Artificial Bee based Optimized Fuzzy c-Means Clustering of Gene Expression Data

Punam Priti Pradhan[1], Debahuti Mishra[2], Sashikala Mishra[3] and Kailash Shaw[4]

[1,2&3] *Institute of Technical Education and Research,*
Siksha O Anusandhan University, Bhubaneswar, Odisha, India
[4]*Gandhi Engineering College, Bhubaneswar, Odisha, India*

[1]punam.pradhan2009@gmail.com
[2] debahuti@iter.ac.in
[3] sashi.iter@gmail.com
[4] kailash.shaw@gmail.com

***Abstract*—Artificial Bee Colony (ABC) algorithm is a swarm based meta-heuristic algorithm that was introduced by Karaboge in 2005 for optimizing numerical problem. Clustering is an important tool for a variety of applications in data mining, statistical data analysis, data compression and vector quantization. The goal of clustering is to organize data into clusters such that the data in each cluster shares a high similarity while being very dissimilar to data from other clusters. Fuzzy clustering extends crisp clustering in the sense that objects can belong to various clusters with different membership degrees at the same time, whereas crisp or deterministic clustering assigns each object to a unique cluster. Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. In this paper, we have used the ABC fuzzy clustering on three different data sets from UCI database. Here we show how ABC optimization algorithm is successful in fuzzy c-means clustering.**

Keywords— **Artificial bee colony, Data normalization, Principal component analysis, Fuzzy c-means**

I. INTRODUCTION

Cluster is a collection of data object that are similar to one another and thus it can be treated collectively one group. The goal of the cluster is to organize the data into cluster such that there is maximum intra cluster similarity and minimum inter cluster similarity. Clustering provides high availability when the data in critical applications to keep running in the event of a failure [1]. It provides a number of advantages over using non-clustered data. There are different types of clustering algorithms such as *k*-means, density based, hierarchical, partitioning based and fuzzy c-means clustering. The most popular class of clustering algorithms is *k*-means algorithm. It is a centre based, simple, and fast algorithm [1]. The aim of the algorithm is to partition the data into *n* objects into *k* clusters in which each object belong to the cluster with the nearest mean. It is unable to handle noisy data and outliers and it is also not suitable for the data belonging to more than one clusters. In fuzzy clustering the data point can belong to more than one cluster. It gives better result for overlapped data set and as comparison then *k*-means algorithm. This is used to improve the accuracy of clustering under noise. FCM was introduced by Bezdek in 1981 is the most popular fuzzy clustering algorithm [2]. FCM is an effective algorithm; the random selection in centre points makes iterative process falling into the local optimal solution easily [1]. Fuzzy clustering problem is a combinatorial optimization problem [3] that is difficult to solve and obtaining optimal solutions to large problems it can be quite difficult. In order to address the issues, different evolutionary algorithm are used such as genetic algorithm (GA), differential evolution (DE), ant colony optimization (ACO), particle swarm optimization (PSO), cat swarm optimization (CSO) have been successfully applied. Since the ABC is a simple in concept, easy to implement and fewer control parameters it has attracted the attention of researchers to solve numerical optimization and engineering optimization problem [4]. The motivation of this paper is to apply ABC algorithm for optimization of clustering by fuzzy *c*-means. The layout of this paper is as follows; section II deals with background study, in section III the preliminary concepts of data normalization, PCA for feature reduction, FCM, and ABC are described. In section IV schematic representations of proposed model is given; in section V experimental evaluations and results are described and finally, section VI deals with conclusion and future work.

II. BACKGROUND STUDY

The problem of optimizing a clustering technique has always been a challenge and an area of interest for the researchers. Changsheng Zhang *et al.* [5] developed an artificial bee colony algorithm to solve clustering problems

