IJEAST

INTERNATIONAL JOURNAL
OF ENGINEERING APPLIED SCIENCE
AND TECHNOLOGY

VOLUME : 3 ISSUE : 02 Print / Issue Publication Date: 07-Sep-2018

ISSN : 2455-2143

Indexed In

Google SCHOLAR OPEN ACCESS doi Crossref

WWW.IJEAST.COM

editor@ijeast.com

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 3, Issue 2, ISSN No. 2455-2143, Pages 7-15
Published Online June 2018 in IJEAST (http://www.ijeast.com)

FUZZY CONTROLLED GRAVITATIONAL SEARCH ALGORITHM FOR CLUSTERING

Himanshu Kaul
Delhi Technological University
Shahbad Daulatpur, New Delhi 110042, India

Neha Dimri
Indian Institute of Technology
Hauz Khas, New Delhi 110016, India

Abstract— **Clustering is a key activity in numerous data mining applications such as information retrieval, text mining, image segmentation, etc. In this paper, a clustering approach, Fuzzy-GSA, based on gravitational search algorithm (GSA) has been proposed. The proposed Fuzzy-GSA algorithm utilizes a fuzzy inference system to effectively control the parameters of gravitational search algorithm. The performance of Fuzzy-GSA algorithm has been evaluated against four benchmark datasets from the UC Irvine repository. The results illustrate that the Fuzzy-GSA algorithm attains the highest quality clustering over the selected datasets when compared with several other clustering algorithms namely, k-means, particle swarm optimization (PSO), gravitational search algorithm (GSA) and, combined gravitational search algorithm and k-means approach (GSA-KM).**

Keywords— **Clustering, Fuzzy-GSA, Gravitational Search Algorithm, Fuzzy Inference System**

I. INTRODUCTION

Clustering or cluster analysis refers to the process of grouping a set of data objects such that objects belonging to the same group are similar, whereas those belonging to different groups are distinct. The final groups are called clusters or classes. It is a major data mining task and is used as a common technique for analysis of statistical data in many fields such as pattern recognition, machine learning, information retrieval etc. In the process of clustering, it is important to define an appropriate similarity or dissimilarity measure over which the data objects are to be clustered. The same set of objects may be partitioned into different groups depending on the choice of similarity or dissimilarity measure. The number of clusters in the final partition may be pre-assigned or may be considered as an internal parameter of the clustering algorithm to be deduced based on the input data.

Clustering is an unsupervised learning task since it groups data objects into clusters without any prior information such as class labels. The clustering techniques, thus, should be able to deduce the structure embedded in data without any extra information. Clustering algorithms have been successfully applied in several fields such as information retrieval (Jardine and Rijsbergen 1971; Tombros et al. 2002), medicine (Liao et al. 2008), biology (Kerr et al. 2008), customer analysis (Saglam et al. 2006), image segmentation (Xia et al. 2007) and many others. Clustering has been an area of active research and many clustering algorithms have been proposed in the literature. The most widely used and the most popular algorithm for clustering is the k-means algorithm, proposed by J. MacQueen in 1967 (MacQueen 1967). K-means algorithm is fairly straightforward, simple to implement and has been employed by several researchers (Jain 2010; Forgy 1965; Kaufman and Rousseeuw 1990). However, it may be easily trapped in a local optimum and fail to achieve a global optimum in several cases since the algorithm's performance is highly dependent on the initial centroids chosen.

To overcome this problem, several heuristic based approaches have been proposed for clustering. Selim and Alsultan (1991) provided a simulated annealing (SA) algorithm for clustering. They have demonstrated that the simulated annealing algorithm converges to a global optimum for the clustering problem. Maulik and Bandyopadhyay (2000) presented a clustering technique based on genetic algorithm, known as GA-clustering. The centers of a pre-defined number of clusters were encoded using chromosomes and the improved performance of GA-clustering over k-means algorithm was demonstrated with the help of three real datasets. A tabu search based method was presented for solving the clustering problem in (Al-Sultan 1995; Sung and Jin 2000).

Shelokar et al. (2004) presented an Ant Colony Optimization (ACO) based technique for optimally assigning objects to a pre-defined number of clusters. The ACO based technique provided very promising results when compared with other heuristic methods such as genetic algorithm, simulated annealing and tabu search. Fathian et al. (2007) proposed an algorithm for clustering based on honeybee mating optimization (HBMO). The performance of HBMO based approach was better compared to SA, GA, tabu search and ACO when evaluated over several well-known datasets. Ching-Yi and Fun (2004) provided a Particle Swarm Optimization (PSO) based approach for clustering. They compared the performance of PSO-based approach with

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 3, Issue 2, ISSN No. 2455-2143, Pages 7-15
Published Online June 2018 in IJEAST (http://www.ijeast.com)

traditional clustering algorithms and demonstrated that the PSO-based approach performed better using four artificial datasets.

