International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 03, 2020
ISSN: 1475-7192

An Improved Gray Wolf Optimization (IGWO) and its Linkage to K-Mean for using in Data Clustering

[1]Ali Al-lami, [2]Adel Ghazikhani, [3]Hussein Al-kaabi

Abstract

The k-mean algorithm remains one of the most well-known and widespread clustering algorithms,whose initial centers are chosen randomly and while being optimally placed, their application can be easily implemented. The meta-heuristic algorithmcan provide data clustering with the optimal solution. This algorithm can also minimize the issue of local minimums. The present study aimsat improvingthe k-mean algorithm's accuracy with use of combined and meta-heuristic techniques.Hence, this paperaddresses an algorithm (Improved Gray Wolf OptimizationK-mean). The optimized form of improved gray wolf is employed for automatically detecting the clusters' number and obtaining the optimal solution as K-mean clustering outcomes and initial K-mean clusteringcenters. The results revealed that the proposed method bears a lesspercentage of errorcompared to the existing methods and reduces by 12%. Additionally, the aggregate distance of intra-cluster was also decreased.

Keywords: Data Clustering, meta-heuristic algorithm, Gray Wolf Optimization (GWO), k-mean.

I. Introduction

Deriving knowledge from numerous data sets is the most overridingcharacteristics of the information era. Advancing in computer technology, particularly the Internet, entailed an "information explosion" [1]. Data frequency availability augments the controllability complexity of these data and affects the constructive decision-making process. Data clustering, therefore, hasplayeda vital role in data mining and helpful clustering technique can improve effective decision-making [2]. The core objective of data clustering lied in arranging optimally all N data things into K clusters and this should be followed in a way in which all architypesunseen in the data become visible[3]. Each clusterconsists of data objects that are the same in terms of size, while the cluster groups differ from each other. The cluster analysis posesacrucialmethod in data mining, machine learning, pattern recognition, neural computing, image segmentation, and other engineering areas. Today, a few clustering algorithms are applied by

[1] Department of computer Engineering, University of Imam reza, Mashhad. Iran.
[2] Department of computer Engineering, University of Imam reza, Mashhad. Iran.
[3] General Directorate of Vocational Education, Ministry of Education in Iraq.

scholars that could be categorizedinto different categories of Segmentation based clustering algorithm, Hierarchical clustering algorithm, Network based clustering algorithm, Density based clustering algorithmand model based clustering algorithm.In segmentation based clustering algorithms, data is classified as K number clusters utilizing Euclidean distance and the treeof clusters is formed in hierarchical clustering algorithm. In the following, two types of dense hierarchical clustering algorithms and hierarchical dividing clustering algorithms stem from the hierarchical algorithms. Using evolutionary algorithms or swarm intelligence for optimal clustering has been recently become a prevalent choice to solve complicate clustering issue [4]. From the optimization standpoint, clustering problems could be formally regardedas a certainkind of solid NP clustering problem. Such algorithms seek to find a proper solution for solving clustering problems and lead to reduced risk of placement in an optimal point. Nonetheless these algorithmsare notbound to genetic algorithms (GA), they cover simulated annealing (SA), Tabu search Artificial Bee Colony (ABCs), Ant colony optimization (ACO), Greedy Randomized Adaptive Search Procedure (GRASP), Iterated Local Search (ILS), Variable Neighborhood Search (ILS), particle swarm optimization (PSO), etc., [5].

II. Previous works

This sections provides a literature review conducted in the area of clustering using meta-heuristic algorithms.

2.1. Clustering using meta-heuristic algorithms

A method of data clustering incorporating the WGC algorithm in determining the optimal centralization has been proposed to carry out the clustering process in [2]. Applying WEGWO with a noble formulated fitness function, the WGC algorithm benefits from the computational steps of the Whale Optimization Algorithm (WOA). Itutilizes the minimum fitness unit for specifyingthe location corresponding to optimal centralization. The minimum unit is dependent on three ranges as follows: inter-cluster distance, intra-cluster distanceand cluster density. The minimum fitness value which are applied to data mining is corresponded to the optimal centralization.

