Set)
Output
Data Set with Clustering

Figure. The proposed WBBA-KM clustering method

Aim

To increase the performance of the clustering analysis by improving the standard bat algorithm.

Design & Methodology

The commonly used UCI datasets are clustered by the proposed weight-based bat algorithm.

Originality

The originality of the proposed WBBA, the equation of the velocity of the standard BA has been modified by adding inertial weight strategy and the better fitness value of the bat.

Findings

The proposed WBBA-KM clustering method outperformed the BA-KM clustering method in all benchmark datasets and in addition, the proposed WBBA-KM clustering method achieved better results than the other clustering methods in 4 out of 6 experiments.

Conclusion

As a result, the proposed WBBA-KM clustering method has been achieved remarkable performance for data mining clustering problems and the WBBA-KM clustering method is able to cluster large scale datasets efficiently in a reasonable amount of time.

Declaration of Ethical Standards

The author of this article declares that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

WBBA-KM: A hybrid weight-based bat algorithm with the k-means algorithm for cluster analysis

Araştırma Makalesi / Research Article

Mohammed H. IBRAHIM[1*]

[1] Department of Computer Engineering, Engineering Faculty, Necmettin Erbakan University, Konya, Turkey

(Geliş/Received : 14.02.2020 ; Kabul/Accepted : 26.08.2020 ; Erken Görünüm/Early View: 23.09.2020)

ABSTRACT

Data clustering is an unsupervised classification method used to classify unlabeled objects into clusters. The clustering is performed by partitioning clustering, hierarchical clustering, fuzzy clustering, and density-based clustering methods. However, the center of the clusters is updated according to local searches with these traditional methods, and finding the best clusters center affects the clustering performance positively. In this study, a variant bat algorithm called weight-based bat algorithm (WBBA) is proposed and the proposed WBBA hybridized with the k-means clustering method (WBBA-KM) to determine the optimal centers of the clusters. The performance of the proposed WBBA-KM has been evaluated by using six different benchmark datasets from the UCI repository and the obtained results are compared with FCM, IFCM, KFCM, KIFCM, PSO-IFCM, GA-IFCM, ABC-IFCM, PSO-KIFCM, GA-KIFCM, ABC-KIFCM, and BA-KM clustering methods in the literature. According to the experimental results, the proposed WBBA-KM clustering method performed better performance from all other clustering methods in 4 of 6 benchmark datasets and achieved better performance from the BA-KM clustering method in all benchmark datasets.

Keywords: Bat algorithm, cluster analysis, optimization algorithms, unsupervised classification.

WBBA-KM: Küme analizi için K-ortalama algoritmalı ve ağırlık tabanlı yarasa algoritmalı hibrit algoritma

ÖZ

Veri kümeleme, etiketlenmemiş nesneleri kümeler halinde sınıflandırmak için kullanılan denetimsiz bir sınıflandırma yöntemidir. Kümeleme, bölümlere ayırarak kümeleme, hiyerarşik kümeleme, bulanık kümeleme ve yoğunluk temelli kümeleme yöntemleri ile gerçekleştirilir. Ancak bu geleneksel yöntemlerle kümelerin merkezi yerel aramalara göre güncellenmekte ve en iyi küme merkezinin bulunması kümeleme performansını olumlu yönde etkilemektedir. Bu çalışmada, ağırlık temelli yarasa algoritması (WBBA) olarak adlandırılan değişken bir yarasa algoritması önerilmiş ve önerilen WBBA, optimum küme merkezlerini belirlemek için k-ortalamalı kümeleme yöntemi (WBBA-KM) ile hibritlenmiştir. Önerilen kümeleme yönteminin performansını değerlendirmek için UCI havuzundan altı farklı veri seti kullanılmış ve elde edilen sonuçlar literatürdeki FCM, IFCM ve BA-KM kümeleme yöntemleri ile karşılaştırılmıştır. Deney sonuçlarına göre, önerilen WBBA-KM kümeleme yöntemi, 6 veri kümesinden 4'ünde tüm kümeleme yöntemlerinden daha iyi performans göstermiştir ve tüm veri setlerinde BA-KM kümeleme yönteminden daha iyi performans elde etmiştir.

Anahtar Kelimeler: Denetimsiz sınıflandırma, küme analizi, optimizasyon algoritmaları, yarasa algoritması.

1. INTRODUCTION

Clustering analysis is the process of separating information in a data set into groups according to certain proximity criteria, each of these groups is called a 'cluster'. The process of clustering analysis is called clustering [1]. The clustering process is an unsupervised classification method and an integral part of data mining [2]. The simple definition of clustering is to distinguish data elements that have similar characteristics among themselves. In clustering analysis for best clustering, the similarity between the objects in the same cluster must be maximized and the similarity between the clusters must be minimized. The clustering aims to partition a set of objects into clusters so that each object belongs to the cluster with the minimum distance to the cluster centroid [3]. The clustering process affects the accuracy of the classification process in the direct proportion when it is used as a pre-processing method. Clustering is performed by more than one clustering method developed by researchers in the literature. These clustering methods are generally classified into four groups: partitioning based, hierarchical based, density-based, grid-based, and model-based clustering [4, 5] and each group contains several clustering methods, the taxonomy of the clustering methods is shown in Figure 1 [6].