Hatamlou et al. (2011) applied the Gravitational Search Algorithm (GSA) to data clustering. The results over four well-known datasets depicted that GSA based approach performed better than several other clustering algorithms namely PSO, HBMO, ACO, GA, SA and k-means. Hatamlou et al. (2012) presented a technique combining the benefits of k-means algorithm with GSA, called GSA-KM, in clustering. In GSA-KM approach, the initial population for GSA was generated with the help of k-means algorithm, which allowed GSA to converge faster. When compared with other well known algorithms, such as k-means, GA, SA, ACO, HBMO, PSO and the conventional GSA approach, GSA-KM approach provided better results over several real datasets.

Gravitational Search Algorithm (GSA) uses a constant value of parameter α for the calculation of gravitational constant. In the beginning, smaller value of α allows for a greater exploration of the search space. Furthermore, higher value of α during the last few iterations enhances the search space exploitation. Therefore, the approach based on GSA can be improved by adapting and controlling the value of parameter α, as the algorithm proceeds.

This paper proposes an algorithm for clustering, called Fuzzy-GSA, based on Gravitational Search Algorithm (GSA). The proposed algorithm uses fuzzy rules for controlling the parameter α in GSA algorithm as the search progresses. In Section II, an overview of the Gravitational Search Algorithm (GSA) proposed by Rashedi et al. (2009) has been presented. Section III describes the proposed clustering algorithm, Fuzzy-GSA, where Section III–A describes the developed fuzzy inference system and Section III–B presents the proposed algorithm for clustering. Section IV discusses the experimental results and comparison with other clustering algorithms. Section V lists the conclusions of the present study.

II. Gravitational Search Algorithm

Gravitational Search Algorithm (GSA) is an optimization algorithm proposed by Rashedi et al. (2009). It is based on the Newton's laws of gravity and motion. The law of gravity states that "Every particle in the universe attracts every other particle with a force that is directly proportional to the product of the masses of the particles and inversely proportional to the square of the distance between them". By this definition, the gravitational force is determined using the following equation (Rashedi et al. 2009):

$$F = G\,\frac{M_1 M_2}{R^2} \tag{1}$$

where, F is the gravitational force acting between two masses $M_1$and M_2, G is the gravitational constant with a value of 6.67259×10^{-11} N m^2/kg^2, and R is the distance between the two masses.

Newton's second law of motion states that when a force acts on a mass, acceleration is produced. The magnitude of acceleration produced is obtained using the equation below (Rashedi et al. 2009):

$$a = \frac{F}{M} \tag{2}$$

where, F and M denote the net force acting on a given particle and its mass, respectively.

The Gravitational Search Algorithm (GSA) employs this physical phenomenon for solving optimization problems. Consider a system with N masses or agents. The position of i^{th} mass is defined as:

$$X_i = (\, x_i^1, \ldots, x_i^d, \ldots, x_i^n) \text{ , for } i = 1,2, \ldots, \text{N} \tag{3}$$

where, x_i^d is the position of i^{th} agent in d^{th} dimension and n is the total number of dimensions in the search space. The positions of agents correspond to the solutions of the problem. The mass of each agent is computed, after evaluating the present population's fitness, using the following equations:

$$m_i(t) = \frac{fit_i(t) - worst(t)}{best(t) - worst\,(t)} \tag{4}$$

$$M_i(t) = \frac{m_i(t)}{\sum_{j=1}^{N} m_j(t)} \tag{5}$$

where, $fit_i(t)$, denotes the fitness value of i^{th} agent at time t, and *best(t)* and *worst(t)* are computed as follows (for minimization problems):

$$best(t) = \min\, fit_j(t)\,,\ j = 1, 2, \ldots, N \tag{6}$$

$$worst(t) = \max\, fit_j(t)\,,\ j = 1, 2, \ldots, N \tag{7}$$

Similarly, for maximization problems *best(t)* and *worst(t)* are computed by taking the maximum and minimum fitness values respectively.

