Resbee Publishers

Multimedia Research (MR)

Received 14 April, Revised 15 June, Accepted 19 July

DIGWO: Hybridization of Dragonfly Algorithm with Improved Grey Wolf Optimization Algorithm for Data Clustering

Amolkumar Narayan Jadhav
Vel-Tech Dr. RR & Dr.SR Technical University
Chennai, Tamil Nadu, India
amolkumarnj@gmail.com

Gomathi N
Vel-Tech Dr. RR & Dr.SR Technical University
Chennai, Tamil Nadu, India

Abstract: Data present in great quantity raises the difficulty of managing them that affects the effectual decision-making procedure. Therefore, data clustering achieves notable significance in knowledge extraction and a well-organized clustering algorithm endorses the effectual decision making. For that reason, an algorithm for data clustering by exploiting the DIGWO method is presented in this paper, which decides the optimal centroid to perform the clustering procedure. The developed DIGWO technique exploits the calculation steps of the Dragonfly Algorithm (DA) with the incorporation of the Improved Grey Wolf Optimization (IGWO) with a novel formulated fitness model. Moreover, the proposed method exploits the least fitness measure to position the optimal centroid and the fitness measure based upon three constraints, such as intra-cluster distance, intercluster distance, and cluster density. The optimal centroid ensuing to the minimum value of the fitness is exploited for clustering the data. Simulation is performed by exploiting three datasets and the comparative evaluation is performed that shows that the performance of the developed method is better than the conventional algorithms such as Grey Wolf Optimization (GWO), Dragonfly and Particle Swarm Optimization (PSO).

Keywords: Data; Clustering; Database; Cluster Distance; Optimum Centroid; Optimization Algorithms

Nomenclature

Abbreviations	Descriptions
CL	Cannot-Link
CS	Cuckoo Search
GA	Genetic Algorithm
BHA	Black Hole Algorithm
ML	Must-Link
ACO	Ant Colony Optimization
CFPA	Chaos Optimization and Flower Pollination
DE	Differential Evolution
AIS	Artificial Immune System
IT2 FSs	Interval Type-2 Fuzzy Sets
ABC	Artificial Bee Colony
FOU	Footprint of Uncertainty ()
PSO	Particle Swarm Optimization
HPSOM	Hybridization of PSO With Mutation Operator
FF	Firefly Algorithm
IoT	Internet of Things

1. Introduction

Generally, data analysis inspires a lot of computing applications, either as a part of their design phase or in an online operation [2]. The process of data analysis can be classified into either exploratory or confirmatory, and it is on the basis of the accessibility of suitable models for the data source. However, a key part of both kinds of processes (whether for decision making or hypothesis formation) is the classification or grouping of measurements. Cluster analysis is the group of a compilation of patterns (typically indicated as a point in a multidimensional space, or a vector of measurements) into clusters on the basis of the similarity. For the grouping of data, the clustering algorithm is considered as an important unsupervised algorithm [20]. Moreover, clustering is an algorithm in that the clusters of items can create, which are someway linked in similar features. The main aspire of the clustering is to present a grouping of related records of data. Clustering is over and over again chaotic with classification;

https://doi.org/10.46253/j.mr.v2i3.a1

however, there is little divergence among the two. An easy measure is an intra-cluster distance that, as in the K-means method, required to be reduced for enhanced clustering outcomes. The phrase clustering is exploited in numerous research communities to portray techniques for the clustering of unlabeled data or untagged data.

Data clustering plays an important role in many areas. Specifically, data in the similar cluster possess maximum similarity, and data in different clusters, possess a minimum similarity. In data mining, clustering is a reasonably demanding algorithm with several business applications. Over the last few years, several strategies with respect to the data clustering were presented and fascinated multitudinous remarkable [4]. In data clustering algorithms, data mining applications maintain notable needs including the capability to obtain rapid clusters, end-user comprehensibility of the outcomes, insensitivity, and scalability to the sequence of input records. Consequential conventional notable clustering techniques possess maximum calculation time or might accost with the issue of placing pattern if the size of the database is broad. In data mining, the clustering algorithm is for determining the significant patterns in big data or large databases and using it for feature behavior prediction. Hence, the effective and efficient clustering algorithm is highly powerful [21].

In machine learning research, data clustering is considered as an issue with a great history. In a setting of the clustering, the aim is to cluster a set of items based on its features [1]. Data clustering has used in numerous applications in an extensive range of fields like business, network analysis, pattern recognition, data mining, health care, and etc. As there is no prior knowledge for the appropriate clusters, the problem of clustering is a challenging issue in the area of machine learning. In recent times, constrained clustering possess fascinated the attraction of numerous researchers by considering the side information into consideration. CL and ML are two renowned kinds of constraints. They indicate pairwise relationships among data hence that points constrained using an ML are compelled to be in a similar group, although a CL constraint indicates grouping into different clusters. Numerous real-time applications, for instance, multiview image correspondence, and content clustering on the basis of the user's feedback can be grouped as a constrained clustering issue and improved from the conventional effectual methods in this domain.

