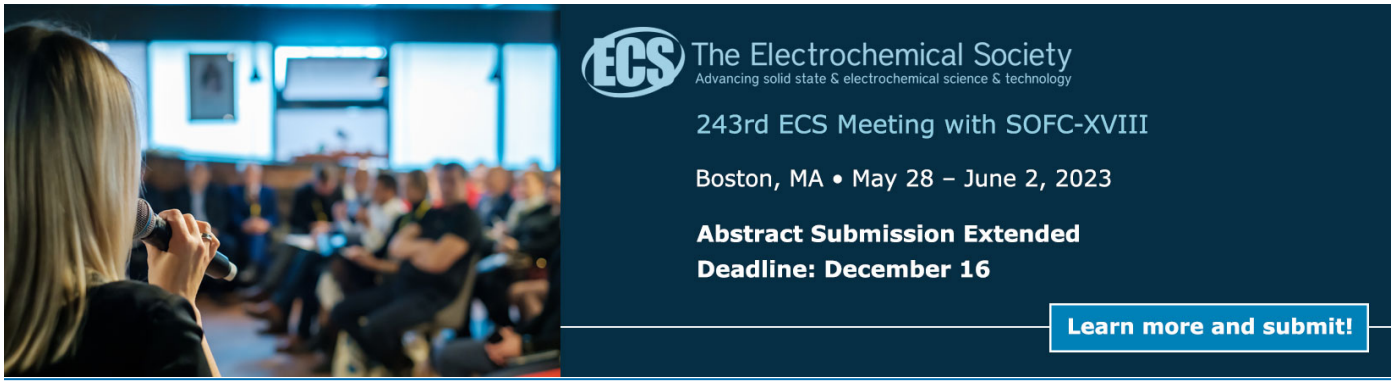
Water Wave Optimization Based Data Clustering Model



Water Wave Optimization Based Data Clustering Model

Arvinder Kaur and Yugal Kumar*
Department of Computer Science & Engineering and Information Technology,
Jaypee University of Information Technology, Waknaghat, Solan, Himachal Pradesh, India.

Email: er.arvinderdhillon@gmail.com, yugalkumar.14@gmail.com

Abstract: This paper presents data clustering model by adopting water wave optimization (WWO) algorithm. In recent times, metaheuristics have gained significance to improve the efficiency of clustering algorithms. Cluster accuracy results express the effectiveness of the clustering algorithm. In this work, WWO is adopted to improve the accuracy for data clustering. On the basis of WWO, clustering model has been proposed. The proposed algorithm aims to improve data clustering accuracy. Several standard datasets from UCI repository are considered for assessing the simulation results and results are evaluated using accuracy and f-score. The Friedman test is applied for statistical analysis to validate the proposed model. Experimental results proved that proposed clustering model succeeds to achieve higher accuracy rate.
Keywords: Metaheuristics, Data Clustering, Water Wave Optimization, Accuracy, Clustering Model.

1. Introduction

As there is prompt expansion in the amount of data being generated at different levels of society, there is always the requirement for getting the accurate and pertinent information from the data. Immense work is being done for data mining to give the appropriate results with greater efficiency for different purposes. One such purpose is the cluster analysis or clustering. In the modern era of computing and technology, there are several fields such as, biology, software engineering, market segmentation, medical imaging etc. that are widely using the cluster analysis. The aim is to identify the groups among huge amounts of data. Many research studies have contributed in the field of data clustering [1-3] but there is always need for the optimized results with less computation time.

The grouping of data objects is characterized in such manner that similar data objects are kept within same group, whereas dissimilar are kept in other groups or clusters is known to be process of clustering or cluster analysis. Similarity of the objects is given using distance metrics. Various distance metrics used for clustering such as Euclidean, Manhattan, Minkowski and Mahalanobis [4-5]. Euclidean is the popular distance measure used for clustering for low dimensional data and Minkowski for large dimension data [6].

Clustering is a step-by-step process as shown in figure 1. The process starts with the data collection where the objects are differentiated based on their trait values. Data cleaning is the next step taken for primary assessment of the data collected from data warehouse. Representation step consists of representing data in a way so that clustering algorithm can be applied to it. It is also checked that whether the data has the tendency to cluster or not during cluster inclination. Clustering approach is followed where initial parameters, clustering algorithm, are selected and validation techniques are used for validating results. The clustering results can be interpreted so that they can be used in future study.

Different types of clusters are formed based on the usefulness of data analysis. Well-separated clusters are formed using some threshold value for a cluster. Prototype-based cluster consists of objects that are similar on the prototype of cluster than prototype of another cluster. it is considered to be center

based if data has continuous attributes where the mean value is considered. On the other hand, if data is categorical prototype is based on the medoid. Data is denoted in the form of graph, where nodes denote objects and connections among nodes are signified by way of links between objects. The clusters formed are based on the connected components with objects connected to each other fall in the same group and exhibit no connection with objects of other group. These are called as graph-based clusters. Another type of cluster is formed based on the density of the region and so are called as density- based clusters. The regions with high density are parted from regions with low density. Clusters can also be formed with shared property known as conceptual clusters. It consists of objects in cluster sharing same property from whole set of objects.

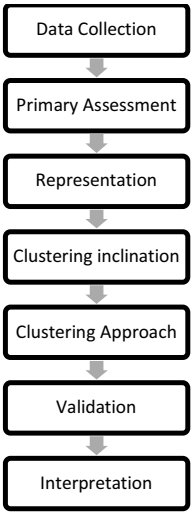


Figure 1. Steps of Clustering.

Clustering is an unsupervised learning method. It is categorized under the combinatorial problems. In order to solve the clustering problems like stuck in local optima, premature convergence, initialization of cluster centroids, handling large amounts of high dimensional data and for improving the efficiency of the clustering algorithm, many mature-inspired clustering algorithms based have been proposed [7-10]. There is an increased concern in the applicability of metaheuristics for clustering algorithms [11-16].