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 3, Issue 2, ISSN No. 2455-2143, Pages 7-15
Published Online June 2018 in IJEAST (http://www.ijeast.com)

The acceleration of an agent is computed next, by considering the total forces from a set of heavier masses using the laws of gravity and motion using Eqs. (8) and (9). The new velocity of an agent is computed next by adding a fraction of its current velocity to its acceleration (Eq. (10)), followed by the calculation of its new position (Eq. (11)).

$$F_i^d(t) = \sum_{j \in kbest, j \neq i} rand_j \, G(t) \frac{M_j(t) M_i(t)}{R_{ij}(t) + \varepsilon} \left(x_j^d(t) - x_i^d(t) \right) \quad (8)$$

$$a_i^d(t) = \frac{F_i^d(t)}{M_i(t)} = \sum_{j \in kbest, j \neq i} rand_j \, G(t) \frac{M_j(t)}{R_{ij}(t) + \varepsilon} \left(x_j^d(t) - x_i^d(t) \right) \quad (9)$$

$$v_i^d(t+1) = rand_i \times v_i^d(t) + a_i^d(t) \quad (10)$$

$$x_i^d(t+1) = x_i^d(t) + v_i^d(t+1) \quad (11)$$

where, $rand_i$ and $rand_j$ are two random numbers uniformly distributed in the range of [0, 1], ε is a small value to prevent division by zero, $R_{ij}(t)$ is the Euclidean distance between agent *i* and agent *j*. *Kbest* is the set of first *K* agents with best fitness values and thus, largest mass. *Kbest* is dependent on time, initialized to K_o at the start and decreases as time progresses. The gravitational constant, $G(t)$, decreases with time to control the search accuracy. The value of $G(t)$ is calculated using the following equation:

$$G(t) = G_o e^{\frac{-\alpha t}{T}} \quad (12)$$

where, G_o is the initial value of gravitational constant, α is a parameter which governs the degree of exploration versus exploitation of the search and *T* is the maximum number of iterations.

III. Fuzzy-GSA Clustering Algorithm

In this section, the proposed algorithm, called Fuzzy-GSA, has been described for data clustering. The proposed approach is based on Gravitational Search Algorithm (GSA), described in section II, and uses fuzzy inference rules for controlling the parameter α as search progresses. This section is divided into two subsections. Section III–A describes the Fuzzy Inference System (FIS) developed, and the section III–B presents the proposed Fuzzy-GSA algorithm for clustering.

A. The Developed Fuzzy Inference System –

The FIS is developed with two input variables and one output variable. The input variables are as follows:

- *IT*: The current iteration number.
- *Fbest*: The best value of fitness achieved till the current iteration.

IT enables us to consider how far we have reached in the search process. During the initial iterations, i.e. when *IT* is low, a lower value of α is desired since lower the value of α, higher the value of gravitational constant, *G(t)*, will be (Eq. (12)) and thus, higher the force, *F*, (Eq. (8)) resulting in a higher acceleration, *a*, (Eq. (9)) and velocity, *v(t)* (Eq. (10)). This allows for higher exploration at the beginning of search. Similarly, towards the final few iterations, i.e. when *IT* is high, a higher value of α is desired to promote higher exploitation. Fig. 1 depicts the membership function for *IT*. The iterations are represented as a fraction of the maximum number of iterations allowed, such that 0.5 means half of the total iterations and 1 represents the maximum iterations.

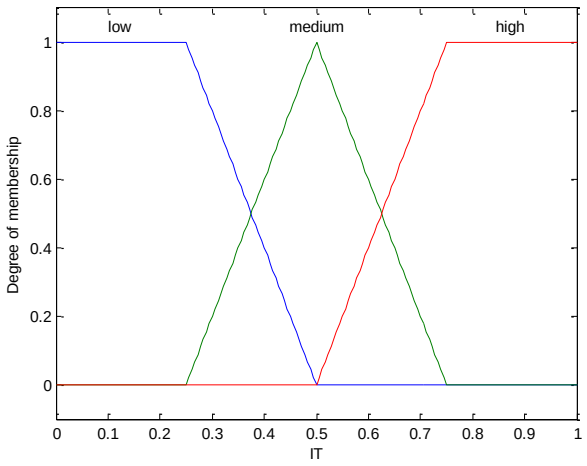

Fig. 1. Membership function for *IT*

Fbest represents the lowest value of fitness, since clustering is a minimization problem with the fitness function as mean square error, achieved till the current iteration. If the value of *Fbest* is high, then we need to reduce α to promote a greater exploration, since higher values for *Fbest* mean we are still far from the solution. However, if *Fbest* is low, we should increase α to allow for a higher exploitation as we are near the

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 3, Issue 2, ISSN No. 2455-2143, Pages 7-15
Published Online June 2018 in IJEAST (http://www.ijeast.com)

solution. Fig. 2 shows the membership function for *Fbest*. Note that the membership function for *Fbest* needs to be tuned as per the input dataset being considered, since the acceptable values of fitness function will vary for different datasets.