which is influence by the bee's foraging behaviour. The ABC algorithm for data clustering can be applied when one data belonging only one cluster. It gives favourable result in terms of the quality of solution found, the average number of function evaluation and the processing time required. Xiaohui Yan *et al*. [6]. Proposed Hybrid artificial bee colony (HABC) algorithm to improve the optimization ability of conical ABC i.e speed of ABC algorithm is decreases when the dimension of the problem increases. In HABC the crossover operator of GA is introduced to increase the information exchange between the bees. The result shows that HABC algorithm gives favourable result in terms of accuracy, robustness, convergence and speed. Dervis karaboge *et al*. [1] used ABC algorithm to fuzzy clustering of different medical datasets i.e cancer, diabetes & heart dataset. The output of ABC algorithm is compared with FCM algorithm the result shows that ABC algorithm is successfully optimization in fuzzy clustering. Salima ouadfel *et al*. [2] used a modified version of ABC algorithm to improve the quality of fuzzy clustering in which a new mutation strategy inspired from differential evaluation is introduced in order to improve the exploitation process. The result shows that Modified ABC with FCM is effective and efficient. Zhi-gang Su *et al*. [7] used an automatic fuzzy clustering technique for novel version of ABC algorithm is called Variable string length Artificial Bee Colony (VABC) algorithm.The important characteristics of the VABC-FCM are that it can automatically develop the number of clusters and find the proper fuzzy partitioning for a wide variety of data sets. To validate the performance of VABC-FCM some artificial data sets and real-life data sets are applied. Anan Banharnsakun *et al*. [8] have investigated hidden patterns that may exist in datasets. In this paper, the Best-so-far ABC with multiple patrilines algorithm, replaced the neighbouring solutions-based approach of the classical ABC with the Best-so-far technique in order to maximize the local search ability of the onlooker bees. This bias the solution selected by onlooker bees towards the optimal solution more quickly. The concept of patrilines, another bee-inspired idea is applied to the Best-so-far ABC in order to increase the global search capability while decreasing the computing time. This algorithm solves the problems that involve a large amount of data, when the number of clusters is known a priori. Bahriye Akay *et al*. [9] developed the performance of classical and modified versions of the ABC algorithm and compared their performances against state-of-the-art algorithm. The result shows that standard ABC algorithm can efficiently solve basic functions while the modified ABC algorithm produces promising results on hybrid functions compared to other algorithms. D. Shanthi *et al*. [10] proposed a novel learning scheme for process neural networks based Gaussian mixture weight functions and Collaborative ABC (C-ABC). The C-ABC has great explorative search features and better convergence compared to the original algorithm and it avoids local minima by promoting exploration of the search space. This is the ability of the C-ABC to perform local and global search simultaneously.

III. Preliminary Concepts

A. *PCA for feature reduction*

Feature reduction transforms the data in the high-dimensional space to a space of lower dimensions [12]. PCA is the most commonly used dimension reduction technique.The data transformation may be linear, as in principal component analysis (PCA), performs a linear mapping of the data to a lower dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. PCA is a way to avoid the dimensionality by projecting the data onto a lower-dimensional space. It is used to compute the eigenvalues and eigenvectors of the matrix. Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables is called principal components. The number of principal components is less than or equal to the number of original variables. PCA is sensitive to the relative scaling of the original variables.

B. *Fuzzy c-means clustering (FCM)*

FCM is a clustering algorithm that was developed by Dunn in 1973 and improved by Bezdek in 1981 [2]. It is a method of clustering algorithm which allows one data may belong to two or more clusters. It is normally used in pattern recognition [13]. It is based on minimization of the following objective function (1):

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\|x_i - c_j\right\|^2, 1 \le m < \infty \qquad (1)$$

Where,

m= is any real number greater than 1

u_{ij} = is the degree of membership of x_i in the cluster j

x_i is the i^{th} of d-dimensional data

c_j is the cluster centre of d-dimension data

$\left\|x_i - c_j\right\|^2$ is the distance measured of similarity between the measured data and the cluster data

Fuzzy partitioning is carried out through an iterative optimization of the objective function with the update of membership u_{ij} and the cluster centres c_j by:

$$u_{ij} = \frac{1}{\sum_{k-1}^{c}\left[\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right]^{\frac{2}{m-1}}} \qquad (2)$$

and

$$c_j = \frac{\sum_{i-1}^{N} u_{ij}^{m} . x_i}{\sum_{i-1}^{N} u_{ij}^{m}} \qquad (3)$$

This iteration will stop when,

$$max_{ij}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < \varepsilon \quad (4)$$

Where ε is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m.

C. Artificial bee colony algorithm

It is a swarm intelligent method which inspired from the intelligent foraging behaviour of honey bee swarms. It strength is its robustness and its simplicity [2]. It is developed by surveying the behaviour of the bees is finding the food source which is called *nectar* and sharing the information of food source the bee which is present in the nest. In the ABC the artificial agents are classify into three types; such as employed bee, the onlooker bee and the scout each of the bee plays different role in the process. The employed bee stays on a food source and in its memory provides the neighbourhood of the food source. Each employed bee carries with her information about the food source and shares the information to onlooker bee. The onlooker bees wait in the hive on the dance area, after getting the information from employed bees about the possible food source then make decision to choose a food source in order to use it. The onlooker bees select the food source according to the probability of that food source. The food source with lower quantity of *nectar that* attracts less onlooker bees compared to ones with a higher quantity of nectar. Scout bees are searching randomly for a new solution (food source).The employed bee whose food source has been abandoned it becomes a scout bee. The goal of the bees in the ABC model is to find the best solution. In the ABC algorithm the number of employed bees is equal to the number of onlooker bees which is also equal to the number of solutions. The ABC algorithm consist of a maximum cycle number (MCN) during each cycle [2], there are three main parts:

I. Sending the employed bees to the food sources and calculate their nectar quantities
II. Selecting the food sources by the onlooker bees
III. Determining the scout bee and discover a new possible food sources

Employed Bee: In the employed bees' phase, each employed bee determines a new solution from the neighbourhood of the current food source (solution). The new food source (new solution) is calculated using (5).

$$x_{ij}(t+1) = \theta_{ij} + \emptyset(\theta_{ij} + \emptyset\left(\theta_{ij}(t) - \theta_{kj}(t)\right) \quad (5)$$

Where x_i represents the position of the i^{th} onlooker bee, t denotes the iteration number, θ_i represents the position of the i^{th} employed bee. θ_k is randomly chosen employed bee, j represents the dimension of the solution and $\varphi(.)$ produces a series of random variables in the range [-1,1]. The employed bee compared the current solution with the new solution and memorizes the best one by apply the greedy selection process. When all employed bees have finished this search process, then they share the fitness value (nectar information) and the position of the food source (solution) to the onlooker bees.

Onlooker Bee: In the onlooker bee phase, after getting the information about the nectar and position of the food source each onlooker bee selects a food source with a probability of higher nectar information. The movement of the onlookers is calculated using (6).

$$P_i = \frac{F(\theta_i)}{\sum_{k=1}^{S} F(\theta_k)} \quad (6)$$

Where θ_i denotes the position of the i^{th} employed bee, S represents the number of employed bees, and P_i is the probability of selecting the i^{th} employed bee. If the selected food source is better than the old solution then it is updated otherwise it keeps the old solution.

Scout Bee: If a food source position cannot be improved through fixed cycles, it is called '*limit'*, it means that the solution has been sufficiently exploited, and it may be removed from the population. In this case, the employed bee becomes scout bee. The scout bees determine a new random food source (solution) position using (7):

$$\theta_{ij} = \theta_{ijmin} + r.\left(\theta_{ijmax} - \theta_{ijmin}\right) \quad (7)$$

Where, r is a random number and $r \in [0, 1]$. If the new food source is better than the abandoned one, then the scout bee become an employed bee. This process is repeated until the maximum number of cycles (MNC) is reached. Based on the better fitness value the optimal solution is determined by the bee.

IV. SCHEMATIC REPRESENTATION OF PROPOSED MODEL

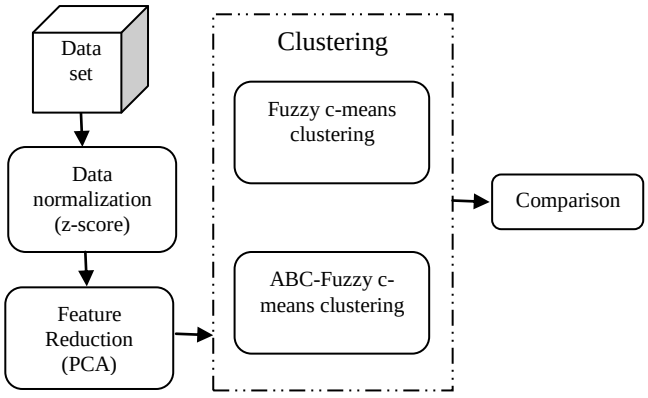

Fig.1: Proposed model

In this proposed model, the dataset is normalized by using z-score normalization. PCA is applied to the normalized dataset for feature reduction, which reduces the dimension of the dataset. FCM and ABC-FCM has been used as clustering methods to those three data sets and finally, the result of the two techniques are measured and compared.

V. EXPERIMENTAL EVALUATION AND RESULT ANALYSIS

We have used three benchmarked datasets downloaded from UCI machine learning repository [14] in our experiment which are described in table 1. In this work, we have used MATLAB version 7.10, release name- R2010a. The experiment was carried on Intel core i3 processor, 2.4 GHZ, 32 bit 1GB RAM, 1GB disk space for MATLAB, 3-4 GB for ideal installation. The total experimental evaluation has been carried out in the following steps:

TABLE I: Description of datasets

Data Sets	**Dimension**
Lung Cancer	181X12533
Leukaemia	72 X 256
Breast cancer	98 X 25

Step1: Collection of datasets: Lung cancer, Leukaemia and Breast Cancer data sets were collected from UCI repository [14].

