Scientific Research and Essays Vol. 5(14), pp. 1899-1902, 18 July, 2010
Available online at http://www.academicjournals.org/SRE
ISSN 1992-2248 ©2010 Academic Journals

Full length Research Paper

Fuzzy clustering with artificial bee colony algorithm

Dervis Karaboga and Celal Ozturk*

Department of Computer Engineering, Faculty of Engineering, Erciyes University, Kayseri, Turkey.

Accepted 23 June, 2010

In this work, performance of the Artificial Bee Colony Algorithm which is a recently proposed algorithm, has been tested on fuzzy clustering. We applied the Artificial Bee Colony (ABC) Algorithm fuzzy clustering to classify different data sets; Cancer, Diabetes and Heart from UCI database, a collection of classification benchmark problems. The results indicate that the performance of Artificial Bee Colony Optimization Algorithm is successful in fuzzy clustering.

Key words: ABC Algorithm, classification, fuzzy clustering.

INTRODUCTION

Clustering is an important tool for a variety of applications in data mining, statistical data analysis, data compression and vector quantization, aims gathering data into clusters (or groups) such that the data in each cluster shares a high degree of similarity while being very dissimilar to data from other clusters (Jain and Dubes, 1998; Sarkar et al., 1997; Han and Kamber, 2001). The goal of clustering is to group data into clusters such that the similarities among data members within the same cluster are maximal while similarities among data members from different clusters are minimal.

The most popular class of clustering algorithms is K-means algorithm, a center based, simple, and fast algorithm, aims to partition *n* objects into *k* clusters in which each object belongs to the cluster with the nearest mean (MacQueen, 1967). However, in real applications there are no sharp boundaries within the clusters so that data objects might partially belong to multiple cluster. In fuzzy clustering, the data points can belong to more than one cluster and membership degrees between zero and one are used instead of crisp assignments of the data to clusters (Jain et al., 1999). The degree of membership in the fuzzy clusters depends on the closeness of the data object to the cluster centers.

Fuzzy c-means (FCM) which is introduced by Bezdek (1981) is the most popular fuzzy clustering algorithm. However, FCM is an effective algorithm; the random selection in center points makes iterative process falling into the local optimal solution easily. To tackle this problem, evolutionary algorithms such as genetic alg-orithm (GA), differential evolution (DE), ant colony optimization (ACO), and particle swarm optimization (PSO) have been successfully applied (Gan et al., 2009; Das et al., 2006; Zhao, 2007; Runkler and Katz, 2006). The motivation of this paper is apply to Artificial Bee

Colony (ABC) algorithm, which is described by Karaboga based on the foraging behavior of honey bees for numerical optimization problems (Karaboga, 2005), in fuzzy clustering. The performance of ABC algorithm is tested on benchmark classification problems of cancer, diabetes and heart obtained from the UCI data (Frank and Asuncion, 2010). The classification task is done by neural networks and clustering, and ABC algorithm is used for training feed-forward neural networks and finding the cluster centers. This work is the first experiment of ABC algorithm on fuzzy clustering. The paper is organized as describing fuzzy clustering and ABC algorithm, and then presenting the results and discussion. Finally, the last section concludes the paper.

FUZZY CLUSTERING

Clustering is the process of recognizing natural groupings or clusters in multidimensional data based on some similarity measures (Jain et al., 1999). Distance measurement is generally used for evaluating similarities

*Corresponding author. E-mail: celal@erciyes.edu.tr. Tel: 352 4374901 # 32581. Fax: 352 4375784.

between patterns. In contrast to clustering data objects in a unique cluster, fuzzy clustering algorithms result in membership values between 0 and 1 that indicate the degree of membership for each object to each of the clusters.

