A Hybrid Clustering Approach using Artificial Bee Colony (ABC) and Particle Swarm Optimization

S. Karthikeyan
Research Scholar
Department of Computer Science
Karpagam University
Coimbatore

T. Christopher, PhD
Assistant Professor & Head
Department of Computer Science,
Government Arts College,
Udumalpet

ABSTRACT

In this paper, Cluster analysis is a group objects like observations, events etc based on the information that are found in the data describing the objects or their relations. The main goal of the clustering is that the objects in a group will be similar or related to one other and different from (or unrelated to) the objects in other groups. In this paper, proposed a hybrid model of PSABC algorithm. The PSABC algorithm is a combination of Particle Swarm Algorithm (PSO) and Artificial Bee Colony (ABC) Algorithm used for data clustering on benchmark problems. The PSABC algorithm is compared with other existing classification techniques to evaluate the performance of the proposed approach. Thirteen of typical test data sets from the UCI Machine Learning Repository are used to demonstrate the results of the techniques. The simulation results indicate that PSABC algorithm can efficiently be used for multivariate data clustering.

Keywords

Clustering, Classification, Artificial Bee Colony, Particle Swarm Algorithm.

1. INTRODUCTION

Cluster analysis is a group objects like observations, events etc based on the information that are found in the data describing the objects or their relations. The main goal of the clustering is that the objects in a group will be similar or related to one other and different from (or unrelated to) the objects in other groups. Clustering is a separation of data into groups of related objects. In support of the data by smaller amount clusters essentially loses certain fine details, but it achieves simplification [1].

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 [2, 3]. 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.

Clustering algorithms are generally classified as hierarchical clustering and partitional clustering [4, 5]. Hierarchical clustering group's data objects with a sequence of partitions, either from singleton clusters to a cluster considering all individuals. Hierarchical procedures can be either agglomerative or divisive: agglomerative algorithms begin with each element as a separate cluster and merge them in successively larger clusters; divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters [6,7]. Partitional procedures that we concerned in this paper, attempt to divide the data set into a set of disjoint clusters without the hierarchical structure. The most popular partitional clustering algorithms are the prototype-based clustering algorithms where each cluster is represented by the center of the cluster and the used objective function (a square- error function) is the sum of the distance from the pattern to the center [8].

In this paper, combined form of PSO algorithm and ABC algorithm is used for clustering purpose. The proposed new swarm algorithm is very simple, accurate and very flexible when compared to the existing swarm based algorithms and replaces the content with your own material.

2. RELATED WORKS

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 [9]. 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. 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 [10] 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 algorithm (GA), differential evolution (DE), ant colony optimization (ACO), and particle swarm optimization (PSO) have been successfully applied [11, 12, 13, 14].

Semi-supervised learning methods construct classifiers using both labeled and unlabeled training data samples. Whereas unlabeled data samples help to improve the accuracy of trained models to definite extent, existing methods still face difficulties when labeled data is not sufficient and biased against the underlying data distribution. In [16], clustering based classification (CBC) approach was introduced. By this approach, training data contains both the labeled and unlabeled data, is clustered initially with the guidance of the labeled data. Certain number of unlabeled data samples are then labeled based on the clusters attained. Discriminative classifiers can consequently be trained with the prolonged labeled dataset. The success of this method is justified analytically. Similar issues such as expanding labeled dataset and interacting clustering with classification are presented in [15].

Genetic algorithm is widely used for mining classification rules. If the data set is of three or four years old, the Artificial Bee Colony (ABC) optimization algorithm, which is described by Karaboga based on the foraging behavior of honey bees for numerical optimization problems [15], is applied to classification benchmark problems (13 typical test databases). The performance of the ABC algorithm on clustering is compared with the results of the Particle Swarm Optimization (PSO) algorithm on the same data sets that are presented in [17]. ABC and PSO algorithms drop in the same class of artificial intelligence optimization algorithms, population-based algorithms and they are proposed by inspiration of swarm intelligence. Besides comparing the ABC algorithm and PSO algorithm, the performance of ABC algorithm is also compared with a wide set of classification techniques that are also given in [17].

