A Simplex Method based Water Flow Optimizer Algorithm for Effective Data Clustering

Prateek Thakral[1*] and Yugal Kumar[1,2]

ARTICLE INFO

Article history:
Received: 1 September 2023
Revised: 22 December 2023
Accepted: 19 March 2024
Online: 31 March 2025

Keywords:
Clustering
Meta-heuristics
Water Flow Optimizer
Simplex method
Evolutionary Algorithm

ABSTRACT

Clustering is a prominent technique that demonstrates its usability in diverse fields such as data analytics, information retrieval, social mining, image analysis, etc. This technique extracts constructive information from a pre-defined collection of data and groups the similar data into the same cluster. Many methods based on various clustering methodologies have been reported in the literature but partitional clustering methods are widely utilized due to their unfussiness and ease of accomplishment. It is seen that traditional algorithms like k-means exhibit several drawbacks like being caught in local optima, dependent on the initial solution, convergence rate and population diversity. These shortcomings of clustering algorithms are also handled through heuristic techniques. This work presents a new clustering algorithm called water flow optimizer. Further, it is noticed that water flow optimizer (WFO) entraps in local optima sometimes and in turn it converges on local solution instead of global one. This issue is resolved through integration of the simplex method into WFO algorithm, called SM-WFO and its performance is appraised using several well-known datasets and results are contrasted with popular measures. The results revealed that SM-WFO gets superior results than other algorithms. SM-WFO algorithm also accomplishes higher average rank (1.58) using all datasets.

1. INTRODUCTION

Data mining is a process that can be used to analyze the massive amounts of data and it comprises of complex algorithms with conventional data analysis techniques. It can also be expressed as a process to extract the knowledge from large volumes of data [1]. It is also considered as one of the important task in knowledge discovery process. The main reason behind the success of data mining techniques is to handle the heterogeneous and complex data, data scalability, data sets with high dimensionality, data ownership and distribution. Data mining tasks can be categorized as either predictive or descriptive tasks. The goal of a descriptive task is to identify patterns such as clusters, correlations, anomalies that explain the underlying relationships among data, while the goal of a predictive task is to calculate the value of a specific attribute based on the value of other attributes. Classification and regression are two popular predictive modeling tasks whereas cluster analysis, association analysis and anomaly detection describes as descriptive task [1, 2]. Further, clustering is a primary approach in the field of unsupervised machine learning that can determine the group of clusters. These groups consist of similar data points based on various features or characteristics they share. The major objective of data clustering is to find patterns and structures within a dataset, enabling insights and understanding of the inherent relationships among data points [3]. Clustering is utilized in diverse applications areas such as marketing, social network analysis, customer segmentation, anomaly detection, image analysis, etc. Broadly, the clustering is having two popular categories known as partitional and hierarchical [3]-[5]. Hierarchical clustering creates a hierarchical representation of the data by repeatedly merging or dividing clusters. It creates a dendrogram, which depicts the order of cluster formations in a tree-like structure. It divides into agglomerative and divisive clustering. The data points are split into a predetermined number of clusters using partitional clustering. Although, the K-Means [5] is the oldest and most prominent partitional clustering algorithm, in literature, several techniques have been presented for partitional clustering. The decision between hierarchical and partitional clustering depends on the type of data, the level of interpretability needed, computational resources that are available, number of clusters. The partitional clustering is effective for larger datasets and clearly defined clusters, while hierarchical clustering is excellent for understanding data structure. It is seen that clustering is an important

[1]*Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat, Solan, Himachal Pradesh, India.*
[2]*School of Technology Management and Engineering (STME), NMIMS Chandigarh Campus, Punjab, India.*
**Corresponding author:* Prateek Thakral; Phone: +91-8083-000-073; Email: 18.prateek@gmail.com, yugalkumar.14@gmail.com.

aspect in the analysis of data that are collected from a variety of realm. The numerous mathematical and classical algorithms have been developed by researchers to address real-world clustering issues, but due to an intrinsic characteristic of these algorithms, these algorithms can exhibit premature converge and local optima issues whereas the traditional clustering techniques have problems like overdependence on the introductory solution quality and trap in local optima. Recently, meta-heuristic algorithms gain wide popularity among research community to get the optimal solution for large number of optimization problems. It is also observed that meta-heuristic algorithms [6] can have specific local search and randomization trade-offs. Numerous meta-heuristic algorithms under various broad categories have been developed in the past decade for cluster analysis. Few are summarized as PSO [7, 8], Magnetic Optimization Algorithm (MOA)[9], Artificial Bee Colony (ABC)[10], BB-BC [11], CSS [12], etc. These algorithms also consist of several additional flaws, including population diversity, trade-off, convergence rate, and occasionally getting stuck in local optima. With the use of additional meta-heuristic algorithms, the aforementioned flaws in meta-heuristic algorithms can be eliminated. It is also observed that weakness of a meta-heuristic algorithm can be replaced with the strength of another meta-heuristic algorithm to produce improved clustering results. In turn there is always a purview to develop a novel clustering algorithm that can produce the optimal clustering results with a variety of datasets. This works also presents an algorithm, called water flow optimizer (WFO) [13] for clustering. This algorithm is energized by the characteristic of water flow in nature. However, it is noticed that that the WFO have good optimization ability, but several shortcomings are related to it like initialization of initial population, lack of balance between search mechanisms and trap in local optima sometimes due to one way search mechanism [14]-[16].It is observed that local optima is one of the prominent issue that can alter the conduct of the WFO algorithm among all issues. Hence this work addresses the local optima issue of WFO algorithm through simplex method and introduces a novel algorithm comprises of simplex method (SM) and WFO for clustering. The main points of this work are listed below:

- To propose a novel meta-heuristic called SM-WFO algorithm of data clustering.
- The local optima affair of the WFO algorithm is handled through simplex method.
- Twelve benchmark datasets are taken for evaluating the performance of SM-WFO. The clustering results are compared to nine well-known algorithms.
- The results are assessed using accuracy rate (AR), detection rate (DR), intra and SD.
- The results showed that SM-WFO gets superior results with most of datasets.

The organization of the paper is expressed as recent works in the field of data clustering are highlighted in section 2. The proposed SM-WFO is presented in section 3. The experimental results of the SM-WFO algorithm are illustrated in section 4. Section 5 concludes the whole work.

2. RELATED WORK

The recent work related to clustering problems using different meta-heuristic algorithms are discussed in this section. In ref. [17], a new algorithm called local neighbor spider monkey optimization (LNSMO) is developed that improves the search process of SMO algorithm. The local leader phase of SMO integrates the neighbor search to narrow the search space. The LNSMO global leader phase is further enhanced using a chaotic element. The LNSMO performance is evaluated using eleven data sets and compared to five conventional methods, including several meta-heuristic algorithms. The findings stated that LNSMO offers a better outcome using popular clustering measures.

Singh et al. [18] introduced the EWO algorithm to handle clustering problems. Two more operational procedures are added in the WOA to improve its performance. To expand the search space and quicken convergence, EWO algorithm is integrated with WWO algorithm position update equations. To deal with the local optima scenario, the Tabu and neighborhood search technologies were implemented. A simulation-based experiment employing eight standard datasets is used to gauge the effectiveness of the proposed EWOA, and the findings are then contrasted with those of seven other clustering algorithms/techniques. Several popular metrics are used to evaluate the performance of each algorithm.

For competent data clustering, the variable neighborhood strategy-based firefly algorithm (VNS-FA) is described [19]. The firefly algorithm (FA) is seen to converge on premature solutions as a result of a deficiency in exploitation ability. Additionally, FA incorporates variable neighborhood strategy (VNS) to address the aforementioned problems. Utilizing eight well-known clustering datasets, the effectiveness of the proposed VNS-FA is computed. Using the intra-cluster distance, internal CH metric, entropy, and F-measure parameters, the results are evaluated. The findings showed that for the majority of the datasets, the proposed VNS-FA approach yields superior outcomes.

For clustering and dynamic social networks, the IGWO (improved Grey Wolves Optimization) method [20] is described. The objective of this effort is to increase clustering problem accuracy rates. A label propagation technique is incorporated into the grey wolf optimization (GWO) algorithm to accomplish the same goal. Six well-known datasets based on normalized mutual information (NMI), intra-cluster distance, and error rate metrics are

used to assess the performance of the IGWO. The findings demonstrated that the IGWO method, when compared to other clustering algorithms, achieves a higher NMI rate. Additionally, it can be noted that the suggested IGWO algorithm outperforms existing algorithms in terms of intra-cluster distance and minimum error rate.

For the purpose of dealing with partitional clustering, a cat-based meta-heuristic approach is reported [21]. Before using this approach, various changes are made to the cat algorithm to address diversity issues, premature convergence, and tradeoff mechanisms between local and global searches. To address the aforementioned problems, enhanced search mechanisms, an accelerated velocity equation, and neighborhood mechanisms are then proposed. On the basis of intra cluster distance and f-measure parameters, the effectiveness of the cat algorithm is evaluated over eight clustering datasets. It is determined that the proposed cat algorithm produces superior f-measure rates than other algorithms while having the smallest intra cluster distance for the majority of datasets.