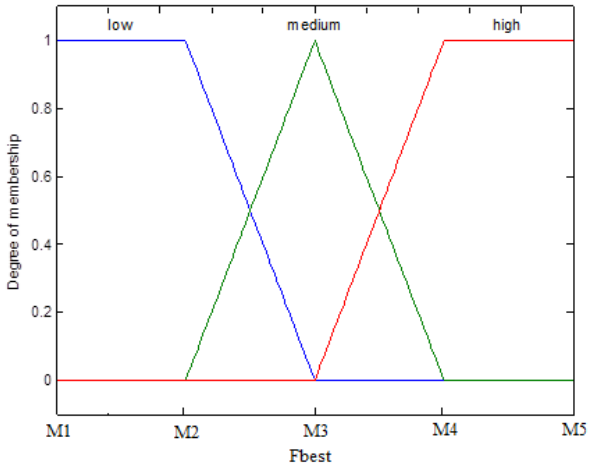

Fig. 2. Membership function for *Fbest*

To obtain the marked value M5 in Fig. 2, we executed five independent runs of GSA for a single iteration and equated M5 to the maximum value of *Fbest* obtained, after adding hundred and then rounding it off to the nearest hundred. For the value of M3, we considered the integer part of the best fitness value obtained using GSA (Hatamlou et al. 2011), for that dataset. M4 was computed by adding one to the value of M3, and M2 was calculated by rounding off M3 to the nearest ten smaller than M3. Finally, M1 was obtained by subtracting ten from M2.

It should be noted that the fitness function, representing the total mean square error or the sum of intra-cluster distances, is computed using the following equation (Yang et al. 2010):

$$f(O,C) = \sum_{l=1}^{k} \sum_{O_i \in C_l} d(O_i, CC_l)^2 \tag{13}$$

where, CC_l represents the centroids of the cluster C_l, $d(O_i, CC_l)$ denotes the distance or dissimilarity between object O_i and cluster centroid CC_l. The most popular and widely used distance metric is the Euclidean distance, which we have used in this work. Euclidean distance between two objects X_i and X_j with d dimensions is calculated as:

$$d(X_i, X_j) = \sqrt{\sum_{p=1}^{d} (x_i^p - x_j^p)^2} \tag{14}$$

where, x_i^p denotes the value of p^{th} dimension for the object X_i and x_j^p denotes the value of pth dimension for the object X_j.

The developed FIS consists of one output variable, i.e. *alpha(t)*, which denotes the value of parameter α in Eq. (12). Fig. 3 shows the membership function for *alpha(t)*. The range of parameter α is taken as [0, 50] to provide a wide range of search on the value of *alpha(t)*.

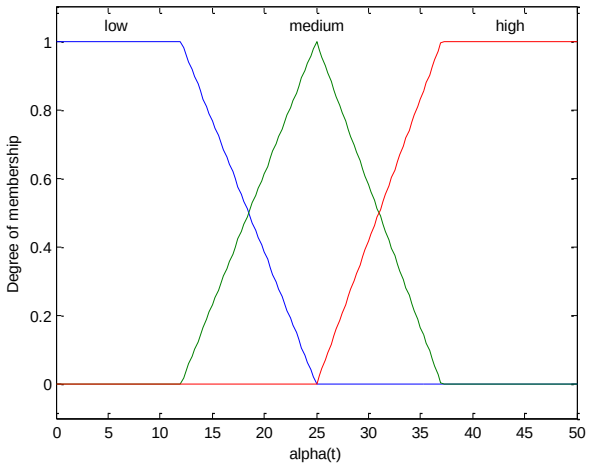

Fig. 3. Membership function for *alpha(t)*

The following eight fuzzy rules were formulated to control the parameter α in the calculation of the gravitational constant (Eq. (12)):

i. If (*IT* is low) and (*Fbest* is low) then (*alpha(t)* is high)
ii. If (*IT* is low) and (*Fbest* is medium) then (*alpha(t)* is medium)
iii. If (*IT* is low) and (*Fbest* is high) then (*alpha(t)* is low)
iv. If (*IT* is medium) and (*Fbest* is high) then (*alpha(t)* is low)
v. If (*IT* is medium) and (*Fbest* is medium) then (*alpha(t)* is medium)
vi. If (*IT* is high) and (*Fbest* is high) then (*alpha(t)* is medium)
vii. If (*IT* is high) and (*Fbest* is medium) then (*alpha(t)* is medium)
viii. If (*IT* is high) and (*Fbest* is low) then (*alpha(t)* is high)

The discussed fuzzy inference system is developed using Mamdani's method, in which the operator for "And" is min and "Or" is max. The implication method is min, aggregation method is max and defuzzification method is centroid.