In recent times, Optimization-based clustering techniques have captured an efficient ability in resolving data clustering issues [9] because of their ability to find out enhanced solutions. Numerous swarm-based optimization techniques such as GA [10], PSO [14], ACO [11], ABC [13], AIS [12], FF [15], CS [16], BHA [17], and others were exploited. Swarm intelligence is considered as the cumulative behavior of agents attains in nature. Swarm technique possess fascinated huge significance in the past two decades [18]. Swarm-based techniques such as BHA, CSA, BA, FFA, PSO, so forth were exploited to issues categorized as NP-Complete or NP-hard and objective is to discover the optimal solution [19]. Aforesaid techniques inspired by nature were advantageous in designing computational intelligence methods efficiently and effectively. Nevertheless, aforesaid techniques can trap to local optima.

The main contribution of this paper is to present an approach called a DIGWO algorithm it is exploited in order to determine the optimal centroid for experiencing optimal data clustering. The most favorable clustering produces more helpful and the precious information needed for experiencing the decision making, which residuals intricate in the complex data environment.

2. Literature Review

In 2019, Arvinder Kaur et al [1], worked on classical clustering techniques such as K-means, which frequently converge to local optima and had sluggish convergence rates for bigger datasets. Swarm based approach had presented to conquer such circumstances in clustering. Swarm based algorithms try to attain the optimal solution for such issues in a sensible time. Here, a technique was developed that integrates CFPA against K-means in order to enhance the effectiveness of reducing the reliability of the cluster.

In 2019, Mohammed Alswaitti et al [2], presented a variance-based DE technique with a possible crossover for data clustering. Furthermore, the proposed method had the ability to improve the clustering solutions quality besides the convergence speed. The proposed technique contemplates the balance among the exploration and exploitation procedures. It was done by establishing a single-based solution demonstration, and a switchable mutation method Moreover, a vector-based calculation for the mutation factor, and finally a possible crossover scheme also established.

In 2019, Amit K. Shukla et al. [3] worked on uncertainties in the gene expression dataset by exploiting IT2 FSs. Hence, the spread of FOU that considers the entire probable noises in the gene expression dataset was developed for both asymmetrically and symmetrically. For the experimentations, the medical science big dataset of microarray gene expression data for the cancer patients was studied. By means of observation, the aspects of uncertainty modeling by exploiting IT2 FSs on the big data

clustering was observed and evaluated. Fuzzy clustering algorithm permits the genes to fit into multiple clusters and therefore permit the genes participation in cellular procedures, subcell deviations, and cellular metabolism. In the big data clustering, the aspect of the induced uncertainty had examined by exploiting several cluster validity measures.

In 2019, Ahmad Ali Abin [4] presented a new concept of tracking the candidate constraints quality in several embedding spaces. In contrast to the previous approach that computes the candidate constraints quality in the input data space. Moreover, the proposed algorithm sets in a distance matrix attainted from an arbitrary walk on the proximity graph into different embedding spaces and allocate every candidate constraint usefulness by tracking its attendance in the clusters' skeleton. The majority of helpful constraints were subsequently selected with respect to their quality and developed to the clustering algorithm.

In 2019, Manju Sharma et al [5], presented a novel sustainable clustering technique using HPSOM for the clustering of the data produced from various networks. In nature, the data produced by such networks were typically heterogeneous and dynamic and a number of clusters were not fixed/ known in the proceeding. Therefore, further, the proposed technique was expanded as AHPSOM for producing and re-adjusting the clusters mechanically against the mobile network devices, and it makes easy the generation of sustainable clusters.

In 2019, Farag Hamed Kuwil et al [6] presented a new distance-based clustering approach named the critical distance clustering technique. This technique based upon the Euclidean distance among data points and few essential mathematical statistics operations. The technique was easy, forceful, and supple; it works with quantitative data, which were real-valued, not qualitative, and definite with different dimensions. Here, 26 experiments were conducted by exploiting various kinds of real and synthetic datasets obtained from various areas. The outcomes show that the novel techniques are better than the few well-known clustering techniques namely K-means, MST based clustering, and Dbscan.

In 2017, Chun-Wei Tsai [7] worked on high-performance data analytics for IoT, which shows potential research areas in current years. Since conventional data mining techniques might not be appropriate for big data for IoT. The major motivations were the data required to be examined might goes over the storage size of a single machine. For a single computer system, the computation cost of data analysis tasks was extremely high, which was another major issue that needs to tackle while verifying data from an IoT system. Hence, a competent data clustering model for meta-heuristic technique on a cloud computing environment was developed for data analytics that portrays how to partition mining tasks of mining techniques into different nodes that were the Map process and subsequently aggregates the mining outcomes from these nodes that were Reduce process.

In 2019, Rosa Altilio [8], addressed the issue of distributed learning whereas the data to be clustered was divided against a set of interconnected agents possess restricted connectivity. In the last, all of these data were transmit in a central node, whereas the conventional techniques may discover the optimal division. At present, for the reason of processing, security, and privacy that a central node holds all of the information should be avoided and obtains a solution which was the same as the centralized one merely with a procedure of communication and association between the agents. A non-convex optimization in a multi-agent network with time-varying connectivity was developed to the renowned Expectation-Maximization algorithm to facade this issue.

3. Proposed DIGWO Algorithm for Data Clustering

The major contribution of the presented algorithm is to calculate the optimum centroid for data clustering, to group the number of data points available in the database. The data clustering procedure facilitates to extract the most effective and important information, which is hidden in a large number of the database that improves the decision-making procedure. Hence, it is much clear that the data clustering process presents helpful information for that an optimum centroid is needed. By exploiting the presented DGWO algorithm, the optimum centroid is determined by means of the investigation of the certain parameters like the intra-cluster distance, inter-cluster distance, and the density of the cluster. Moreover, the optimum centroid is adapted by exploiting the proposed method that is subsequently used for clustering a large number of data available in the database.