An enhanced water wave optimization (WWO) technique was presented by Kaur and Kumar [22] in order to produce more effective and encouraging clustering outcomes. These enhancements are explained in terms of a revised decay operator and global search mechanism. The purpose of the decay operator is to solve the WWO algorithm's premature convergence problem. Thirteen standard clustering datasets are used to assess the effectiveness of the WWO method in terms of parameters such as accuracy and F-score. The simulation results are compared to a number of available clustering algorithms, and it is found that the suggested WWO clustering method outperforms the current clustering algorithms in terms of accuracy and F-score rates for the majority of clustering datasets.

In order to address the problems with the k-mean methodology, Kushwaha et al. [23] introduced an electromagnetic field optimization (EFO) method. Poor initial centroid selection causes the k-mean method to become trapped in local optimums. To overcome this challenge, the optimal initial centroid for the K-mean technique is identified using the EFO algorithm. Additionally, it is claimed that due to mechanisms of attraction and repulsion, the EFO algorithm may not remain in local optima. The strength of the suggested clustering approach based on NMI, rand index (RI), and purity is evaluated using a number of well-known datasets. The findings indicated that the suggested algorithm attains significantly better clustering results than techniques in the same class.

Hashemi et al. [24] created an upgraded PSO algorithm to carry out the clustering in large data. The proposed method uses the multi-start pattern reduction strategy to reduce the calculation time. A reduction operator is used to shorten the clustering process, and the multi-start operator is used to guarantee population diversity and local minima. In order to assess the performance of the suggested method in terms of precision and execution time, six clustering datasets are taken into account. The outcomes demonstrated that the proposed PSO approach yields superior clustering results.

Kuo et al. [25] proposed the FPCOM fuzzy c-means and fuzzy c-ordered means technique to mitigate the impact of outliers in clustering. Additionally, the SCA's SCA-FPCOM is used to set the parameters and initial centroids. Ten clustering datasets chosen from the UC Irvine (UCI)repository are used to assess the efficacy of the suggested SCA-FPCOM based on the measurements of the rand index and the Silhouette coefficient. The outcomes showed that SCA-FPCOM outperforms other algorithms. It is clear that the bulk of clustering algorithms validate the clusters using single-featured datasets, distance, density, and features. The concept of created clusters may be violated by the inclusion of semantic data, though. However, substantial clusters can be produced by evolutionary and statistical operators using integrative approaches.

A multi-objective clustering algorithm knows as MOVPS, which is based on a vibrating particle system, was presented by Kaur and Kumar [26] for effective data clustering. The current scenario considers two different objective functions named as intra-cluster variance and connectedness. To achieve favorable clustering outcomes, the VPS algorithm is employed to optimize the preceding objectives. By contrasting the clustering results with those of several multi-objective and single-objective clustering algorithms from the literature, the effectiveness of the MOVPS algorithm is examined using a number of benchmark datasets. The outcomes of the simulation shown that, when compared to current multi-objective and single-objective clustering algorithms, the suggested MOVPS algorithm greatly lifts the clustering results.

A clustering method named GWO-TS, which is presented in [27], relies on the Tabu search (TS) and the grey wolf optimizer (GWO) for efficient cluster analysis. The suggested method addresses GWO issues including premature convergence and local optima using the tabu search strategy. Thirteen datasets are used to evaluate the effectiveness of the proposed GWO-TS. Utilizing the settings for purity, SSE, and entropy, the simulation results are assessed. It is asserted that the proposed clustering method generates results of superior quality when compared to existing techniques.

For the purpose of handling partitional clustering issues, an enhanced gradient-based clustering algorithm (IGE) is proposed in [28]. Instead of concentrating on a single objective function, the suggested algorithm has two objective functions. Based on Pareto optimality, the dominant and non-dominant set of solutions is produced. SSE, SSB and accuracy parameters are used to evaluate the

simulation results. They are compared to PSO, GA, ABC, and DE algorithms. IGE allegedly yields more encouraging clustering outcomes than other clustering algorithms.

For effective cluster analysis, Duan et al. [29] integrated the gradient-based strategy with the elephant herding optimization technique, known as GBEHO. The initial centroids are chosen using a variety of algorithms. Additionally, the Gaussian chaos map is used to handle the elephant herding optimization (EHO) trade-off problem. The population is updated using the wandering and variation operators. The effectiveness of the suggested GDEHO algorithm is assessed based on accuracy, detection rate, specificity, and f-measure parameters. According to reports, GBEHO delivers more reliable findings than other algorithms.

Kuo and Wang [30] created a hybrid k-prototype clustering approach based on improved SCA. The SCA algorithm is used to determine the best attribute weighting and initial attribute selection. To further improve clustering results, the k-prototype method integrates many mutation strategies, including Gaussian, Cauchy, and single-point. Ten datasets are selected from the UCI repository to test the accuracy and Cohen kappa of the k-prototype algorithm. The results demonstrated that the suggested k-prototype algorithm offers superior clustering outcomes to other algorithms.

Singh et.al introduced a meta-heuristic algorithm called as artificial chemical reaction optimization [31] (ACRO) algorithm inspired from the chemical reaction process. This method uses reactants, which are uniformly determined from the search space to find the best solution. Reactants can also be used to symbolize the best possible solution to the difficulties. The ACRO algorithm's primary task is to calculate the ideal cluster centroid for partitional clustering issues. The suggested algorithm's performance is evaluated on a number of real-world clustering applications and contrasted with cutting-edge clustering techniques. It can be shown from the simulation results that, when compared to existing clustering methods, the suggested technique produces superior clustering results. Table 1 summarizes details of clustering algorithm using meta-heuristic algorithms.

3. PROPOSED SM-WFO ALGORITHM

This section presents the simplex method based WFO algorithm for cluster analysis. Recently, a new meta-heuristic algorithm, called water flow optimizer (WFO) is developed for solving different kind of optimization problems [13, 14]. The working of the algorithm is described by two operators i.e. laminar and turbulent flows. The aim of these operators is either to minimize or maximize the objective function. It is noticed that WFO achieves good optimization capability, but several shortcomings are associated with it [15, 16]. Some of these are listed as selection of initial population for WFO algorithm, lack of balance between the search mechanisms, sometimes get stuck in local optima due to one side search mechanism and convergence issue etc. Further, it is identified that local optima and convergence are more promising ones among aforementioned issues. This work considers the local optima issue of WFO and this issue is addressed through simplex method. This method is extensively adopted for improving the performance of the meta-heuristic algorithms [34] – [36]. In this work, the simplex method is utilized to generate the new position of water particles for avoiding the local optima.

3.1. Simplex Method

This subsection discusses the simplex method to overcome the local optima issue. In WFO algorithm, the laminar phase consists of one-way search strategy to generate the new position of water particles. The working of this phase is described through the laminar flow and as per this flow, the water particles move in straight parallel lines. It is also revealed that the particles far away from walls and obstacles can move faster compared to close to walls and obstacles. In turn, an optimal position of water particles is not generated sometimes and algorithm can converge on local solution instead of global one, called local optima. This issue of the WFO algorithm is handled through the simplex method. This method was developed by Spendley et al. [35] and can be described as number of points equal to one more than the number of dimensions in the search space. This method [36]-[38] is widely adopted in the field of meta-heuristic for generating the optimal position and also to get rid of local optima issue of algorithms. This study also examines the worth of the simplex method in resolving the local optima issue of WFO algorithm. The schematic process of the simplex method is depicted into Fig. 1.

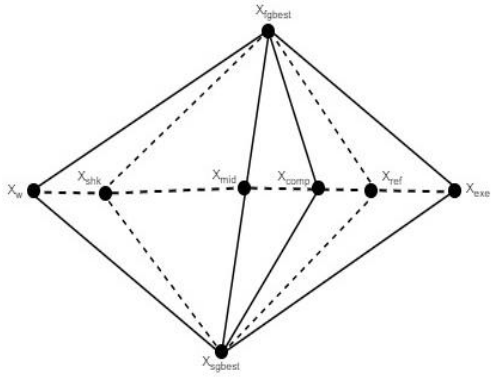

Fig. 1. Schematic process of the simplex method.