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 3, Issue 2, ISSN No. 2455-2143, Pages 7-15
Published Online June 2018 in IJEAST (http://www.ijeast.com)

B. Proposed algorithm –

The proposed algorithm, Fuzzy-GSA, comprises of two steps. The first step is to generate an initial population for GSA. We have generated the initial population by considering three agents (or candidate solutions) corresponding to the maximum, minimum and median values for all features in a given dataset, respectively. This provides a better initial population, which would allow for a higher exploration since a wide range of values, including maximum, minimum and median, are present while searching the solution space. The rest of the agents are generated randomly by considering the range of features in the given dataset.

The second step involves application of GSA, described in Section II, to the given dataset and using the fuzzy inference system developed to control the parameter α while searching for the solution. The flow diagram for the proposed Fuzzy-GSA algorithm is depicted by Fig. 4.

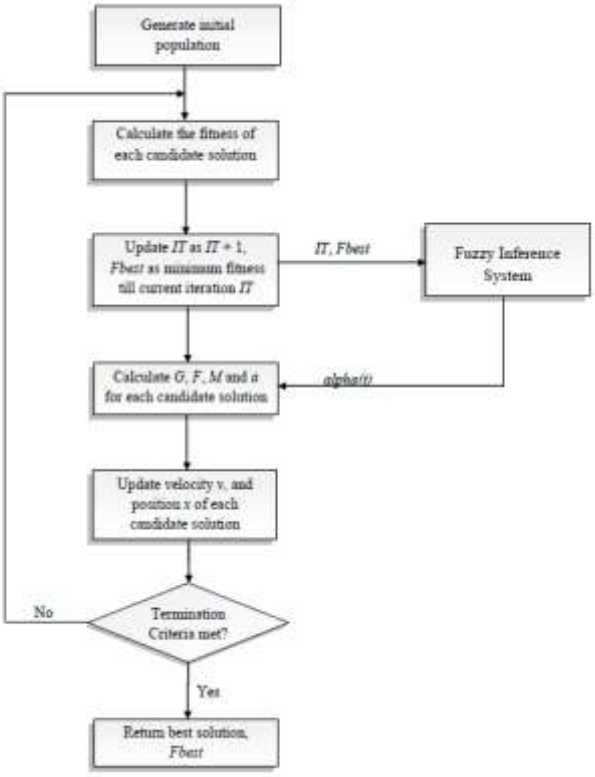

Fig. 4. Flow diagram for the proposed Fuzzy-GSA clustering algorithm

The step-by-step algorithm for the proposed Fuzzy-GSA clustering is stated next. Let N denote the population size, C_i be the i^{th} candidate solution or agent, k be the number of clusters, d be the number of features in a given dataset.

Step 1: Generate initial population, $P = \{C_1, C_2, \ldots, C_N\}$.

- Generate C_1 consisting of maximum values of all the features.
- Generate C_2 consisting of minimum values of all the features.
- Generate C_3 consisting of median values of all the features.
- Generate the remaining N-3 candidates randomly within the range of minimum to maximum values for all features.

Step 2: Apply GSA and use the developed FIS, described in Section III–A, for parameter adaptation.

- Calculate the fitness function, as per Eq. (13), for all the candidate solutions.
- Feed the values of *IT*, current iteration number, and *Fbest*, best fitness achieved, as inputs to the developed FIS, and obtain the value of parameter α.
- Calculate G, F, M and a for all the candidate solutions using Eqs. (5), (8), (9) and (12), respectively.
- Update the velocity and position of each candidate solution as per Eq. (10) and (11) respectively.
- Check if termination criteria, i.e. maximum number of iterations allowed is reached or fitness function is not exhibiting a minimum improvement, are met. If yes, then return the best value of fitness function achieved as the final solution, else reiterate through step 2.

The final solution consists of the best value of fitness function, i.e. the minimum mean square error, achieved by running the proposed Fuzzy-GSA algorithm.

IV. Results and Discussion

The proposed Fuzzy-GSA algorithm for clustering is implemented in MATLAB R2014b. Each candidate solution, in the population, consists of cluster centers for each of the k clusters, and each cluster center comprises of values for each feature in a dataset. Fig. 5 illustrates the representation of the i^{th} candidate solution C_i. CC_{ij} denotes the j^{th} cluster center of the i^{th} candidate solution and F_{ij} represents the value of j^{th} feature for i^{th} cluster center. Therefore, each candidate solution consists of ($d \times k$) values. The design parameters used in numerical computations have been defined in Table 1.