The database comprises mass data points and these data points are grouped as clusters on the basis of the optimal centroids. Moreover, the data points have many attributes, which are the tradition points to group the data points and thus the clusters are created on the basis of the similarities of these attributes available in the data points. In a group, the data points are the same as each other in contrast to the attributes of the different clusters that vary from each other. Assume n as the number of data points available in the database and that can be indicated as eq. (1).

$$\mathrm{Db} = \left\{\mathrm{Db}_1, \mathrm{Db}_2, ..., \mathrm{Db}_g, ...\mathrm{Db}_n\right\} \tag{1}$$

In eq. (1), Db represents the database that conveys total data points that are indicated as Db_1, Db_2, and hitherto. Db_g represents g^{th} data points available in the database. Each data point conveys total data attributes that can be also referred to as data features. Eq. (2) indicated as the attributes equivalent to the g^{th} data point.

$$at = \{at_{g1}, at_{g2}, ..., at_{gv}\} \tag{2}$$

In eq. (2), v indicates the total amount of attributes and at_{gv} indicates the v^{th} attribute equivalent to the g^{th} data point. The data points are grouped on the basis of the centroids and the centroids portray the number of clusters. Consider the centroids which can be indicated as eq. (3).

$$CR = \{CR_1, CR_2, ..., CR_s, ...CR_i\} \tag{3}$$

In eq. (3), i correlates to the total number of clusters and CR_s indicates the s^{th} cluster.

Data clustering is the procedure of grouping the data points and it is performed using the proposed DGWO method, which is shown in Fig. 1. The main aim of exploiting the DGWO method is to decide the optimum centroid for clustering the data points that are available in the database. The DGWO method exploits a fitness function for clustering, which is based upon the intra-cluster distance, inter-cluster distance, and the density cluster. The fitness measure is exploited in order to discover the optimum centroid point and on the basis of this optimum centroid point, the data are grouped into clusters. The efficient grouping of the data points offers the effectual information available in the database that is important for efficient decision making.

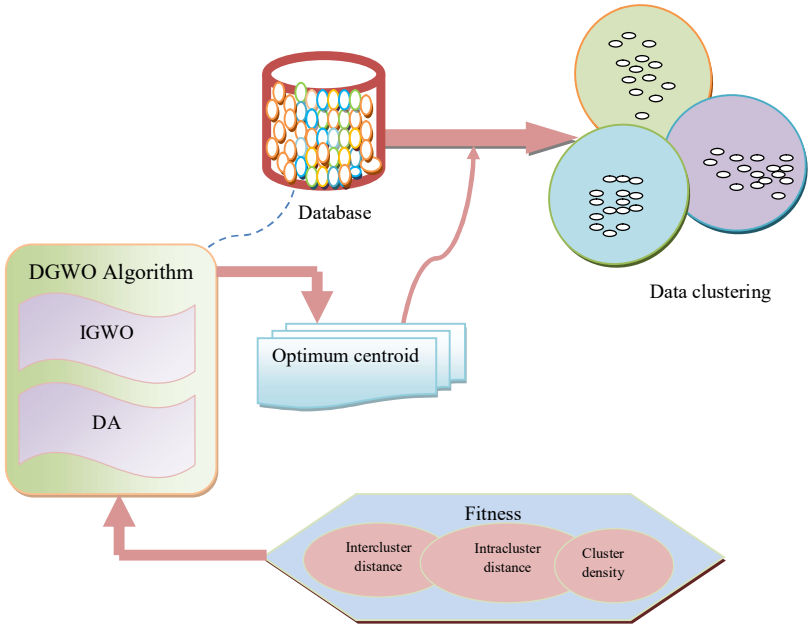

Fig. 1. *Bock diagram for data clustering using the proposed method*

3.1 Calculation of Fitness Measure

The fitness measure is important for determining the fitness of the clusters that facilitates the optimum centroid determination so that valuable data clustering is performed. The fitness measure is based upon three constraints, namely the intra-cluster and inter-cluster distance and the cluster density. The fitness measure is done least by exploiting these three constraints and the centroid which produces the least fitness value is done as the optimum cluster.

$$F_v = \frac{[I_c + (1 - I_R) + (1 - I_Y)]}{3} \tag{4}$$

In eq. (4), I_R indicates the intra-cluster distance, I_c indicates the inter-cluster distance and I_Y indicates the cluster density. By exploiting these fitness constraints the fitness measure is performed least and the centroid that obtains the least fitness measure, which is implemented as the data classification and the optimal centroid, is based upon the optimal centroid. The parameter I_c must be undoubtedly least and the other parameters $(1 - I_R)$ indicate the intra-cluster distance and $(1 - I_Y)$ indicate the cluster density which should be high.

a) Distance of Inter-cluster:

The inter-cluster distance is considered as the most significant fitness constraints. Here, the distance matrix is created by exploiting the centroid point and the data point which exhibits the distance among the centroid points and data points. The least distance among the data points and the centroids are used and the least distances are added together. The summation of the least distance among the centroid point and the data point is needed to endure as the least value.

$$I_c = \frac{\sum_{\substack{g=1 \\ g\in s}}^{n} \left\| Db_g - CR_s \right\|^2}{n \times Max_{\substack{g=1 \\ g\in r}}^{n} \left\| Db_g - CR_s \right\|^2} \tag{5}$$