The fuzzy clustering of objects is described by a fuzzy matrix μ with *n* rows and *c* columns in which *n* is the number of data objects and *c* is the number of clusters. μ_{ij}, the element in the i^{th} row and j^{th} column in μ, represents the degree of membership function of the i^{th} object with the j^{th} cluster. The characteristics of μ are as follows:

$\mu_{ij} \in [0,1]$, i = 1,2,...,n; j = 1,2,...,n (1)

$$\sum_{j=1}^{c} \mu_{ij} = 1 \quad i = 1, 2, \ldots, n \tag{2}$$

$$0 < \sum_{i=1}^{n} \mu_{ij} < n \quad j = 1, 2, \ldots, c \tag{3}$$

The objective function of the fuzzy clustering is to minimize the equation:

$$\sum_{j=1}^{c} \sum_{i=1}^{n} \mu_{ij}^{m} \left\| x_i - z_j \right\|^2 \tag{4}$$

where, z_j is the j^{th} and m is the fuzzy index that governs the influence of membership grades and m is set to 2.

$||x_i - z_j||^2$ is the Euclidean distance from sample points x_i to cluster center z_j.

ARTIFICIAL BEE COLONY ALGORITHM

Artificial Bee Colony (ABC) algorithm is a new swarm intelligence method which simulates intelligent foraging behavior of honey bees. The first studies of ABC algorithm is testing the performance of the algorithm on constrained and unconstrained problems and comparing with those of other well-known modern heuristic algorithms such as Genetic Algorithm (GA), Differential Evolution (DE), Particle Swarm Optimization (PSO) (Karaboga and Basturk, 2007). The classification performance of the ABC algorithm is tested on training neural networks (Karaboga and Ozturk, 2009) and on clustering (Karaboga and Ozturk, 2010) with benchmark classification problems and the results are compared with those of other widely-used techniques. In the model of ABC algorithm, there are three groups of bees; employed bees, onlooker bees and scout bees in the colony of artificial bees (Karaboga, 2010). Firstly, half of the colony consists of the employed bees and the second half consists the onlookers. Employed bees go to the food sources, and then they share the nectar and the position information of the food sources with the onlooker bees which are waiting on the dance area determine to choose a food source. The employed bee whose food source has been abandoned by the bees becomes a scout bee that carries out random search in the simulating model. The goal of bees in the ABC model is to find the best solution, the position of a food source represents a possible solution to the optimization problem and the nectar amount of a food source corresponds to the quality (fitness) of the associated solution.

In the ABC algorithm, the number of employed bees is equal to the number of food sources which is also equal to the number of onlooker bees. There is only one employed bee for each food source whose first position is randomly generated. At each iteration of the algorithm, each employed bee determines a new neighboring food source of its currently associated food source by Equation (5), and computes the nectar amount of this new food source:

$$v_{ij} = z_{ij} + \theta_{ij} (z_{ij} - z_{kj}) \tag{5}$$

where θ_{ij} is a random number between [-1,1]. If the nectar amount of this new food source is higher than that of its currently associated food source, then this employed bee moves to this new food source, otherwise it continues with the old one. After all employed bees complete the search process, they share the information about their food sources with onlooker bees. An onlooker bee evaluates the nectar information taken from all employed bees and chooses a food source with a probability related to its nectar amount by Equation (6). This method, known as roulette wheel selection method, provides better candidates to have a greater chance of being selected:

$$P_i = {fit_i}/{\sum_{n=1}^{SN} fit_n} \tag{6}$$

where fit_i is the fitness value of the solution *i* which is proportional to the nectar amount of the food source in the position *i* and *SN* is the number of food sources which is equal to the number of employed bees. Once all onlookers have selected their food sources, each of them determines a new neighboring food source of its selected food source and computes it's nectar amount. Providing that this amount is higher than that of the previous one, and then the bee memorizes the new position and forgets the old one. The employed bee becomes a scout bee when the food source which is exhausted by the employed and onlooker bees is assigned as abandoned. In other words, if any solution cannot be improved further through a predetermined number of cycles which is called limit parameter, the food source is assigned as an abandoned source and employed bee of that source becomes a scout bee. In that position, scout generates

Table 1. Mean CEP and SD values.