Table 1: Related work reported on clustering using meta-heuristic algorithms

Ref.	Meta-heuristic Techniques	Objective Function	Datasets
[17]	Spider Monkey Optimization (2023)	Sum of within cluster distance (SWCD)	Iris, Glass, Cancer, CMC, Wine, Seed, Heart, Bupa, Magic, HTRU2, Haberman
[18]	Enhanced Whale Optimization algorithm, Tabu Search (2023)	Sum of Squared Error	Iris, Wine, Cancer, CMC, LR, Glass, ISOLET, Thyroid
[19]	Firefly Algorithm (2022)	Euclidean Distance	Obesity, , Vehicle, Ecoli, Glass, CMC, Segment, Hepatitis and Mammographic
[20]	Grey Wolf Optimizer algorithm, Label Propagation algorithm (2022)	Homogeneity Criterion	Heart, Ecoli, Horse, Cancer, Balance, Dermatology, Credit, Cancer-Int, Diabetes
[21]	Enhanced Cat Swarm Optimization (2022)	Sum of Squared Error	Iris, Wine, Glass, CMC, LD, Cancer, Vowel, Thyroid
[22]	Water wave optimization (WWO) Algorithm (2022)	Euclidean Distance	Iris, Wine, Glass, CMC, LD, Vowel, Thyroid, BC, Balance, Heart, WDBC, Diabetes, Dermatology
[23]	Electromagnetic Clustering Algorithm (2022)	Intra Cluster Distance	Gas, Human Activity Recognition, Vowel, Thyroid, Iris, IONO, Crude Oil, CMC
[24]	Improved Particle Swarm Optimization (2022)	Sum of Squared Error	Iris, Wine, Brest Cancer, Car Evaluation, Statlog, Yeast
[25]	Sine Cosine Algorithm-Based Fuzzy Possibilistic C-ordered Means Algorithm (2021)	HUB Loss Function	Vertebral, Wine, Iris, Aggregation, Breast Tissue, Compound, R15, E.coli, Glass, Stamps
[26]	Vibrating Particle System (2022)	Intra Cluster Variance and connectedness	Iris, Wine, Glass, Vowel, WBC, Dermatology, Ionosphere, Zoo
[27]	Grey wolf optimizer, Tabu Search (2020)	Sum of Squared Error	Iris, Blood, Breast Cancer, Glass, Seeds, Wine, Australian, Diabetes, Haberman, Heart, Liver, Planning Relax, Tic-tac-toe
[28]	Gradient evolution Algorithm (2020)	Sum of Squared Error, Sum of Distance within Cluster	Iris, Wine, Glass, R15, D311, Libras Movement, Bank Note Authorization, User Knowledge Modeling, Yeast, Aggregation
[29]	Elephant Optimization Algorithm, Gradient-based Algorithm (2021)	Sum of Squared Error	Iris, Wine, Heart, CMC, Vowel, Two-moon, Aggregation, Seeds, Breast
[30]	sine-cosine algorithm, k-prototypes algorithm (2022)	Euclidean Distance, Hamming Distance	CMC, Statlog, Thyroid Disease, Teaching Assistant Evaluation, Credit Approval
[31]	Artificial chemical reaction optimization (ACRO) algorithm (2019)	Intra Cluster Distance	Artificial dataset 1, Artificial dataset 2

It is observed that the simplex method considers the two global best positions to compute the optimal position of water particles. It is also stated that a limit operator is also used to ensure local optima issue in laminar phase. The best position of the water particles is determined using the simplex method, which is applied if the fitness of the particles does not improve within a predefined limit. The steps of the simplex method are mentioned in Algorithm 1 for computing the optimal position of water particles.

3.2. Proposed SM-WFO Algorithm for Clustering

The proposed SM-WFO algorithm for effective data clustering is described in this section. This algorithm aims to identify the optimal centroids for effective data clustering. Further, the issue of local optima is handled through simplex method which is described in subsection 3.1. The working procedure of the proposed SM-WFO is mainly segregated into following steps (i) Upload the dataset and initialize user defined parameters, (ii) Compute the objective Function, (iii) Allocation of the data objects, (iv) Laminar Flow, (v) Turbulent Flow, (vi) Update the position of water particles (cluster centroids) (vii) Performance evaluation. The steps of the proposed SM-WFO algorithm are mentioned below.

Algorithm 1: Steps of the Simplex Method

1. *Compute all of solutions (water particles position), and select first global best position* (X_{fgbest}) *and the second global best position* (X_{sgbest}). *It is assumed that* X_w *denotes the position of current water particles and* $f(X_{fgbest}), f(X_{sgbest})$, *and* $f(X_w)$ *represent the fitness of* X_{fgbest}, X_{sgbest} *and* X_w *particles.*

2. *Compute the mid position* (X_{mid}) *based on the* X_{fgbest}, *and* X_{sgbest} *which is expressed as*

$$X_{mid} = \frac{X_{fgbest} + X_{sgbest}}{2} \tag{1}$$

3. *Compute the reflection point* (X_{ref}) *using the equation 2.*

$$X_{ref} = X_{mid} + \alpha(X_{mid} - X_w) \tag{2}$$

α *is a reflection coefficient whose value set equal 1.*

4. *Compute the fitness value of* X_{ref} *and* X_{fgbest}.
If $\left(f(X_{ref}) < f(X_{fgbest})\right)$ *execute the extension operation using the equation 3.*

$$X_{exe} = X_{mid} + \gamma(X_{ref} - X_{mid}) \tag{3}$$

In equation 3, can be defined as extension coefficient, and its value is set to 2. Now, the fitness value of the extension operation (X_{exe}) *and first global best* (X_{fgbest}) *is compared. If* $\left(f(X_{exe}) < f(X_{fgbest})\right)$, *then*, X_w *replace by* X_{exe}, *otherwise,* X_{ref} *replace the* X_w.

5. *Comparing the fitness value of* X_{ref} *and* X_w. *If* $\left(f(X_{ref}) > f(X_w)\right)$, *then compression operation can utilized using equation 4.*

$$X_{comp} = X_{mid} + \beta(X_w - X_{mid}) \tag{4}$$

In equation 4, β *can be defined as compression coefficient, and its value is set to 0.5. Now, the fitness values of compression operation* (X_{comp}) *and water particle* (X_w). *If* $\left(f(X_{comp}) < f(X_w)\right)$, *then* X_w *replace by* X_{comp}, *otherwise* X_{ref} *replace the* X_w.

6. $If\left(f(X_{fgbest}) < f(X_{ref}) < f(X_w)\right)$, *utilizethe shrinking operation using the equation 5.*

$$X_{shk} = X_{mid} + \varphi(X_w - X_{mid}) \tag{5}$$

In equation 5, φ *can be defined as shrinking coefficient, and its value is set to 0.5. Now, the fitness values of compression operation* (X_{shk}) *and water particle* (X_w). $If\left(f(X_{shk}) < f(X_w)\right)$, *then* X_w *replace by* X_{shk}, *otherwise* X_{ref} *replace the* X_w.

(i) Upload the dataset and initialize the user-defined parameters: This step corresponds for uploading the dataset and initializing the user defined parameters of proposed SM-WFO algorithm. After uploading the dataset, the dimension (d), lower bound (lb) and upper bound (ub) of the dataset should be defined. Further, the user defined parameters like population of water particles, laminar probability (pl), eddying probability (pe), clusters (K), limit operator (lt), and maximum iteration (max_iter)should be defined. This step is responsible for computing the initial centroids from the dataset. The number of water particles used to define the centroids is equal to the number of clusters included in the given dataset. So, the initial position of water particles (initial centroids) is computed using the equation 6.

$$C_k = lb + rand(pop, d).[ub - lb] \tag{6}$$

In equation 6, C_k defines the initial position of water particles (initial centroids), lb stands for the lower bound, ub indicates upper bound of the given dataset, pop indicates the population of water particles, and d indicates the dimension of the dataset.

(ii) Compute the objective function: This step corresponds to compute the problem dependent objective function. For clustering problems, a distance measure is used for evaluating the compactness between data objects and cluster centroids, and acted as objective function. Euclidean distance is considered here as an objective

function and task of this function is to compute the distance between all data objects to each cluster centroid. This function is expressed in the equation 7.

$$D(X_i, C_k) = \sqrt{\sum_{s=1}^{d}(X_{is}, C_{ks})^2} \quad (7)$$

In equation 7, $D(X_i, C_k)$defines the Euclidean distance between data object (X_i) and cluster centroid (C_k), d describes the dimension such that $s = 1,2,3, \ldots . d$ .The aforementioned distance measure is used to compute the similarity between data objects.

(iii) Allocation of the data objects and compute fitness function: In this step, data objects are allocated to the respective cluster centroid based on the similarity. The similarity can be defined as similar data objects allocate to same cluster. The similarity among data objects are identified by using the Euclidean distance which is computed in previous step. The previous step computes the distance among all data objects and cluster centroids. It is noticed that a data object consists of different Euclidean distances with different cluster centroids (water particles). So, it is quite tough task to identify the appropriate cluster centroid for allocation of data objects. Finally, a data object can be allocated to a cluster centroid for which it have minimum Euclidean distance. Further, fitness function (potential energy) is also computed to evaluate the goodness of the cluster centroids (water particles). This fitness function (potential energy) is expressed using equation 8.

$$f(C_k) = \sum_{k=1}^{K} \frac{SSE(C_k)}{\sum_{i=1}^{n} SSE(C_k)} \quad (8)$$

In equation 8, $f(C_k)$ denotes the fitness function (potential energy) of the water particle (C_k), and $SSE(C_k)$ indicates the sum of squared error of the k^{th} water particle (cluster centroid).