In eq. (5), Db_g indicates the g^{th} data point and CR_s indicates the s^{th} cluster. For clustering, n indicates the total number of data points available in the database. $\left\| Db_g - CR_s \right\|^2$ indicates the Euclidean distance among the cluster and the data point and the inter-cluster distance based upon this Euclidean distance. The summation of the Euclidean distances of the data point's equivalent to the clusters decides the inter-cluster distance value and this value of the distance are normalized. The normalization is done by exploiting the product of the number of data points and the maximum distance among the cluster and the data point.

b) Distance of Intra-cluster:

The second constraint is the Intra-cluster distance that determines the fitness measure. The distance among the clusters is computed on the basis of the Euclidean distance measure and this measure is normalized by exploiting the maximum distance attained when computing the intra-cluster distance of the m number of clusters. The intra-cluster distance computed among the two clusters is needed to be a high value to attain the least fitness measure value.

$$I_R = \frac{2 \times \sum_{s=1}^{i} \sum_{k=s+1}^{i} \left\| CR_s - CR_k \right\|^2}{s(s-1) \times Max_{\substack{s=1 \\ k=s+1}}^{i} \left\| CR_s - CR_k \right\|^2} \tag{6}$$

In eq. (6), CR_s indicates the s^{th} cluster and I_R indicates the intra-cluster distance and i indicates a total number of the clusters. The value of the sum of the intra-cluster distance is computed using the Euclidean distance among the centroids and the distance value is summed hence the summed distance is surely least. The I_R value is normalized by exploiting the least value of the distance among the centroids.

c) cluster density:

The cluster density is considered as another parameter which indicates the number of data points available in a cluster. It is important to have a large number of data points in the cluster hence to least the fitness function. Besides with the least inter-cluster distance, and the large intra-cluster distance, the cluster density must be high.

$$CR_Y = \frac{Min_{s=1}^{i} \left| CR_s \right|}{Max_{s=1}^{i} \left| CR_s \right|} \tag{7}$$

In eq. (7), $\left| CR_s \right|$ indicates the number of data points available in the s^{th} cluster. Hence, high intra-cluster distance, the centroid with the least inter-cluster distance, and the high number of the data points in the cluster compose the least fitness function. The centroid that fulfills the fitness measure is the optimal possibility of the optimal centroid to carry out clustering.

3.2 Conventional GWO Algorithm

GWO method is inspired by the hunting behavior and social nature of wolves. Moreover, to improve the method, the wolves' behavior is mathematically performed, and the leader indicated as α wolf. Successively, resultant to better outcomes called as β, and γ wolves. The residual results attained can be calculated as ω wolves. The approach of optimization is going after by α, β, and γ in the GWO algorithm [23]. The ω wolves go after the α, β, and γ wolves to search the worldwide optimization. The eq. (8) and (9) indicates the hunting behavior of wolves.

$$\vec{B} = \left| \vec{A} \cdot \vec{Y}_p(t) - \vec{Y}(t) \right| \tag{8}$$

$$\vec{Y}(t+1) = \vec{Y}_p(t) - \vec{C} \cdot \vec{B} \tag{9}$$

In eq. (8) and (9), t indicate the number of iteration, $\vec{B}$ and $\vec{C}$ indicate a coefficient vector. $\vec{Y}_p$ indicates the position of prey, $\vec{Y}$ represents a wolves position.

$$\vec{C} = 2\vec{c} \cdot \vec{r}_1 - \vec{c} \tag{10}$$

$$\vec{A} = 2 \cdot \vec{r}_2 \tag{11}$$

The elements in eq. (10) and (11) are incessantly minimizing the value 2 to 0 above the sequence in iterations.

Here, r_1 and r_2 denotes the unsystematic vectors[0,1]. The followers and leadership ways of grey wolves are stated as follows:

$$\vec{B}_\alpha = \left| \vec{A} \cdot \vec{Y}_\alpha - \vec{Y} \right| \tag{12}$$

$$\vec{B}_\beta = \left| \vec{A} \cdot \vec{Y}_\beta - \vec{Y} \right| \tag{13}$$

$$\vec{B}_\delta = \left| \vec{A} \cdot \vec{Y}_\delta - \vec{Y} \right| \tag{14}$$

$$\vec{Y}_1 = \vec{Y}_\alpha - \vec{C}_1 \cdot \left(\vec{B}_\alpha \right) \tag{15}$$

$$\vec{Y}_2 = \vec{Y}_\alpha - \vec{C}_2 \cdot \left(\vec{B}_\alpha \right) \tag{16}$$

$$\vec{Y}_3 = \vec{Y}_\alpha - \vec{C}_3 \cdot \left(\vec{B}_\alpha \right) \tag{17}$$

$$\vec{Y}(t+1) = \frac{\vec{Y}_1 + \vec{Y}_2 + \vec{Y}_3}{3} \tag{18}$$

IGWO algorithm exhibits three better outcomes, which indicates α, β, and γ and these are indicated as leaders. To recognize the optimal solution near to the worldwide optimization aforesaid leaders aids to explore the agents to attain the selected areas in the search space. The proposed IGWO method different developments are complete hence discovering a solution to issues of effectiveness and convergence rate. Here, the $\vec{a}$ parameter is studied as an arbitrary vector in the range [0, 1] that values are vital in given that balance among exploitation and exploration. Adaptive values of $\vec{a}$ parameter sustain exploration to put off obtaining trapped in local optima and deal with the precise issue. It is a significant parameter in refining the ensued vectors and economically used in operating the rate convergence in the algorithm. With this purpose that imitates the hunting behavior of grey wolves a postulation is done in which the fittest wolf α possesses enhanced knowledge on the well-organized position of the prey, and for this cause, only the fittest solution is stored. The global optimal solution attained from the complete population aids in attaining the global optimum. The IGWO method is presented to present improved performance regarding evading obtaining trapped in convergence rate, pre-mature convergence, and precision.