These clustering methods are different from each other according to the applied principles and there are advantages and disadvantages of each method according to these principles [7]. Nevertheless, clustering methods are successfully used in many applications such as economic science [8], document classification [9], cluster weblog data [10], pattern recognition [11], spatial data

**Sorumlu Yazar (Corresponding Author)*
e-posta: mibrahim@erbakan.edu.tr

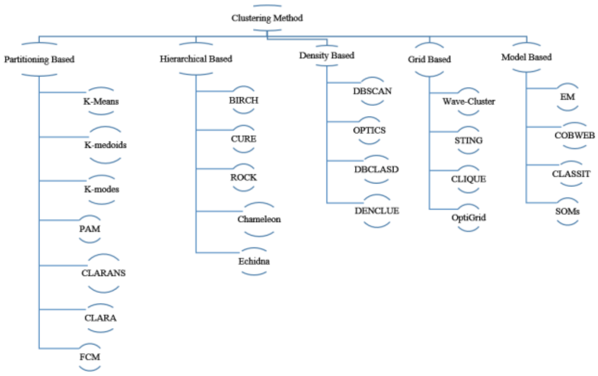


Figure 1. The hierarchy of the clustering methods

analysis [12], and image segmentation [13]. The clustering process is generally performed with the clustering methods shown in Figure 1, but these clustering methods search the clustering solution in a local search space [7]. In this paper, the optimization algorithm with global search is proposed for the clustering solution. In the literature, several optimization algorithms for clustering have been applied for different application areas. Kuo et al combined a kernel intuitionistic fuzzy c-means method with different three optimization algorithms: particle swarm optimization (PSO), genetic algorithm (GA), and artificial bee colony (ABC) algorithms. The performance of the proposed methods was evaluated on six benchmark datasets and compared the obtained results with other clustering methods. According to the experimental results the GA achieved better accuracy on four datasets [14]. Tripathi et al proposed a novel variant of bat algorithm for clustering named dynamic frequency-based parallel k-bat algorithm (DFBPKBA) and they used the proposed bat algorithm to clustering process. They evaluated the variant bat algorithm based clustering method on five datasets commonly used from UCI. The test results showed that the proposed variant bat algorithm based clustering method performed better results from k-means PSO and standard bat algorithm-based clustering methods [15]. Tang et al, integrated the nature-inspired optimization algorithms including the wolf algorithm (WA), firefly algorithm (FA), cuckoo algorithm (CA), bat algorithm (BA), and ant colony optimization algorithm (ACO) into the k-means method so that the clustering process can have a more global solution. The nature-inspired based clustering proposed method improved the performance of the clustering process. According to the experimental results the developed method gives very well results on the test datasets [7]. Maulik and Bandyopadhyay used a genetic algorithm for the clustering process. Have tested the proposed genetic algorithm-based clustering technique on data clusters such as iris, vowel, and crude oil with deferent cluster and iteration numbers. They compared the performance of the proposed algorithm with the k-means cluster method. The proposed method showed better results than the k-means cluster method [16]. Kalyani and Swarup used the PSO algorithm as a hybrid with the k-means cluster method and used the proposed PSO-based k-means clustering method to classify the security systems of power systems. And tested the performance of the proposed PSO-based k-means in standard test systems such as IEEE 30 Bus, 57 Bus, 118 Bus, and 300 bus and compared the test results with the unsupervised k-means clustering method. According to simulation results, the proposed method has given high accuracy classifiers with less misclassification rate than the k-means clustering method [17]. Karaboga and Ozturk used the ABC optimization algorithm for data clustering in benchmark problems and the performance of the ABC algorithm-based clustering method is compared with nine clustering methods in the literature. They used thirteen commonly used data sets from the UCI Machine Learning Repository to demonstrate the performance of the proposed method. Depending on the simulation results, the ABC algorithm performs the clustering process more effectively than other methods [18].

In this study, a variant bat algorithm is proposed and has been used to determine the optimum centers of the clusters. The paper is organized as follows: in section 1, general information about the clustering and literature review is given. In section 2, the clustering process is explained. In section 3, the bat optimization algorithm is explained. In section 4, the steps of the proposed clustering method are detailed. In section 5, the experimental results of the proposed clustering method and other clustering methods are discussed and finally, the conclusion of the study is given in section 6.

2. MATERIAL and METHOD

2.1. Datasets

In this study, the performance of the proposed WBBA-KM clustering method has been evaluated on the commonly used UCI datasets. The characteristics of the UCI datasets are given in Table 1.