Akernel exponential gray wolf optimizer (KEGWO) for evaluating rapid centralization in data clustering has been developed by authors in [4]. Recently, KEGWO has been offered to carry out rapid centralization search by a new target evaluation, considering two parameters of the difference between two upper clusters and logarithmic kernel function. According to the new objects function and ModifiedKEGWO Algorithm, access to status vectors and optimal centralization to the last clustering can lead to decoding the centralization.

Relying onthe recently developed meta-heuristic optimization algorithm, named the Multivariate Optimization Algorithm (MOA), the authors proposed a new clustering technique for automatically finding the optimal solution via global and local replacement search using global exploration team and a few local exploration groups [6]. The scholars also suggested methods helpful for data clustering with the use of cuckoo optimization algorithm, which is also known as COAC or Fuzzy Cuckoo Optimization Algorithm (FCOAC) [7]. The Cuckoo Optimization Algorithm (COA) which incorporates the cuckoo bird’s nature of life in problem-solving, aims at solving continuousissues. Clustering an enormous amount of data, this algorithm ranksthe number of known

clustersusing the meta-heuristic algorithm and optimizes the outcomesby optimalambiguous logic. A multiple kernel algorithm relying on Dynamic Fractional Lion Optimization has been proposed for data clustering by researchers in [8]. The proposed method considers the database as the input. The current researchtakes two aspects into consideration, the proposed Adaptive Dynamic Directive Operative Fractional Lion (ADDOFL) algorithm and the fitness function and. The former applies the dynamic directive operation search algorithm to the algorithm of the adaptive fractional lion. This algorithm also employs the new fitness function for assessingthe search agents' value. The algorithm, chiefly inspired by the lion's prideful behavior,makes the updated female lion identify the optimal value on the basis of the dynamic directive operational search scheme. Consequently, the proposed ADDOFL algorithm has been applied to acquire the optimal clustering center so as toconduct data clustering.The latter is developed by four kernel functions namelylogical quadratic, Gaussian, tangential and inverse multilevel quadratic functions. Therefore, ambiguous WLI clustering is used to compute distance on the basis of the new kernel function namedmulti-kernel WLI (MKWLI). A new technique of clustering-based gene selection, in which multiplecalculations were attainedbykernel functions was proposed in [9]. This method looks for simultaneously optimizing the most suitableclustering-associated weights of genes for improving target function. This process uses adaptive distance to either discover the weight of genes over clustering, or amelioratethe algorithm's performance. This algorithm is straightforward and needs no parameter correction or optimization for each dataset.

Three meta-heuristic kernel intuitionistic fuzzy c-means(KIFCM) algorithmsincluding PSO-KIFCM, GA-KIFCM and ABC-KIFCM algorithms are proposed in [10],tryingin their population to achieve the optimal cluster centralization. In PSO-KIFCM and GA-KIFCM algorithms, KIFCM algorithm is carried out solely once and the lastclusters are achievedvia KIFCM algorithm with the use of PSO and GA algorithms as their preliminary solution. Whereas, other initial solutions are choen from the dataset randomly. It means that, in the ABC-KIFCM algorithm, every onlooker beeuses the KIFCM algorithm to upgrade its position. To do so, the onlooker bee's current location remains the final centralization which becomes available by the KIFCM algorithm.

A modified BCO (MBCO) scheme has been used for data clustering[11], through with bee-forgotten characteristics as well as similar chance for both reliable and unreliable bees become applicable.The MBCO proposes a selection-based likelihood approach for selecting unallocated data points in each iteration. The proposed MBCO is combined with the K-mean algorithm in order to improve the MBCO'sperformance and reach a global and disparate optimal solutions. The hybrid MBCO, k-means (MKCLUST) and k-means and MBCO (KMCLUST), on average, outperform the proposed Modified Bee Colony Optimization (MBCO).