3.3 Conventional Dragonfly Algorithm

Generally, Dragonfly Algorithm (DA) is theoretical to be a minute predator in nature that pursues chiefly the other small insects. An interesting fact on a dragonfly is its characteristic and astonishing swarming behavior. The chief principle of the swarm of Dragonflies is exploited to migrate and hunt. The previous motivation is recognized as a static or feeding swarm, the last one is called as the dynamic or migratory swarm [24]. The leading inspiration of the Dragonfly method starts in static as the preliminary aim and the energetic swarming behaviors of a dragonfly. Both the swarming behaviors seem similar in both the two type phases in optimization during Metaheuristic algorithms, that is, exploration and exploitation. In the swarm of static, all the dragonflies form sub-swarms to blast-off over the various areas, namely the major reason for the exploration phase.

a) Steps for Dragonfly Algorithm:

Step 1: The dragonflies' behavior can be explained regarding five steps, such as departure, consistency, arrangement, magnetism in the way of a food source and interruption in the direction of the outside an enemy [28]. The eq. (19) states the calculation of departure.

$$r_j = -\sum_{i=1}^{N} Y - Y_i \tag{19}$$

In eq. (19), Y indicates the position of a particular individual Y_i indicates the location of i^{th} close to the individual, N is a number of the close to individuals.

Step 2: Eq. (20) denotes the calculation of the position, where, U_j indicates the velocity of i^{th} close to the individual.

$$C_j = \frac{\sum_{i=1}^{N} U_i}{N} \tag{20}$$

Step 3: Eq. (21) is used to calculate the consistency

$$a_j = \frac{\sum_{i=1}^{N} Y_i}{N} - Y \tag{21}$$

Step 4: In eq (22), Y and Y^+ denotes the position of current individual equivalent to the source of food, correspondingly.

$$FS_j = Y^+ - Y \tag{22}$$

Step 5: Eq. (23) is used to calculate the interruption noticeable of the enemy.

$$EN_j = Y^- + Y \tag{23}$$

Eq. (23), Y and Y^- represents the position of the conventional individual and an enemy, correspondingly $EN_j = Y^- + Y$ indicates updating the position of the simulated dragonfly in the research space to imitate the advancements of dragonflies, both the vectors, that is step (ΔY) vector with location (Y) vector are measured. The phase vector exploited in DA is the same as the vector velocity of PSO. In fact, the DA is developed by exploiting the model of the PSO algorithm. In the DA algorithm, the phase vector ΔY presents the way of development of the dragonflies. Eq. (24) is used to calculate the step vector.

$$\Delta Y_{f+1} = (rR_j + cC_j + aA_j + eE_j + fF_j) + q\Delta Y_f \tag{24}$$

In eq. (24), R_j states parting of j^{th} entity, C_j states j^{th} individual configuration, A_j states j^{th} individual consistency, E_j denotes food resource in j^{th} entity, F_j denotes the position of the enemy of j^{th} entity, c denotes configuration mass, a denotes consistency mass, r denotes parting mass, e denotes food aspect, f denotes enemy source, q denotes indolence weight, t denotes iteration count.

$$\Delta Y_{f+1} = Y_f + levy(d) * Y_f \tag{25}$$

Once the step vector ΔY is calculated, the subsequent step is to calculate the location vectors Y using eq. (26)

$$Y_f = Y_f + \Delta Y_{f+1} \tag{26}$$

Whereas there is no close solution, the location of dragonflies is updated using an arbitrary walk (Levy flight). Hence, the location vectors Y are computed using eq. (27).

$$Y_{f+1} = Y_f + levy(d) * Y_f \tag{27}$$

The neighborhood area is improved and eventually, at the last stage of the optimization procedure, the swarm turns out to be merely one group. Food resources with the adversary are chosen from superior beside with the deprived outcomes attained in every swarm at any instantaneous. This tends the convergence towards the show's potential areas of research space and at the related time, it tends discrepancy outward the non-promising areas in search space.

3.4 Proposed DIGWO Algorithm

GWO algorithm is on the basis of the Swarm Intelligence approach of optimization which is similar to the leadership of grey wolves regarding its hunting behavior. This algorithm comprises of two modifications in conventional GWO. Mainly, increases in the diversity with the adversary learning method basis. Subsequently, the value of the parameter, which oscillates among the value [2, 0] for nearly 75% iteration and it attained some continued values for the remaining period. The conventional DA presents the exploration and exploitation stage. While integrating these two stages of the Dragonfly algorithm it aims final outcomes that could be attained. Steps in DGWO algorithm is stated as follows:

Eq, (28) represents the initial search agents which adopt the position often using several sets of uniform number distributing ranges of the boundaries.

