Journal of Computer Science, 9 (4): 404-412, 2013
ISSN 1549-3636

doi:10.3844/jcssp.2013.404.412 Published Online 9 (4) 2013 (http://www.thescipub.com/jcs.toc)

ARTIFICIAL BEE COLONY ALGORITHM INTEGRATED WITH FUZZY C-MEAN OPERATOR FOR DATA CLUSTERING

M. Krishnamoorthi and A.M. Natarajan

Department of Computer Science and Engineering,
Bannari Amman Institute of Technology, Erode, Tamil Nadu, India

Received 2012-07-25, Revised 2012-09-04; Accepted 2013-05-06

ABSTRACT

Clustering task aims at the unsupervised classification of patterns in different groups. To enhance the quality of results, the emerging swarm-based algorithms now-a-days become an alternative to the conventional clustering methods. In this study, an optimization method based on the swarm intelligence algorithm is proposed for the purpose of clustering. The significance of the proposed algorithm is that it uses a Fuzzy C-Means (FCM) operator in the Artificial Bee Colony (ABC) algorithm. The area of action of the FCM operator comes at the scout bee phase of the ABC algorithm as the scout bees are introduced by the FCM operator. The experimental results have shown that the proposed approach has provided significant results in terms of the quality of solution. The comparative study of the proposed approach with existing algorithms in the literature using the datasets from UCI Machine learning repository is satisfactory.

Keywords: Clustering, Optimization, ABC Algorithm, FCM Algorithm, FCM Operator

1. INTRODUCTION

Clustering is a data mining technique that is widely studied in several research fields such as statistical pattern recognition, machine learning, information retrieval and data mining. Clustering deals with unsupervised classification of patterns into clusters. Clustering approaches can be divided as, partitioning methods, hierarchical methods, (Yu *et al*., 2010), (Krinidis and Chatzis, 2010), fuzzy clustering (Dervis and Ozturk, 2010). Suguna (2011) density based clustering, artificial neural clustering, statistical clustering, grid based, mixed and more (Cheng-Fa and Yen, 2007). Among these approaches, partitional and hierarchical clustering algorithms are the two important approaches in research areas. Partitional clustering aspires directly to obtain a single partition of the set of items into groups. In partitional clustering algorithms, the datasets are partitioned into a specific number of clusters and then it is evaluated based on certain criterion. The hierarchical clustering is considered as a technique of cluster analysis, which tries to construct a hierarchy of clusters. Hierarchical clustering analyzes all the database items individually and treats each of them as separate clusters. The method repeatedly joins the clusters by changing the intercluster distances. Among these algorithms, partitional clustering uses less memory and time for execution.

As far as clustering is considered, Fuzzy C means that it has a major role as it has been used since earlier days. Fuzzy clustering, as a soft clustering method, has been widely studied and successfully applied in clustering and classification. Among the fuzzy clustering methods, Fuzzy C-Means (FCM) algorithm (Dervis and Ozturk, 2010) is the most popular method used in data clustering. Many researches are involved in data clustering using different methods. A genetic algorithm based approach to decide the clustering problem by (Mualik and Bandyopadhyay, 2000) was experimented to evaluate the clustering performance.

Corresponding Author: M. Krishnamoorthi, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Erode, Tamil Nadu, India

(Krishna and Murty, 1999) have proposed an approach called genetic K-means algorithm for clustering analysis, which expresses a basic mutation operator controlled clustering named as distance-based mutation. The main challenges faced in the clustering algorithm are that there are no optimization functions for optimizing the clusters. When redundant data collections are considered, optimization is highly essential for efficient clustering.