3. CLUSTERING PROBLEM

Clustering is the process of recognizing natural groupings or clusters in multidimensional data based on some similarity measures. Distance measurement is usually used for evaluating similarities between patterns. In particular the problem is stated as follows: given N objects, assign each object to one of K clusters and minimize the sum of squared Euclidean distances between each object and the center of the cluster belonging to every such allocated object. The clustering problem minimizing Eq. (1) is described as in [20]:

$$J(w, z) = \sum_{i=1}^{N} \sum_{j=1}^{k} w_{ij} \left\| x_i - z_j \right\|^2 \qquad (1)$$

where K is the number of clusters, N the number of patterns, x_i (i = 1, . . . , N) the location of the ith pattern and z_j (j = 1, . . . , K) is the center of the j^{th} cluster, to be found by Eq. (2):

$$z_j = \frac{1}{N_j} \sum_{i=1}^{N} w_{ij}\, x_i \qquad (2)$$

Where N_j is the number of patterns in the jth cluster, w_{ij} the association weight of pattern x_i with cluster j, which will be either 1 or 0 (if pattern i is allocated to cluster j; $w_{ij}$is 1, otherwise 0).

The clustering process, separating the objects into the groups (classes), is realized by unsupervised or supervised learning. In unsupervised clustering which can also be named automatic clustering, the training data does not need to specify the number of classes. However, in supervised clustering the training data does have to specify what to be learned; the number of classes. The data sets that tackled contain the information of classes. Therefore, the optimization goal is to find the centers of the clusters by minimizing the objective function, the sum of distances of the patterns to their centers.

In this paper, the adaptation is carried out by minimizing (optimizing) the sum on all training set instances of Euclidean distance in N-dimensional space between generic instance x_j and the center of the cluster z_j . The cost function for the pattern i is given by Eq. (3), as in [17, 18]:

$$f_i = \frac{1}{D_{Train}} \sum_{j=1}^{D_{Train}} d(x_j, p_i^{CL_{known}(x_j)}) \qquad (3)$$

Where D_{Train} is the number of training patterns which is used to normalize the sum that will range any distance within [0.0, 1.0] and ($p_i^{CL_{known}(x_j)}$) defines the class that instance belongs to according to database.

4. METHODOLOGY

4.1 Particle Swarm Optimization

PSO [19] algorithm is forced by the social behavior of a collection of migrating birds trying to reach your destination that is an unknown destination. In PSO, each solution is a 'bird' in the flock and is known to as a 'particle'. A particle is equivalent to a chromosome (population member) in Genetic Algorithms (GAs) [20]. The PSO does not produce new birds from parent ones. Instead of that the birds in the population only evolve their social behavior and as a result their movement towards a destination [21].

A group of birds communicate together when they fly. Each bird appears in a particular direction, and they communicating collectively and recognize the bird that is in the best location. Consequently, each bird speeds in the direction of the best bird through a velocity that is based on its current position. Each bird, examines the search space from its new local location, and the process repeats until the flock arrives at a preferred destination. It is to be observed that the procedure comprises both social interaction and intelligence so that birds discover from their own experience (local search) and also from the experience of others around them (global search).

The process is initiated with a collection of random particles, N. The ith particle is denoted by its position as a point in S-dimensional space, where S denotes the number of variables. All through the process, each particle i observes three values namely its current position (X_i), the best position it arrived in previous cycles (P_i), its flying velocity (V_i). These three values are denoted as follows:

Current position $X_i = (x_{i1}, x_{i2}, \dots, x_{iS})$

Best previous position $P_i = (p_{i1}, p_{i2}, \dots, p_{iS})$

Flying velocity $V_i = (v_{i1}, v_{i2}, \dots, v_{iS})$

In each time interval (cycle), the position (P_g) of the best particle (g) is computed as the best fitness of all particles. Thus, each particle updates its velocity V_i to get closer to the best particle g, as follows [22]:

$$New\ \ V_i = \omega \times current\ V_i + c_1 \times rand() \times (P_i - X_i) + c_2 \times Rand() \times (P_i - X_i) \qquad (4)$$

