Cooperative Swarm based Evolutionary Approach to find optimal cluster centroids in Cluster Analysis

Bighnaraj Naik[1] **, Sarita Mahapatra**[1] **, Subhra Swetanisha**[2] **, Swadhin Kumar Barisal**[1]

[1]**Dept of Information Technology,**
Institute of Technical Education & Research,
Siksha 'O' Anusandhan University,
Bhubanewar, Odisha, India

[2]**Dept of Computer Science and Engineering**
Trident Academy of Technology
Bhubaneswar, Odisha, India

Abstract

Centroid-based clustering is a NP-hard optimization problem, and thus the common approach is to search for cluster centers only for approximate solutions. Well-known centroid-based clustering methods are k-means, k-medoids and fuzzy c-means. In this paper we proposed swarm intelligence based nature-inspired center-based clustering method using PSO optimization. PSO searches the optimized solution from available solutions in multidimensional search space. So PSO is capable to search best cluster with maximum fitness using social-only model and cognition-only model, such that the square distances from the cluster are minimized. In this article, it is shown that how PSO based clustering can be used to find N number of cluster specified by the user in a dataset. Our suggested method has been tested with artificial dataset and several real multidimensional dataset from UCI repository. Effectiveness of the method is demonstrated by comparing fitness of proposed method with effectiveness of K-means and Fuzzy c-means technique. Results shows that, this method is quite simple, effective and has much potential to search best cluster centers in multidimensional search space.

Keywords : *Centroid-based clustering; Cluster Analysis; Swarm intelligence; Particle swarm optimization; Fuzzy C-means clustering; K-Means clustering; Euclidean distance;*

1. Introduction

1.1 Swarm intelligence:

Swarm Intelligence (SI) [12][13] is an artificial intelligence technique inspired by nature, based around on the study of collective behavior in centralized, self-organized systems. SI was introduced by Beni & Wang in 1989, in the context of cellular robotic system. A swarm has been defined as a set of agents which are liable to communicate directly or indirectly with each other, and which collectively carry out to solve an optimization problem. Swarm Intelligence is defined as property of the system whereby the collective behaviors of agents [12] (swarm) interacting locally with their environment causes coherent functional global patterns.SI provides a basis with which it is possible to explore collective problem solving without centralized control or the provision of a global model. Example of systems like this can be found in nature, including ant colonies, bird flocking, bee swarming, animal herding, bacteria molding and fish schooling.Two of the successful swarm intelligence techniques currently in existence are Ant Colony Optimization (ACO) [18] and Particle Swarm Optimization (PSO)[12].

1.1.1 Ant colony optimization:

Ant Colony Optimization [18] is a class optimization algorithm modeled on the actions of an Ant Colony, proposed by Marco Dorigo in 1992. The main idea behind this is loosely inspired by behavior of real ants, is that of parallel search over several constructive computational threads based on local problem data and containing information from previously obtained result. The collective behavior and the interaction of different threads have used effectively to solve optimization problems.

1.1.2 Particle swarm optimization:

PSO is originally attributed to Kennedy, Eberhart and Shi[12] and was first intended for simulating social behavior, as a stylized representation of the movement of organisms in a bird flock or school. Particle swarm optimization (PSO) is a computational method that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. PSO optimizes a problem by having a population of candidate solution and moving these particles around in the search-space according to simple mathematical formulae over the particle's position and velocity. Each particle's movement is influenced by its local best known position and is also guided toward the best known positions in the search-space, which are updated as better positions are found by other particles. This is expected to move the swarm toward the best solutions. PSO is a meta-heuristic approach as it makes few or no assumptions about the problem being optimized and can search very large spaces of candidate solutions. Beside this, PSO do not guarantee an optimal solution always.

1.2 Centroid-based clustering:

Center based clustering is more efficient for clustering large databases and high dimensional databases .Center based clustering is more efficient for clustering with distance function instead of similarity function, so that the more similar two items are when shorter their distance is. Each data item is placed in the cluster whose corresponding center it is closer to. Center is the representative of cluster. Center in a cluster travels a little distance as possible to reach the center of cluster. It means that each cluster is tightly and closely associated as possible around the corresponding center. The most well known and commonly used centroid-based methods are k-means, k-medoids, fuzzy c-mean and their variations.