Table 1. Datasets characteristics

Datasets	Number of		
	attributes	instances	clusters
Iris	4	150	3
Wine	13	178	3
Tae	5	151	3
Flame	2	240	2
Glass	9	214	6
Wbc	10	699	2

2.2. Clustering and Clustering Validity

The process of grouping a set of objects into classes of similar objects is called clustering [19]. A cluster is a collection of data objects that are similar to each other in the same cluster and different from objects in the other clusters. In clustering, the similarity of the elements in the cluster should be high and the similarity between the clusters should be low. Clustering falls from data mining techniques to descriptive models, namely unsupervised classification. In unsupervised classification, the aim is to cluster the data that was initially given and not yet classified, to form meaningful subsets [3]. The clustering process is performed by clustering methods such as k-means, k-medoids FCM, BIRCH, and DBSCAN. The performance of clustering methods depends on updating the cluster centers and each method updates the cluster centers differently. Updating the cluster centers is known as an NP-complete problem and it is an optimization problem [20]. An N-dimensional with n points dataset and k-number of clusters, the closeness of all points in the dataset is calculated according to the distance equation such as Euclidean Distance, Manhattan Distance, Squared Euclidean Distance, Mahalanobis Distance, Minkowski Distances, Chebyshev Distance, Cosine Distance, Bray Curties distance and Canberra Distance [21]. In this study, the Euclidean distance equation is used to find the similarity between the points, Euclidean distance equation is given in Equation 1.

$$E^D(X_i, C_j) = \sqrt{\sum_{l=1}^{p}(X_{il} - C_{jl})^2} \qquad (1)$$

Where $X_i; i = 1,2, \dots n$ where n a set of objects, each object with p attributes and $C_j; j = 1,2, \dots k$; where k is the number of clusters. In clustering, the distance of each object to all cluster centroid is calculated by Equation 1 and the object is assigned to the cluster centroid with the smallest value. After the clustering process, the accuracy of the clustering method is usually calculated by the internal cluster validation or external cluster validation. The clustering process is generally performed by the following algorithm [19].

```
Inputs:
    k: the number of clusters,
    D: a data set containing n objects without clusters.
Output:
    A set of objects with k clusters and each object belongs to the cluster with the minimum distance to the cluster centroid.
Method:
    1. Choose k objects from D as the initial cluster centroid,
    2. Repeat
    3. Calculate the distance of each object to all cluster centroids and assign the object to the cluster centroid with the smallest value,
    4. Update the cluster centroid, until no change.
```

The term cluster validation is used to evaluate the performance of the clustering method and it helps us to choose a good clustering method among clustering methods [22]. In general, clustering validation can be labeled into three classes: Internal cluster validation, External cluster validation, and Relative cluster validation [5]. Internal cluster validation evaluates the structure of the clustering method by using internal information without reference to external information. Also, it is used to estimate the number of clusters and an appropriate clustering method. External cluster validation includes comparing the results of a clustering method to an externally known result, such as class labels. It measures the extent to which the results from the clustering method match externally supplied actual class labels. Relative cluster validation evaluates the clustering method by changing the number of cluster parameter values for the same method and it's used for finding the optimal number of clusters [23].

In this study, the sum of the within-cluster distances (S_w) index is used as internal cluster validation [23] and the clustering accuracy (CA) is used as external cluster validation [14], the sum of the within-cluster distances index and clustering accuracy are given in Equations 2 and 3 respectively. The standard deviation (SD) is used to show the distribution of the S_w and CA values obtained from the clustering method in N times run, SD is given in Equation 4 [14].

$$S_w = \sum_{k=1}^{q} \sum_{i,j \in C_k \; and \; i<j} d(x_i, x_j) \qquad (2)$$

$$CA = \frac{\# \; of \; correct \; examples}{size \; of \; dataset} * 100 \qquad (3)$$

$$SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(CA_i - \overline{CA})^2} \qquad (4)$$

In Equation 2, q represents the number of clusters and $d(x_i, x_j)$ represents the distance between x_i and x_j points in the cluster C_k. The sum of the within-cluster distances index is used as the fitness function and minimized on the other hand the clustering accuracy equation is used to compare the proposed clustering method with the clustering methods proposed in the literature. The minimum of the within-cluster distances index and the maximum of the clustering accuracy indicates that the clustering method is sturdy. In Equation 4, the SD is the standard deviation of the CA values, N is the number of CA values, CA_i is the i'th CA and $\overline{CA}$ is the arithmetic mean of the CA.

2.3. Bat Optimization Algorithm

The bat algorithm is an optimization algorithm that is based on the echolocation behaviors of bats in nature that search to locate food and prey. This algorithm was developed by Yang [24] and used in various optimization problems in literature. Yang has made some of the bat's echolocation characteristics ideal and based on the following rules for the formation of the algorithm [24].

- All bats use echolocation to sense distance, and they also know the difference between food/prey and background barriers in some magical way;
- Each bat will have a velocity v_i, a position x_i, a frequency value f_i, a wavelength (λ), a loudness value L_i, and a pulse emission (r).
- Although the loudness can vary in many ways, we assume that the loudness varies from a large (positive) L_0 to a minimum constant value A_{min}.