The performance of the proposed approach, Fuzzy-GSA, has been measured by calculating the sum of intra-cluster distances as defined by Eq. (13) by considering four benchmark datasets namely, Iris, Wine, Breast Cancer Wisconsin and Contraceptive Method Choice (CMC) for

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 3, Issue 2, ISSN No. 2455-2143, Pages 7-15
Published Online June 2018 in IJEAST (http://www.ijeast.com)

evaluation. The selected datasets for evaluation have been obtained from UC Irvine repository of machine learning databases (Blake and Merz 2016), and are the benchmark datasets used by many researchers to validate the performance of clustering algorithms.

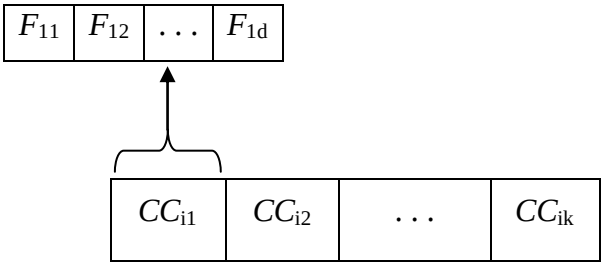

Fig. 5. Representation of i^{th} Candidate Solution, C_i

Table -1 Design parameters used in the proposed Fuzzy-GSA clustering algorithm

Parameters	**Value**
N (population size)	50
maximum number of iterations	300
G_o	100
minimum acceptable improvement	1×10^{-6}

A description of each benchmark dataset is provided below:

- *Iris Dataset*: It consists of three classes with 50 instances each, where each class refers to a species of iris flower. There are four features in the dataset namely, petal length, petal width, sepal length and sepal width which report certain characteristics of iris flower. The dataset comprises of a total of 150 instances. There no missing feature values in this dataset.

- *Wine Dataset*: It consists of three classes representing different types of wine. The data is a result of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivators. There are 13 features which represent quantities of different constituents found in each of the three types of wines. The dataset consists of 178 instances with no missing values.

- *Breast Cancer Wisconsin Dataset*: This dataset comprises of two classes namely, malignant and benign representing the severity of cancer. There are a total of 683 instances, without missing values. It has 9 attributes or features namely, clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses.

- *Contraceptive Method Choice (CMC) Dataset*: It consists of three classes namely, no-use, long-term and short-term. There are 1473 instances in this dataset, without any missing values. It contains 9 features or attributes namely, wife's age, wife's education, husband's education, number of children ever born, wife's religion, whether wife's working, husband's occupation, standard of living and media exposure.

Sum of intra-cluster distances has been calculated over each of the four benchmark datasets considered, using Eq. (13). Further, the performance of the proposed Fuzzy-GSA algorithm has been compared with existing clustering algorithms such as conventional GSA (Hatamlou et al. 2011), combined k-means and GSA (Hatamlou et al. 2012), PSO (Ching-Yi and Fun 2004) and k-means (MacQueen 1967) algorithms on selected datasets. Due to the stochastic nature of these algorithms, 20 independent runs for each algorithm over each dataset have been considered. The results are then compared in terms of best, average and worst solutions over 20 independent simulations. Moreover, the standard deviation of the achieved solutions through each clustering algorithm is calculated. It should be noted that, a lower value of the sum of intra-cluster distances denotes a higher quality clustering.

Fig. 6 illustrates the three clusters assigned by the proposed Fuzzy-GSA algorithm over the Iris dataset. It shows a 3-D plot considering the three dimensions namely, petal length, sepal width and sepal length of the Iris dataset. The remaining dimension, petal width, is highly correlated with the dimension, petal length, and thus, can be ignored without losing the quality of cluster representation. The three clusters, corresponding to the three types of iris flower, are depicted by three different colour coded symbols namely, blue squares, green circles and red triangles. The X-axis represents the dimension sepal length, the Y-axis represents the dimension sepal width and the Z-axis represents the dimension petal length. The respective cluster centers are represented by black coloured circles.

The best, average, worst and standard deviation of the obtained solutions over 20 independent simulations by different clustering algorithms on the selected datasets are shown in Table 2.

International Journal of Engineering Applied Sciences and Technology, 2018
Vol. 3, Issue 2, ISSN No. 2455-2143, Pages 7-15
Published Online June 2018 in IJEAST (http://www.ijeast.com)

Fig. 6. Cluster plot by Fuzzy-GSA algorithm over Iris Dataset

As can be seen from Table 2, the proposed algorithm, Fuzzy-GSA, demonstrates the highest quality solutions in terms of best, average and worst intra-cluster distances over all the four benchmark datasets. Furthermore, the standard deviation of Fuzzy-GSA is lower, which indicates that it can locate a near-optimal solution in most of the cases when compared with other clustering algorithms.