This paper presents a WWO based data clustering model. WWO algorithm is considered for cluster formation segment of the proposed model. This algorithm aims to improve clustering accuracy. The remaining structure of paper as follows. Section 2 reports the related works on data clustering. Section 3 presents motivation and WWO algorithm. The proposed WWO based data clustering model is demonstrated in section 4. The results and performance measure are discussed in section 5. The findings are summarized in section 6.

2. Related Works

This section describes related works reported for data clustering.

Oliveira & Lorena proposed [17] clustering Search (CS) algorithm by combining metaheuristics with clustering. It helped in finding promising areas of local search through cluster of solutions. The cluster center was updated through interaction with neighboring-solutions. If the search metaheuristic being used is the evolutionary algorithm, then approach is known to be Evolutionary Clustering Search (ECS). Best setting of ECS was computed from experiments. The results were computed for best setting of ECS. Comparison of ECS with other approaches reveals that it generates similar and sometimes superior results for considered applications.

To deal with prior determination of clusters, Maulik and Bandyopadhyay [18] proposed GA-

clustering algorithm. Appropriate cluster centers were found in feature space through search ability of GA. Another variable-string-length Genetic Algorithm (GA) was applied to clustering problems by Maulik and Bandyopadhyay [19]. This was done with real encoding of cluster centers coordinates. The proposed algorithm was capable of giving good clustering results and finding proper clustering.

Ghosal et al. discussed about various clustering approaches along with their applications in the various fields [20]. Pandove et al. [21] presented an inclusive comparison of clustering approaches related to big data and high dimensional data. Latest trends in current times, various challenges for handling large dimensional data and application areas have been discussed so as to make it more significant. Wang et al. [22] presented critical review of data mining techniques for knowledge discovery from multiples sources like classification, pattern analysis, clustering and fusion. Challenges for the quality mining of data from multiple sources were discussed.

A comparison of well-known clustering methods was performed by Rodriguez et al. [23] for normally distributed data. Variable sized artificial datasets helped in regulating various attributes and evaluating sensitivity of clustering means. Experiments revealed that simple means with random selection of parameters helps in improving performance as the default configuration was not always accurate. Erdogmus & Kayaalp presented a study of GA, PSO, BBO and GWO for clustering [24]. These swarm-based techniques were applied to Iris dataset and sum of distance values was used as performance evaluation measure. GWO and PSO were having lesser parameters than GA and BBO. So, the results pointed out to be stable and faster for GWO and PSO as compared to GA and BBO. It was concluded that special adaptation can be applied to increase the performance of clustering algorithms.

For reaching global optimum solution Lakshmi et al. [25] combined Crow Search Algorithm with K-means (KM). The proposed algorithm was applied to benchmark datasets-Iris, Breast Cancer, Wine, Glass, Haber- man's Survival and Contraceptive Method Choice (CMC). Results were compared with KM, Genetic-KM, KM++, and PSO-KM. The efficiency of the proposed algorithm was verified using internal, external measure and statistical test. Rafi et al. embedded KM in black hole algorithm for getting optimal result in document clustering [26]. Local and global search were used for adjustment of parameters. Experiments were conducted using datasets- NEWS20, WebKB, Reuter and DOC50. The performance evaluation was done using Silhouette Index and Purity. The results show that Black Hole algorithm gives optimal solution and performs better than K-Mean algorithm.

To deal with data clustering problems, Singh [27] proposed a novel chaotic Harris hawks optimizer (CHHO). This approach begins with identification of candidate solutions as Harris' hawks. Best candidate solution is considered to be the intended prey or optimal solution. To solve the problems of getting trapped in local search domain along with non-linear objective function; chaotic sequences were applied. With this the global and local search capability of HHO was improved that generated the updated solutions called offspring. Objective values from the parent solutions were used to check the quality of solution generated. The proposed approach was evaluated on twelve benchmarks datasets that consisted of four UCI datasets (glass, iris, wine, yeast) and eight shape datasets (flame, jain, R15, D31, Aggregation, Compound, Path Based, Spiral). Six evolutionary algorithms were considered for assessing the proposed approach along with the statistical tests namely Friedman, Iman-Davenport and Holm's test. The experimental results of proposed approach show satisfactory clustering performance. The proposed approach can be used in real-world applications and multiobjective problems. For more effectiveness, Cauchy or gaussian distribution or learning-based approach can be used in conjunction with original HHO algorithm.

Kuwil et al. proposed a new hard partitional clustering algorithm named as Gravity Center Clustering (GCC) algorithm [28]. It depends on the critical distance for defining threshold among clusters. There is no requirement of specifying the initial parameters beforehand for the implementation of GCC. Rather it makes use of two coefficients λ and η, an indicator sigma σ to deals with challenges of noise, outliers and overlapping. GCC provides the robust result and depends on gravity center for cluster formation. Synthetic, and real datasets are adopted for simulation results and datasets are categorized into three groups. The GCC performance was evaluated with KM, K-medians, and K-medoids on the basis of execution time. Limitation of GCC is its computational complexity for large datasets. As a future work,

this limitation can be addressed by developing a model that could handle matrices in smaller parts.

In order to improve the efficiency, hybrid algorithm was proposed using black hole optimization, and K-means by Pal [29]. K-means was used for initializing half of the population was initialized using the values obtained from K-means by multiple runs and remaining population was initialized arbitrarily. After that black hole optimization was applied. The experiment was done using Pima Indian Diabetes, Iris, Lower Back Pain Symptoms, Red Wine Quality, Wine and Glass dataset. The results of proposed hybrid algorithm were compared three popular clustering algorithms. It was pointed out from the results that proposed algorithm did not give the worse results than algorithms in comparison. The advantage of proposed method is that the best object does not move out even if the iteration has given the worst result. The work can be extended for further improvements using some other hybridization methods and by considering more datasets.