Yang obtained the following formulas to update the velocity and position of the bat from the above information [24].

$$f_{iglobal} = f_{min} + (f_{max} - f_{min})a \tag{5}$$

$$v_i^t = v_i^{t-1} + (x_i^t - x^*)f_i \tag{6}$$

$$x_i^t = x_i^{t-1} + v_i^t \tag{7}$$

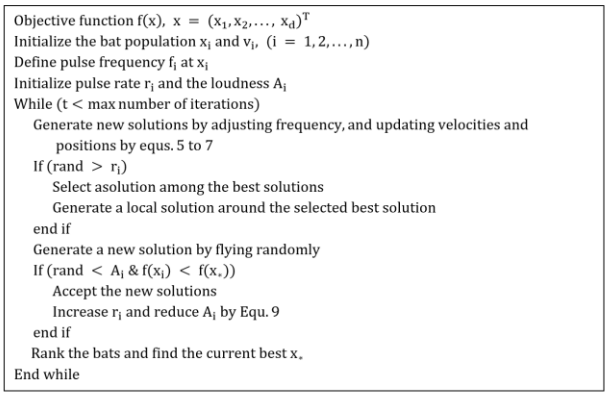
Where α is a random number between 0 and 1, f_i is the frequency value of i'th bat, f_{min} and f_{max} are the minimum and maximum frequency values respectively, and x^* represents the best solution value in the population. After selecting the best solution value from the available solution values, a new solution value is generated using local random walk by Equation 8 [24].

$$x_{new} = x_{old} + \varepsilon L^t \tag{8}$$

where ε represents a random number value in the range [1, -1] and L^t represents the average loudness of all bats in the t time. When the bat finds the prey, the signal propagation rate (r) is generally increased although the loudness (L) decreases.

$$L_i^{t+1} = \beta L_i^t, \;\; r_i^{t+1} = r_i^0 \;\; [1 - exp(-\gamma t)] \tag{9}$$

where β is a constant number between [0, 1], and γ is a constant positive number. When t → ∞, the loudness is $L_i^t \to 0$ and $r_i^t \to r_i^0$. The pseudo-code of the BA is shown in Figure 2 [25].



```
Objective function $f(x)$, $x = (x_1, x_2, \ldots, x_d)^T$
Initialize the bat population $x_i$ and $v_i$, $(i = 1, 2, \ldots, n)$
Define pulse frequency $f_i$ at $x_i$
Initialize pulse rate $r_i$ and the loudness $A_i$
While ($t <$ max number of iterations)
   Generate new solutions by adjusting frequency, and updating velocities and
      positions by equs. 5 to 7
   If (rand $> r_i$)
      Select asolution among the best solutions
      Generate a local solution around the selected best solution
   end if
   Generate a new solution by flying randomly
   If (rand $< A_i$ & $f(x_i) < f(x_*)$)
      Accept the new solutions
      Increase $r_i$ and reduce $A_i$ by Equ. 9
   end if
   Rank the bats and find the current best $x_*$
End while
```

Figure 2. The pseudo-code of the BA

3. THE PROPOSED APPROACH FOR CLUSTERING

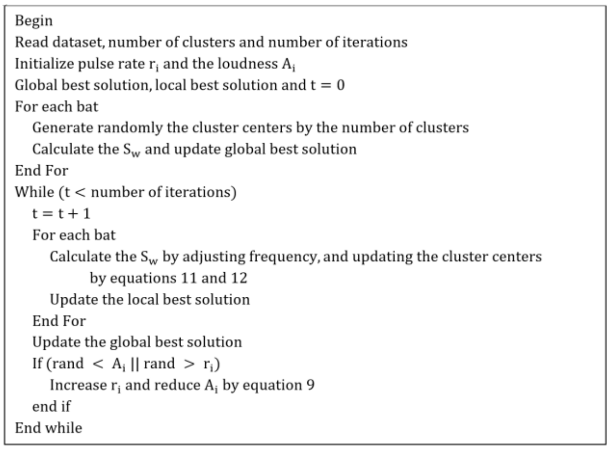
In this section, the proposed variant of the bat algorithm and its hybridization with the k-means algorithm has been explained. The proposed variant of the bat algorithm is named a weight-based bat algorithm (WBBA). In the standard bat algorithm, updating the position is done only according to the best solution of all bats. In this case, the best solution is obtained with a higher number of iterations. To eliminate this problem in the bat algorithm, in the proposed WBBA both the best solution of all bats and the best solution of the bat update the position of the bat in an iteration. In addition, the inertial weight strategy is added in the proposed WBBA and experimentally found that this strategy increased the convergence of the bat algorithm in the early iterations. In the proposed WBBA, the velocity and position of the bats in each generation are updated according to Equations 11 and 12, respectively.

$$\begin{gathered} f_{iglobal} = f_{min} + (f_{max} - f_{min})a \\ f_{ilocal} = (f_{max} - f_{min})a \end{gathered} \tag{10}$$

$$v_i^t = w * v_i^{t-1} + (x_i^t - x^*)f_{ilocal} + (x_i^t - x^{**})f_{iglobal} \tag{11}$$

$$x_i^t = x_i^{t-1} + v_i^t \tag{12}$$

In Equation 11, v_i^t is the new velocity, t is the number of iteration, w is inertia weight, i is a bat's index, x_i^t position of bat i, v_i^{t-1} the previous velocity of bat i, x^* is a better fitness value of the bat i, x^{**} is a better fitness value of all bats. The pseudo-code of the proposed WBBA is shown in Figure 3 below.



```
Begin
Read dataset, number of clusters and number of iterations
Initialize pulse rate $r_i$ and the loudness $A_i$
Global best solution, local best solution and $t = 0$
For each bat
   Generate randomly the cluster centers by the number of clusters
   Calculate the $S_w$ and update global best solution
End For
While ($t <$ number of iterations)
   $t = t + 1$
   For each bat
      Calculate the $S_w$ by adjusting frequency, and updating the cluster centers
            by equations 11 and 12
      Update the local best solution
   End For
   Update the global best solution
   If (rand $< A_i$ || rand $> r_i$)
      Increase $r_i$ and reduce $A_i$ by equation 9
   end if
End while
```