For Iris dataset, the best, average and worst solutions by the proposed Fuzzy-GSA algorithm are 96.5403, 96.5425 and 96.5581, respectively with a standard deviation of 0.0054. For Wine dataset, the best, average and worst solutions achieved by the Fuzzy-GSA algorithm are 16292.23, 16293.369 and 16294.35 respectively, with a standard deviation of 0.82. For Breast Cancer Wisconsin dataset, the achieved best, average and worst solutions are 2964.38, 2964.38 and 2964.39, respectively with a standard deviation of 0.003. Lastly, for CMC dataset, the best, average and worst solutions obtained by Fuzzy-GSA algorithm are 5532.2, 5532.6 and 5533.7, respectively with a standard deviation of 0.5831.

To summarize, the proposed Fuzzy-GSA algorithm achieves the best quality clustering when compared with several popular clustering algorithms, depicted in Table 2, over four benchmark datasets considering 20 independent runs.

Table -2 Simulation Results for different clustering algorithms over 20 independent simulations

Dataset	Criteria	K-means	PSO	GSA	GSA-KM	Fuzzy-GSA
Iris	Best	97.2046	96.7170	96.6700	96.6173	96.5403
	Avg	101.2562	97.8962	96.6952	96.6687	96.5425
	Worst	124.2155	99.7773	96.8961	96.6989	96.5581
	Std	9.8954	0.9306	0.0485	0.0227	0.0054
Wine	Best	16555.6794	16340.1288	16319.0752	16300.0862	16292.23
	Avg	16990.4711	16378.4879	16351.3308	16301.6686	16293.369
	Worst	18294.8465	16505.4147	16481.6366	16302.5723	16294.35
	Std	772.6452	44.7783	34.3939	0.6280	0.8200
Cancer	Best	2988.4278	2974.6453	2965.1822	2965.0778	2964.38
	Avg	2988.4278	3078.4729	2975.0247	2965.6777	2964.38
	Worst	2988.4278	3336.6453	2997.7815	2966.7573	2964.39
	Std	0	113.8010	8.8999	0.4191	0.003
CMC	Best	5543.5119	5710.8682	5544.6439	5543.5119	5532.2
	Avg	5543.7652	5840.8038	5627.3252	5544.5250	5532.6
	Worst	5545.2005	5987.0105	5697.1460	5545.2005	5533.7
	Std	0.6186	82.3954	48.8495	0.8487	0.5831

V. CONCLUSIONS

This paper proposes an algorithm, *Fuzzy-GSA*, for clustering. *Fuzzy-GSA* algorithm is based on the conventional Gravitational Search Algorithm (GSA) with a provision for adapting the value of parameter α used in the calculation of the gravitational constant. In the beginning, a smaller value of α is desired to achieve a higher exploration, whereas towards the end of search, a relatively higher value of α helps in achieving a higher exploitation. *Fuzzy-GSA* algorithm incorporates Fuzzy Inference System (FIS) into the conventional GSA to allow for parameter adaptation. The parameter α is controlled by using eight formulated fuzzy inference rules, in *Fuzzy-GSA*. Also, *Fuzzy-GSA* algorithm generates a better quality initial population for GSA by considering the nature of dataset being considered. It generates three candidate solutions consisting of maximum, minimum and median values, respectively in a dataset thereby building an initial population, which covers a wider range. This helps in achieving a higher exploration. The performance of *Fuzzy-GSA* is evaluated by comparing its best, average and worst solutions with several other clustering algorithms over four selected benchmark datasets namely, Iris, Wine, Breast Cancer Wisconsin and CMC, considering 20 independent runs. The results show that *Fuzzy-GSA* achieves the highest quality clustering with very small standard deviation, when compared with several other clustering algorithms.

VI. REFERENCES

Al-Sultan K.S. (1995) A tabu search approach to the clustering problem, Pattern Recognition, vol. 28, issue 9, pp. 1443–1451.

Blake C.L., Merz C.J. (2016) UCI repository of machine learning databases. Available from: http://www.ics.uci.edu/-mlearn/MLRepository.html. Accessed 20 June 2016.

Ching-Yi C., Fun Y. (2004) Particle swarm optimization algorithm and its application to clustering analysis, in IEEE International Conference on Networking, Sensing and Control.