Wu et al. developed an adaptive Differential Evolution (DE) algorithm Ant Colony Optimization (ACO), named as ACODE [30]. Main focus was to deal with clustering problems. Proposed algorithm reconfigured four inter-dependent components that is mutation, crossover, scaling factor value and crossover rate through ACO and modelled directed acyclic graph. The reasonable path was optimized using ACO so that four inter-dependent components of DE can be constructed automatically. This helped in search behaviour of DE. The results demonstrated that proposed ACODE gave similar or better results when compared to Original DE, CSO, SL-PSO, DSDE and EPSDE using the eight datasets (Iris, New thyroid disease, Glass identification, Wine, Balance Scale, Lung Cancer, Heart, Landsat). For the evolutionary optimization, time overheads may be decreased in future for clustering data streams.

A new hybrid ACO-ALO algorithm was proposed by Kumar et.al proposed, to escape from local minima problem and reduce intra-cluster distance [31]. Cauchy's mutation operator, and iterated local search algorithm were used in proposed algorithm. The results showed that ACO-ALO outperforms when compared with K-means, and ACO using datasets zoo, iris, wine and glass. This algorithm can be hybridized using neural networks.

Aljarah et. al. hybridized GWO with Tabu search (TS) known as GWOTS, to enhance the performance of original GWO for data clustering problems [32]. GWOTS utilized the idea of adaptive memory from TS in the neighbourhood discovery by avoiding the recently visited solutions. Proposed hybrid algorithm was tested over thirteen various clustering datasets. Overall GWOTS outperforms when compared with other metaheuristic algorithms on the basis of measures namely; SSE, entropy and purity. In future, GWOTS can be used for solving other problems and spatial applications. It can also be investigated for synthetic datasets and perform the parallel computation for reducing run time.

For solving the data clustering problems, an AMADE optimization algorithm was presented by Mustafa et. al. [33]. Adaptive differential evolution (mutation) operator can be hybridized with memetic-algorithm (MA). The proposed algorithm resulted in faster convergence and stabilizing local and global search. The experiment was conducted using six datasets. Results pointed out that AMADE performs better when compared with HyDE, HyGA, DE and GA based on accuracy, average intra-cluster and F-measure. This work can be extended for mixed and categorical datasets by using other data clustering objective functions. Also, it can be used to find the association of different validity measures in multiobjective approaches.

Various evolutionary algorithms namely Biogeography-Based Optimization, GA, GWO and PSO were used over varying size of datasets by Kayaalp & Erdogmus for clustering [34]. The performance of these evolutionary algorithms was compared with K-means using several clustering indexes. The results have pointed out that all these evolutionary algorithms are suitable for small and medium sized datasets but gives promising results for large scale datasets. It was also concluded that K-means performance lies on the number of clusters and data. For large datasets Minkowski distance can be considered rather than eculidean distance. This can be applied to medical datasets in future for case study in clustering.

A study investigated the four hybrid firefly algorithms (FAABC, FAIWO, FAPSO & FATLBO) in [35]. The main focus of study was automatic clustering and unlabelled large datasets. The proposed hybrid algorithms determine the number of clusters automatically. Performance was evaluated using

Compact-Separated, and Davis-Bouldin indices over twelve datasets. Results demonstrated that among the hybrid algorithms FAPSO outperforms and FAIWO emerged to be the least superior method. In future, proposed hybrid FA algorithm is applied for solving different problems

To handle the limitations of FCM, Pantula et. al. presented a Neuro-Fuzzy C-means by adopting ANN [36]. The proposed method constructs a functional map for reducing number of decision variables. This map is constructed between data points and membership function values. ANN helped NFCM for finding optimal number of clusters. NFCM was tested on nine data sets and clustering results were superior.

Zhu [37] proposed new Swarm Clustering Algorithm (SCA) based on PSO. It helped in detecting the number of clusters automatically. Data points were represented as particles and their movement was depended on intrinsic data distribution. By time, particles gathered in numerous areas. Particles within same neighborhood form a cluster. Experiments were conducted for the proposed algorithm using five datasets that is aggregation, flame, R15, D31 and DS850. The performance metrics F-measure, ARI and NMI were used for evaluation of SCA, and results were compared with K-means, DBSCAN, HAC and BIRCH. This can be extended for reducing time complexity by calculating density of particles. The proposed algorithm can be tested for high-dimensional datasets and for automatic estimation of parameters in datasets.

A novel metaheuristic framework including Edge Recombination Operator (ERX) was proposed by Moussa et. al. [38]. GA, artificial immune system and immune-GA were used for identifying number of clusters. The proposed technique was tested on basketball, bolts, pollution, stock and stulong datasets using different sizes and dimensions. The comparison of proposed techniques was done to find the best solution. Mann–Whitney–Wilcoxon rank-sum test was used for statistical validation of number of clusters.

3. Motivation

The traditional clustering algorithms suffer from the various problems like getting stuck in the local optima, premature convergence [39]. Initialization of cluster centers is required to be done while performing clustering operation. These are not able to handle large dimensional data [40,41]. Local and global search are required to be equalized while exploring for optimal solution [42,43]. So, there is always need for improving the proficiency of clustering algorithms. From related works, it is noticed that different metaheuristics methods have been employed for solving various clustering problems. Metaheuristics algorithms do not require preconditions for the objective functions and can also lead to good optimal solutions thereby enhancing the efficiency of clustering algorithm [44].

3.1 Water Wave Optimization (WWO) Method:

This subsection describes about method of WWO.

WWO is a metaheuristic based on water wave motion used to deal with global optimization problems [45]. In WWO, seabed acts as solution-space to a problem and population consists of the waves. Fitness of each wave is measured by distance to sea level. Higher fitness of wave indicates lesser distance to still water level. There is major three operators that is propagation, refraction and breaking operator, used in WWO for finding the optimal solution.