Figure 3. The pseudo-code of the proposed WBBA

To increase the performance of the clustering process, the proposed WBBA is used as a hybrid with the k-means clustering method to update the cluster centers. The proposed WBBA based clustering method is named the WBBA-KM clustering method. Figure 4 shows the basic components of the proposed WBBA-KM clustering method and for a simple understanding, the proposed WBBA-KM clustering method is explained in steps format.

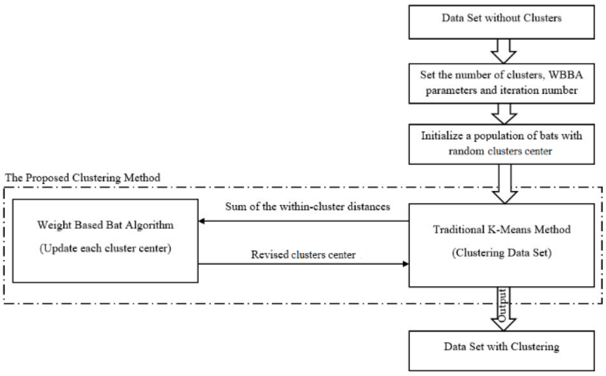

Data Set without Clusters
Set the number of clusters, WBBA parameters and iteration number
Initialize a population of bats with random clusters center
The Proposed Clustering Method
Weight Based Bat Algorithm
(Update each cluster center)
Sum of the within-cluster distances
Revised clusters center
Traditional K-Means Method
(Clustering Data Set)
Output
Data Set with Clustering

Figure 4. The basic components of the proposed WBBA-KM

Step 1: Set the number of clusters, WBBA parameters, and iteration number,

Step 2: Initialize a population of bats with random cluster centers,

Step 3: Start the iterative procedure and set the iteration count t = 1,

Step 4: For every bat in the population,

- Calculate Euclidean distance measure, between x_i^t cluster center and i'th data point using Equation (1),
- Assign each i'th data point to the nearest cluster center x_i^t,

Step 5: Evaluate the fitness function which is the minimization of fitness value S_w as given by Equation (2) and update the local best cluster center x^* of bat according to compare evaluation with the bat's previous best value, x^*, in terms of fitness value S_w. If the current cluster center is better than x^*, then assign the current cluster center to x^*, else retain x^*at its old value. This process is carried out for each bat in the population,

Step 6: Updating the velocity and cluster centers of each bat using Equations (11) and (12) respectively,

Step 7: Updating the global best cluster center x^{**} of the population according to compare evaluation with the x^{**}, in terms of fitness value S_w. If the current cluster center is better than x^{**}, then assign the current cluster center to x^{**}, else retain x^{**}at its old value,

Step 8: Check the convergence criterion, maximum number of iterations. If converged, return the global best solution as the optimal cluster's center and calculate the clustering accuracy according to Equation (3) for the proposed WBBA-KM clustering method, else increment the iteration count, t = t + 1 and loop to Step 4.

5. RESULTS AND DISCUSSION

To evaluate the performance of the proposed WBBA-KM clustering method is applied to Iris, Wine, Tae, Flame, Glass, and Wisconsin-Breast Cancer (Wbc) which are the most frequently used UCI datasets in the literature [14]. All the experiments are conducted on a machine with an Intel Core i7@2.00 GHz processor and 8 GB memory, running on Microsoft Windows 10 OS. The proposed method is implemented in C#.Net using visual studio 2017. The best parameters of the BA and the proposed WBBA are f_{min} is 0, f_{max} is 10, β is 0.9 and γ is equal to 0.01. The main goal of the proposed ASD clustering method is to increase the similarity of the objects in the same cluster and decrease the similarity between the clusters by sum square error (SSE). In the SSE, X represents the objects, C represents clusters set, and $\|.\|$ represents Euclidean distance, and the SSE is given in Equation 13.

$$SSE(X,C) = \sum_{i=1}^{k} \sum_{x_j \in C_i} \left\| x_j - c_i \right\|^2 \tag{13}$$

To compare the performance of the proposed WBBA-KM with the standard BA-KM, the SSE has been used as a fitness function. The SSE results of the datasets obtained from the standard BA-KM and the proposed WBBA-KM method are given in Table 2.