III. The proposed method

Due to the characteristics of K-mean clustering algorithm,it can be integrated in many algorithms. The gray wolf optimization algorithm is resulted from combining an optimization algorithm with clustering algorithms. For solvingsuchissues, this algorithm likethe K-mean clustering algorithm no longer requires information and labeling in excess, therefore it can make use ofintegrating two algorithms. Additionally, the combination of two algorithms can

assist to solve one of the clustering problems. The underlying problem pertains to a big inter-cluster distance the K-mean clustering has. This distance can be reduced using Gray Wolf Optimization K-Means (GWO-K-means).

3.1. Initial population by GoodPointset method

An important factor in intelligent computing techniques is initialization as it influences the speed and accuracy of convergence. No information on the optimal global location exits prior to solving the optimization problem. There is a need that the initial candidate members getcompletely scattered in the solution space, therefore the algorithm could look for the solution space wholly. It needed to bear in mind that, the faster the searches of the available optimal global area are, the better the proposed solutions should search the entire flexible space in comparison with initial time in the IGWO algorithm. Suitable dataset methodswerebroadlyapplied to produce initial proposed solutionsfor collectingthe information of flexible solution space in order to increase population diversity [16]. Hence, in current study, an appropriate dataset method is proposed to develop an initial population's appropriate set in the IGWO algorithm. Figure 1 demonstrates the flowchart of the proposed hybrid method (IGWO-K-means).

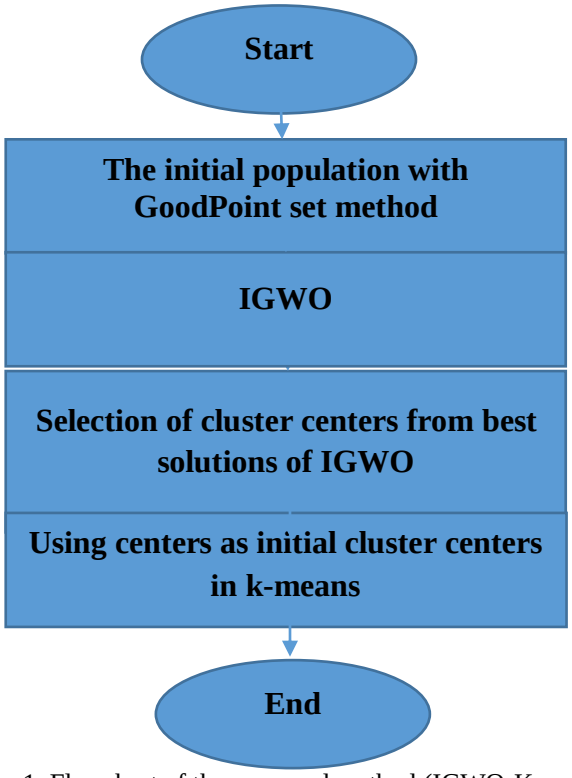


Figure 1. Flowchart of the proposed method (IGWO-K-means)

As cited in the base paper of the K-Means centers, there has been used an evolutionary algorithm to acquire optimal values for it owing to the sensitivity of K-Means to the centers' initial value; therefore if not chosen properly, the convergence remains unfulfilled. Because of this, it is going to employ an evolutionary approach. The algorithm applied in this researchcorresponds to a type of GWO algorithm. The first, second, and third bests in the

gray wolf optimization algorithm, are taken into account to update each new wolf's position; despite this paper includes a fourth best in Equation (1) as it is divided into four sectionswhile updating (Equation 2).

$$\vec{x}1 = \vec{x}\alpha - \vec{a}1.(\vec{d}\alpha)$$

$$\vec{x}2 = \vec{x}\beta - \vec{a}2.(\vec{d}\beta)$$

$$\vec{x}3 = \vec{x}\gamma - \vec{a}3.(\vec{d}\gamma)$$

$$\vec{x}4 = \vec{x}\delta - \vec{a}4.(\vec{d}\delta) \quad (1)$$

$$\frac{\vec{x}1 + \vec{x}2 + \vec{x}3 + \vec{x}4}{4} \quad (2)$$

Modification in blockade behavior and updating equation of the gray wolf optimization algorithm helped develop this technique, aimed at improving the convergence, efficiency, speed and meta-heuristic accuracy of the gray wolf optimizer. The modified Gray Wolf optimization divides the population into five different groups like alpha, beta, gamma, delta and omega using whichthe leadership hierarchy is simulated [see Figure 2].