As such, using the new velocity V_i, the particle's updated position becomes:

$$New\ position\ X_i = current\ position\ X_i + \ New\ \ V_i\ V_{max} \geq V_i \geq -V_{max} \qquad (5)$$

where c_1 and c_2 represent two positive constants named learning factors (usually $c_1 = c_2 = 2$); rand ( ) and Rand ( ) denotes two random functions in the range [0, 1], V_{max} is an upper limit on the maximum change of particle velocity, and ω denotes an inertia weight employed as an enhancement proposed by Shi and Eberhart [21] to manage the influence of the previous history of velocities on the current velocity. The ω balances the global search and the local search; and it is introduced to minimize linearly with time from a value of 1.4–0.5 [21]. For itself global search is initiates with a large weight and then decreases with time to favor local search over global search.

It is observed that the second term in equation (2) indicates cognition or the private judgment of the particle when comparing its current position to its own best position. The third term in equation (2), denotes the social collaboration among the particles, compares a particle's current position to that of the best particle. Furthermore, in order to control the change of particles velocities, upper and lower bounds for velocity change is limited to a user-specified value of $\mathrm{V_{max}}$. Once the new position of a particle is computed using equation (3), the particle, then, flies towards it [21]. Therefore, the main parameters used in the PSO are the population size (number of birds); number of generation cycles; the maximum change of a particle velocity $\mathrm{V_{max}}$ and ω.

Detailed pseudo-code of PSO algorithm:

```
1)  A population of agents is created randomly.
        $X_i = (P_1, P_2, P_3, \dots\dots.., P_N)$
2)  Evaluate each particle's position according to the objective function. In this case it is the total operational cost given by C for each particle and evaluate their fitness (i.e minimization of the objective function)
3)  Cycle =1
4)  Repeat
5)  Update the velocity of the particles according to the formula ,
        $V_i(t) = V_i(t-1) + C_i r_i\big(pbest(t) - x_i\,(t-1)\big) + C_2 r_2\big(gbest(t) - x_i\,(t-1)\big)$    (6)
    c = acceleration factor. r = random values between 1 and 0
6)  Evaluate the velocity to ascertain if it is the range of $V_{max} \le V_i \le V_{min}$
7)  Move particles to their new position
        $X_i(t) = X_i(t-1) + V_i(t)$    (7)
8)  Evaluate to ensure that limits have not been exceeded.
9)  Compare the particle's fitness evaluation with its previous pbest. If the current value is better than the previous pbest, then set the pbest value equal to the current value and the pbest location equal to the current location in the N dimensional search space.
10) Compare the best current fitness evaluation with the population gbest. If the current value is better than the population gbest, then reset the gbest to the current best position and the fitness value to current fitness value.
11) Check if stopping criterion had been met. If not update the cycle and go back to step (5).
12) End when the stopping criterion, which here is the number of iterations, has been met.
```

4.2 Artificial Bee Colony

Artificial Bee Colony (ABC) algorithm was proposed by Karaboga for optimizing numerical problems in [23]. The algorithm simulates the intelligent foraging behavior of honey bee swarms. It is a very simple, robust and population based stochastic optimization algorithm.

In ABC algorithm, the solution of the optimization problem is represented by the location of a food source and the quality of the solution is represented by the nectar amount of the source (fitness). In the first step of ABC, the locations for the food source are produced randomly. In other words, for SN (the number of employed or onlooker bees) solutions, a randomly distributed initial population is produced. In the solution space, each solution $(\mathrm{X_i} = (\mathrm{X_{i1}}, \mathrm{X_{i2}}, \dots\dots, \mathrm{X_{iSN}}))$ is a vector on the scale of its number of optimization parameters.