When the results in Table 2 are analyzed, the improvement in the proposed WBBA-KM clustering method is visible. The proposed WBBA-KM clustering method obtained better SSE results from the standard BA in all datasets. In addition, the proposed WBBA-KM clustering method is compared with the FCM, IFCM, KFCM, KIFCM, PSO-IFCM, GA-IFCM, ABC-IFCM,

Table 2. The SSE results of the datasets

Methods	SSE	Iris	Wine	Tae	Flame	Glass	Wbc
BA-KM	Average (%)	96.97	16293.52	1493.16	1424.87	239.69	2966.73
	Best (%)	96.97	16292.74	1492.93	1423.96	229.22	2966.73
	Worst (%)	96.97	16294.93	1493.95	1425.83	247.83	2966.73
	SD (%)	3.86E-6	1.74	0.48	0.57	8.68	1.48E-07
WBBA-KM	Average (%)	96.35	16291.42	1491.03	1423.91	238.59	2964.38
	Best (%)	96.35	16291.19	1490.92	1423.82	226.22	2964.38
	Worst (%)	96.35	16293.86	1491.97	1425.72	247.79	2964.38
	SD (%)	1.66E-10	0.51381	0.33	0.42	8.11	1.48E-09

PSO-KIFCM, GA-KIFCM, ABC-KIFCM, and BA-KM clustering methods in the literature [14]. The common parameters such as the number of iterations, execution time, and bats are used as in the literature [14]. Table 3 summarizes the accuracy results of the proposed WBBA-KM and other clustering methods in the literature [14].

Table 3. The accuracy results of the datasets

Methods	Accuracy	Iris	Wine	Tae	Flame	Glass	Wbc
K-means	Average (%)	79.246	91.195	46.004	83.944	39.034	96.066
	Best (%)	88.670	94.940	56.291	85.833	45.327	96.193
	Worst (%)	57.330	67.416	38.411	83.750	33.178	96.047
	SD (%)	0.146	0.095	0.051	0.006	0.038	0.001
FCM	Average (%)	89.330	94.940	49.669	85.000	42.056	95.608
	Best (%)	89.330	94.940	49.669	85.000	42.056	95.608
	Worst (%)	89.330	94.940	49.669	85.000	42.056	95.608
	SD (%)	0.000	0.000	0.000	0.000	0.000	0.000
IFCM	Average (%)	91.333	71.178	46.203	84.167	44.860	96.340
	Best (%)	91.333	75.280	47.682	84.167	44.860	96.340
	Worst (%)	91.333	65.170	42.384	84.167	44.860	96.340
	SD (%)	0.000	0.035	0.013	0.000	0.000	0.000
KFCM	Average (%)	89.467	94.233	50.177	85.292	46.012	96.003
	Best (%)	96.000	94.944	51.656	87.083	48.598	96.047
	Worst (%)	84.667	92.700	42.384	85.000	43.925	95.754
	SD (%)	0.019	0.005	0.015	0.005	0.016	0.001
KIFCM	Average (%)	91.556	93.205	51.832	85.319	46.573	96.389
	Best (%)	96.000	94.944	56.954	87.083	49.065	97.365
	Worst (%)	89.333	92.700	50.331	85.000	43.925	96.193
	SD (%)	0.019	0.009	0.023	0.006	0.016	0.003
PSO-IFCM	Average (%)	80.667	65.506	44.768	85.583	41.869	95.666
	Best (%)	90.000	84.831	51.656	88.333	48.598	96.633
	Worst (%)	60.000	52.809	34.437	84.167	37.383	94.290
	SD (%)	12.009	14.605	7.646	1.990	4.341	0.980
GA-IFCM	Average (%)	84.933	65.169	44.238	85.000	39.346	95.959
	Best (%)	97.333	80.899	45.695	87.500	50.000	97.218
	Worst (%)	74.000	41.011	42.384	84.167	28.972	94.583
	SD (%)	9.691	17.465	1.510	1.413	7.912	1.094

ABC- IFCM	Average (%)	68.933	64.157	44.768	84.667	46.729	97.482
	Best (%)	78.667	71.348	47.682	85.833	51.869	97.804
	Worst (%)	61.333	60.112	42.384	81.250	40.187	97.218
	SD (%)	6.229	4.304	2.369	1.941	4.494	0.217
PSO- KIFCM	Average (%)	91.822	95.206	52.958	85.167	47.118	96.960
	Best (%)	94.667	96.629	56.954	87.083	49.065	97.365
	Worst (%)	90.000	94.382	50.331	84.167	45.327	96.340
	SD (%)	0.014	0.008	0.025	0.007	0.014	0.005
GA- KIFCM	Average (%)	91.844	**95.243***	52.936	85.486	47.103	**97.057***
	Best (%)	96.000	97.191	55.629	87.083	49.065	97.365
	Worst (%)	90.000	94.382	51.656	85.000	46.729	96.633
	SD (%)	0.013	0.007	0.009	0.007	0.005	0.002
ABC- KIFCM	Average (%)	91.733	95.187	52.848	85.292	47.134	96.911
	Best (%)	96.000	96.067	55.629	86.667	48.131	97.365
	Worst (%)	90.000	94.382	52.318	85.000	46.262	96.633
	SD (%)	0.015	0.005	0.008	0.003	0.004	0.002
BA-KM	Average (%)	90.523	93.187	53.641	85.427	45.637	92.638
	Best (%)	93.860	95.086	55.962	86.762	48.526	95.625
	Worst (%)	88.370	91.672	51.583	85.837	43.371	91.383
	SD (%)	0.019	0.007	0.009	0.005	0.006	0.006
WBBA-KM	Average (%)	**94.190***	95.216	**56.735***	**85.753***	**53.832***	96.571
	Best (%)	96.786	96.826	57.637	88.064	60.170	97.347
	Worst (%)	91.594	94.957	55.623	85.381	47.494	95.896
	SD (%)	0.013	0.008	0.009	0.007	0.071	0.121

* Best result.