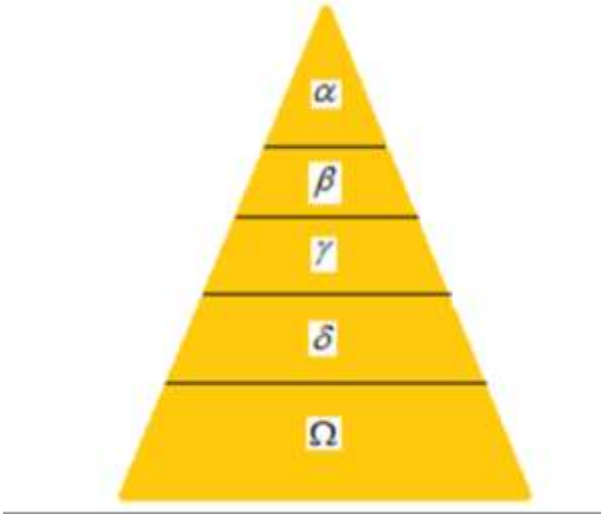


Figure 2: Gray Wolf Hierarchy (Domination reduces from top to bottom)

The remaining operations follow the gray wolf optimization algorithm.

3.2. K-Means Clustering Method by Selecting Initial Centers

The K-Means clustering algorithm captures either the data or the number of clusters, and accepts the initial centers as inputs. In a normal mode, K-means randomly selects the initial centers from the data and then examines which data is near to the centers and assigns those data to the centers, finallyit averages the data for obtaining the new center. In this study, the K-means algorithm works according to the normal mode, if the value of the initial

centers remains unassigned to K-means.For instance, consider that there are randomly selected three initial centers,then the calculation ofthe data distance from the initial centers in each iteration is done. Accordingly, to discover that which data belong to which centers, the distance is minimized in each part. The calculation is performed according toeach class of the average data where a certain cluster is placedfor obtaining new centers. To have intra-cluster distance obtained, the data interval associated to that center based on the number of classes is calculated and summarized. Currently, if the initial centers of K-meansare chosen based on the IGWO evolutionary algorithm, thosecenters will be input toK-means as initial centers; as a result, the initial data is not randomly selected and accordingly clustering efficiency is improved.

3.3. Selection of the initial Centers with IGWO

In order toachieve the initial centers with applying this algorithm, vectors or wolves indeed, their dimensionstimes the number of dimensions per centerand the number of centers. Given the aim is derive the initial centers between the minimum and maximum data, the upper bound (U) and the lower bound (L) of the data were defined; the minimum data corresponded tolower bound and so did themaximum data on the upper bound; therefore the initial centerswere placedbetween these values. In the following, the population size and dimensions are introduced to the GoodPoint set method algorithm for obtaining the initial centers.

$$X = \text{GoodPoint}(pop\_size , d*numClass) .* (U - L) + L \quad (3)$$

GoodPoint set technique is widely applied to produce initial proposed solutions; the technique also produces flexible solution space information to enhance population diversity. In conclusion, the current study benefits from this method to choose the appropriate population for IGWO.In the following,the fitness for the initial set is computed, and the sorting task is implemented accordingly.

The fitness function calculates the intra-cluster distance. Therefore, if the purpose is to have good wolves, the intra-cluster distance must be low. Initially, the distances are computed and the minimum amount is achieved. If there were a cluster that contained no data, it would not be a good sign. Hence, we set the fitness infinitely, otherwise the distance of those data belongingto that cluster is calculated according to the number of classes and their sum is computedwhich is regarded as fitness. The obtained fitness is dependent on the level of center closeness to data, and consequently the smaller, the better.