(iv) Laminar Flow: This step is related to update the position of water particles (initial centroids). The water particles are updated by evaluating the laminar probability(pl) with respect to the $rand(\,)$function such as $if(\text{rand}\,() < P_{lam})$. Hence, it is stated that if laminar probability is higher than the random function value, then flow is considered as laminar flow and the position of water particles (initial centroids) are updated using equation 9 which is expressed as below.

$$C_{k,new} = C_k + \text{rand}(\,) \times \vec{V} \quad \forall k = \{1,2,3,4, \ldots . K\} \quad (9)$$

In equation 9, $C_{k,new}$ denotes the new position of water particles, C_k denotes the current position of water particles, $\text{rand}()$ denotes water coefficient in the range of $[0, 1]$, $\vec{V}$ is described as a vector value associated with motional direction and K denotes total number of water particles. The common motional direction is expressed by equation 10.

$$\vec{V} = C_e - C_f \text{ such that } \left(e \neq f, \left(f(C_e) \leq f(C_f)\right)\right) \quad (10)$$

In equation 10, $f(C_e)$ and $f(C_f)$ indicate the potential energy of e^{th} and f^{th} water particles such that $e \neq f$. Further, the potential emery is utilized for chosen the best water particle in each successive iteration.$if\left(f(C_e) \leq f(C_f)\right)$, e^{th} particle energy is less than f^{th} particle energy, then the best particle is C_e, otherwise C_f. If the fitness function (potential energy) of the water particles remains same in predefined limit operator, then call the Algorithm 1 for generating the optimal position of water particles.

(v) Turbulent Flow: If laminar probability is less than the random function value, then, as turbulent flow phase is invoking and current position of water particle can be mutated. The mutation process can be accomplished through eddying parameter (pe) and random function i.e. $\text{rand}()$ such as $if(\text{rand}() < pe)$. If value of eddying parameter is higher than $\text{rand}()$ function, the new position of water particle (cluster centroid) can be generated by mutating the randomly chosen dimension of the current water particle. This behavior is expressed using equations11-15.

$$C_{k,new} = (C_k^s,\ p,\ C_k^{s+1}, p,\ C_k^{s+2}, p, \ldots \ldots C_k^{s+n}, p) \quad \text{where } s = \{1,2,3, \ldots . d\} \quad (11)$$

In equation 11, the variable p can be defined as jostling operator (mutation operator), C_k^s denotes the value of the current water particle using sth dimension, $C_{k,new}$denotes the new mutated position of kth water particle. The mutation process (p) can be expressed using equation 12.

$$p = \begin{cases} \gamma(C_k^s,\ C_l^s), & \text{if rand}() < pe \\ \vartheta(C_k^s\ , C_l^{s+1}), & \text{otherwsie} \end{cases} \quad (12)$$

In equation 12, l indicates the randomly chosen water particle such that $l \in \{1, 2, 3, \ldots . K\}andk \neq l$, 's+1' denotes the randomly chosen dimension. pe denotes the eddying parameter and $\text{rand}(\,)$ denotes a random number in the between [0, 1]. The eddying parameter is computed using the equation 13 which is expressed as below.

$$\gamma(C_k^s,\ C_l^s) = C_k^s + \varphi \times \theta \times \cos(\theta) \quad (13)$$

In equation 13, θ denotes a random value in the range of $[-\pi, \pi]$ and φ indicates the shearing force between k^{th} and l^{th}water particles. The shearing force (φ)can be expressed using the equation 14.

$$\varphi = \left|C_k^s,\ C_{l,}^s\right| \quad (14)$$

Moreover, a transformation function is employed to describe the behavior of water particles. This behavior is expressed using equation 15.

$$\vartheta\left((C_k^s\ , C_l^{s+1})\right) = (ub^s - lb^s)\frac{C_k^s - lb^s}{ub^s - lb^s} \quad (15)$$

In equation 15, ub^s indicates the upper bound in s^{th} dimension, lb^s indicates the lower bound in s^{th} dimension. C_k^sindicates the s^{th} dimension of k^{th} water particle.

(vi) Update the position of water particles (cluster centroids): In this step, the position of water particles is updated. For updating process, the fitness function (potential energy) of the new position of water particles and old position of water particles is compared i.e. $if\ \left(f(C_{k,new}) > f(C_{k,old})\right)$. If, the fitness of the new position of water particles is higher than old ones, then the current position is replaced by the new position of water particles such as $C_k \leftarrow C_{k,new}$. Otherwise, the old position can be acted as current position of water particles such as $C_k \leftarrow C_{k,old}$. The steps (ii-vi) are rerun until the maximum number of iteration is not reached. Finally, the optimal cluster centroids in terms of position of water particles are obtained.

(vii) Performance Evaluation: The performance of the proposed SM-WFO algorithm is evaluated in this step. After obtaining the optimal position of water particles (cluster centroids), the performance of the algorithm is evaluated using intra cluster distance (Intra), accuracy rate (AR), and detection rate (DR) measures.

The steps of the proposed SM-WFO algorithm are specified in Algorithm 2, whereas the flowchart of the same is mentioned in the Fig. 2.

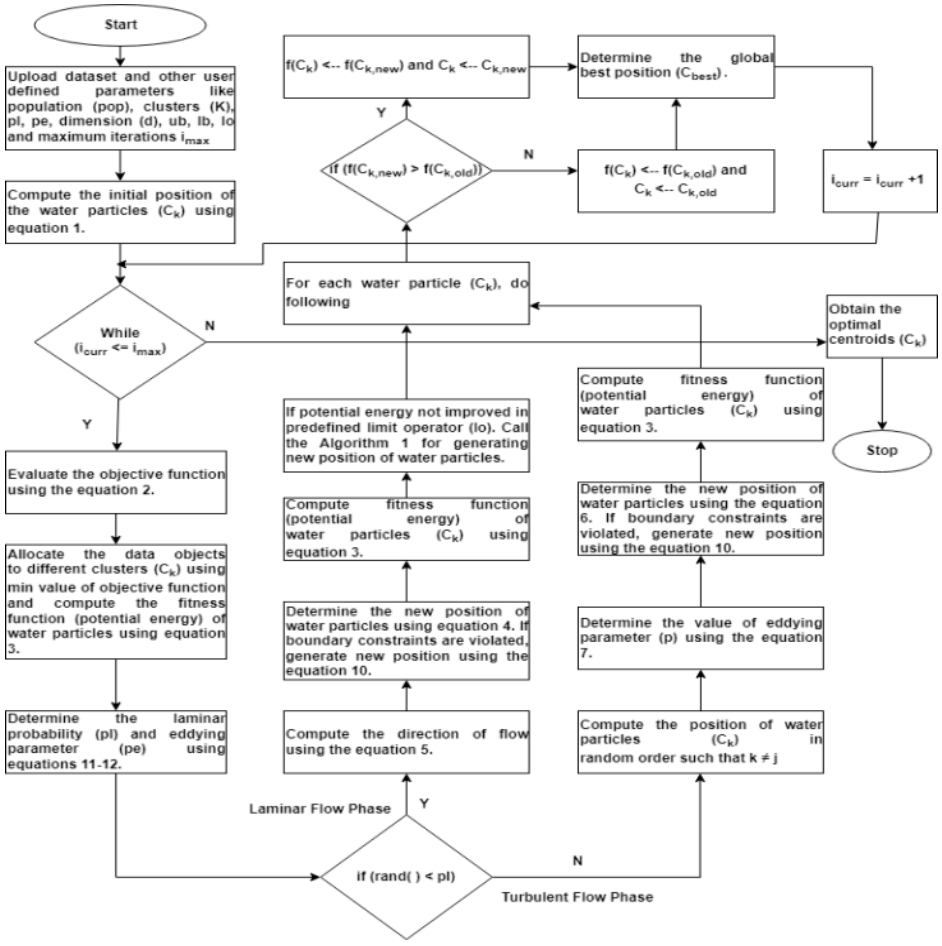

Fig. 2. Flowchart of the proposed SM-WFO algorithm.

4. EXPERIMENTAL RESULTS

The experimental results of the proposed SM-WFO are discussed in this section. Several well-known clustering datasets such as Iris, Glass, Wine, Balance, Vowel, LD, WBC, etc. are considered for evaluating the efficiency of the proposed algorithm. The experimental results of the proposed algorithm are evaluated using intra, AR and DR parameters. The intra parameter computes compactness between the data objects and cluster centroids. It is defined as total distance between data objects and cluster centers. The intra parameter results are also validated using standard deviation (SD). Further, the accuracy rate (AR) can be defined as correctly allocated data objects to clusters divided by total numbers of data objects. The results are also compared with existing renowned clustering algorithms such as ABC, DE, GA, CS, GSA, and LION, etc.

Algorithm 2: Steps of SM-WFO Algorithm for cluster analysis

Input: Dataset and Number of Cluster(K)

Output: Optimal Cluster Centroids

1: Upload the dataset and other user-defined parameters like population of water particles (pop), total number of clusters(K), dimension of dataset (d), laminar probability (pl), eddying parameter (pe), lb, ub, limit operator (lo) and maximum number of iteration (i_{max}).

2: Compute the initial position of the water particles (C_k) using equation 6.

3: While ($i_{curr} \leq i_{max}$), do following

4: Evaluate the objective function using the equation 7.