Later, different optimization algorithms such as genetic algorithm, particle swarm optimizations are arrived for cluster optimization. Recently, Davis Karaboga and Basturk (2008) have proposed an algorithm called Artificial Bee Colony Algorithm for cluster optimization. The Artificial Bee Colony Algorithm is based on the foraging behavior of the honey bees. Honey-bee is among the most closely studied social insect. Their foraging behavior, knowledge, remembrance and information sharing features have recently been one of the most stimulating study areas in swarm intelligence. The ABC algorithm is recently introduced in the cluster optimization process. So, it is bound with some defaults, which may affect in some particular data optimizations. In order to improve the performance, a new algorithm is proposed to replace the existing ABC algorithm (Dervis and Ozturk, 2011; Changsheng *et al*., 2010; Bahriye and Karaboga, 2012; Suguna, 2011; Visu *et al*., 2012). In this approach, an ABC algorithm with FCM operator is introduced to improve the optimization efficiency of ABC algorithm.

In the normal ABC algorithm, the optimization contains three different stages. They are employed as bee phase, onlooker bee phase and the scout bee phase. The input datum is categorized into any of the employed bee or the onlooker bee. The employed bee and the onlooker bee are processed according to the fitness of the solutions. Scout bee is one of the attracting elements in the ABC algorithm, which is introduced to the bee colony when an abandon solution is developed. In such a case, a random bee is introduced into the bee colony. The conventional approach is slightly modified in the proposed approach using FCM operator. Instead of random assignment, the scout bee is introduced according to a fuzzy function. In this approach, the scout bee is introduced after every cycle.

The main contributions in this study are:

- A modified ABC algorithm is used for optimizing the clusters
- Processing is concentrated on the onlooker bees
- At the end of every cycle, a scout bee is introduced
- The scout bee is introduced using the FCM operator

1.1. Modified ABC Algorithm for Clustering (AB-FF)

The modified ABC algorithm stands for Artificial Bee Colony with FCM function. The proposed approach is a hybrid algorithm, to incorporate the FCM operator into the ABC algorithm. So, developing a method which will provide effectiveness of both the algorithms is a tedious task. The proposed algorithm is well designed to obtain all the features of the above mentioned algorithms. A modification is proposed in the ABC algorithm with the help of FCM function. The ABC algorithm consists of three phases i.e., the employed bee phase, the onlooker bee phase and the scout bee phase. In the three phases, employed bee phase and onlooker phase are inevitable phases whereas the scout bee phase is a random phase. So, in the proposed work, the FCM operator is incorporated in scout bee phase.

In ABC algorithm (Dervis and Ozturk, 2011), a random solution is generated for the scout bees, when an abandon solution occurs. The random solution is expected to deliver the best solution, but it is not dependable. The proposed approach comes with an alternative, in each cycle, a solutions for the scout bees is introduced by the FCM operator. The new solution from the FCM operator is generated based on the solutions of the employed bee and onlooker bee phases for the better optimization results. The fitness function used in the proposed approach is given below Equation 1:

$$fit_i = \begin{cases} \frac{1}{1+f_i}, \text{if } f_i \geq 0 \\ 1+\text{abs}(f_i), \text{if } f_i \leq 0 \end{cases} \quad (1)$$

The data are initially processed according to the number of clusters. The centroids are defined for the processing according to the ABC algorithm. The dataset can be considered as the following set with ‘n’ elements:

$$D = \{d_1, d_2, d_3 ....., d_n\} \quad (2)$$

Assuming that the two clusters are generated from the given dataset. Two centroids for the clusters are selected from the dataset D:

	d_1	d_2	.....	d_n
c_1	i_1	i_2	.....	i_n
c_2	j_1	j_2	.....	j_n

where, c_1 and c_2 represent the two centroids of the two clusters respectively. i and j be the representation of distance values between the centroids and the data points. After the distance calculation, the data are moved into the cluster, which has the least distance value when compared with other distance values of the data point. Finally, the data points are grouped into two clusters according to their least distance value. The f_i values of the clusters are calculated from the distance value of the data points. Consider the following clusters:

Table 1 represents the clusters with its distance value. Now, the proposed approach finds the f_1 values from the distances according to the following formula Equation 3:

$$f_{c1} = i_1 + i_2; f_{c2} = j_1 + j_2 + j_3$$

i.e., In general:

$$f_i = \sum_{i=0}^{n} \text{distance}_i \quad (3)$$

The fitness function is calculated as the sum of all the f_i values Equation 4:

$$\text{fitness} = \sum_{i=0}^{n} f_i \quad (4)$$

The fitness value is calculated for finding the most relevant data for the next population. The solution with better fitness value among the population survives and the solutions with worst fitness value are rejected.