1.2.1 k-means clustering

K-means clustering generates a specific number (n) of disjoint clusters. The K-Means method is a numerical, unsupervised, non-deterministic and iterative method. K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume n clusters). The main idea is to define n centroids, one for each cluster. At first centroids should be placed randomly. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed. At this point we need to re-calculate n number of new centroids from the previous step. After we have these n number of new centroids, a new binding has to be done between the same data set points and the nearest new centroids. A loop has been generated. As a result of this loop we may notice that, a centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

1.2.2 Fuzzy c-means clustering

Fuzzy c-means (FCM) is a data clustering technique in which a dataset is grouped into n number of clusters with every data point in the dataset belonging to every cluster to a certain degree. For example, a certain data point that lies close to the center of a cluster will have a high degree of belonging or membership to that cluster and another data point that lies far away from the center of a cluster will have a low degree of belongingness [1][19][20] or membership to that cluster. FCM starts with an initial guess for the cluster centers, which are intended to mark the mean location of each cluster. The initial guess for these cluster centers is most likely to be incorrect. Next, FCM assigns every data point a membership grade for each cluster. By iteratively updating the cluster centers and the membership grades for each data point, FCM iteratively moves the cluster centers to the right location within a data set. This iteration is based on minimizing an objective function that represents the distance from any given data point to a cluster center weighted by that data point's membership grade.

2. Particle Swarm Optimization

Particle swarm optimization (PSO)[12] is a stochastic based search algorithm widely used to find the optimum solution introduced by Kennedy and Eberthart[1] in 1995. PSO is a effective optimization technique to search for global optimized solution[17][9] but time of convergence[14] is uncertain. Like other population based optimization[13][16] methods the particle swarm optimization starts with randomly initialized population[16] for individuals.PSO works on the social behavior[12] of particle. It finds the global best solution by adjusting each individual's positions[12] with respect to global best position of particle of the entire population. Each individual is adjusting by altering the velocity[12] according to its own experience and by observing the experience of the particles in search space. According to the used fitness function, local best (lbest) and global best (gbest) will be calculated. The positions and velocities of the particles initially in search space are denoted by V and X respectively. Then the new velocities and positions of the particles for next iterations [15] can be evaluated by using the equations 1 and 2.

$$V_{id}(t+1)= \underbrace{V_{id}(t) + c_1* rand() * (lbest_{id} - X_{id})}_{\text{Social influence the particles}} + \underbrace{c_2* rand()*(gbest_{id} - X_{id})}_{\text{Cognition influence of particles}} \qquad (1)$$

$$X_{id}(t+1)= X_{id}(t) + V_{id}(t+1) \qquad (2)$$

Where C1 and C2 are the constants and rand() is random function which generates random number in between 0 and 1. In above equation ‘i’ is the instance number,‘d’ is the dimensions of instances and ‘t’ is the iteration number. ‘gbest’ is the particle in the neighborhood with the best fitness and ‘lbest’ is the position for a particle's best fitness yet encountered. Equation-1 is responsible for social influence of the particles and cognition model [12] of particles in the search space. Basis concept of PSO can be used for cluster analysis [2][5][3] and classification[6].

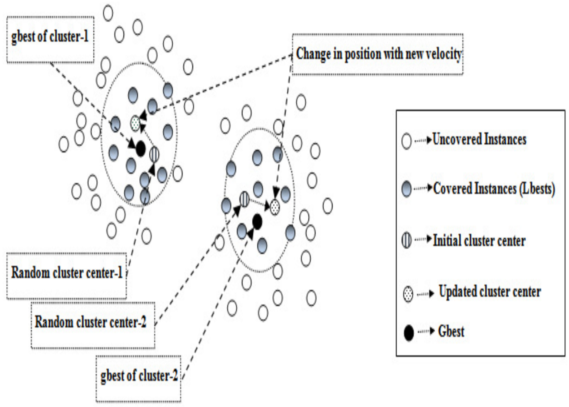
3. Cluster Analysis Using PSO

Cluster analysis is a collection of statistical methods, which identifies groups of objects (instances) that have similar characteristics. Cluster analysis (or clustering) is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar to each other than to those in other clusters. In general, it is also called look-a-like groups. The simplest mechanism is to partition the objects using measurements that capture similarity or belongingness or distance between objects. In this way, clusters and groups are interchangeable words. Often in market research studies, cluster analysis is also referred to as a segmentation method. In neural network concepts, clustering method is called unsupervised learning. Typically in clustering methods, all the objects with in a cluster is considered to be equally belonging to the cluster. Clustering can be achieved by various algorithms [7][8] that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with low distances among the cluster members, dense areas of the data space [4], intervals or particular statistical distributions [10]. The appropriate clustering algorithm and parameter settings depend on the individual data set and intended use of the results [11]. Cluster analysis is not an automatic task, but an iterative process of knowledge discovery that involves both trial and failure.In this paper, we have proposed a cluster analysis model based on most popular nature-inspired swarm intelligent-based PSO technique. It has following steps-