Step2: Normalization of datasets: In z-score normalization, the values for and attribute A are normalized based on the mean and standard deviation of A. A value v of A is normalized to v' by computing (8):

$$v' = ((v - \bar{A})/\sigma_A) \quad (8)$$

Where, $\bar{A}$ and σ_A are the mean and the standard deviation respectively of attribute A. This method of normalization is useful when the actual minimum and maximum of attribute A are unknown. Here, we have used z-score for normalization of all three data sets.

Step3: Feature reduction: Principal components analysis (PCA) is a very popular technique for dimensionality reduction. Given a set of data on n dimensions, PCA aims to find a linear subspace of dimension d lower than n such that the data points lie mainly on this linear subspace. Such a reduced subspace attempts to maintain most of the variability of the data. After normalizing the data sets, PCA has been used for reduction purpose as given in table II.

TABLE II: Description of datasets after reduction

Data Sets	Dimension	After PCA
Lung Cancer	181X12533	181X27
Leukaemia	72 X 256	72 X 16
Breast cancer	98 X 25	98 X 12

Step4: Fuzzy c-means clustering: In this step, the PCA reduced data sets have been clustered using FCM. This algorithm works by assigning membership to each data point corresponding to each cluster centre on the basis of distance between the cluster centre and the data point. More the data is near to the cluster centre more is its membership towards the particular cluster centre. Clearly, summation of membership of each data point should be equal to one. After each iteration, membership and cluster centres are updated according to (2) and (3) as shown in fig.2. In lung cancer dataset when fuzzy c-means has been applied, the data is assigned to membership of each cluster centre as a result the data belongs to more than one cluster centre but while using ABC-FCM it reduces the noisy data and the number of cluster centre the performance of ABC represent their effectiveness and efficiency as shown in fig. 3 and fig.4.

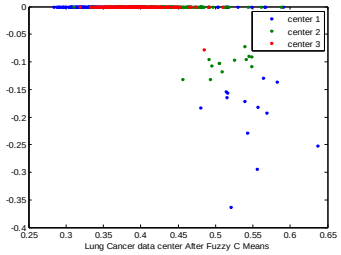

Fig. 2: Center adjustment for Lung cancer data set

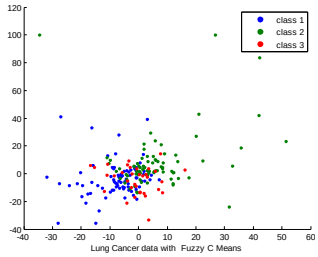

Fig. 3: FCM clustering for Lung cancer data set

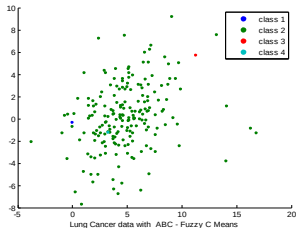

Fig. 4: ABC-FCM clustering for Lung cancer data set

The Leukaemia data set contain two classes. When we apply FCM in the data set having lower value we get the better result but at the expense of more number of iteration thus the processing time is slow. It cannot identify the data is belong to which class but in case of ABC-FCM has few control parameter, so we can easily identify the data and its processing time is faster than FCM thus saving time. The centre updation, clustering after applying FCM and ABC-FCM is shown in fig. 5, fig.6 and fig.7 respectively.

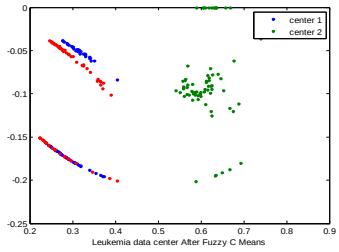

Fig. 5: Center adjustment for Leukemia data set

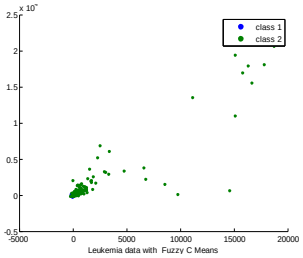

Fig. 6: FCM clustering for Leukemia data set

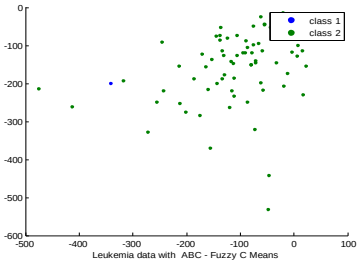

Fig. 7: ABC-FCM clustering for Leukemia data set

In breast cancer data the result of FCM algorithm do not appear vary stable because of not been able to easily escape from the local optimal solution easily. ABC algorithm has fast convergence. The result of ABC-FCM is very successful on optimization of fuzzy clustering as shown in fig. 8, fig.9 and fig.10 for centre updation, FCM clustering and ABC-FCM clustering.