In the three phases of the ABC algorithm, the employed bee and the onlooker phase are same as that described in the basic ABC algorithm and the employed bee is considered as the initial population. The major modification is made in the scout bee phase.

1.2. Employed Bee Phase

The ABC algorithm consists of a multidimensional search space, in which there are employed bees and onlooker bees. Both the bees stated above are characterized by their experience in finding the food source. The data available are selected and their corresponding solutions are randomly created using the uniform distribution. The initial population is then selected for the employed bee phase. Consider the solution in **Table 2**.

This is the generated solution after the initialization process. This solution is changed in the employed bees using Equation 5. These employed bees possess the food location. Here, the position values are marked with the notation 'I'.

Table 1. Clusters and distances

C1	C2
i_1	j_1
i_2	j_2
	j_3

Table 2. Sample solution

I_1	I_2
I_5	I_6
I_3	I_4
I_7	I_8

1.3. Onlooker Bee Phase

In the above table, E stands for the employed bee variants and O stands for the onlooker bee variants. In the employed bee phase, only the E variants in the **Table 2** are considered. i.e., the solution for the first onlooker bee can be calculated using the following formula:

$$v_{i,j} = E_{i,j} + \Phi_{i,j}(E_{i,j} - E_{k,j})$$
$$i, e, v_0 = < I_1 I_2 > + \text{value}[0,1](< I_1 I_2 > - < I_k I_j >) \quad (5)$$

where, k and j are random indices and Φ_{ij} is a randomly produced number in the range [-1, 1].

As per the equation, a new solution is generated. The solution is treated with the fitness function in order to obtain the fitness value. The new fitness value is compared with the previous best value **Table 3**. If the new fitness is better than the old, the new solution will be selected and the old one be rejected. This process will continue until all the employed bees are processed. The employed bee phase can be elaborated by means of a numerical example.

Example 1:

Consider the following distance values:

1.4112	–2.5644
0.4756	1.4338
–0.1824	–1.0323

The fitness of the above randomly generated solution is:

f(x)	fitness
8.5678	11.9488
2.2820	
1.0990	

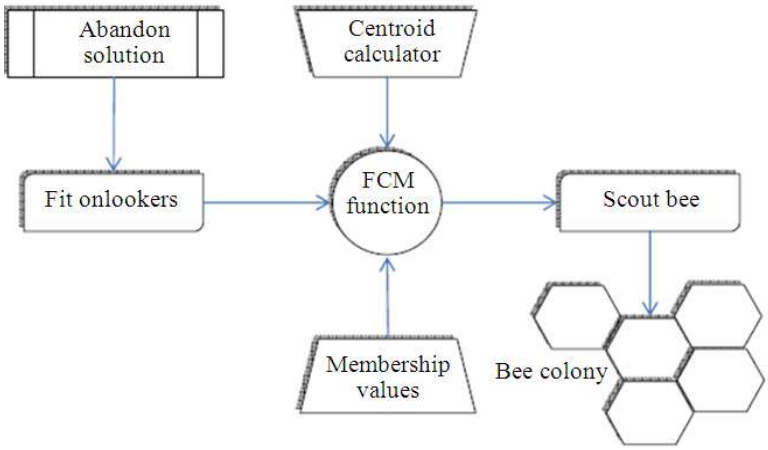

Fig. 1. The scout bee phases

AB-FF Algorithm

/* Input: Dataset */
/* Output: Clusters */
Step 1. Read dataset
Step 2. Initialize the colony.
/* Cycle starts */
Step 3. Apply employed bees.
Step4.Calculate Distance matrix