Detailed pseudo-code of ABC algorithm

```
1)  Initialize the population of solutions $X_i; i = 1,2, \dots\dots, SN$.
2)  Evaluate the population.
3)  Cycle = 1.
4)  Repeat
5)  Produce new solutions $V_i$for the employed bees by using below for evaluation.
        $V_{ij} = X_{ij} + \varphi_{ij}(X_{ij} - X_{kj})$    (8)
6)  Apply the greedy selection process for the employed bees.
7)  Calculate the probability values of $P_i$ for the solutions of $X_i$ by:
        $P_i = \frac{F(X_j)}{\sum_{j=1}^{SN} F(X_j)}$    (9)
8)  Produce the new solutions of $V_i$ for the onlookers from the solutions of $X_i$ selected depending on $P_i$ and evaluating them.
9)  Apply the selection process for the onlookers.
10) Determine the abandoned solution for the scout, if it exists, and replace it with a new randomly produced solution $X_{\,i}$ by:
        $X_{ij} = X_j^{min} + \left(X_j^{max} - X_j^{min}\right) * r$    (10)
        $j \in \{1,2 \dots D\}$
11) Memorize the best solution achieved so far.
12) Cycle = cycle + 1.
13) Until the cycle = MCN (maximum cycle number)
```

4.3 ABC-PSO Hybrid Algorithm (PSABC)

In this method of hybridization, ABC runs till its stopping criterion, which in this case is the maximum number of iterations, is met. Then the optimal values of individuals generated by the ABC are given to the PSO as its starting point. Ordinarily the PSO randomly generates its first individual sets, but in this case of hybridization that is taken care of by providing the starting point for the Particle Swarm Optimization who is the final values for individuals generated by the Artificial Bee Colony.

Detailed pseudo-code of PSABC algorithm

```
1)  Initialize the PSABC
2)  Generate the initial population $X_i; i = 1,2, \dots\dots, SN$.
3)  Select half part of bees as employed bee with PSO
4)  Evaluate the fitness ($\mathrm{f_{i=}\,P_i}$) of the population
5)  Set cycle to 1
6)  Repeat
```

```
7)  For each employed bee Do
8)  Produce new solution $V_i$
9)  Calculate the value $f_i$
10) Apply greedy selection process
11) Calculate the probability values pi for the solutions $X_i$
12) For each onlooker bee
13) Select a solution $X_i$depending on pi
14) Produce new solution $V_i$
15) Calculate the values $f_i$ Apply greedy selection process
16) If there is an abandoned solution for the scout Then
17) replace it with a new solution which will be randomly produced
18) Memorize the best solution so far
19) cycle = cycle + 1
20) until cycle=MCN
```

In a robust search process, exploration and exploitation processes must be carried out together. In the ABC algorithm, while onlookers and employed bees carry out the exploitation process in the search space, the scouts control the exploration process. The local search performance of ABC algorithm depends on neighborhood search and greedy selection mechanisms performed by employed and onlooker bees. The global search performance of the algorithm depends on random search process performed by scouts and neighbor solution production mechanism performed by employed and onlooker bees.

5. EXPERIMENTATION RESULTS

In this work, 13 classification problems from the UCI database [23] which is a well-known database repository, are used to evaluate the performance of the hybrid method called Artificial Bee Colony algorithm and particle swarm optimization (PSABC). The data sets and their features: the # of patterns, the # of inputs and the # of classes are presented in Table 1. These 3 benchmark problems are chosen exactly the same as in [23], to make a reliable comparison. From the database, the first 75% of data is used in training process as a train set, and the remaining 25% of data is used in testing process as a test set. Although, some data sets' (glass, thyroid, and wine) classes are given in sequential list, they are shuffled to represent every class both in training and in testing as in [23]. The sizes of the train and test sets can be found in Table 1.