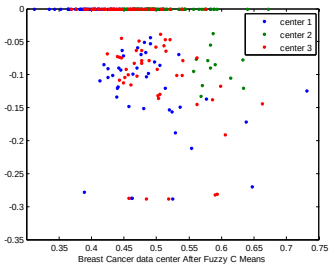

Fig. 8: Center adjustment for Breast cancer data set

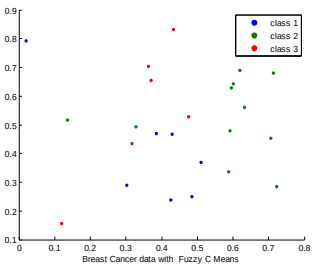

Fig.9: FCM clustering for Breast Cancer data set

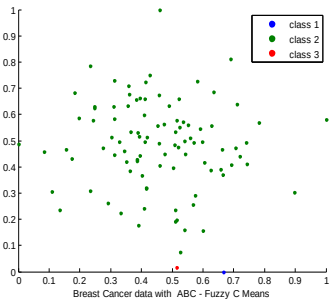

Fig. 10: ABC-FCM clustering for Breast cancer data set

VI. CONCLUSIONS

Swarm intelligence is an emerging area for problem solving and purpose of search, modelling the behaviour of social insects such as ants, cats, birds and bees. Honey bees are among the most closely studied social insects. In this paper, ABC algorithm is developed to solve clustering problems. ABC algorithm is an optimization algorithm used in fuzzy clustering of different datasets which are widely used in benchmark problems. The result of ABC-FCM algorithm compared with FCM algorithm and the result shows that the ABC algorithm is successful on optimization of fuzzy clustering. Furthermore, applying the proposed algorithm to solve other optimization problems is also possible in future research.

REFERENCES

[1] Dervis Karaboge & Celal Ozturk, ”Fuzzy clustering with Artificial Bee colony Algorithm”, *Scientific Research and Essays*, vol.5, no. 14, pp. 1899-1902, 2010

[2] Salima Ouadfel, Souham Meshoul, ”Handling fuzzy image clustering with a modified ABC algorithm”, *International Journal Intelligent System and Application*, pp.65-74, 2012

[3] J. G. Klir and B. Yuan, “Fuzzy sets and fuzzy logic, theory and applications”, *Prentice-Hall Co.*, 2003.

[4] D. Karaboga & B. Basturk, ” Artificial bee colony (ABC) optimization algorithm for solving constrained optimization problems”, *Lec. Notes Comput. Sci*, vol. 4529, 789–798, 2007

[5] Changsheng Zhang, Dantong Ouyang, Jiaxu Ning, ”An artificial bee colony approach for clustering”, *Expert Systems with Applications*, pp.4761–4767,2010

[6] Xiaohui Yan, YunlongZhu, WenpingZou, Liang Wang,” A new approach for data clustering using hybrid artificial bee colony algorithm ”, *Neurocomputing*, pp. 241–250, 2012

[7] Zhi-gang Su, Pei-hong Wang, Jiong Shen, Yi-guo Li, Yu-fei Zhang, En-jun Hu,” Automatic fuzzy partitioning approach using Variable string length Artificial Bee Colony (VABC) algorithm”, A*pplied soft computing*, pp.3421-3441, 2012

[8] Anan Banharnsakun, Booncharoen Sirinaovakul, Tiranee Achalakul,” The Best-so-far ABC with Multiple Patrilines for clustering problems” N*eurocomputing*, 2012

[9] Bahriye Akay, Dervis Karaboga,” A modified Artificial Bee Colony algorithm for real-parameter optimization”, I*nformation Sciences*, pp.120-142, 2012

[10] D.Shanthi, R.Amalraj, ”Collaborative Artificial Bee Colony Optimization Clustering Using SPNN”, *Procedia Engineering*, pp. 989 – 996, 2012

[11] Luai Al Shalabi,Zyad Shaaban, Basel Kasasbeh, ”Data Mining: A preprocessing Engine”, *Journal of Computer Science*, vol.2, no.9, pp.735-739, 2006

[12] Yeung Ka Yee and Ruzzo Walter L, ”An empirical study on principal component analysis for clustering gene expression data”, *Tech. Report, University of Washington*, 2000

[13] M.S Yang, ”A Survey of fuzzy clustering”, *Mathl. Comput. Modelling* vol. 18, no. 11, pp. 1-16, 1993

[14] http://archive.ics.uci.edu/ml