$$y_j = y_j^{min} + v\left(y_j^{max} - y_j^{min}\right) \tag{28}$$

Where, y_j indicates the j^{th} wolf following to the j^{th} direct variable, y_j^{max} beside y_j^{min} are the superior and smaller limits of the j^{th} direct variable beside with v indicates the same chance in the time [0, 1]. In addition, generate a repository for Pareto–set. By fitness function of data, clustering is used to calculate the basic process in the course of iteration most of the wolves do the complete NR (Newton-Raphson) investigation. The fundamental of the attained outcome from the NR data clustering the value of the basic process is computed. The approach of Pareto is stated below:

$$\forall_j = \{1,2,....,n\}, F_{j(y1)} \le F_{j(y2)} \tag{29}$$

$$\Xi_i = \{1,2,....,n\}, F_{i(y1)} \le F_{i(y2)} \tag{30}$$

Saved in the most important set of Pareto is subsequently attained following the application. Comparison of the scheme in supremacy is adapted to update the records is stated below:

- ❖ It is significant too, they attained outcomes might not be stored in the repository if the attained value is feeble or conquered by the neighboring number in the colony.
- ❖ The attained outcomes can be stored in the repository while the non dominated neighbour might not be dominated by any outcomes in the conventional depositary.
- ❖ The other complete dominated outcome in the conventional depositary will be eradicated by the non dominated individual.

The main content from the Pareto set in all iteration must be updated; inserting the conventional non dominated individual in the depositary is performed. All the dominated places are discarded in the procedure by the depositary.

The eq. (31) and (32) followed by prey Encircle: during the hunt, all the grey wolves encircle the prey.

$$C = 2c \cdot r_1 - c \tag{31}$$

$$A = 2 \cdot r_2 \tag{32}$$

Where, a minimized successively from 2 to 0, r_1 and r_2 are distributed arbitrarily within the range [0, 1].

The boundary limits confirmation has a definite case, obliteration in the difference limits, the position of the resultant wolves have to be fixing by using eq. (33). Fig 2 illustrates the flow chart of the proposed DGWO model.

$$y^{lim} = \begin{cases} y^{max} & \text{if } y > y^{max} \\ y^{min} & \text{if } y > y^{min} \end{cases} \tag{33}$$

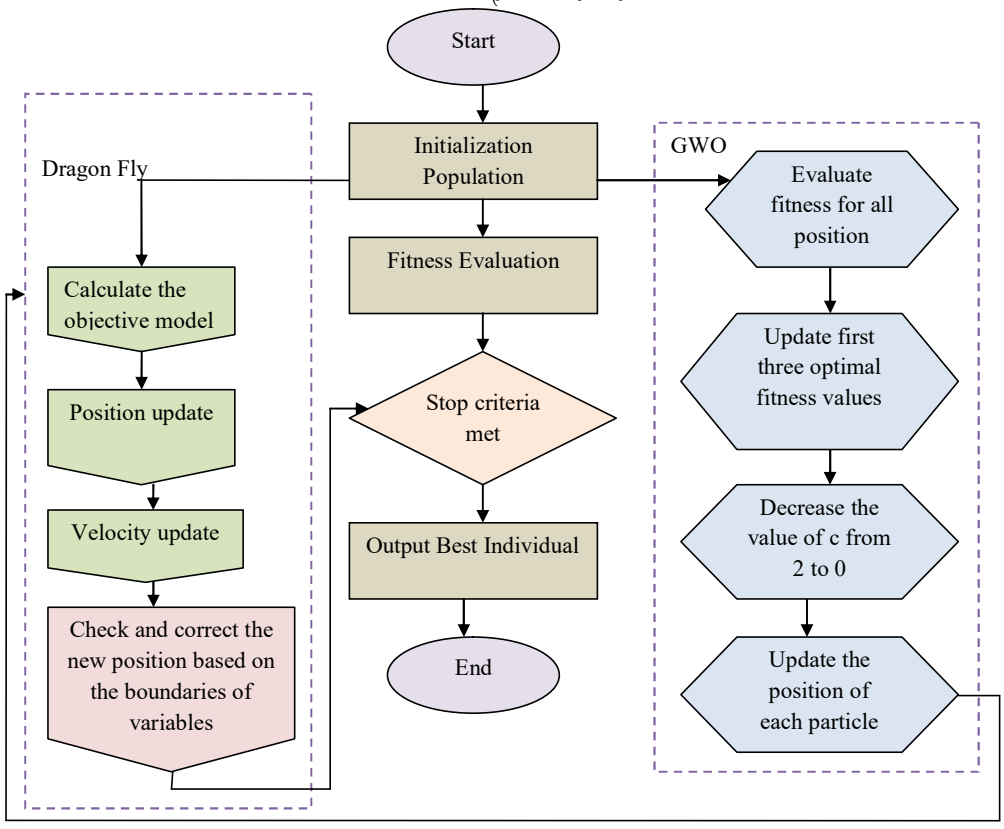

Fig. 2. *Flow chart of the proposed DGWO model*

4. Results and Discussions

4.1 Experimental Procedure

In this section, the outcomes of the proposed DGWO approach are presented. Here, the performance evaluation of the proposed algorithm was performed regarding the performance metrics, which highlights the advantage of the proposed algorithm. Moreover, this work exploits 3 databases for the investigational evaluation which were obtained from the UCI machine learning repository [24] and the 3 datasets were stated as follows:

a) Dataset 1: Banknote authentication dataset

In this dataset, the images are structured from a real and forged banknote-like and a camera of size 400×400 pixels is exploited for print inspection is utilized for digitalizing the image. The wavelet transformation is exploited and the attributes comprise the variance (continuous), skewness (continuous), curtosis (continuous) from the wavelet image and the entropy of image (continuous) and class (integer) for feature extraction.

b) Dataset 2: Iris dataset

In this dataset, 50 instances for every class is present that totally comprises 150 data instances with 3 classes. The total number of attributes contemplated in the iris dataset is 4 with one class.

c) Dataset 3: Wine dataset

In data clustering, this dataset is most frequently exploited that comprises the chemical analysis of wines developed in a similar region in Italy although derived from 3 different cultivars.