5: Allocate the data objects (X_l)to different clusters (C_k) using minimum value of objective function and compute the fitness function (potential energy) of water particles using equation 8.

6: Determine the laminar probability (pl) and eddying parameter(pe) using equations 16-17.

$$pl = (c_1 - c_2) \times \left(\frac{1}{1 + e^{-t_{curr}/t_{max}}}\right) \tag{16}$$

$$pe = c_2 \times \left(\frac{2}{1 + e^{-2t_{curr}/t_{max}}} - 1\right) + c_1 \tag{17}$$

In equations 16-17, c_1 and c_2 are cognitive parameter whose values are 0.5 and 0.3 respectively.

7: If (rand () < pl), do following */*call laminar Flow Phase */*

8: Compute the direction of flow using the equation 10.

9: Determine the new position of water particles using equation 9. If boundary constraints are violated, generate new position using the equation 15.

10: Compute fitness function (potential energy) of water particles (C_k)using the equation 8.

11: If fitness (potential energy) i.e. $f(C_k)$ of water particles is not improved in predefined limit operator (lo). Call the Algorithm 1 for generating the new position of water particles.

12: Else */*call turbulent Flow Phase */*

$if(\text{rand}() < pe)$

13: Compute the position of water particles (C_k) in random order such that that$k \neq j$

14: Determine the value of eddying parameter(pe) using the equation 13.

15: Determine the new position of water particles using the equation 11. If boundary constraints are violated, generate new position using the equation 15.

16: Compute fitness function (potential energy) of water particles (C_k) using the equation 8.

17: End if

18: End if

19: For, each water particle (C_k), do following

20: If $\left(f(C_{k,new}) > f(C_{k,old})\right)$
21: $f(C_k) \leftarrow f(C_{k,new})$ and $C_k \leftarrow C_{k,new}$
22: Else
23: $f(C_k) \leftarrow f(C_{k,old})$ and $C_k \leftarrow C_{k,old}$
24: End If
25: End For
26: Update the position of water particles and determine the global best position(C_{best})..
27: $i_{curr} = i_{curr} + 1$
28: End While
29: Obtain the optimal centroids (C_k).

The user-defined parameters of the proposed SM-WFO are listed as total number of clusters(K), dimension of dataset (d), population of water particles (pop), laminar probability $(pl \in (0.2, 0.5))$,), eddying parameter $(pe \in (0.5, 0.9))$, lb indicates minimum value of each dimension, up indicates maximum value of each dimension, limit operator $(lo) = 8$, and maximum number of iteration (i_max=100).The maximum iteration (stopping criteria) is set to 100. The population of water particles is similar to cluster numbers (K). The parameter settings of other algorithms are taken same as reported in concerned literature. Additionally, the proposed SM-WFO algorithm is implemented using the MATLAB environment on a corei5 processor with 32 GB of RAM.

4.1. Simulation Results

This subsection discusses the experimental results of the SM-WFO. The results of the SM-WFO are also compared with popular clustering algorithms. Table 2 illustrates the experimental results of the proposed SM-WFO algorithm and other algorithms using intra, SD and rank parameters. It is analyzed that SM-WFO algorithm obtains least intra distance with glass (2.18E+02), control (2.38E+04), vowel (1.57E+05), crude oil (2.42E+02), CMC (5.57E+03), and cancer (2.69E+03) among all algorithms, whereas, the proposed SM-WFO algorithm obtains second minimum distance with iris (9.43E+01), ionosphere (1.06E+03), balance (5.80E+04), LD (1.37E+03), and thyroid (1.37E+03) among all algorithms. It is analyzed that K-means achieves minimum distance (9.20E+01) with iris dataset, ACO algorithm achieves minimum intra distance (9.38E+02) with ionosphere dataset, CS algorithm achieves minimum intra distance (5.79E+04) with balance dataset, and DE algorithm achieves minimum intra distance (1.30E+03) with thyroid dataset. Moreover, SM-WFO algorithm achieves third minimum intra distance (1.63E+04) with wine dataset, while DE algorithm obtains minimum intra distance (1.58E+04) with wine dataset among all algorithms and ACO algorithm achieves second minimum intra distance (1.62E+04) with wine dataset among all algorithms. By analyzing the SD parameter, it is revealed that proposed SM-WFO algorithm achieves minimum SD rate with majority of the datasets. This parameter specifies dispersion of the intra parameter in each run. Further, the rank parameter is also computed to validate the efficacy of SM-WFO algorithm. Rank parameter is computed of each algorithm with respect to each dataset based on intra parameter. It's revealed that SM-WFO gets highest average rank (i.e. 1.58) compared to other algorithms. The second highest rank (3.75) is obtained by the LION algorithm while K-means gets lower rank (8) among all algorithms. Hence, it is said that proposed SM-WFO algorithm gets superior results in context of intra, SD and rank parameters.

The effectiveness of the proposed SM-WFO algorithm is also assessed using AR (acc. rate) and DR (det. rate) parameters. Table 3 illustrates the findings of the SM-WFO based on AR (acc. rate) and DR (det. rate) parameters. It is marked that SM-WFO algorithm gets higher AR rate for iris (92.8%), glass (70.7%), ionosphere (77.4%), control (89.1%), vowel (87.8%), CMC (60.4%), LD (67.1%), cancer (81.7%), and thyroid (72.9%) datasets. It is also noticed that SM-WFO gets second higher AR (acc. rate) rates with wine (87.3%), crude oil (79.8%), and balance (91.3%) datasets among all algorithms. It is acknowledged that LION algorithm gets higher AR (acc. rate) (87.8 and 80.5%) with wine and crude oil datasets, CS algorithm gets higher acc. rate (91.7%) with balance dataset. By analyzing the DR (det. rate), it is revealed that SM-WFO algorithm gets higher det. rate for iris (93.2%), glass (71.6%), ionosphere (78.1%), control (89.4%), vowel (86.6%), CMC (61.9%), LD (68.4%), cancer (83.1%), and thyroid (74.1%) datasets. It is also noticed that SM-WFO gets second higher det. Rate (DR) with wine (87.9%), balance (90.4%) and crude oil (81.4%) datasets among all algorithms.

Table 2: Simulation results of SM-WFO and other existing algorithms based on distance parameters

Datasets	Measure	Clustering Algorithms									
		K-means	**PSO**	**ACO**	**ABC**	**DE**	**GA**	**CS**	**GSA**	**LION**	**SM-WFO**
Iris	Intra	9.20E+01	9.86E+01	1.01E+02	1.08E+02	1.21E+02	1.25E+02	9.64E+01	9.79E+01	9.76E+01	9.43E+01
	SD	2.67E+01	4.67E-01	1.31E+00	6.63E+00	5.23E+00	1.46E+01	3.25E+00	3.19E+00	4.03E+00	2.56E+00
	Rank	1	6	7	8	9	10	3	5	4	2
Glass	Intra	3.79E+02	2.76E+02	2.19E+02	3.29E+02	3.62E+02	2.82E+02	2.41E+02	2.39E+02	2.38E+02	2.18E+02
	SD	7.05E+01	1.86E+01	3.36E+00	1.14E+01	1.21E+01	4.14E+00	1.12E+01	9.87E+00	1.07E+01	1.39E+01
	Rank	10	6	2	8	9	7	5	4	3	1
Wine	Intra	1.81E+04	1.64E+04	1.62E+04	1.69E+04	1.58E+04	1.65E+04	1.65E+04	1.70E+04	1.65E+04	1.63E+04
	SD	9.06E+02	8.55E+01	3.69E+01	4.74E+02	5.60E+01	7.84E+01	2.64E+01	2.74E+01	2.66E+01	5.61E+01
	Rank	10	4	2	8	1	7	6	9	5	3
Ionosphere	Intra	2.42E+03	1.08E+03	9.38E+02	1.11E+03	1.13E+03	1.07E+03	1.23E+03	2.83E+03	1.42E+03	1.06E+03
	SD	4.55E+02	3.34E+02	4.48E+02	2.61E+02	3.17E+02	4.13E+02	3.46E+01	4.04E+01	3.38E+01	2.79E+02
	Rank	9	4	1	5	6	3	7	10	8	2
Control	Intra	1.01E+06	4.18E+04	2.39E+04	5.12E+04	5.23E+04	4.62E+04	3.03E+04	3.13E+04	2.75E+04	2.38E+04
	SD	5.05E+03	1.02E+03	1.71E+02	1.32E+03	9.16E+02	1.58E+03	6.18E+01	5.37E+01	4.47E+01	7.87E+01
	Rank	10	6	2	8	9	7	4	5	3	1
Vowel	Intra	1.60E+05	1.58E+05	1.89E+05	1.70E+05	1.81E+05	1.59E+05	1.59E+05	1.60E+05	1.59E+05	1.57E+05
	SD	4.52E+03	2.88E+03	2.58E+03	4.64E+03	2.86E+03	3.11E+03	3.42E+01	3.87E+01	1.74E+01	4.23E+01
	Rank	7	2	10	8	9	5	4	6	3	1
Balance	Intra	1.20E+05	6.20E+04	5.94E+04	6.61E+04	6.78E+04	6.91E+04	5.79E+04	5.81E+04	5.86E+04	5.80E+04
	SD	9.28E+03	4.01E+03	7.56E+02	6.79E+02	5.25E+03	5.62E+03	2.19E+02	3.02E+02	2.98E+02	2.22E+02
	Rank	10	6	5	7	8	9	1	3	4	2
Crude oil	Intra	2.91E+02	2.86E+02	2.47E+02	2.81E+02	3.69E+02	2.83E+02	2.74E+02	2.71E+02	2.69E+02	2.42E+02
	SD	2.63E+01	1.14E+01	4.71E+01	1.09E+01	2.33E+01	8.14E+00	1.24E+01	1.75E+01	1.39E+01	1.95E+01
	Rank	9	8	2	6	10	7	5	4	3	1
CMC	Intra	5.59E+03	5.85E+03	5.83E+03	5.94E+03	5.95E+03	5.76E+03	5.85E+03	5.67E+03	5.70E+03	5.57E+03
	SD	4.68E+01	4.89E+01	1.23E+02	1.31E+02	8.69E+01	5.04E+01	2.35E+02	6.32E+01	9.47E+01	2.96E+01
	Rank	2	8	6	9	10	5	7	3	4	1
LD	Intra	1.17E+04	3.24E+03	3.24E+03	9.85E+03	1.15E+04	2.54E+03	1.39E+03	1.73E+03	1.33E+03	1.37E+03
	SD	6.68E+02	2.88E+01	1.64E+01	8.20E+02	2.07E+03	4.18E+01	5.62E+01	4.79E+01	2.68E+01	1.88E+01
	Rank	10	6	7	8	9	5	3	4	1	2
WBC	Intra	1.93E+04	4.26E+03	3.37E+03	3.50E+03	3.73E+03	3.00E+03	3.72E+03	2.92E+03	2.89E+03	2.69E+03
	SD	5.14E-12	2.08E+02	4.17E+01	2.12E+02	1.84E+02	2.25E+02	1.73E+01	8.36E+00	7.67E+00	1.88E+02
	Rank	10	9	5	6	8	4	7	2	3	1
Thyroid	Intra	2.39E+03	1.11E+04	1.99E+03	1.98E+03	1.30E+03	1.22E+04	1.43E+03	1.86E+03	1.54E+03	1.37E+03
	SD	2.46E+02	2.71E+01	3.09E+01	2.23E+02	2.06E+01	3.26E+01	2.39E+01	1.90E+01	2.46E+01	1.48E+01
	Rank	8	9	7	6	1	10	3	5	4	2