Step 5.Find the fitness, $fit_i = \sum_{i=1}^{n} f_i$

$$f_i = \frac{1}{N}\sum_{j=1}^{N} mind^2(z_j, c_i)$$

Step 6. Start Onlooker bee phase.
Step 7. Find new solution for onlookers,

$$v_{i,j} = x_{i,j} + \Phi_{i,j}(x_{i,j} - x_{k,j})$$

Step 8. Find fitness of onlooker bee
Step 9. Select the bee with Least fit value.
Step 10. Apply FCM operator in the selected bee to find Scout bee.
Step 11. Add Scout bee into Employed bee.
/* Cycle ends */
Step 12. Repeat Step 3- Step 11.
Step 13. End

Fig. 2. Pseudocode

Table 3. Sample results

Current solution	Current fitness	New solution	New fitness
(1.4112,-2.5644)		(2.1644,-2.5644)	
(0.4756, 1.4338)	11.9488	(0.4756, 1.6217)	11.9988
(-0.1824,-1.0323)		(-0.0754,-1.0323)	

Here, the ABC algorithm first starts processing the employed bee in the cycle.

The distance values are iteratively changed by changing the indices k and j in the Equation (2). Thus, a set of data with different fitness values are generated. Among those set of distance values, the set of values with higher fitness value is selected for the scout bee phase.

1.4. Scout Bee Phase

The scout bee is a randomly assigned bee in the ABC algorithm, if an abandon solution occurs. In other words, if there is no new solution obtained at the end of the cycle, a bee position will be randomly assigned to get a new solution. In the proposed approach (as shown in **Fig. 1**), a new method is used to introduce the scout bee, in the case of an abandon solution. Instead of adding a random position to the bee colony, the proposed approach uses the FCM function to introduce the scout bee. The scout bee is produced from the onlookers with the highest fitness, though they have not possessed best fitness than the previous one. The onlookers with high fitness are sorted in their ascending order of their fitness values:

$$O_i = [p_i, p_2, ..., p_n]$$

Here, O_i is the set of onlooker bee positions, which has the highest fitness. A scout bee is generated from the processing of the above set with the help of the FCM function. Thus, the new scout bee can be generated from the following function Equation 6 and 7:

$$C_j = \frac{\sum_{i=1}^{n} m_{ij}^{n}.p_i}{\sum_{i=1}^{n} m_{ij}^{n}} \quad (6)$$

$$m_{ij} = \frac{1}{\sum_{i=1}^{C} (|p_i - C_j|)^{\frac{2}{x-1}}} \quad (7)$$

Where:

m_{ij} = The fuzzy membership value,
C_i = The centroid calculated for the set

The FCM function generates a new position for the scout from the membership values and the centroid values. As mentioned above, only one centroid is defined since a single scout bee has to be introduced to the context. Unlike the FCM algorithm, in the proposed approach a single iteration is conducted to obtain the new solution. The main advantage in the proposed approach is that the scout bee is added to the solution by processing from the best fitness values obtained from the last solution in the cycle instead of randomly assigning a scout bee. Moreover the scout bee is introduced in every cycle, which improvises the optimal solution and the overall execution time. The Psuedocode for Modified ABC Algorithm for Clustering is given in **Fig. 2**.

1.5. Results and Performance Evaluation

The proposed approach deals with the clustering of data based on the ABC algorithm. The method we have proposed incorporates the FCM function with the ABC algorithm for obtaining better efficiency. The performance of the proposed approach is evaluated in the following section under different evaluation criteria. The algorithm is implemented in the JAVA language and executed on a core i5 processor, 2.1MHZ, 4 GB RAM computer.

1.6. Dataset Description

The proposed hybrid clustering algorithm is tested on three different datasets and compared with other optimization algorithms in the literature. The three datasets are namely Iris, Thyroid and Wine datasets taken from UCI Machine Learning Laboratory.

The Iris dataset: This data set contains 3 categories of 50 objects, in which each class refers to a type of iris plant. There are 150 instances with four numeric features and no missing attribute value. The attributes of the iris data set are sepal length in cm, sepal width in cm, petal length in cm and petal width in cm.