(i) Propagation Operator: Consider height (h) of each wave W_i and λ as wavelength of each wave. h is set as H (maximum height) for each wave and λ is taken as 0.5 during initial stage. Propagation operator is applied to W_i for creating new wave W_i' by shifting dimension(d) of original wave W_i by equation 1.

$$W_i'(d) = W_i(d) + rand(-1,1).\ \lambda . \mathrm{L}(d) \quad (1)$$

$L(d)$ represents length of dimension. If position moves outside feasible range then it is reset to arbitrary position in the range. The fitness value f of W_i and W_i' are compared. If f(W_i') > f(W_i), then old wave is replaced by new wave and height is set to H. If not, old wave remains in the

population and its height is decreased by 1. The step is repeated at each iteration. λ is also updated after each iteration using equation 2.

$$\lambda = \lambda + \alpha^{-(f(w_i)-f_{min}+\varepsilon)/(f_{max}-f_{min}+\varepsilon)} \quad (2)$$

where f_{max} and f_{min} indicates the maximum and minimum fitness values for current population, α represents wavelength decay coefficient and ε is used to avoid divide by zero.

(ii) Refraction Operator: It is applied to waves that tend to or decay to zero after propagation operation shows no improvement. The position of new wave is a Gaussian-random number calculated using mean along with standard deviation by using equation 3.

$$W_i'(d) = N\left(\frac{W_i^*(d) + W_i(d)}{2}, \frac{|W_i^*(d) - W_i(d)|}{2}\right) \quad (3)$$

where W^* represents best known solution, N is a Gaussian-random number. The wave-height is reset to H; and new wavelength is computed by equation 4.

$$\lambda' = \lambda \frac{f(W)}{f(W')} \quad (4)$$

(iii) Breaking Operator: This breaks a wave into series of waves after reaching below certain threshold value. So, main task of breaking operator is to break wave W, when it reaches better position than current best position Wbest. Solitary wave W′ is chosen by adding offset to original position using equation 5.

$$W_d' = W_d + Gaussian(0,1).\beta L(d) \quad (5)$$

Here β indicates breaking coefficient, Gaussian (0,1) is random number generator between 0 and 1. If the solitary wave W' is no better than W_{best}, than W_{best} remains else it is replaced by fittest among the solitary waves.

Many research studies have used WWO for addressing different varieties of problems [46-51]. Different steps and strategies required for adapting WWO were proposed by Zheng et. al. [52]. The adaptation of propagation operator and wavelength according to the problem can assist for efficient problem solving. In this research work, WWO metaheuristic approach is used for improving the efficiency of data clustering.

4. Proposed WWO based Data Clustering Model

This section presents a WWO based data clustering model. It is divided among four segments as i) dataset pre-processing (ii) cluster formation, and (iii) performance evaluation. Figure 2 illustrates the proposed WWO based data clustering model.

(i) *Dataset Pre-processing Segment:* Dataset is loaded into proposed model for performing the clustering operation. Pre-processing and data cleaning are performed by asserting missing values, detecting and removing erroneous records. Information regarding attributes of dataset and class labels is processed from raw dataset. Class label information is removed from pre-processed dataset and handed to subsequent segment for cluster formation.

(ii) *Cluster Formation Segment:* This segment groups the data with same attributes into the same cluster. WWO based approach is used for data clustering. It is efficient and gives more robust and accurate.

(iii) *Performance Evaluation Segment*: This segment evaluates the performance of proposed WWO based data clustering. The main task is to divide data among clusters efficiently and accurately.

Accuracy and F-score are computed for performance evaluation of proposed data clustering model.

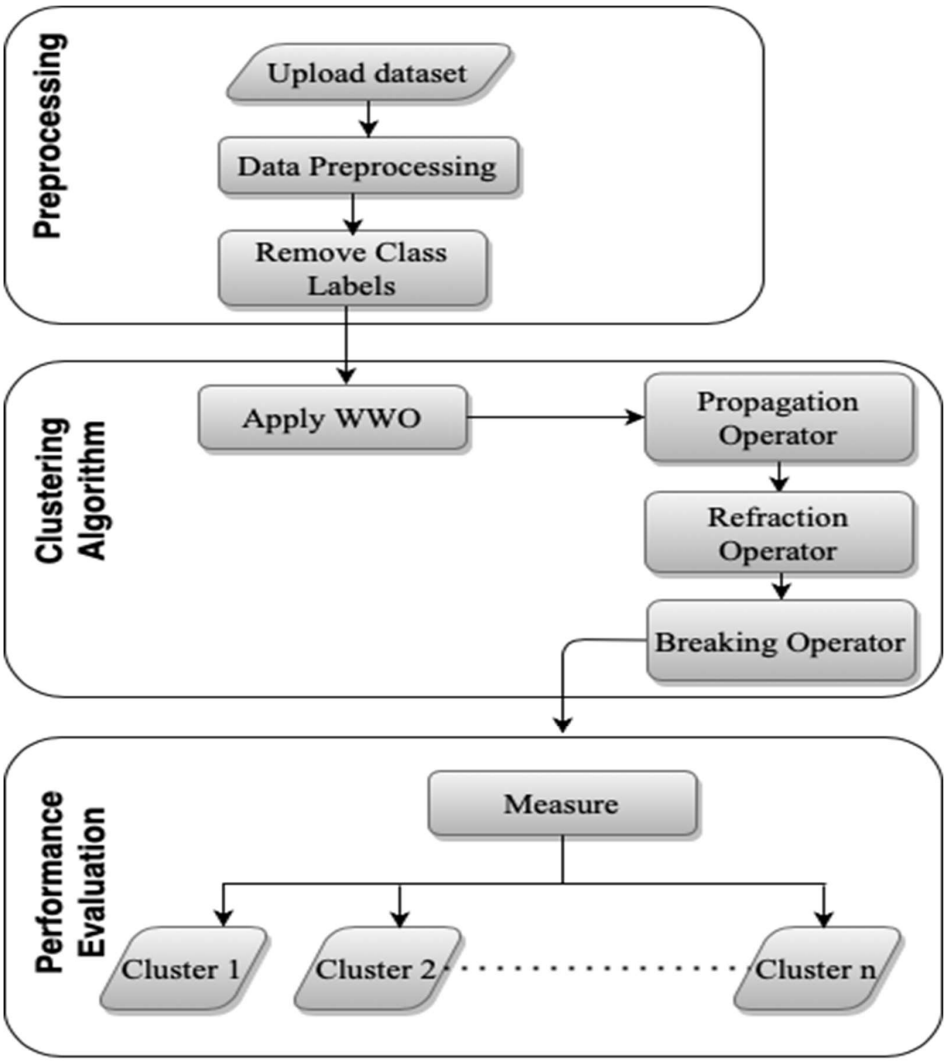


Figure 2. Proposed WWO based Data Clustering Model.