Table 3: Simulation results of SM-WFO and other existing algorithms based on accuracy rate and detection rate parameters

Dataset	Parameter	KM	PSO	ACO	ABC	DE	GA	CS	GSA	LION	SM-WFO
Iris	Acc. Rate	67.30%	83.30%	78.90%	88.70%	84.20%	74.10%	88.50%	78.30%	85.10%	92.80%
	Det. Rate	69.60%	85.70%	79.40%	89.20%	86.80%	77.30%	89.30%	77.40%	81.10%	93.20%
Glass	Acc. Rate	51.90%	53.70%	37.40%	48.90%	48.10%	49.00%	68.90%	66.40%	68.10%	70.70%
	Det. Rate	53.80%	57.20%	38.40%	50.90%	49.20%	51.10%	72.60%	68.40%	71.30%	71.60%
Wine	Acc. Rate	73.90%	71.10%	74.60%	77.30%	74.10%	72.90%	80.30%	79.00%	87.80%	86.90%
	Det. Rate	75.20%	73.70%	78.10%	79.30%	76.20%	74.90%	82.40%	81.80%	90.80%	87.90%
Ionosphere	Acc. Rate	71.20%	64.80%	60.70%	64.40%	63.00%	60.10%	73.50%	75.20%	76.90%	77.40%
	Det. Rate	72.80%	64.70%	61.20%	66.40%	65.30%	61.80%	74.60%	76.20%	77.90%	78.10%
Control	Acc. Rate	59.70%	41.20%	39.50%	35.60%	39.30%	46.70%	69.80%	67.40%	73.80%	89.10%
	Det. Rate	61.40%	45.20%	43.70%	39.60%	41.70%	49.20%	72.20%	69.40%	76.20%	89.40%
Vowel	Acc. Rate	76.30%	75.30%	77.50%	79.60%	69.80%	74.50%	83.50%	84.70%	85.10%	86.40%
	Det. Rate	84.60%	79.50%	80.60%	83.20%	74.70%	79.00%	86.60%	85.80%	89.70%	86.60%
Balance	Acc. Rate	85.00%	89.80%	74.30%	76.70%	75.00%	78.00%	91.70%	84.90%	85.50%	89.60%
	Det. Rate	86.30%	90.40%	77.20%	78.30%	77.60%	82.40%	93.80%	88.70%	89.60%	90.40%
Crude oil	Acc. Rate	65.50%	76.50%	59.10%	56.80%	66.50%	63.20%	70.40%	76.10%	80.50%	79.80%
	Det. Rate	68.40%	77.30%	64.50%	58.70%	68.10%	65.40%	71.30%	79.20%	83.40%	81.40%
CMC	Acc. Rate	35.70%	51.40%	36.90%	41.60%	43.70%	40.30%	57.10%	54.70%	53.40%	60.40%
	Det. Rate	45.50%	59.80%	46.50%	48.90%	46.60%	43.50%	61.40%	58.20%	55.80%	61.90%
LD	Acc. Rate	52.20%	54.10%	52.90%	49.90%	52.00%	49.30%	65.90%	63.10%	66.20%	67.10%
	Det. Rate	63.50%	58.70%	56.40%	54.90%	64.80%	59.00%	68.90%	65.40%	71.30%	68.40%
Cancer	Acc. Rate	69.80%	72.10%	74.90%	78.60%	65.40%	69.10%	82.00%	72.80%	78.70%	81.70%
	Det. Rate	75.10%	74.80%	77.30%	82.40%	67.80%	71.50%	84.10%	78.60%	84.20%	83.10%
Thyroid	Acc. Rate	63.80%	68.90%	64.90%	64.40%	65.80%	63.20%	69.10%	67.80%	72.40%	72.90%
	Det. Rate	66.10%	69.20%	65.40%	66.10%	69.70%	64.30%	74.80%	72.90%	75.90%	74.10%

It is revealed that LION algorithm gets higher DR rate (0.908 and 0.834) with wine and crude oil datasets, CS algorithm gets higher DR rate (0.938) with balance dataset. Hence, it is said that SM-WFO acquires better clustering results using AR and DR parameters with most of datasets.Fig.3 (a-l) presents the clusters in different dataset using the proposed SM-WFO algorithm. Fig.3 (a) shows the iris dataset clusters based on the petal length, petal width and sepal width attributes. The proposed SM-WFO separates the data into three clusters such as cluster-1, cluster-2 and cluster-3. I.Setosa, I.Versicolour, and I.Virginica clusters. It is observed that the data belonging to cluster1 linearly separable than clusrer-2 and cluster-3. While, the data belongs to cluster-2 and cluster-3 are non-linear in nature. The clusters of glass dataset are presented into Fig. 3(b). It is revealed that proposed SM-WFO algorithm divides the glass dataset into six clusters, but data are non-linear in nature. Fig. 3(c) depicts the distinct clusters of wine dataset. The wine dataset comprises of three clusters: cluster1, cluster2, and cluster3 based on ash, malic acid and alcohol. It is also noticed that data is non-linear and cannot be easily separable. Fig. 3(d) displays the clusters of ionosphere dataset using attribute3, attribute4 and attribute5. The findings demonstrated that this dataset is non-linear separable, but proposed SM-WFO algorithm satisfactory allocate data to corresponding clusters. Fig. 3(e) shows the clusters in control dataset using feature2, feature59 and feature60. The control dataset consists of sixty attributes, but data is not linear in nature. Fig. 3(f) demonstrates the clusters of vowel dataset and this dataset comprises of six clusters. It is analyzed that clusters are not linearly separable. The clustering in balance dataset is presented in Fig. 3(g). This dataset comprises of three clusters. The findings stated that data are linear in nature. Fig. 3(h) depicts the crude oil dataset clusters using the

price open, price high and volume attributes. The data is separated into three clusters, and SM-WFO algorithm effectively allocates the data to appropriate clusters. The clusters in liver disease (LD) dataset are shown in Fig. 2(i). The ALP, age and aspartate attributes are used to compute the clusters. The finding stated that SM-WFO correctly divides the data into cluster1 and cluster2. The clusters of cancer dataset based on cell size, clam thickness and adhesion are reported into in Fig. 3(j). The proposed SM-WFO categorizes the data into clusters effectively. The thyroid dataset clusters are shown in Fig. 2(k) using TT4, FT1 and age attributes. This dataset consists of three clusters. It is discovered that the cluster-1 and cluster-2 are non-linear, but the cluster-3 is linear in nature. The CMC dataset clusters are shown in Fig. 3(l) using living index, wife education and number of children. SM-WFO separates the data into three clusters effectively. Finally, it is claimed that SM-WFO precisely assigns the data to relevant clusters.