Fathian M., Amiri B., Maroosi A. (2007) Application of honey-bee mating optimization algorithm on clustering, Applied Mathematics and Computation, vol. 190, issue 2, pp. 1502–1513.

Forgy E.W. (1965) Cluster analysis of multivariate data: efficiency versus interpret- ability of classifications, Biometrics, vol. 21, issue 2.

Hatamlou A., Abdullah S., Nezamabadi-pour H. (2011) Application of gravitational search algorithm on data clustering, Rough Sets and Knowledge Technology, Springer, pp. 337–346.

Hatamlou A., Abdullah S., Nezamabadi-pour H. (2012) A combined approach for clustering based on K-means and gravitational search algorithms, Swarm and Evolutionary Computation, vol. 6, pp. 47–52.

Jain A.K. (2010) Data clustering: 50 years beyond K-means, Pattern Recognition Letters, vol. 31, issue 8, pp. 651–666.

Jardine N., Rijsbergen C.J. van (1971) The use of hierarchic clustering in information retrieval, Information Storage and Retrieval, vol. 7, issue 5, pp. 217–240.

Kaufman L., Rousseeuw P.J. (1990) Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, New York.

Kerr G., Ruskin H.J., Crane M., Doolan P. (2008) Techniques for clustering gene expression data, Computers in Biology and Medicine, vol. 38, issue 3, pp. 283–293.

Liao L., Lin T., Li B., (2008) MRI brain image segmentation and bias field correction based on fast spatially constrained kernel clustering approach, Pattern Recognition Letters, vol. 29, issue 10, pp. 1580–1588.

MacQueen J. (1967) Some methods for classification and analysis of multivariate observations, In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281-297.

Maulik U., Bandyopadhyay S. (2000) Genetic algorithm-based clustering technique, Pattern Recognition, vol. 33, issue 9, pp. 1455–1465.

Rashedi E., Nezamabadi-pour H., Saryazdi S. (2009) GSA: a gravitational search algorithm, Information Sciences, vol. 179, issue 13, pp. 2232–2248.

Saglam B., Türkay M., Salman F.S., Sayın S., Karaesmen F., Örmeci E.L. (2006) A mixed-integer programming approach to the clustering problem with an application in customer segmentation, European Journal of Operational Research, vol. 173, issue 3, pp. 866–879.

Selim S.Z., Alsultan K. (1991) A simulated annealing algorithm for the clustering problem, Pattern Recognition, vol. 24, issue 10, pp. 1003–1008.

Shelokar P.S., Jayaraman V.K., Kulkarni B.D. (2004) An ant colony approach for clustering, Analytica Chimica Acta, vol. 509, issue 2, pp. 187–195.

Sung C.S., Jin H.W. (2000) A tabu-search-based heuristic for clustering, Pattern Recognition, vol. 33, issue 5, pp. 849–858.

Tombros A., Villa R., Rijsbergen C.J. van (2002) The effectiveness of query-specific hierarchic clustering in

information retrieval, Information Processing & Management, vol. 38, issue 4, pp. 559–582.

Xia Y., Fenga D., Wangd T., Zhaob R., Zhangb Y. (2007) Image segmentation by clustering of spatial patterns, Pattern Recognition Letters, vol. 28, issue 12, pp. 1548–1555.

Yang S., Wu R., Wang M., Jiao L. (2010) Evolutionary clustering based vector quantization and SPIHT coding for image compression, Pattern Recognition Letters, vol. 31, issue 13, pp. 1773–1780.

IJEAST

INTERNATIONAL JOURNAL
OF ENGINEERING APPLIED SCIENCE
AND TECHNOLOGY

ABOUT IJEAST

International Journal of Engineering Applied Science and Technology (IJEAST) is a peer-reviewed, open access journal that publishes high-quality research papers in the field of Engineering, Applied Science and Technology.

IJEAST aims to provide a platform for researchers, academicians, and professionals to share their innovative ideas, research findings, and practical experiences with the global scientific community.

FOCUS AREAS

- Engineering
- Applied Science
- Technology
- Innovation & Development
- Interdisciplinary Studies

PEER REVIEWED
All submissions are rigorously peer reviewed to ensure quality.

OPEN ACCESS
Free and unrestricted access to research for all.

GLOBAL REACH
Connecting researchers and professionals worldwide.

TIMELY PUBLICATION
We ensure a swift and efficient publication process.

For more information, visit our website
www.ijeast.com

INTERNATIONAL JOURNAL
OF ENGINEERING APPLIED SCIENCE
AND TECHNOLOGY

editor@ijeast.com
www.ijeast.com
India

2455-2143