Method problem	ABC-FC	FCM
Cancer	2.65 (0.47)	3.52 (0.49)
Diabetes	31.04 (1.02)	32.55 (2.56)
Heart	15.26 (1.11)	17.19 (1.20)

randomly a new solution by Equation (7). Assume that z_i is the abandoned source and j $\in$ {1,2,...,*D*} where *D* is the solution vector, the scout discovers a new food source which will be replaced with z_i:

$$z_i^j = z_{min}^j + \text{rand}(0,1)(z_{max}^j - z_{min}^j) \quad (7)$$

where *j* is determined randomly which is different from *i*. The steps of ABC fuzzy clustering (ABC-FC) is:

Generate initial population z_i, i=1...*SN*
Evaluate the population
Set cycle to 1
Repeat
FOR each employed bee
Produce new solution v_i by using (5)
Calculate the fitness
Apply the greedy selection process
FOR each onlooker bee
Choose a solution z_i depending on p_i
Produce new solution v_i
Calculate the fitness
Apply the greedy selection process
If there is abandoned solution then
Replace that solution with a new randomly produced solution by (7) for the scout.
Memorize the best solution achieved yet
Assign cycle to cycle + 1
Until cycle = *MCN*

There are three control parameters in the ABC algorithm: The first parameter is the number of food sources which is equal to the number of employed and also onlooker bees (*SN*), the second one is the value of limit parameter (limit), and the third one is the maximum cycle number (MCN).

RESULTS AND DISCUSSION

In this work, three classification problems from the repos itory of UCI data repository (Frank and Asuncion, 2010) are used to evaluate the performance of Artificial Bee Colony Algorithm. The selected problems are Cancer, Diabetes, Heart medical data sets whose file names are breast-cancer, pima-diabetes, and Cleveland-Heart respectively. From the database, the first examples of each three problem are used in training. The first 75% of the data is used as train set, and remaining 25% data is used as test data. Cancer is the diagnosis of breast cancer (classify a tumor as either benign or malignant). The data set contains 569 patterns; each has 30 inputs and 2 outputs. First 427 of the patterns as training set and the remaining 142 patterns as test set are used. Diabetes is the diagnosis of diabetes (whether an individual is diabetes positive or not). The data set contains 768 patterns; each pattern has 8 inputs and 2 outputs. We used the first 576 patterns as training set and the remaining 192 as test set. Heart is the diagnosis of a heart condition (decides to whether or not at least one of four major vessels is reduced in diameter by more than 50%). The data set contains 303 patterns; each pattern has 35 inputs and 2 outputs. We used the first 227 of them as training set and the remaining 76 as test set.

In order to compare the performance of ABC-FC algorithm Fuzzy C-Means (FCM) algorithm which is widely-used, fuzzy clustering algorithm is used with matlab FCM function. Experiments are repeated 30 times for each case started with a random population. The colony size is chosen 20 and the limit value is set to 300. Training process is stopped when the maximum generation is reached, set to 1000 cycles where the cluster centers are decided.

Statistical results of the algorithms for Cancer, Diabetes, and Heart problems are given in Table 1. We report the mean Classification Error Percentage (CEP) which is the percentage of incorrectly classified patterns, is obtained for the test patterns for each problem. We classified each pattern of test data by assigning it to the class whose center is closest, using the Euclidean distances, to the center of the clusters. This assigned class is compared with the desired class and if they are not exactly the same, the pattern is separated as incorrectly classified. It is calculated for all test data and the total incorrectly classified pattern number is percentages to the size of test data set, which gives us CEP value. In Table 1, the standard deviation (SD) of CEP values 30 runs are presented in parenthesis.

The results demonstrate that ABC-FC method is able to reach a generalization performance in fuzzy clustering. Looking closer to each data set; Cancer appears to be the easiest data set among the three that we tackled. ABC algorithm outperforms FCM clustering, with 2.65% mean CEP value less than 75% test data correctly where ABC is 31.04% and FCM has 3.52% mean classification error. Diabetes is the most difficult problem, both algorithms can classify and FCM is 32.0%. For the Cleveland Heart problem, the performances of the algorithms are 15.26% for ABC and 17.19% for FCM clustering.