5. CONCLUSION

This work presents a simplex method based water flow optimizer (SM-WFO) algorithm for clustering problems. WFO algorithm is inspired by flow of water i.e. highland to lowland and comprises of laminar flow and turbulent flow. It is revealed that WFO algorithm obtains superior outcomes for most of the optimization problems, but sometimes it can stick in local optima and converges on immature results. This issue of the WFO algorithm is addressed through well-known simplex method. The simplex method utilizes the two global best positions of particles for generating the new optimal location of water particles. Further, local optima are handled through a limit operator. Both modifications are integrated into laminar flow phase of the WFO. The effectiveness of the SM-WFO is validated by a set of twelve datasets. Moreover, the results are assessed by distance measures (intra, SD), AR, DR and rank parameters. The results demonstrated that SM-WFO algorithm obtains superior in terms of intra parameter with most of datasets. Further, SM-WFO also achieves better AR (acc. rate) and DR (det. rate) results with most of datasets. The rank parameter also supports the existence of the SM-WFO algorithm in the field of clustering. SM-WFO obtains the minimum average rank (1.58) among all datasets. Hence, it is claimed that the SM-WFO outperforms than other algorithms. Finally, it is concluded that SM-WFO is an efficient to solve clustering problems. The efficiency of the SM-WFO will be explored in some other fields like feature extraction and reduction, image processing, classification, constrained optimization problems and outlier detection.

ACKNOWLEDGEMENTS

Authors have not received any funding to carry out this work.

REFERENCES

[1] Tan, P., Steinbach, M.S., & Kumar, V. (2022). Introduction to Data Mining. Data Mining and Machine Learning Applications.

[2] Han, J., Kamber, M., & Pei, J. (2012). Data mining concepts and techniques third edition. University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University.

[3] Aggarwal, C. C., & Reddy, C. K. (2014). Data clustering. Algorithms and applications. Chapman & Hall/CRC Data mining and Knowledge Discovery series, Londra.

[4] Gan, G., Ma, C., & Wu, J. (2020). Data clustering: theory, algorithms, and applications. Society for Industrial and Applied Mathematics

[5] Vishwakarma, S., Bhardwaj, S. K., Bihari, A., Tripathi, S., Agrawal, S., & Joshi, P. (2024) Cancer Gene Clustering Using Computational Model. GMSARN International Journal, 18(2), 252-257.