4.1 Steps of WWO algorithm for data clustering model:
The algorithmic steps of proposed WWO algorithm for data clustering are stated in Algorithm 1.

Algorithm 1: Pseudocode of WWO algorithm for data clustering

Step 1: Begin with population of wave (C) such as $C_j \in (i = 1, 2, \ldots, n)$

Step 2: Compute the objective function value with equation 6.

$$D(X_i, C_j) = \sqrt{\sum_{k=1}^{d} (X_{ik}, C_{jk})^2} \quad (6)$$

X_i and C_j denote data points and cluster centers i.e., wave.

Step 3: Allot the data instance to different waves with minimum objective-function value and determine best wave (C_{best}).

Step 4: While (stopping condition is not met), perform following

Step 5: For-each wave $x \in C$

Step 6: Propagate the wave (x) to new position x' using equation 1.

Step 7: If $f(x') > f(x)$, then

Step 8: If $f(x') > f(x^*)$, then

Step 9: Break the wave x' using equation 5.

Step 10: Update the x^* with x' and

Step 11: Replace x with x'.

Step 12: Else, Refract the wave (x) to new x' using equations 3 and 4.

Step 13: Apply wavelength updating step using equation 2.

Step 14: Determine the best wave (C_{best})

Step 15: End while

Step 16: Compute the optimum position of waves

5. Experimental Result

The various experimental results are presented in this section. The proposed WWO based data clustering model is assessed over various datasets. Table 1 presents descriptions of various datasets. Further, accuracy and f-score metrics are employed for assessing the proposed model. The experimental results are compared with various existing models/techniques.

Table 1. Descriptions of different clustering datasets.

Sr. No.	Datasets	Clusters (K)	Instances	Dimension
1	Iris	3	150	4
2	Wine	3	178	13
3	Vowel	6	871	3
3	Balance	3	625	4
5	Glass	7	214	9

5.1 Performance Metrics:
The performance metrics used to evaluate the proposed WWO based diagnostic model are described in this subsection. Accuracy and f-score are chosen as performance metrics.

(i) Accuracy: It determines the correctness of the model as compared to true class labels. Accuracy can be described as the true label of an object "i" to cluster "c" is matched with cluster label

using the map function. Clustering results are accurate when a high value of accuracy is obtained.

$$\text{Accuracy} = \sum_{i=1}^{n} \delta(\text{Truelabel}, \text{map}(c))/n \quad (7)$$

(ii) F-Score: Harmonic mean of precision and recall computes F-score for testing the accuracy of the model.

$$\text{F} - \text{Score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (8)$$

5.2 Experiment:
Simulation and results of the proposed model are discussed in this section.

5.2.1 Simulation Results and Discussion:
The proposed WWO based model is evaluated on five datasets taken from UCI repository. These datasets are (i) iris, (ii) wine, (iii) vowel, (iv) Balance, and (iv) glass. Table 2 displays the experimental results of the proposed WWO based model and various other models/techniques reported in the literature using non-healthcare datasets. It illustrates experimental results using accuracy metric. The proposed clustering model obtains a higher accuracy rate in contrast to other models/ techniques except on Iris and Glass datasets. FCM algorithms provide a higher accuracy rate on Iris and balance datasets respectively, but in contrast, the proposed model performed significantly better on all other datasets. The proposed model results are significantly better in contrast to other algorithms on datasets reported in the literature on accuracy.

Table 2. Demonstrate experimental results of the proposed WWO model in contrast to other models/techniques based on accuracy performance metric.

Dataset	FCM	Fuzzy-PSO	KFCM	PSO	K-means	GA	WWO
Iris	**99.33**	67.33	83.33	84.13	78.53	78.34	85.23
Wine	70.22	70.25	71.91	67.94	67.61	65.73	**71.82**
Vowel	78.68	76.27	63.31	84.04	73.45	84.7	**85.11**
Balance	**86.64**	84.12	71.12	85.76	84.99	78.62	85.93
Glass	60.35	60.29	50	58.02	62.45	57.27	**64.73**

F-score is also considered an important performance measure to validate the proposed model. It considers precision along with recall to evaluate performance comparison to accuracy as a metric. Table 3 illustrate experimental results based on the F-score metric. The proposed diagnostic model obtains a higher F-score rate in contrast to other models/ techniques. The fuzzy-PSO and K-means algorithms provides some higher results on the F-score rate for vowel and balance datasets, but in contrast, the proposed model performed significantly better on all other datasets. The proposed model results are significantly better in contrast to algorithms on datasets reported in the literature for F-score. From observation, it is clear that proposed WWO based model delivers significant accurate results in contrast to other models/techniques in literature for benchmark datasets. Hence, the proposed WWO algorithm is robust, viable, and efficient algorithm.

Table 3. Demonstrate experimental results of proposed WWO model in contrast to other models/techniques based on F-score performance metric.

Dataset	FCM	Fuzzy-PSO	KFCM	PSO	K-means	GA	WWO
Iris	0.778	0.790	0.783	0.782	0.778	0.776	**0.784**
Wine	0.520	0.523	0.521	0.518	0.521	0.515	**0.522**
Vowel	0.649	0.651	0.646	0.647	**0.652**	0.647	0.651
Balance	0.734	**0.746**	0.727	0.727	0.724	0.716	0.730
Glass	0.548	0.568	0.493	0.573	0.563	0.561	**0.576**

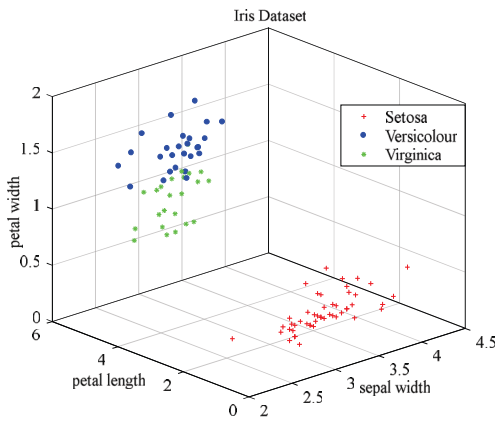


Figure 3. (a)