The Thyroid dataset: This data set contains three types of 215 patients suffering from human thyroid diseases. Thyroid diseases are tested based on 5 different tests and no missing attribute value.

The Wine dataset: This Data Set is the results of a chemical analysis of 178 wines grown in the same region in Italy, but derived from three different cultivars. Wine type is based on 13 attributes derived from chemical analysis of the wine.

The performance of the proposed hybrid method (AB-FF) is plotted here. The evaluation factors considered for the experimentation is a number of clusters and the number of cycles for the execution. The proposed approach has selected three test cases for the datasets, the best case, worst case and average case. The best case includes the cycle value in which the algorithm's optimum performance and the worst case include the algorithm's worst performance producing cluster cycles. The average case includes the cluster cycles, which gives a satisfactory output by the Hybrid algorithm.

1.7. Evaluation Based on Time

The performance of the proposed approach is evaluated on the basis of time in the iris dataset. Three cases have been selected that are the time of execution on worst case, average case and the best case. The performance of the proposed approach is plotted in the graph given below.

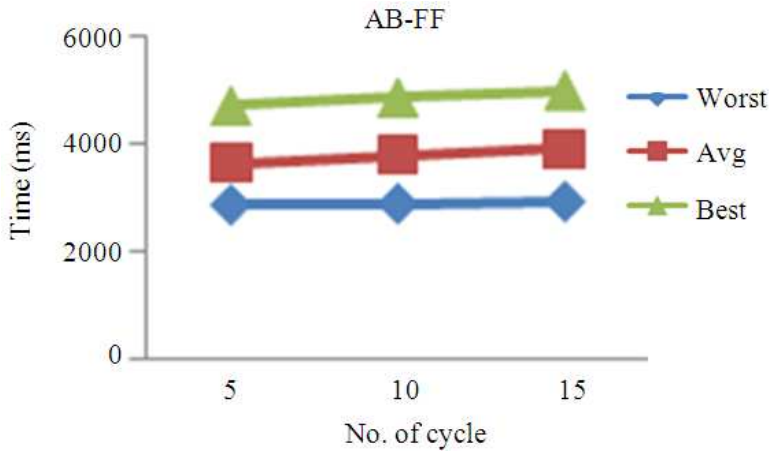

Fig. 3. Time for execution iris data

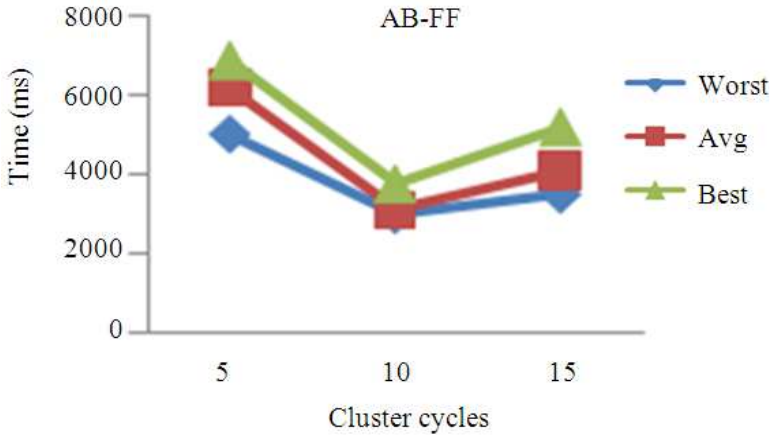

Fig. 4. Time for execution wine data

1.8. Iris Dataset

Figure 3 shows that as the number of iteration increases, the time for execution also increases proportionally. The analysis from the graph shows that higher execution time is needed for the best case, for executing the data at different cycle iterations. The responses of the Iris data are impressive in the case of time for execution. The figure illustrates that with the application of the FCM function, remarkable reduction in time for execution is determined.