When the results in Table 3 were analyzed, the proposed WBBA-KM clustering method in the iris dataset outperformed the other clustering methods with a 94.190 average accuracy value. The proposed WBBA-KM clustering method in the tae dataset had an average accuracy rate of 56.735 and with this value had better accuracy than other clustering methods. In the flame dataset, the proposed WBBA-KM clustering method performed better performance than the other clustering methods with an 85.753 average accuracy value. The proposed clustering method in the glass dataset had an average accuracy rate of 53.832 and with this value had better accuracy than other clustering methods. On the other hand, the GA-KIFCM clustering method showed better accuracy in the wine and wbc datasets than the proposed and other clustering methods with values of 95.243 and 97.057, respectively. A clustering process is successful if it has high cluster accuracy or a low sum of the within-cluster distances. The proposed WBBA-KM clustering method is increased the cluster accuracy value and decreased the sum of the value of the within-cluster distance more efficiently than other clustering methods when applied to the datasets. To determine the best clustering method, the ranking test is performed according to the clustering accuracy of the clustering methods. The order of clustering methods according to Friedman's ranks test and Wilcoxon signed ranking are given in Table 4 and Table 5 for clustering accuracy, respectively.

Table 4. Friedman's ranks test result for clustering accuracy

Methods	Mean Rank	Methods	Mean Rank
K-means	10.83	GA-IFCM	11.08
FCM	8.92	ABC-IFCM	9.08
IFCM	8.83	PSO-KIFCM	3.83
KFCM	7.25	GA-KIFCM	2.67
KIFCM	5.83	ABC-KIFCM	4.25
PSO-IFCM	9.58	BA-KM	7.00
WBBA-KM	1.83		

Table 5. Wilcoxon signed ranking result for clustering accuracy

Methods	Value	Methods	Value
WBBA-KM - K-means	0.027	WBBA-KM - GA-IFCM	0.028
WBBA-KM - FCM	0.027	WBBA-KM - ABC-IFCM	0.058
WBBA-KM - IFCM	0.026	WBBA-KM - PSO-KIFCM	0.131
WBBA-KM - KFCM	0.027	WBBA-KM - GA-KIFCM	0.395
WBBA-KM - KIFCM	0.026	WBBA-KM - ABC-KIFCM	0.058
WBBA-KM - PSO-IFCM	0.026	WBBA-KM - BA-KM	0.027

According to the results of Friedman's ranks test, the proposed WBBA-KM clustering method performed better ranks test than all other clustering methods with a 1.83 value. In the Wilcoxon signed-rank test, the proposed clustering method yields statistically significantly better results than the other clustering methods without PSO-KIFCM and GA-KIFCM clustering methods. In general, the proposed WBBA-KM clustering method achieved better results than the other clustering methods in 4 out of 6 experiments. The proposed WBBA-KM clustering method outperformed the BA-KM clustering method in all benchmark datasets, with this good result, the proposed WBBA-KM clustering method enhances the efficiency and performance of the BA-KM clustering method. According to this result, the proposed WBBA-KM performed clustering on datasets while the cluster centers update process is performed very successfully.

6. CONCLUSION

In this study, a variant of the bat algorithm named WBBA is proposed and the proposed WBBA hybridized with the k-means clustering method and used in clustering problems. To evaluate the performance of the proposed WBBA-KM clustering method, WBBA-KM has been applied on six datasets and the obtained results are compared with FCM, IFCM, KFCM, KIFCM, PSO-IFCM, GA-IFCM, ABC-IFCM, PSO-KIFCM, GA-KIFCM, ABC-KIFCM, and BA-KM clustering methods. The results show that the proposed WBBA-KM clustering method has outperformed FCM, IFCM, KFCM, KIFCM, PSO-IFCM, GA-IFCM, ABC-IFCM, PSO-KIFCM, GA-KIFCM, ABC-KIFCM, and BA-KM clustering methods in 4 of 6 datasets in terms of the quality of the clustering. On the other hand, the proposed WBBA-KM clustering method has achieved better performance than the BA-KM clustering method in all data clustering. According to all experimental results, the proposed WBBA-KM clustering method achieved remarkable performance for data mining clustering problems and the WBBA-KM clustering method is able to cluster large scale datasets efficiently in a reasonable amount of time. As a result, the proposed WBBA-KM clustering method can be used as a useful alternative clustering method for the clustering process. In future work, the proposed WBBA-KM clustering method can be used efficiently on real-world clustering problems of different domains such as economic science, document classification, cluster weblog data, pattern recognition, spatial data analysis, and image segmentation.