4.2 Performance Analysis

In this section, the investigational evaluation of the proposed technique is presented, which is done by exploiting datasets used from the UCI machine learning repository [24] from that 3 datasets are extracted for analysis. Here, the proposed method is compared with conventional algorithms such as GWO, DA, and PSO.

Fig 3 states the performance analysis of the proposed method with respect to the data set 1(i.e.,) Bank note authen. Here, the analysis is done with the aid of the performance metrics such as Rand coefficient, F-measure, and Jaccord coefficient, and MSE. Here, the proposed algorithm obtains a greater value for the metrics rand coefficient, F-measure, and the jaccord coefficient but lesser values for the MSE.

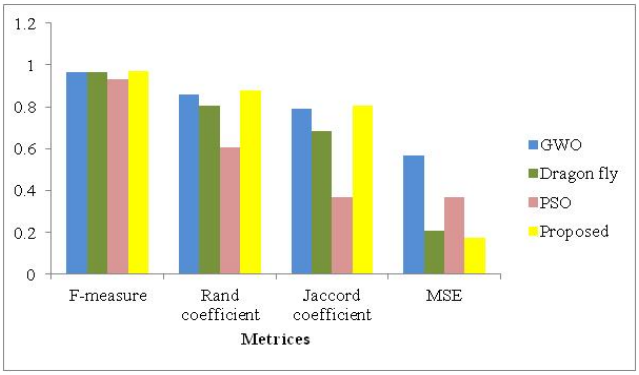

Fig. 3. *Performance analysis of the proposed method for data set 1 (Bank note authen)*

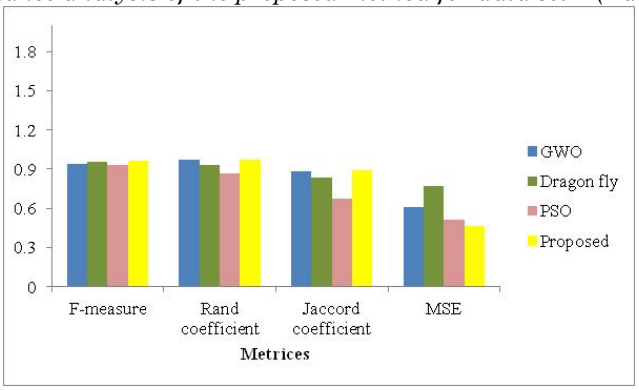

Fig. 4. *Performance analysis of the proposed method for data set 2 (Iris)*

Fig 4 summarizes the performance analysis of the proposed technique regarding the data set 2 (i.e.,) Iris. Here, the analysis is performed with the help of the performance metrics such as Rand coefficient, F-measure, Jaccord coefficient, and MSE. Here, the proposed DGWO obtains a higher value of the rand coefficient, F-measure, and jaccord coefficient while comparing with the conventional algorithms such as GWO, DA, and PSO. Moreover, the MSE value attained by exploiting the proposed algorithm is very less while comparing with the conventional algorithms.

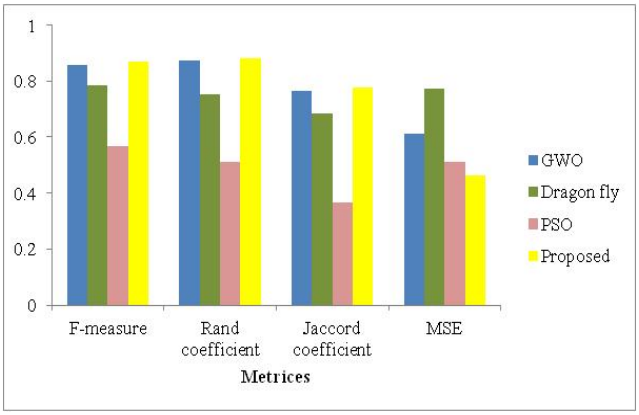

Fig. 5. *Performance analysis of the proposed algorithm for data set 3 (Wind dataset)*

Fig 5 explains the performance analysis of the proposed algorithm about the data set 3(i.e.,) Wind dataset. Here, the analysis is performed with the assist of the performance metrics like Rand coefficient, F-measure, Jaccord coefficient, and MSE. From the comparative analysis of the proposed DGWO algorithm with conventional GWO, DA, and PSO algorithms, it is recognized that the Proposed DGWO algorithm has enhanced performance concerning F-measure, Rand coefficient, Jaccord coefficient, and MSE.