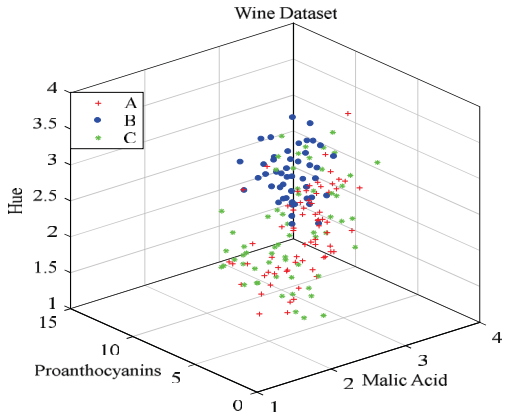


Figure 3. (b)

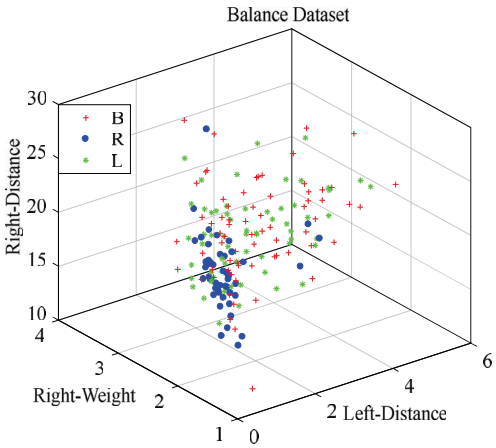


Figure 3. (c)

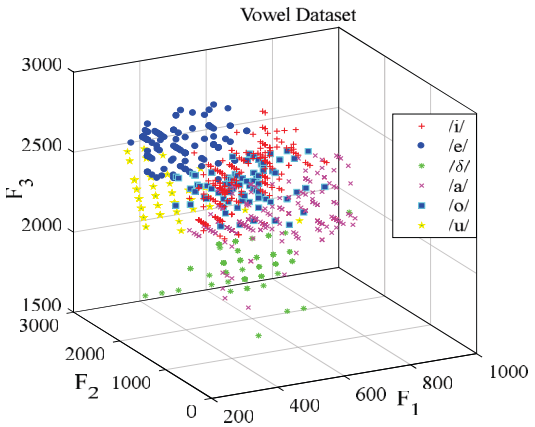


Figure 3. (d)

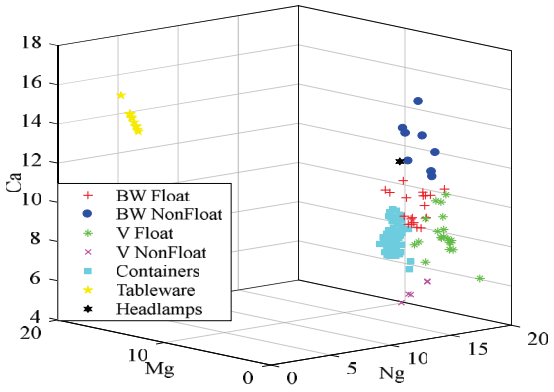


Figure 3 (e)

Figure 3. (a-e). Demonstrate clustering of data objects using WWO based Model.

Figure 3 (a-e) demonstrates the clustering of data objects based on the WWO based model on datasets. Figure 3 (a) considers the iris dataset and proposed WWO model groups data into 3 clusters (i) setosa (ii) versicolour and (iii) virginica. Figures 3 (b & c) depict the clustering of wine and balance datasets. The proposed model categorizes the wine dataset into three clusters i.e. (i) A, (ii) B and (iii) C. Balance dataset is divided among three clusters that is B, L and R. Figure 3(d) demonstrates the clustering results of vowel dataset. The proposed WWO model divides the data objects into seven clusters i.e. (i) /i/, (ii) /e/, (iii) /δ/, (iv) /a/, (v) /o/, and (vi) /u/. It is also observed that the proposed model separates data objects of the vowel dataset effectively. Figure 3(e) illustrates considers glass dataset. The proposed model divides data objects into seven different clusters. It is seen that one cluster is linearly separable from the other six clusters. Whereas, the rest of the six are non-linearly separable. The proposed model effectively performs the clustering task and it is an effective model for the clustering of data objects.

5.2.2 Statistical Results:

Statistical analysis is performed in this subsection. It is done to validate the performance of the proposed WWO algorithm using non-healthcare datasets. Tables 4-5 illustrate the results of Friedman statistical test via accuracy parameter. Table 4 shows the average ranking of the proposed WWO algorithm and the rest of the clustering algorithms. It is observed that the proposed WWO clustering model gets the first rank with value 1.4 as compared to other algorithms. Whereas, GA algorithm achieves the lowest rank with value 5.4 among all algorithms. Table 5 presents the statistics of Friedman test. The critical-value of Friedman test (0.05, 6) is 14.057143 whereas, the p-value is 0.029004. The null hypothesis (H0) is rejected at the confidence level of 0.05. The critical value is 12.591587. So, there is a significant change for the performance of proposed WWO algorithm and clustering algorithms in comparison.

Table 4. Average ranking of clustering algorithms using an accuracy metric based on the Friedman statistical test.

FCM	Fuzzy-PSO	KFCM	PSO	K-means	GA	WWO
2.6	4.8	5.2	4	4.6	5.4	1.4

Table 5. Friedman test results of accuracy metric.