As mentioned in the introduction, the results of FCM algorithm do not appear very stable because of not been able to easily escape from the local optimal solution easily. Furthermore, when the standard deviations of the results of each problem are examined, ABC algorithms'

SD values are less than FCM algorithms' SD values. they are 0.47, 1.02 and 1.11 of ABC algorithm and 0.49, 2.56 and 1.20 of FCM algorithm for cancer, diabetes and heart problems, respectively. Therefore, it can be claimed that ABC has good results in terms of classification error and also it is a robust algorithm as indicated by the standard deviations of the results.

Conclusions

In this work, Artificial Bee Colony algorithm which is a recently introduced optimization algorithm is used to fuzzy clustering of medical data which are widely used benchmark problems. The results of ABC algorithm are compared with Fuzzy C-Means (FCM) algorithm and the experiments show that the Artificial Bee Colony algorithm is very successful on optimization of fuzzy clustering. As a future plan of clustering, we are planning to apply ABC algorithm on more problems and compare the performance of the algorithm not only with FCM algorithm but also with other well-known optimization techniques.

ACKNOWLEDGEMENTS

This project is supported as Graduate Research Project with Project ID:FBD-09-1004 by Scientific Research Project Foundation of Erciyes University. The databases used in the experiments are taken from the repository of UCI data repository.

REFERENCES

Bezdek JC (1981). Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York.

Das S, Konar A, Chakraborty UK (2006). Automatic Fuzzy Segmentation of Images with Differential Evolution, In IEEE Congress on Evolutionary Computation, pp. 2026-2033.

Frank A, Asuncion A (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Gan G, Wu J, Yang Z (2009). A genetic fuzzy k-Modes algorithm for clustering categorical data, Expert Syst. Appl., 36: 1615-1620.

Han J, Kamber M (2001). Data Mining: Concepts and Techniques, Academic press.

Jain A, Dubes R (1998). Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ.

Jain AK, Murty MN, Flynn PJ (1999). Data Clustering: A review, ACM Comput. Surveys. 31(3): 264-323.

Karaboga D (2005). An Idea Based On Honey Bee Swarm For Numerical Optimization, Technical Report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department.

Karaboga D (2010). Artificial Bee Colony Algorithm, www.scholarpedia.org/article/Artificial_bee_colony_algorithm, Scholarpedia. 5(3):6915.

Karaboga D, Basturk B (2007). Artificial Bee Colony (ABC) optimization algorithm for solving constrained optimization problems, LNCS: Advances in Soft Computing: Foundations of Fuzzy Logic and Soft Computing, Springer–Verlag, 4529: 789-798.

Karaboga D, Basturk B (2007). A powerful and efficient algorithm for numerical function optimization: Artificial Bee Colony (ABC) algorithm, J. Global Optim, 39(3): 459-171.

Karaboga D, Ozturk C (2009). Neural networks training by Artificial Bee Colony Algorithm on pattern classification, Neural Network World, 19(3): 279-292.

Karaboga D, Ozturk C (2010). A Novel Clustering Approach: Artificial Bee Colony Algorithm, Applied Soft Computing, In Press.

MacQueen J (1967). Some methods for classification and analysis of multivariate observations, 5th Berkeley Symp. Math. Stat. Probability. 281-297.

Sarkar M, Yegnanafayana B, Khemani D (1997). A Clustering Algorithm using an Evolutionary Programming based Approach, Pattern Recognit. Lett., 18: 975-986.

Runkler TA, Katz C (2006). Fuzzy Clustering by Particle Swarm Optimization, IEEE International Conference on Fuzzy Systems, Canada. 601-608.

Zhao B (2007). An Ant Colony Clustering Algorithm, Sixth International Conference on Machine Learning and Cybernetics, Hong. Kong. pp. 3933-3938.