1.9. Wine Dataset

Figure 4 shows that as the number of iteration increases, the time for execution also increases proportionally. The analysis from the graph shows higher execution time is needed for the best case for executing the data at different cycle iterations. Also the case of Wine dataset is not different from the IRIS data. The time for execution proportionally increases as the number of clusters increase. Considering the wine data set, there is much difference in the iris data. The time of execution is remarkably changes for wine dataset.

1.10. Thyroid Dataset

Figure 5 shows that as the number of iteration increases, the time for execution also increases proportionally. The analysis from the graph shows that higher execution time is needed for the best case for executing the data at different cycle iterations. When considering the case of thyroid data, in spite of the large dataset, the response to time execution is remarkable. The AB-FF algorithm took less time for execution for thyroid dataset at different levels.

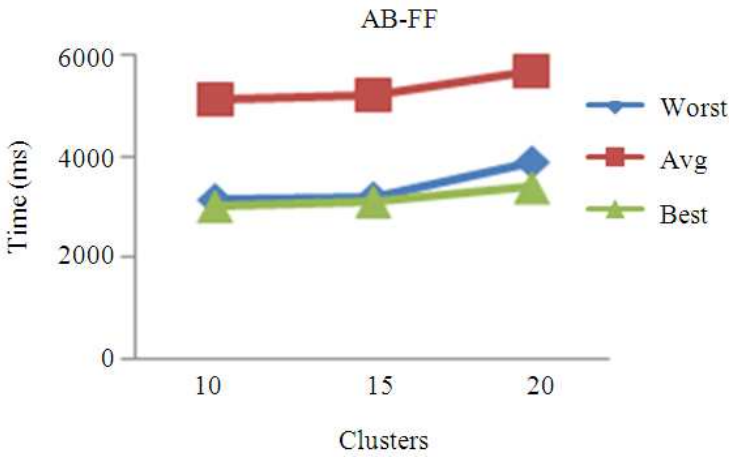

Fig. 5. Time for execution thyroid data

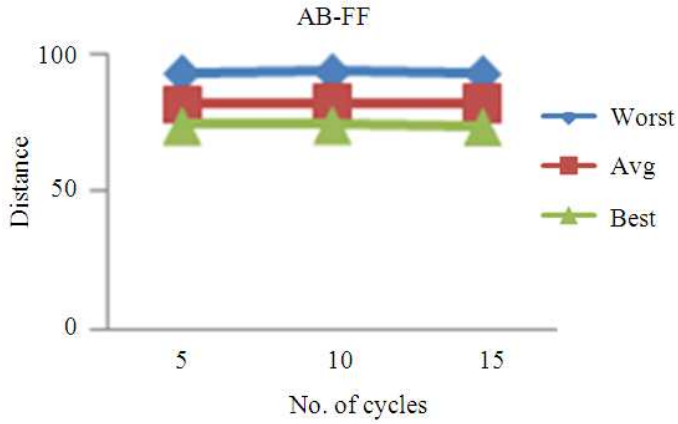

Fig. 6. Intra-cluster distance of iris data

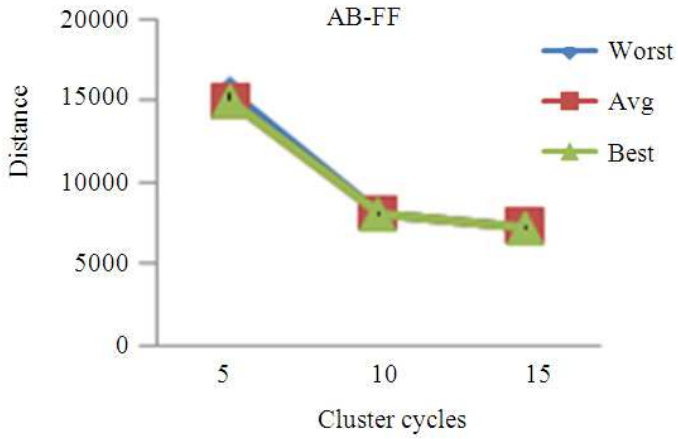

Fig. 7. Intra cluster distance of wine data

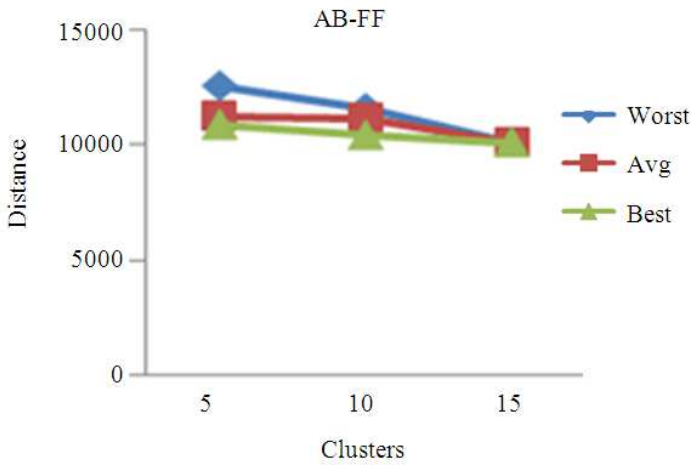

Fig. 8. Intra-Cluster distance of thyroid data

Table 4. Comparative analysis

Dataset	Criteria	GA	TS	SA	ACO	K-NM-PSO	ABC	AB-FF
Iris	Average	125.19	97.86	97.13	97.17	96.67	78.94	74.4
	Worst	139.78	98.57	97.26	97.81	97.01	78.94	74.6
	Best	113.98	97.36	97.10	97.10	96.66	78.94	74.1
Thyroid	Average	10128.82	10354.31	10114.04	10112.13	10109.70	10104.03	9987.0
	Worst	10148.39	10438.78	10115.93	10114.82	10112.86	10108.24	9542.0
	Best	10116.29	10249.73	10111.82	10111.82	10108.56	10100.31	8866.0
Wine	Average	16530.53	16785.46	16530.53	16530.53	16293.00	16260.52	8101.0
	Worst	16530.53	16837.54	16530.53	16530.53	16295.46	16279.46	15521.0
	Best	16530.53	16666.22	16530.53	16530.53	16292.00	16257.28	7289.0

1.11. Evaluation Based on Intra-Cluster Distance