IV. The evaluation of results

This study involves 4 datasets, selected from standard and actual UCI datasets, in the area of clusteringto be tested [12]. In this study, the objective is to choose the datasets for analysis incorporatingdifferent characteristics of the problem spacelikesample design, sample dimension, feature diversity, boundary samples, shared samples, sample size, amplitude changes in different dimensions, the number of classes and population classes.A summaryof

the data set such as name, number of samples, and number of sample attributes in each class has been provided in Table 1.

Dataset	Number of features	Number of clusters	Number of data objects
Wine	13	3	178(48,71,59)
Glass	9	6	214(70,17,76,13,9,29)
CMC	9	3	1473(629,334,510)
Hill-Valley (HV)	101	2	606(305,301)

Table 1. Attributes of the datasets.

4.1. Parameter setting

For quality investigation of the proposed algorithm, the proposed technique was set out 10 times. The implementation process was carried out by 2018b MATLAB on a seven-core 2.1 GHz Pentium with 8 GB of RAM. The parameters were adjusted based onthosementioned inthe base paper [11]. The process was implemented in a way that the maximum parameter was amounted to T, 200, population size was 20 and the size of the maximum and minimum data was corresponded to upper and lower bounds, respectively. Table 2 presents the settings for the parameters in the proposed method (K-means + IGWO).

Parameter	Parameter value
T	200
pop_size	20
L	(min(data))
U	(max(data))

Table 2. The parameter settings related to IHd-ABC

4.2.The criterion of evaluation

Efficiency evaluation of the proposed method has been performed by comparing it with clustering algorithm of the base paper (MBCO + K-means). The comparison criteria of the proposed method and the base paper method include the percentage of error (PE) and the sum of intra-cluster distance (SICD).

4.2.1. The percentage of error

The percentage of error (PE) is computed for each problem, representingthe percentage of incorrectly classified objects in the experimental dataset. The classification is followed by choosing each pattern for each class whose center is the nearest oneas well asemploying Euclidean distance from the center of clusters. Then, thecomparison between the selected class (output) and the ideal output is done, therefore if it isselected incorrectly and is not accuratelyidentical, the object is separated. The percentage of error is computed from the incorrectly chosen data with the total experimental data. The percentage of error is given as:

$$PE = \frac{s}{n} \times 100 \tag{4}$$

Where s stands for the number of falsely classified objects and n represents the size of the experimental dataset.

4.2.2. The sum of intra-cluster distance

The measurements of efficiencyshow various intervals that for clustering problem, one interval can be selected and applied as a target function. There are fewextensivelyemployed target functions such assum of the Distance between Centroid of the Clusters (SDCC) or inter-cluster intervals, the sum of intra-cluster distance (SICD) and so on. The proximity or compression of data points in a cluster can beexcellently defined by measuring SICD. Thus, this study involved SICD value as the proposed algorithm's objective function for analyzing cluster solutions qualitatively. The SICD values are computed for the clusters determined by all the wolves producing a set of partial solutions. $SICD_w$can be defined as:

$$SICD_w(s,t) = \sum_{i=1}^{s} \sum_{j=1}^{n} D\left(o_j, c_i^w\right) \tag{5}$$

Where,c_i^wdetermines i center chosen by the gray wolf. $D\left(o_j, c_i^w\right)$represents the Euclidean distance between the object $o_j$and the center of c_i. SICD value also evaluates the quality of the solution.

4.3. Test results

Table 3 compares the efficiency of K-means + IGWO with one of the existing hybrid optimization methods (base paper). As can be seen in table 3, the K-means + IGWO hybrid algorithm is approximatelyequal to some examples and a solution withmore improved quality is acquiredcompared to the MBCO + K-means hybrid algorithm.