Method	Statistical Value	*p*-Value	Hypothesis
Friedman Test	14.057143	0.029004	Rejected

Results of Friedman statistical test using the F-score parameter are reported in Tables 6-7. Table 6 depicts the average ranking of proposed WWO algorithm and the rest of the clustering algorithms. The proposed WWO algorithm acquires first rank i.e., 1.7 in contrast to other algorithms. It is also noted that the GA algorithm achieves the lowest rank i.e., 6.3 among all algorithms. Table 7 presents the statistics of the Friedman test. The critical value for Friedman test (0.05, 6) is 14.404412; whereas p-value is 0.025431. So, the null hypothesis (H0) is rejected at the confidence level of 0.05. The critical value is 12.591587. It is stated that the performance of proposed WWO algorithm significantly differs than other compared clustering algorithms

Table 6. Average ranking of clustering algorithms by employing F-score metric based on the Friedman statistical test.

FCM	Fuzzy-PSO	KFCM	PSO	K-means	GA	WWO
4.5	2.7	4.7	4.2	3.9	6.3	1.7

Table 7. Friedman test results of F-score metric.

Method	Statistical Value	*p*-Value	Hypothesis
Friedman Test	14.404412	0.025431	Rejected

6. Conclusion

In this work, WWO based data clustering model is proposed for dividing the data among clusters. The functioning of proposed model is distributed among three segments (i) preprocessing, (ii) cluster formation, and (iii) performance evaluation segment. In the cluster formation segment, WWO algorithm is adopted for data clustering where different classes are determined for the datasets. The model helps to improve clustering accuracy and efficiency. The proposed model is implemented on five benchmark datasets and performance is evaluated as measure of accuracy and F-score. The experimental results of proposed WWO algorithm are compared with popular metaheuristic models/techniques from literature. It is observed that the proposed clustering model achieves higher accuracy and F-score rate. It is concluded that the proposed WWO model obtains better clustering results for most of the datasets. It is a competent and efficient model for clustering.

References

[1] Wu W Xiong H Shekhar S 2013 Clustering and Information Retrieval. *Springer Science & Business Media*, 2013;. https://doi.org/10.1007/978-1-4613-0227-8
[2] Mu¨ller H Hamm U 2013 Stability of market segmentation with cluster analysis—A methodological approach. Food Qual Prefer. 2014; 34: 70–78. https://doi.org/10.1016/j.foodqual.2013.12.004
[3] Abbasi AA Younis M 2007 A survey on clustering algorithms for wireless sensor networks. Comput Com-mun.; 30: 2826–2841. https://doi.org/10.1016/j.comcom.2007.05.024
[4] Greche L Jazouli M Es-Sbai N Majda A Zarghili 2017 A Comparison between Euclidean and Manhattan distance measure for facial expressions classification. In: 2017 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), Fez. pp. 1-4
[5] Ramadas M & Abraham A 2019 Metaheuristics and Data Clustering. In Metaheuristics for Data Clustering and Image Segmentation (pp. 7-55). Springer, Cham
[6] Jain A K Murty M N & Flynn P J 1999 Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323.
[7] Sharma M & Chhabra J K 2019 An efficient hybrid PSO polygamous crossover based clustering algorithm. Evolutionary Intelligence, 1-19.

[8] Gupta Y & Saini A 2019 A new swarm-based efficient data clustering approach using KHM and fuzzy logic. Soft Computing, 23(1), 145-162.
[9] Mageshkumar C Karthik S & Arunachalam V P 2019 Hybrid metaheuristic algorithm for improving the efficiency of data clustering. Cluster Computing, 22(1), 435-442.
[10] Zhu E & Ma R 2018 An effective partitional clustering algorithm based on new clustering validity index. Appl. Soft Comput., 71, 608-621.
[11] José-García A & Gómez-Flores W 2016 Automatic clustering using nature-inspired metaheuristics: A survey. Applied Soft Computing, 41, 192-213.
[12] Nanda S J & Panda G 2014 A survey on nature inspired metaheuristic algorithms for partitional clustering. Swarm and Evolutionary computation, 16, 1-18.
[13] Blum C Puchinger J Raidl G R & Roli A 2011 Hybrid metaheuristics in combinatorial optimization: A survey. Applied soft computing, 11(6), 4135-4151.
[14] Elavarasi S A Akilandeswari J & Sathiyabhama B 2011 A survey on partition clustering algorithms. International Journal of Enterprise Computing and Business Systems, 1(1).
[15] Dokeroglu T Sevinc E Kucukyilmaz T & Cosar A 2019 A survey on new generation metaheuristic algorithms. Computers & Industrial Engineering, 137, 106040.
[16] Kumar A Kumar D & Jarial S K 2017 A review on artificial bee colony algorithms and their applications to data clustering. Cybernetics and Information Technologies, 17(3), 3-28.
[17] Oliveira A C & Lorena L A 2007 Hybrid evolutionary algorithms and clustering search. In HYBRID evolutionary algorithms (pp. 77-99). Springer, Berlin, Heidelberg.
[18] Maulik U & Bandyopadhyay S 2000 Genetic algorithm-based clustering technique. Pattern recognition, 33(9), 1455-1465.
[19] Bandyopadhyay S & Maulik U 2001 Nonparametric genetic clustering: comparison of validity indices. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 31(1), 120-125.
[20] Ghosal A Nandy A Das A K Goswami S & Panday M 2020 A short review on different clustering techniques and their applications. In Emerging Technology in Modelling and Graphics (pp. 69-83). Springer, Singapore.
[21] Pandove D Goel S & Rani R 2018 Systematic review of clustering high-dimensional and large datasets. ACM Transactions on Knowledge Discovery from Data (TKDD), 12(2), 1-68.
[22] Wang R Ji W Liu M Wang X Weng J Deng S ... & Yuan C A 2018 Review on mining data from multiple data sources. Pattern Recognition Letters, 109, 120-128.
[23] Rodriguez M Z Comin C H Casanova D Bruno O M Amancio D R Costa L D F & Rodrigues F A 2019 Clustering algorithms: A comparative approach. PloS one, 14(1), e0210236.
[24] Erdogmus P & Kayaalp F 2020 Introductory Chapter: Clustering with Nature-Inspired Optimization Algorithms. In Introduction to Data Science and Machine Learning. IntechOpen.
[25] Lakshmi K Visalakshi N K & Shanthi S 2018 Data clustering using k-means based on crow search algorithm. Sādhanā, 43(11), 190.
[26] Rafi M Aamer B Naseem M & Osama M 2018 February Solving document clustering problem through meta heuristic algorithm: black hole. In Proceedings of the 2nd International Conference on Machine Learning and Soft Computing (pp. 77-81).
[27] Singh T 2020 A chaotic sequence-guided Harris hawks optimizer for data clustering. Neural Computing & Applications.
[28] Kuwil, F. H., Atila, Ü., Abu-Issa, R., & Murtagh, F. (2020). A novel data clustering algorithm based on gravity center methodology. Expert Systems with Applications, 156, 113435.
[29] Pal S S & Pal S 2020 Black Hole and k-Means Hybrid Clustering Algorithm. In Computational Intelligence in Data Mining (pp. 403-413). Springer, Singapore.
[30] Wu G Peng W Hu X Wang R & Chen H 2020 Configuring differential evolution adaptively via path search in a directed acyclic graph for data clustering. Swarm and Evolutionary Computation, 100690.
[31] Mageshkumar C Karthik S & Arunachalam V P 2019 Hybrid metaheuristic algorithm for improving the efficiency of data clustering. Cluster Computing, 22(1), 435-442.
[32] Aljarah I Mafarja M Heidari A A Faris H & Mirjalili S 2020 Clustering analysis using a novel locality-informed grey wolf-inspired clustering approach. Knowledge and Information Systems, 62(2), 507-539.
[33] Mustafa H M Ayob M Nazri M Z A & Kendall G 2019 An improved adaptive memetic differential evolution optimization algorithms for data clustering problems. PloS one, 14(5), e0216906.
[34] Kayaalp F & Erdogmus P 2020. Benchmarking the Clustering Performances of Evolutionary Algorithms: A Case Study on Varying Data Size. IRBM.
[35] Ezugwu A E S Agbaje M B Aljojo N Els R Chiroma H & Abd Elaziz M 2020 A Comparative Performance Study of Hybrid Firefly Algorithms for Automatic Data Clustering. IEEE Access, 8, 121089-121118.
[36] Pantula P D Miriyala S S & Mitra K 2020 An Evolutionary Neuro-Fuzzy C-means Clustering Technique. Engineering Applications of Artificial Intelligence, 89, 103435.
[37] Zhu W Luo W Ni L. & Lu N 2018 November Swarm clustering algorithm: Let the particles fly for a while. In 2018 IEEE Symposium Series on Computational Intelligence (SSCI)(pp. 1242-1249). IEEE.
[38] Moussa D A Eissa N S Abounaser H & Badr A 2018 Design of Novel Metaheuristic Techniques for Clustering. IEEE Access, 6, 77350-77358.
[39] Yu H Wen G Gan J Zheng W & Lei C 2020 Self-paced learning for k-means clustering algorithm. Pattern Recognition