Here, the performance of the proposed approach is evaluated on the basis of intra-cluster distance of the clusters. Intra-cluster distance on worst case, average case and the best case are selected as three parameters for the evaluation. The performance of the proposed approach is plotted in the graph given below.

1.12. Iris Dataset

Figure 6 shows that unlike the time for execution, as the number of iteration increases the intra-cluster distance decreases proportionally. The intra-cluster distance is.

1.13. Wine Dataset

Figure 7 shows that unlike the time for execution, as the number of iteration increases the intra-cluster distance decreases proportionally. The intra-cluster distance is lower for the best case, which is evident from the plotted graph. The wine data become more responsive, when the FCM function is introduced to the ABC algorithm. The enhancement has got improved results regarding the intra-cluster distance.

1.14. Thyroid Dataset

Figure 8 shows that unlike the time for execution, as the number of iteration increases the intra-cluster distance decreases proportionally. The intra-cluster distance is lower for the best case, which is evident from the plotted graph. The case of thyroid is not different from the other two datasets, as the number of cycles increases the intra-cluster distance reduces. The responses are plotted in the figures mentioned below.

The reduced number of intra-cluster distance is because the FCM operator gives importance to the membership function. The membership function evaluates values each minute and specifies the centroid. This feature of the FCM operator gives emphasis to the proposed ABC algorithm.

1.15. Comparative Analysis and Discussion

Here, we describe the comparative study of the proposed approach with different similar algorithms.

The comparative evaluation is done based on the intra cluster distances of the proposed approach and that of the other comparing methods. The methods used for the comparison are Genetic algorithm, Tabu search algorithm, Simulated annealing, Ant colony algorithm, K-NM-PSO and Artificial bee colony algorithm. The detailed comparison is plotted below in **Table 4**. The comparison table shows the response of AB-FF data with the other similar algorithms. The **Table 4** is plotted in reference to the research conducted by Changsheng *et al*. (2010).