5.1 Performance Evaluation of the Clustering Methods

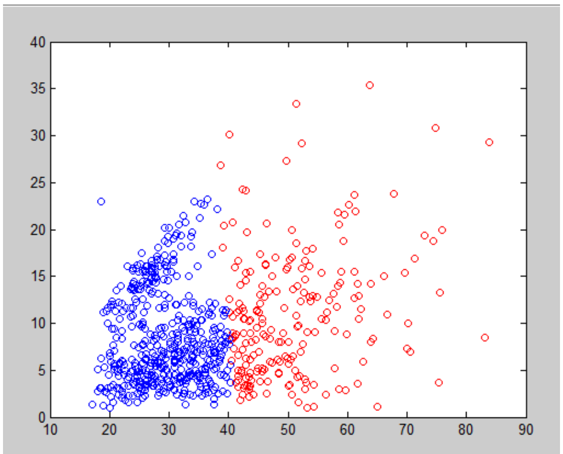

Fig. 1: Cluster formation using ABC Technique

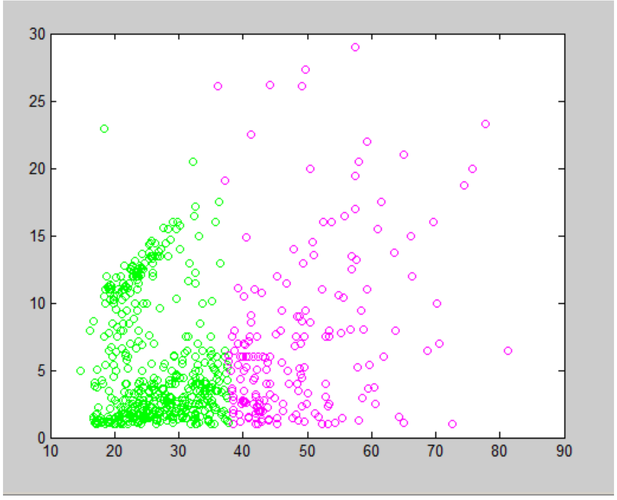

Fig. 2: Cluster formation using PSO Technique

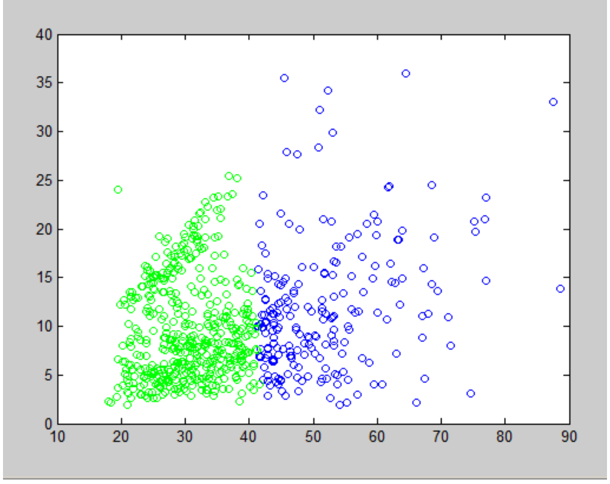

Fig. 3: Cluster formation using PSABC Technique

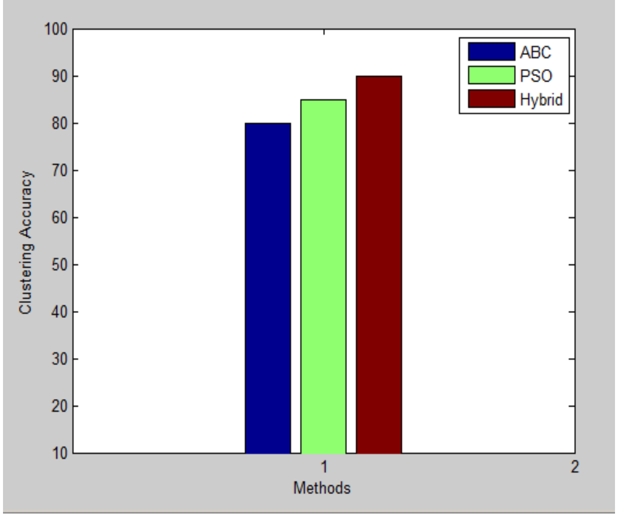

Fig. 4: Comparison of Clustering Accuracy of the various techniques

5.2 Test problems

The problems considered in this work can be described briefly as follows. Balance data set was generated to model psychological experimental results. Each example is classified as having the balance scale tip to the right, tip to the left, or be balanced. The data set includes 4 inputs, 3 classes and there are 625 examples which are split into 469 for training and 156 for testing. Cancer data sets are based on the "breast cancer Wisconsin - Diagnostic" and "breast cancer Wisconsin - Original" data sets, respectively. They are diagnosis of breast cancer, with 2 outputs (classify a tumor as either benign or malignant). The former one contains 569 patterns, 30 inputs and the latter one contains 699 patterns, 9 inputs.

The diabetes data set, a two class problem which is the diagnosis of diabetes (whether an individual is diabetes positive or not), has 768 patterns. We used the first 576 patterns as training set and the remaining 192 as test set. There are 8 inputs for each pattern. Wine data which was obtained from a chemical analysis of wines were derived from three different cultivators. Therefore, the data analysis determines the three types of wines. There are 178 instances of wine samples with 13 inputs.

Table 1. Properties of the problems

Dataset	Data	Train	Test	Input	Class
Balance	625	469	156	4	3
Cancer	569	427	142	30	2
Diabetes	768	576	192	8	2
wine	178	133	45	13	3

5.3 Results and discussion

For each problem, we report the Classification Error Percentage (CEP) which is the percentage of incorrectly classified patterns of the test data sets. We classified each pattern by assigning it to the class whose center is closest, using the Euclidean distances, to the center of the clusters. This assigned output (class) is compared with the desired output 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 percentage to the size of test data set, which is given by Eq. (11).