Letters, 132, 69-75.
[40] Assent I 2012 Clustering high dimensional data. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(4), 340-350.
[41] Mittal M Goyal L M Hemanth D J & Sethi J K 2019 Clustering approaches for high-dimensional databases: A review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), e1300.
[42] Gupta Y & Saini A 2019 A new swarm-based efficient data clustering approach using KHM and fuzzy logic. Soft Computing, 23(1), 145-162.
[43] Ilango S S Vimal S Kaliappan M & Subbulakshmi P 2019 Optimization using artificial bee colony based clustering approach for big data. Cluster Computing, 22(5), 12169-12177.
[44] Sahoo G 2017 A two-step artificial bee colony algorithm for clustering. Neural Computing and Applications, 28(3), 537-551.
[45] Zheng Y J 2015 Water wave optimization: a new nature-inspired metaheuristic. Computers & Operations Research, 55, 1-11.
[46] Ibrahim A M Tawhid M A & Ward R K 2020 A binary water wave optimization for feature selection. International Journal of Approximate Reasoning, 120, 74-91.
[47] Shao Z. Pi D & Shao W 2018 A novel discrete water wave optimization algorithm for blocking flow-shop scheduling problem with sequence-dependent setup times. Swarm and Evolutionary Computation, 40, 53-75.
[48] Siva M Balamurugan R & Lakshminarasimman L 2016 Water wave optimization algorithm for solving economic dispatch problems with generator constraints. International Journal of Intelligent Engineering and Systems, 9(4), 31-40.
[49] Zhao F Zhang L Liu H Zhang Y Ma W Zhang C & Song H 2019 An improved water wave optimization algorithm with the single wave mechanism for the no-wait flow-shop scheduling problem. Engineering Optimization, 51(10), 1727-1742.
[50] Manshahia M S 2017 Water wave optimization algorithm based congestion control and quality of service improvement in wireless sensor networks. Transactions on Networks and Communications, 5(4), 31-31.
[51] Singh G Rattan M Gill S S & Mittal N 2019 Hybridization of water wave optimization and sequential quadratic programming for cognitive radio system. Soft Computing, 23(17), 7991-8011.
[52] Zheng Y J Lu X Q Du Y C Xue Y & Sheng W G 2019 Water wave optimization for combinatorial optimization: Design strategies and applications. Applied Soft Computing, 83, 105611.
[53] Soltanian A Derakhshan F & Soleimanpour-Moghadam M 2018 March MWWO: Modified water wave optimization. In 2018 3rd Conference on Swarm Intelligence and Evolutionary Computation (CSIEC) (pp. 1-5). IEEE.
[54] Alireza A L F I 2011 PSO with adaptive mutation and inertia weight and its application in parameter estimation of dynamic systems. Acta Automatica Sinica, 37(5), 541-549.
[55] Rana S Jasola S & Kumar R 2011 A review on particle swarm optimization algorithms and their applications to data clustering. Artificial Intelligence Review, 35(3), 211-222.
[56] Hematabadi A A & Foroud A A 2019 Optimizing the multi-objective bidding strategy using min–max technique and modified water wave optimization method. Neural Computing and Applications, 31(9), 5207-5225.