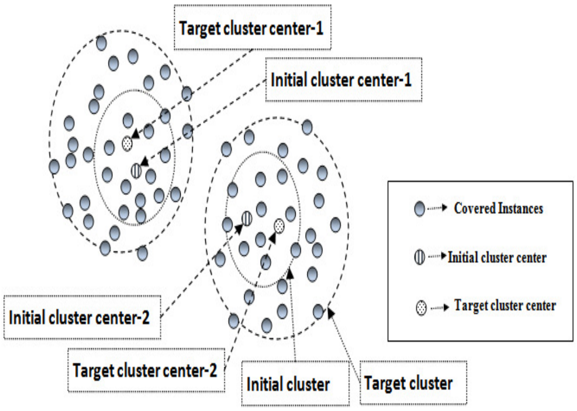
Algorithm psoBasedClustering (X, n)

```
1. load dataset X and set number of clutser ‘n’ to be found.
2. Set initial random clutser center vector <$C_1$,$C_2$,….$C_n$>.
3. Set random velocity V=<$V_1$,$V_2$,….$V_n$>, where each
     V1=<v1,v2,….vk>. here n is the number of cluster and k is
     dimension of dataset.$V_1$,$V_2$,$V_3$ ,…$V_n$ are initial random
     velocity vector for $C_1$, $C_2$, $C_3$,…$C_n$ respectively.
4. Compute Euclidian distance from all
     clusters<$C_1$,$C_2$,….$C_n$> to all the instances of X.
5. Create clusters based on Euclidian distances computed at
     step-4.
6. Calculate fitness of all instances ($F_{xi}$) of clusters by using
     the equation-3 and generate lbest.
7. The instance having highest fitness in each cluster is
     chosen as gbest of that cluster. Generate n number of
     gbest, where ‘n’ is the number of cluster.
8. Compute new velocity $V_{NEW}$ out of initial velocity, lbest and
     gbest by use of equation-1.
9. Update the position of all cluster centers (centroid) with
     new velocity $V_{NEW}$ and generate $C_{NEW}$ by using equation-2.
10. if(Euclidian distance(C,$C_{NEW}$) <= s)
11.          goto step-4
12. else display final clusters
13.          goto step-14
14. Compute the performance of PSO ($F_{CT}$) using equation-2.
15. stop
```

In this algorithm, X is the dataset, C=<C_1,C_2,….C_n> is the cluster centers vector, C_i is the ith cluster center, n is the expected total number of clusters in X, V=<V_1,V_2,….V_n> is a vector of random velocities. V_i is the velocity vector of C_i.V_{new} and C_{new} is new velocity and next cluster center position respectively.



(Fig-1: Generation of initial cluster centers, bests and gbest)



(Fig-2: Formation of target cluster)

The proposed algorithm works upon the dataset to compute the cluster, where number of clusters (n>1) is to be calculated is provided by the user. Initial cluster centers will be selected randomly. Fig-1 demonstrates the selection of two cluster centers. Based on randomly selected cluster centers, initial cluster-1 and cluster-2 is generated by calculating Euclidian distances. Fitness of all instances of generated clusters has been calculated as it is used as lbest. Two vector lbest-1 and lbest-2 is generated from computed fitness. Instance with best fitness of cluster-1 and cluster-2 are selected as gbest of cluster-1 and cluster-2 respectively. Next velocity vectors have been computed by using initial velocity, lbest and gbest. By the use of new velocity, next positions of cluster centers are generated. These steps will be repeatedly executed until and unless the target clusters (fig-2) are found.

The complete method can be visualized with the help of flowchart (fig-3). The positions and velocities of the particles initially in search space denoted by V and X. Then the new velocities and positions of the particles for next iterations [5] can be evaluated by using the equations 1 and 2.

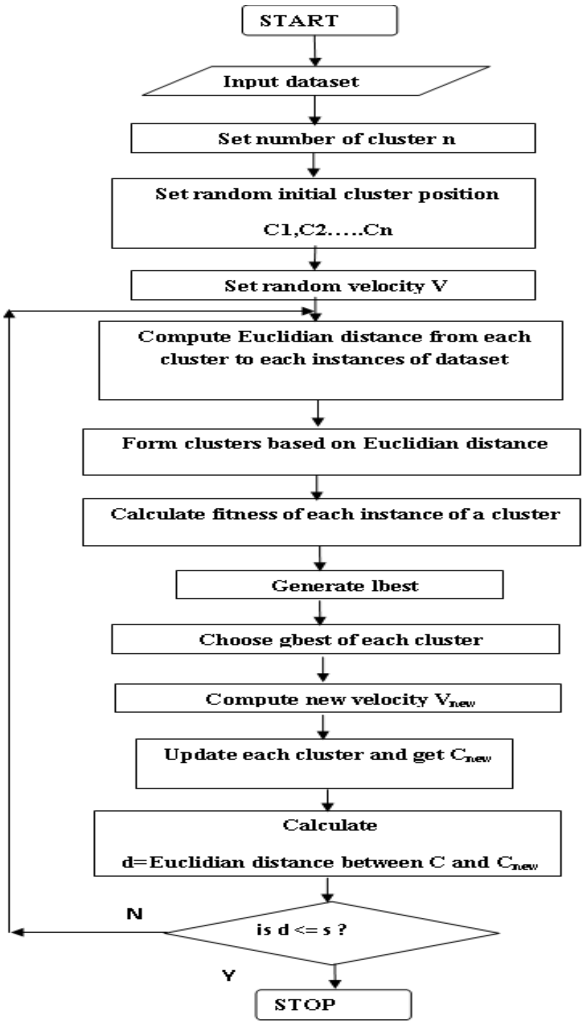
$$F_{X_i} = \frac{1}{\sum_{k=1}^{N} |X_i - X_k|^2} \qquad (3)$$