$$CEP = 100 \times \frac{\#\ of\ the\ misclassified\ samples}{size\ of\ the\ test\ data\ set} \qquad (11)$$

As described above, the data is given in two pieces: the training set (the first 75%) and the test set (the last 25%). The results of the algorithms Proposed PSABC, ABC and PSO for the problems are given in Table 2 where classification error percentages (CEP values) are presented.

Table 2. Average classification error percentages for proposed PSABC and other existing techniques

Dataset	ABC	PSO	Proposed PSABC
Balance	15.38	25.47	13.12
Cancer	2.81	5.80	1.98
Diabetes	22.39	22.50	21.19
wine	0.00	2.22	0.00

In Table 2, the classification error percentages of PSABC algorithm and other algorithm that are presented. From the above table, the proposed PSABC outperforms the other existing techniques.

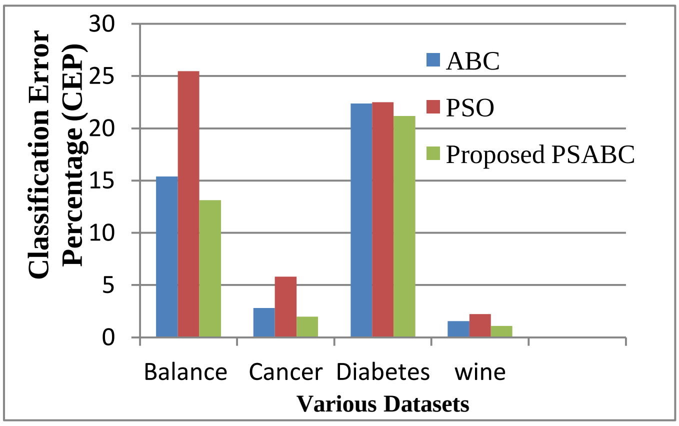

Fig.5. Comparison of Average classification error percentage

6. CONCLUSION

In this work, a hybrid algorithm of combination of Particle Swarm Algorithm (PSO) and Artificial Bee Colony (ABC) Algorithm, simple and robust optimization technique is used in clustering of the benchmark classification problems for classification purpose. Clustering is an important classification technique that gathers data into classes (or clusters) such that the data in each cluster shares a high degree of similarity while being very dissimilar from data of other clusters. The performance of the PSABC algorithm is compared with Particle Swarm Optimization algorithm and other nine techniques which are widely used by the researchers. The results of the experiments show that the PSABC algorithm can successfully be applied to clustering for the purpose of classification. There are several issues remaining as the scopes for future studies such as using different algorithms in clustering and comparing the results of PSABC algorithm to the result of those algorithms.