DECLARATION OF ETHICAL STANDARDS

The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

AUTHORS' CONTRIBUTIONS

Mohammed H. IBRAHIM: Performed the experiments and analyse the results. Wrote the manuscript.

CONFLICT OF INTEREST

There is no conflict of interest in this study.

REFERENCES

1. Han, J., Kamber, M., Pei, J., "Data mining concepts and techniques third edition", ***The Morgan Kaufmann Series in Data Management Systems***, 83-124 (2011).
2. Ngo, T., "Data mining: practical machine learning tools and technique", by ian h. witten, eibe frank, mark a. hell, ***ACM Sigsoft Software Engineering Notes*** 36(5), 51-52 (2011).
3. Zhao, Q., "Cluster validity in clustering methods", ***Publications of the University of Eastern Finland*** (2012).
4. Nagpal, A., Jatain, A., Gaur, D., "Review based on data clustering algorithms", ***In: 2013 IEEE Conference on Information & Communication Technologies***, pp. 298-303 (2013).
5. Halkidi, M., Batistakis, Y., Vazirgiannis, M., "On clustering validation techniques", ***Journal of intelligent information systems***, 17(2-3), 107-145 (2001).
6. Gulati, H., Singh, P., "Clustering techniques in data mining: A comparison", ***In: 2015 2nd international conference on computing for sustainable global development (INDIACom)***, pp. 410-415, (2015).
7. Tang, R., Fong, S., Yang, X.-S., Deb, S., "Integrating nature-inspired optimization algorithms to K-means clustering", ***In: Seventh International Conference on Digital Information Management (ICDIM 2012)***, pp. 116-123, (2012).

8. Gupta, M.D., "Socio-economic status and clustering of child deaths in rural Punjab", ***Population Studies***, 51(2), 191-202 (1997).
9. El-Hamdouchi, A., Willett, P., "Hierarchic document classification using Ward's clustering method", ***In: Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval***, pp. 149-156, (1986).
10. Wong, W.-C., Fu, A.W.-C., "Incremental document clustering for web page classification", ***In: Enabling Society with Information Technology,*** pp. 101-110. Springer, (2002).
11. Ebrahimzadeh, A., Addeh, J., Rahmani, Z., "Control chart pattern recognition using K-MICA clustering and neural networks", ***ISA transactions*** 51(1), 111-119 (2012).
12. Yu, J., Wu, J., Sarwat, M., "Geospark: A cluster computing framework for processing large-scale spatial data", ***In: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems***, p. 70 (2015).
13. Uslan, V., Bucak, I., "Microarray image segmentation using clustering methods", ***Mathematical and Computational Applications,*** 15(2), 240-247 (2010).
14. Kuo, R., Lin, T., Zulvia, F.E., Tsai, C., "A hybrid metaheuristic and kernel intuitionistic fuzzy c-means algorithm for cluster analysis", ***Applied Soft Computing***, 67, 299-308 (2018).
15. Tripathi, A.K., Sharma, K., Bala, M., "Dynamic frequency based parallel k-bat algorithm for massive data clustering (DFBPKBA)", ***International Journal of System Assurance Engineering and Management,*** 9(4), 866-874 (2018).
16. Maulik, U., Bandyopadhyay, S., "Genetic algorithm-based clustering technique", ***Pattern recognition,*** 33(9), 1455-1465 (2000).
17. Kalyani, S., Swarup, K.S., "Particle swarm optimization based K-means clustering approach for security assessment in power systems", ***Expert systems with applications***, 38(9), 10839-10846 (2011).
18. Karaboga, D., Ozturk, C., "A novel clustering approach: Artificial Bee Colony (ABC) algorithm", ***Applied soft computing***,11(1), 652-657 (2011).
19. Han, J., Pei, J., Kamber, M., "Data mining: concepts and techniques", ***Elsevier***, (2011)
20. Fathian, M., Amiri, B., Maroosi, A., "Application of honey-bee mating optimization algorithm on clustering", ***Applied Mathematics and Computation***,190(2), 1502-1513 (2007).
21. Giancarlo, R., Bosco, G.L., Pinello, L., "Distance functions, clustering algorithms and microarray data analysis", ***In: International Conference on Learning and Intelligent Optimization***, pp. 125-138 (2010).
22. Hämäläinen, J., Jauhiainen, S., Kärkkäinen, T., "Comparison of internal clustering validation indices for prototype-based clustering" ***Algorithms***,10(3), 105 (2017).
23. Charrad, M., Ghazzali, N., Boiteux, V., Niknafs, A., "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set| Charrad|", ***Journal of Statistical Software***, In. (2014)
24. Yang, X.-S., "A new metaheuristic bat-inspired algorithm", ***In: Nature inspired cooperative strategies for optimization (NICSO 2010)***, pp. 65-74. Springer, (2010).
25. Yılmaz, S., Küçüksille, E.U., "A new modification approach on bat algorithm for solving optimization problems", ***Applied Soft Computing***, 28, 259-275 (2015).