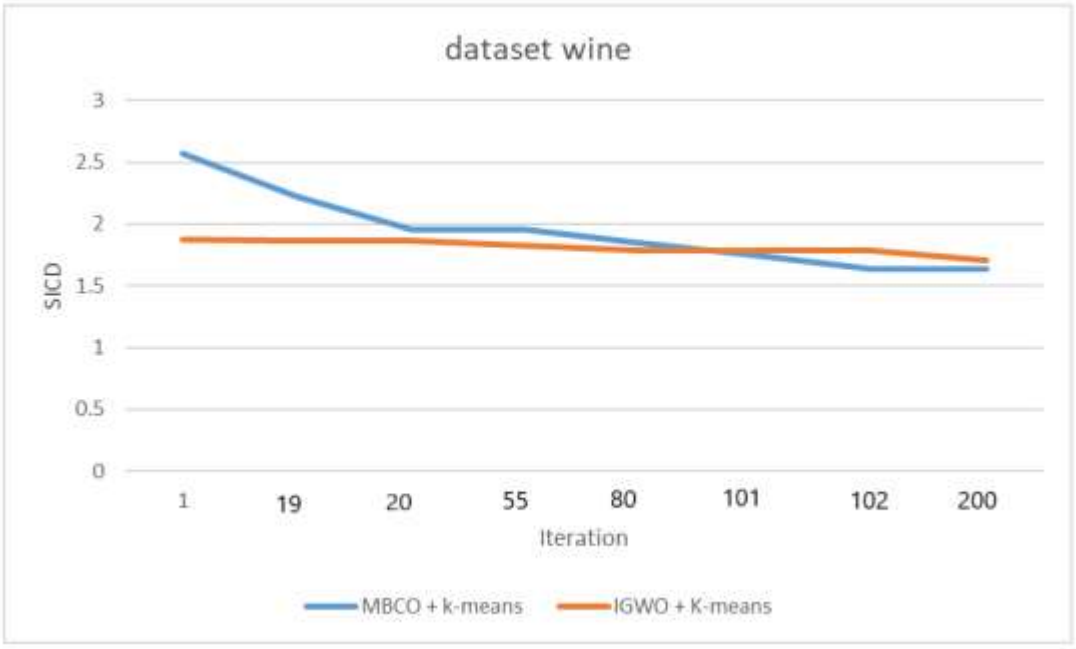
K-means + IGWO presents a less SICD value for all datasets because ofselecting better initial centers.

Table 3. Comparative SICD Analysis of Proposed Hybrid Algorithm with Basic Article Hybrid Algorithm.

Dataset	Measurement	MBCO +	Kmeans+IG

		Kmeans	WO
Wine	Mean SICD	21488.00	16555.6794
	MIN SICD	16400.00	16555.6794
	MAX SICD	21939.00	16555.6794
	STD SICD	41.67	2.2082e-12
Glass	Mean SICD	220.00	369.1198
	MIN SICD	215.00	369.1198
	MAX SICD	333.00	369.1198
	STD SICD	13.10	1.7496e-13
CMC	Mean SICD	5684.80	1029.2787
	MIN SICD	5678.20	1018.8234
	MAX SICD	5790.21	1018.8234
	STD SICD	1.97	14.7748
Hill-Valley (HV)	Mean SICD	66214038.78	34528442.6725
	MIN SICD	63929295.21	34528442.6725
	MAX SICD	71811560.40	34528442.6725
	STD SICD	358276.93	7.6441e-09

For investigating the convergence of the hybrid algorithm, the data analysis of Wine is presented in Figure 3. In Figure 3 also gives the convergence of the iris and wine datasets. The results reveal that the convergence of the proposed method will happen through choosing the best initial centers in the wine dataset after 20 iterations; additionally if the number of iterations risesin the wine dataset,the convergence becomes more imminent.



erated several times. In this step, other parameters like the number of wolves and irritationsbecomes fixed and changes in SICD values can be seen by using the chart. It wasseen that the convergence is achieved more quickly when using the proposed method.