7. REFERENCES

[1] Han, J. and Kamber, M. 2001. Data Mining: Concepts and Techniques, Academic Press.

[2] Jain, A. and Dubes, R. 1998 .Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ.

[3] Sarkar, M., Yegnanafayana, B. and Khemani, D. 1997. A Clustering Algorithm using an Evolutionary Programming based Approach, Pattern Recognit. Lett., Vol.18 ,pp.975-986.

[4] Frigui, H. and Krishnapuram, R. 1999. A robust competitive clustering algorithm with applications in computer vision, IEEE Trans. Pattern Anal. Mach. Intell. 21, pp.450–465.

[5] Leung, Y., Zhang, J. and Xu, Z. 2000.Clustering by scale-space filtering, IEEE Trans. Pattern Anal. Mach. Intell. 22,pp.1396–1410.

[6] Jain, A.K., Murty, M.N. and Flynn, P.J. 1999. Data clustering: a review, ACM Comput. Surveys ,Vol.31 ,No.3 ,pp.264–323.

[7] Data Mining and Knowledge Discovery Handbook, Springer, New York, pp. 321–352,2005.

[8] Mirkin, B. 1996. Mathematical Classification and Clustering, Kluwer Academic Publishers, Dordrecht, The Netherlands.

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

[10] Bezdek, JC. 1981. Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York.

[11] Gan, G., Wu, J. and Yang, Z. 2009. A genetic fuzzy k-Modes algorithm for clustering categorical data, Expert Syst. Appl., Vol.36, pp.1615-1620.

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

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

[14] Runkler, TA. and Katz, C. 2006. Fuzzy Clustering by Particle Swarm Optimization, IEEE International Conference on Fuzzy Systems", Canada. 601-608.

[15] Hua-Jun Zeng, Xuan-Hui Wang, Zheng Chen and Wei-Ying Ma. 2003. CBC: Clustering Based Text Classification Requiring Minimal Labeled Data, IEEE International Conference on Data Mining - ICDM , pp. 443-450.

[16] Karaboga, D. 2005. An idea based on honey bee swarm for numerical optimization, Technical Report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department.

[17] De Falco, I., Della Cioppa, A. and Tarantino, E. 2007. Facing classification problems with Particle Swarm Optimization, Appl. Soft Comput,Vol.7 No.3 ,pp. 652–658.

[18] Marinakis, Y., Marinaki, M., Doumpos, M., Matsatsinis, N. and Zopounidis, C. 2008. A hybrid stochastic genetic—GRASP algorithm for clustering analysis, Oper. Res. Int. J.(ORIJ) ,Vol.8 ,No.1pp. 33–46.

[19] Kennedy, J and Eberhart, R. 1995. Particle swarm optimization, Proceedings of the IEEE international conference on neural networks (Perth, Australia), pp. 1942–1948. Piscataway, NJ: IEEE Service Center.

[20] Al-Tabtabai, H. and Alex, PA. 1999. Using Genetic Algorithms to Solve Optimization Problems in Construction, Eng Constr Archit Manage,Vol. 6,No.2,pp.121–32.

[21] Shi, Y. and Eberhar, R. 1998. A modified particle swarm optimizer, Proceedings of the IEEE international conference on evolutionary computation. Piscataway, NJ: IEEE Press. pp. 69–73.

[22] Chen, H., Jin, H., Sun, J., Liao, X. and Deng,D 2003. A new proxy caching scheme for parallel video servers", Computer Networks and Mobile Computing, pp.438–441.

[23] Karaboga, D. and 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, Vol. 4529, Springer–Verlag pp. 789–798.