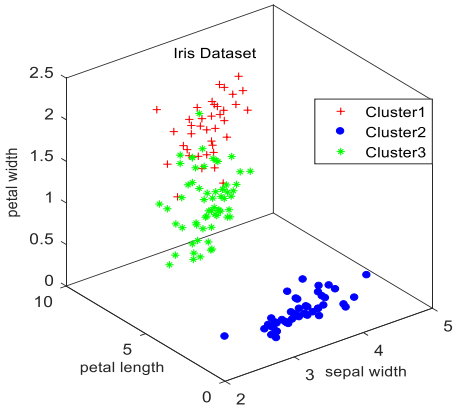

Fig 3(a) Iris dataset

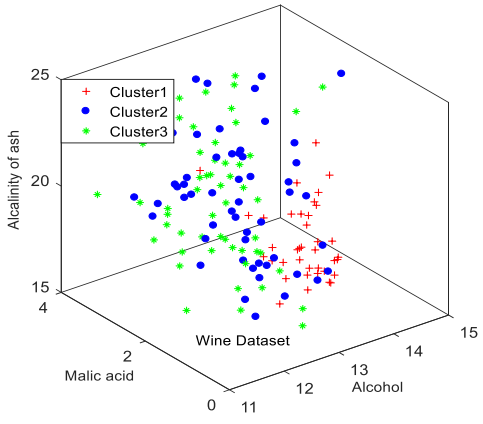

Fig 3(c) Wine dataset

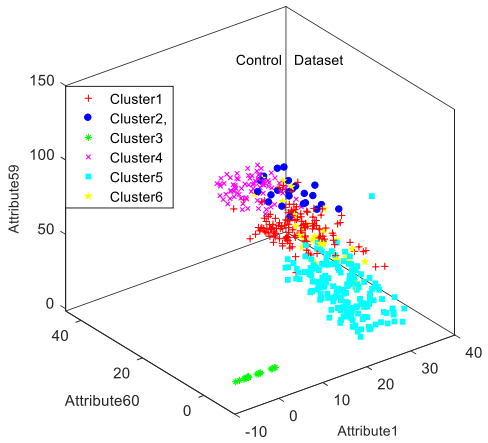

Fig 3(e) Control dataset

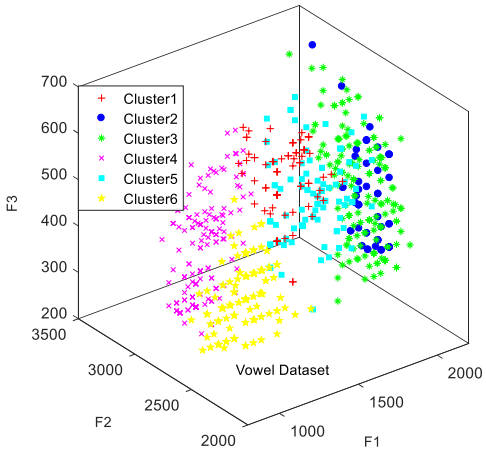

Fig 3(f) Vowel dataset

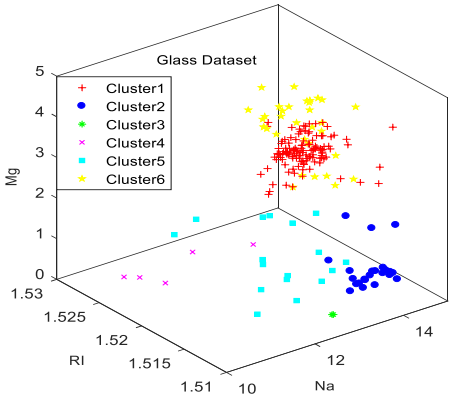

Fig 3(b) Glass dataset

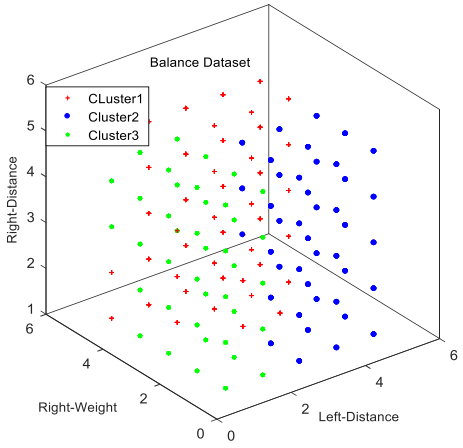

Fig 3(g) Balance dataset

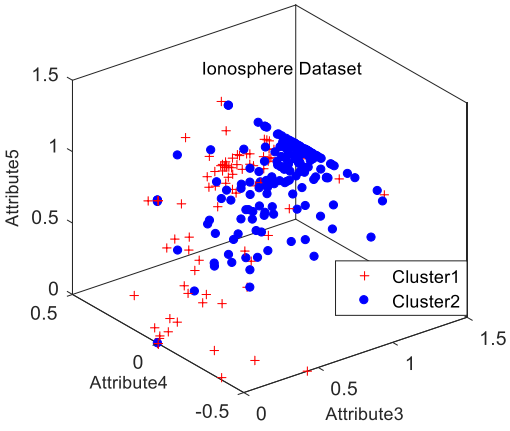

Fig 3(d) Ionosphere dataset

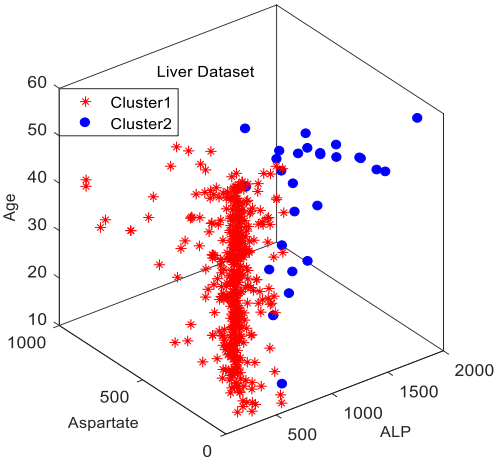

Fig 3(i) Liver dataset

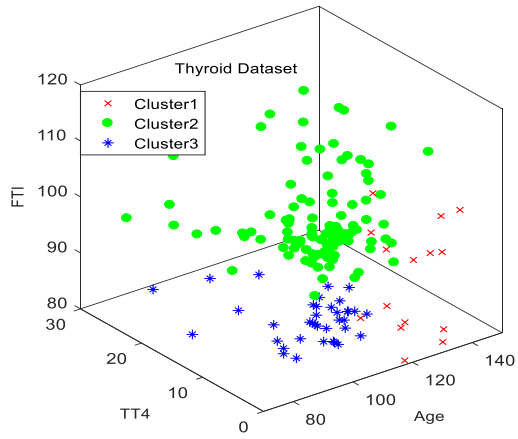

Fig 3(k) Thyroid dataset

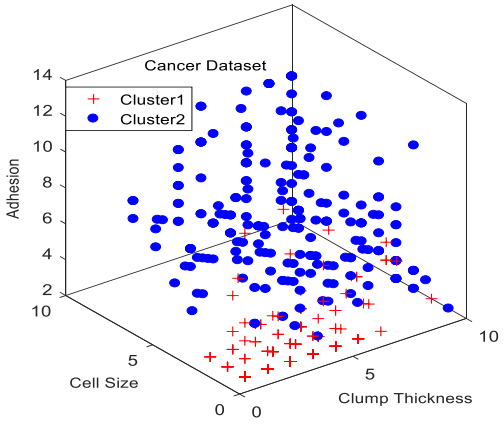

Fig 3(j) Cancer dataset

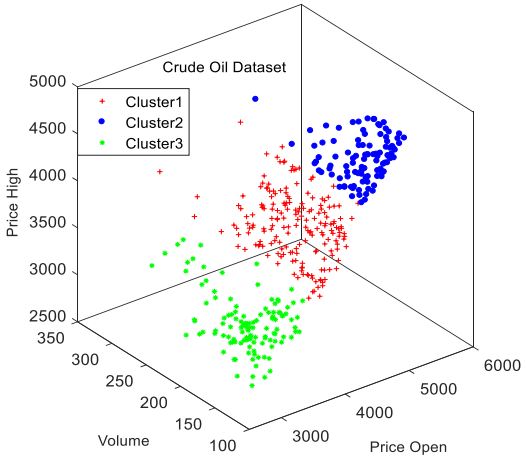

Fig 3(h) Crude Oil dataset

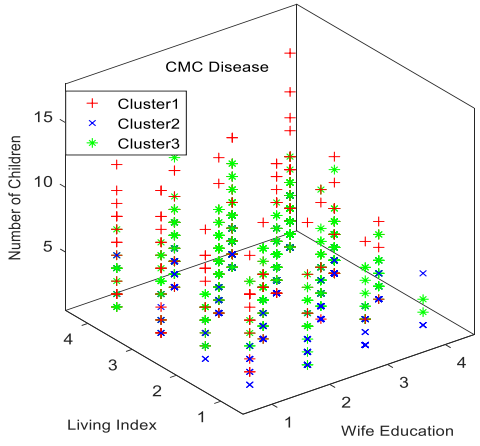

Fig. 3(l) CMC dataset

Fig. 3(a-l) Illustrates the clustering results based on SM-WFO algorithm.

[6] Nanda, S. J., & Panda, G. (2014). A survey on nature inspired metaheuristic algorithms for partitional clustering. Swarm and Evolutionary computation, 16, 1-18.

[7] Cura, T. (2012). A particle swarm optimization approach to clustering. Expert Systems with Applications, 39(1), 1582-1588.

[8] Ha, P. T., Tran, D. T., & Nguyen, T. T. (2021) Optimal Generation for Hydrothermal System with Pumped Storage Hydroelectric Plants Using Six Particle Swarm Optimization Algorithms. GMSARN International Journal, 16(4), 451-460

[9] Kushwaha N, Pant M, Kant S, Jain VK (2018) Magnetic optimization algorithm for data clustering. Pattern Recogn Lett 115:59–65

[10] Karaboga, D., & Basturk, B. (2007). A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. Journal of global optimization, 39(3), 459-471.

[11] Erol, O. K., & Eksin, I. (2006). A new optimization method: big bang–big crunch. Advances in Engineering Software, 37(2), 106-111.

[12] Kumar, Y., & Sahoo, G. (2014). A charged system search approach for data clustering. Progress in Artificial Intelligence, 2(2), 153-166.

[13] Luo, K. (2021). Water flow optimizer: a nature-inspired evolutionary algorithm for global optimization. IEEE Transactions on Cybernetics.

[14] Matos Macêdo, F. J., & da Rocha Neto, A. R. (2022). A Binary Water Flow Optimizer Applied to Feature Selection. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 94-103). Springer, Cham.

[15] Said, M., Shaheen, A. M., Ginidi, A. R., El-Sehiemy, R. A., Mahmoud, K., Lehtonen, M., &Darwish, M. M. (2021). Estimating parameters of photovoltaic models using accurate turbulent flow of water optimizer. Processes, 9(4), 627.

[16] Cheng, M. M., Zhang, J., Wang, D. G., Tan, W., & Yang, J. (2023). A Localization Algorithm Based on Improved Water Flow Optimizer and Max-Similarity Path for 3D Heterogeneous Wireless Sensor Networks. IEEE Sensors Journal.

[17] Patel, V. P., Rawat, M. K., & Patel, A. S. (2023). Local neighbour spider monkey optimization algorithm for data clustering. Evolutionary Intelligence, 16(1), 133-151.
[18] Singh, H., Rai, V., Kumar, N., Dadheech, P., Kotecha, K., Selvachandran, G., & Abraham, A. (2023). An enhanced whale optimization algorithm for clustering. Multimedia Tools and Applications, 82(3), 4599-4618.
[19] Al-Behadili, H. N. K. (2022). Improved firefly algorithm with variable neighborhood search for data clustering. Baghdad Science Journal, 19(2), 0409-0409.
[20] Besharatnia, F., Talebpour, A., &Aliakbary, S. (2022). An improved Grey Wolves optimization algorithm for dynamic community detection and data clustering. Applied Artificial Intelligence, 36(1), 2012000.
[21] Singh, H., & Kumar, Y. (2022). An Enhanced Version of Cat Swarm Optimization Algorithm for Cluster Analysis. International Journal of Applied Metaheuristic Computing (IJAMC), 13(1), 1-25.
[22] Kaur, A., & Kumar, Y. (2022). A new metaheuristic algorithm based on water wave optimization for data clustering. Evolutionary Intelligence, 15(1), 759-783.
[23] Kushwaha, N., Pant, M., & Sharma, S. (2022). Electromagnetic optimization based clustering algorithm. Expert Systems, 39(7), e12491.
[24] Hashemi, S. E., Tavana, M., & Bakhshi, M. (2022). A New Particle Swarm Optimization Algorithm for Optimizing Big Data Clustering. SN Computer Science, 3(4), 1-16.
[25] Kuo, R. J., Lin, J. Y., & Nguyen, T. P. Q. (2021). An application of sine cosine algorithm-based fuzzy possibility c-ordered means algorithm to cluster analysis. Soft Computing, 25(5), 3469-3484.
[26] Kaur, A., & Kumar, Y. (2022). A multi-objective vibrating particle system algorithm for data clustering. Pattern Analysis and Applications, 25(1), 209-239.
[27] 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.
[28] Kuo, R. J., & Zulvia, F. E. (2020). Multi-objective cluster analysis using a gradient evolution algorithm. Soft Computing, 24(15), 11545-11559.
[29] Duan, Y., Liu, C., Li, S., Guo, X., & Yang, C. (2022). Gradient-based elephant herding optimization for cluster analysis. Applied Intelligence, 1-32.
[30] Kuo, T., & Wang, K. J. (2022). A hybrid k-prototypes clustering approach with improved sine-cosine algorithm for mixed-data classification. Computers & Industrial Engineering, 108164.
[31] Singh, H., Kumar, Y.& Kumar, S. (2019). A new meta-heuristic algorithm based on chemical reactions for partitional clustering problems. Evolutionary Intelligence, 12(2), 241-252
[32] Kao, Y. T., Zahara, E., & Kao, I. W. (2008). A hybridized approach to data clustering. Expert Systems with Applications, 34(3), 1754-1762.
[33] García, M. L. L., García-Ródenas, R., & Gómez, A. G. (2014). Hybrid meta-heuristic optimization algorithms for time-domain-constrained data clustering. Applied Soft Computing, 23, 319-332.
[34] Zhao, R., Wang, Y., Xiao, G., Liu, C., Hu, P., & Li, H. (2021). A selfish herd optimization algorithm based on the simplex method for clustering analysis. The Journal of Supercomputing, 77, 8840-8910.
[35] Spendley, W. G. R. F. R., Hext, G. R., & Himsworth, F. R. (1962). Sequential application of simplex designs in optimization and evolutionary operation. Technometrics, 4(4), 441-461.
[36] Fan, S. K. S., Liang, Y. C., & Zahara, E. (2004). Hybrid simplex search and particle swarm optimization for the global optimization of multimodal functions. Engineering optimization, 36(4), 401-418.
[37] Zhu, C., Zhang, Y., Pan, X., Chen, Q., & Fu, Q. (2022). Improved Harris Hawks optimization algorithm based on quantum correction and Nelder-Mead simplex method. Math. Biosci. Eng, 19(8), 7606-7648.
[38] Yang, X., Zhou, H., Alathamneh, M., & Nelms, R. M. (2023). An Evolutionary Annealing–Simplex Method for Inductance Value Selection for LCL Filters. Energies, 16(10), 4192.