Table 4 shows the comparison of the proposed approach with some similar algorithms, which are related to the optimization of the clustering process. We have selected three different datasets for comparison, which includes Iris dataset, Thyroid dataset and Wine dataset. The study of the comparison data from three datasets has stated that in the best cases, our proposed approach has given a minimum intra cluster distance when compared to the other methods. The analyses have shown the significance of our approach in optimization of the clusters.

2. CONCLUSION

In this study, a cluster optimization methodology is proposed by highlighting the Artificial Bee Colony (ABC) algorithm. The proposed approach deals with a modified ABC algorithm for cluster optimization. The major modification is made on the scout bee phase of the ABC algorithm. In the scout bee phase, rather than applying a random position to the scout bee, the position is assigned with the help of the Fuzzy C- Means operator (FCM). The scout bee is introduced after every cycle, which results in the reduced number of cycle iterations. The experimentation is done with three different datasets namely, the Iris dataset, the Thyroid dataset and the Wine dataset. The comparative analysis has shown that the computational result obtained from the modified algorithm is very encouraging in terms of the quality of the solution and the execution time.

3. REFERENCES

Bahriye, A. and D. Karaboga, 2012. A modified artificial bee colony algorithm for real-Parameter optimization. Int. J. Inform. Sci., 192: 120-142. DOI: 10.1016/j.ins.2010.07.015

Changsheng, Z., D. Ouyang and J. Ning, 2010. An artificial bee colony approach for clustering. Expert Syst. Applic., 37: 4761-4767. DOI: 10.1016/j.eswa.2009.11.003

Cheng-Fa, T. and C.C. Yen, 2007. ANGEL: A new effective and efficient hybrid clustering technique for large databases. Comput. Sci., 4426: 817-824. DOI: 10.1007/978-3-540-71701-0_90

Dervis, K. and C. Ozturk, 2010. Fuzzy clustering with artificial bee colony algorithm. Sci. Res. Essays, 5: 1899-1902.

Dervis, K. and C. Ozturk, 2011. A novel clustering approach: Artificial Bee Colony (ABC) Algorithm. Applied Soft Comput., 11: 652-657. DOI: 10.1016/j.asoc.2009.12.025

Karaboga, D. and B. Basturk, 2008. On the performance of Artificial Bee Colony (ABC) algorithm. Applied Soft. Comput., 8: 687-697. DOI: 10.1016/j.asoc.2007.05.007

Krinidis, S. and V. Chatzis, 2010. A robust fuzzy local information c-means clustering algorithm. IEEE Trans. Image Proc., 19: 1328-1337. DOI: 10.1109/TIP.2010.2040763

Krishna, K. and N.M. Murty, 1999. Genetic K-means algorithm. IEEE Trans. Syst. Man Cybernetics B Cybernetics, 29: 433-439. DOI: 10.1109/3477.764879

Mualik, U. and S. Bandyopadhyay, 2000. Genetic algorithm-based clustering technique. Patt. Recogn., 33: 1455-1465. DOI: 10.1016/S0031-3203(99)00137-5

Suguna, N., 2011a. An independent rough set approach hybrid with artificial bee colony algorithm for dimensionality reduction. Am. J. Applied Sci., 8: 261-266. DOI: 10.3844/ajassp.2011.261.266

Visu, P., S. Koteeswaran and J. Janet, 2012. Artificial bee colony based energy aware and energy efficient routing protocol. J. Comput. Sci., 8: 227-231. DOI: 10.3844/jcssp.2012.227.231

Yu, F., H. Xu, L. Wang and X. Zhou, 2010. An improved automatic FCM clustering algorithm. Proceedings of the 2nd International Workshop on Database Technology and Applications, Nov. 27-28, IEEE Xplore Press, Wuhan, pp: 1-4. DOI: 10.1109/DBTA.2010.5659043