Likewise, four datasets were included to investigate the error percentage of the proposed method with K-means algorithm. The objective of this study is to examine the impacts of evolutionary algorithm on choosing initial centers when not applying evolutionary algorithm. As can be seen from Table 4, the outcomesof the proposed method's percentage of errorremain lower than the K-means one. Table 4 presents PE (percentage of error) analysis, indicating the performance of the proposed algorithm in comparison to the K-means algorithm. Table 4 shows thatthe proposed algorithm has better average PE values than the classical K-means. It is clearly obvious that, the better the quality of the solution, the better the PE will be, and consequently the lowerthe PE levels in the classification. Based on this analysis, the proposed algorithm has clearly shownlower classification errors than the present K-means algorithm.

Dataset	Kmeans	Kmeans+IGWO
Wine	5.618	3.3708

glass	22.8972	10.7477
cmc	57.1623	56.5513
hv	49.505	49.505

Table 4.The error percentage comparison

V. Conclusion

The present paper aims at improving the accuracy of the k-mean algorithm with the use of hybrid and meta-heuristic techniques. Consequently, the algorithm (improved Gray Wolf Optimization k-mean) has been proposed in this research, which automatically detects the number of clusters and obtains the optimal solution as the K-mean clustering results or initial K-mean clustering centers. According to the experimental results, it can be found that the proposed method exhibitshigher efficiencyand less error percentage thanthe existing methods. Additionally, in comparison to the base paper, the proposed method has the improved total distance of intra-cluster.

References

[1] Kumar, Dhiraj. "Study On Clustering Techniques And Application To Microarray Gene Expression Bioinformatics Data." PhD diss., 2009.

[2] Jadhav, Amolkumar Narayan, and N. Gomathi. "WGC: Hybridization of exponential grey wolf optimizer with whale optimization for data clustering." Alexandria engineering journal 57, no. 3 (2018): 1569-1584.

[3] Kumar, Yugal, and Gadadhar Sahoo. "Hybridization of magnetic charge system search and particle swarm optimization for efficient data clustering using neighborhood search strategy." Soft Computing 19, no. 12 (2015): 3621-3645.

[4] Jadhav, Amolkumar Narayan, and N. Gomathi. "Kernel-based exponential grey wolf optimizer for rapid centroid estimation in data clustering." Jurnal Teknologi 78, no. 11 (2016).

[5] Han, XiaoHong, Long Quan, XiaoYan Xiong, Matt Almeter, Jie Xiang, and Yuan Lan. "A novel data clustering algorithm based on modified gravitational search algorithm." Engineering Applications of Artificial Intelligence 61 (2017): 1-7.

[6] Zhang, Qin-Hu, Bao-Lei Li, Ya-Jie Liu, Lian Gao, Lan-Juan Liu, and Xin-Ling Shi. "Data clustering using multivariant optimization algorithm." International Journal of Machine Learning and Cybernetics 7, no. 5 (2016): 773-782.

[7] Amiri, Ehsan, and Shadi Mahmoudi. "Efficient protocol for data clustering by fuzzy Cuckoo Optimization Algorithm." Applied Soft Computing 41 (2016): 15-21.

[8] Chander, Satish, P. Vijaya, and Praveen Dhyani. "Multi kernel and dynamic fractional lion optimization algorithm for data clustering." Alexandria engineering journal 57, no. 1 (2018): 267-276.

[9] Chen, Huihui, Yusen Zhang, and Ivan Gutman. "A kernel-based clustering method for gene selection with gene expression data." Journal of Biomedical Informatics 62 (2016): 12-20.

[10] Kuo, R. J., T. C. Lin, Ferani E. Zulvia, and C. Y. Tsai. "A hybrid metaheuristic and kernel intuitionistic fuzzy c-means algorithm for cluster analysis." Applied Soft Computing 67 (2018): 299-308.

[11] Das, Pranesh, Dushmanta Kumar Das, and Shouvik Dey. "A modified Bee Colony Optimization (MBCO) and its hybridization with k-means for an application to data clustering." Applied Soft Computing 70 (2018): 590-603.

[12] M. Lichman, Uci machine learning repository, URLhttp://archive.ics.uci.edu/ml/ 8 (2013).