5. Conclusion

In this paper, the DGWO method was used for clustering the data for that the proposed method exploits the combination of the methods, such as the GWO and the IGWO. The proposed method calculates the optimal centroid to do the data clustering which was on the basis of the minimum fitness function. The minimum fitness function was examined on the basis of the three fitness constraints, such as the intra-cluster distance, inter-cluster distance, and cluster density. Moreover, the fitness measure is calculated as the least value for that the intra-cluster distance, inter-cluster distance, and cluster density must be less. Once the optimal centroid is calculated by exploiting the DGWO algorithm, this optimal centroid was exploited to cluster the data points available in the data. For the decision making the procedure, the clustered data were the optimally clustered data and it presents all the important information. Finally, the performance analysis was performed by exploiting the three datasets, which shows that the proposed DGWO algorithm was better than all the conventional algorithms by obtaining a less value of MSE.

Compliance with Ethical Standards

Conflicts of interest: Authors declared that they have no conflict of interest.

Human participants: The conducted research follows the ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

References

[1] Arvinder Kaur, Saibal Kumar Pal, Amrit Pal Singh,"Hybridization of Chaos and Flower Pollination Algorithm over K-Means for data clustering", Applied Soft Computing, In press, corrected proof, Available online 24 May 2019.

[2] Mohammed Alswaitti, Mohanad Albughdadi, Nor Ashidi Mat Isa,"Variance-based differential evolution algorithm with an optional crossover for data clustering", Applied Soft Computing, Volume 80, July 2019, Pages 1-17.

[3] Amit K. Shukla, Pranab K. Muhuri,"Big-data clustering with interval type-2 fuzzy uncertainty modeling in gene expression datasets", Engineering Applications of Artificial Intelligence, Volume 77, January 2019, Pages 268-282.
[4] Ahmad Ali Abin,"Querying informative constraints for data clustering: An embedding approach",Applied Soft Computing, Volume 80, July 2019, Pages 31-41.
[5] Manju Sharma, Jitender Kumar Chhabra,"Sustainable automatic data clustering using hybrid PSO algorithm with mutation", Sustainable Computing: Informatics and Systems, Volume 23, September 2019, Pages 144-157.
[6] Farag Hamed Kuwil, Fadi Shaar, Ahmet Ercan Topcu, Fionn Murtagh,"A new data clustering algorithm based on critical distance methodology", Expert Systems with Applications, Volume 129, 1 September 2019, Pages 296-310.
[7] Chun-Wei Tsai, Shi-Jui Liu, Yi-Chung Wang,"A parallel metaheuristic data clustering framework for cloud", Journal of Parallel and Distributed Computing, Volume 116, June 2018, Pages 39-49.
[8] Rosa Altilio, Paolo Di Lorenzo, Massimo Panella,"Distributed data clustering over networks", Pattern Recognition, Volume 93, September 2019, Pages 603-620.
[9] B.K. Elfarra, T.J. El Khateeb, W.M. Ashour, BH-centroid: A new efficient clustering algorithm, Int. J. Artif. Intell. Appl. Smart Dev. 1 (2013) 15–24.
[10] C. Grosan, A. Abraham, M. Chis, Swarm intelligence in data mining, Stud. Comput. Intell. 34 (2006) 1–20.
[11] D. Martens, B. Baesens, T. Fawcett, Editorial Survey: Swarm Intelligence for Data Mining Machine Learning, vol. 82, Springer, 2011, pp. 1–42.
[12] R. Younsi, W. Wang, A new artificial immune system algorithm for clustering, in: Z.R. Yang (Ed.), in: LNCS, vol. 3177, Springer, Berlin, 2004, pp. 58–64.
[13] D. Karaboga, C. Ozturk, A novel cluster approach: Artificial bee colony (ABC) algorithm, Appl. Soft Comput. 11 (1) (2010) 652–657.
[14] X.S. Yang, Nature Inspired Metaheuristic Algorithms, second ed., Luniver Press, 2008.
[15] T. Hassanzadeh, A new hybrid approach for data clustering using firefly algorithm and k-means, in: The 16th CSI International Symposium on Artificial Intelligence and Signal Processing, 2012.
[16] D. Binu, M. Selvi, A. George, MKF-cuckoo: hybridization of cuckoo search and multiple kernel-based fuzzy C-means algorithm, in: AASRI Conference on Intelligent Systems and Control, vol. 5, Elsevier, 2013, pp. 243–249.
[17] B.K. Elfarra, T.J. El Khateeb, W.M. Ashour, BH-centroid: A new efficient clustering algorithm, Int. J. Artif. Intell. Appl. Smart Dev. 1 (2013) 15–24.
[18] X.S. Yang, Nature Inspired Metaheuristic Algorithms, second ed., Luniver Press, 2008.
[19] A. Vattani, K-means requires exponentially many iterations even in the plane, Discrete Comput. Geom. 45 (2011) 596–616.
[20] J.S. Anita and J.S. Abinaya,"Impact of Supervised Classifier on Speech Emotion Recognition",Multimedia Research, Volume 2, Issue 1, January 2019.
[21] P.Jegatheeswari and T. Angelin Deepa,"Fuzzy Weighted Least Square Filter for Pansharpening in Satellite Images",Multimedia Research, Volume 2, Issue 1, January 2019.
[22] Byamakesh Nayak, Alivarani Mohapatra, Kanungo Barada Mohanty,"Parameter estimation of single diode PV module based on GWO algorithm", Renewable Energy Focus, Volume 30, September 2019, Pages 1-12.
[23] Sree Ranjini K.S., S. Murugan,"Memory based Hybrid Dragonfly Algorithm for numerical optimization problems", Expert Systems with Applications, Volume 83, 15 October 2017, Pages 63-78.
[24] Datasets from <http://archive.ics.uci.edu/ml/>.