$$F_{CT} = \frac{k}{\left({}^{1}\!/_{\sum_{i=1}^{N} \sum_{j=1}^{n} |C_j - X_i|^2} \right) + d} \qquad (4)$$

$$Fc = \frac{1}{\sum_{i=1}^{N} \sum_{j=1}^{n} |C_j - X_i|^2} \qquad (5)$$

F_{Xi} represents fitness of an instance, where X is the dataset, N is the number of instances in X, X_i is the ith instance of X.F_C represents fitness of cluster center vector, where X is the dataset, N is the number of instances in X, X_i is the ith instance of X.F_{CT} represents fitness of particular clustering method of technique, where X is the dataset used, N is the number of instances in X, X_i is the ith instance of X, k is a positive constant and d is a small-valued constant.

The above flowchart (Fig.3) describes the working principle of proposed PSO based clustering. Most of time the PSO algorithm stops in two conditions, 1st – if the velocity exceeds the given maximum range and 2nd – if it reached the specified maximum number of iterations. Our proposed model will stop in neither of these conditions. It will stop when difference between old cluster center and new cluster center is less than or equal to s. Here s is small valued constant. Value of s depends upon dataset being used. Values of s has been chosen for different datasets and listed at table-4.


START
Input dataset
Set number of cluster n
Set random initial cluster position
C1,C2.....Cn
Set random velocity V
Compute Euclidian distance from each cluster to each instances of dataset
Form clusters based on Euclidian distance
Calculate fitness of each instance of a cluster
Generate lbest
Choose gbest of each cluster
Compute new velocity V_{new}
Update each cluster and get C_{new}
Calculate
d=Euclidian distance between C and C_{new}
N
is d <= s ?
Y
STOP

(Fig-3: Flowchart for cluster analysis using PSO)

4. Simulation Result

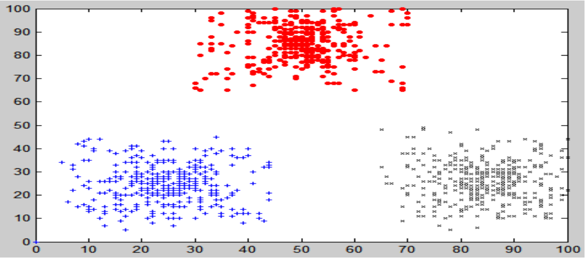
The above clustering method has been implemented on a system with MATLAB with the configuration specified in table-1. It has been tested with one artificial dataset and ten real multidimensional datasets from UCI repository (iris, lense, haberman, balance scale, wisconsin breast cancer, contraceptive method choice, hayes-roth, robot navigation, spect heart and wine). Better configurations are required to run the program faster.

Table-1: System configuration

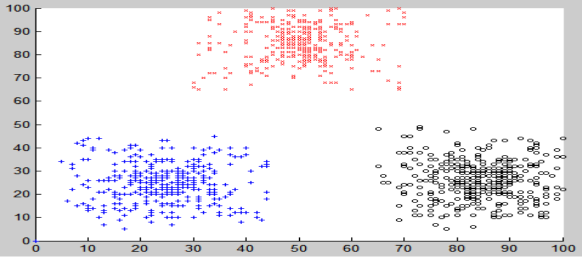
Platform	MATLAB-10
Operation System	Window-7 (64-bit)
Processor	Intel(R) Core(TM) i3 CPU M380 @ 2.53 GHz
RAM	3.00 GB

The proposed clustering method using PSO has been tested with one artificial dataset and ten real multidimensional datasets (iris, lense, haberman, balance scale, wisconsin breast cancer, contraceptive method choice, hayes-roth, robot navigation, spect heart and wine). Effectiveness of proposed model is tested with various dataset having multiple clusters with multiple dimensions.

After clustering, the cluster centers of different datasets are listed at table-2 and table-3.The results shows that , a little change in cluster center vector have significant effect of total fitness of clustering. Performance of proposed PSO based clustering is compared with K-Mean, Fuzzy C-Mean and simulation result has been demonstrated in table-4.Clusters formed after applying PSO based clustering and K-Means clustering on artificial 2d dataset is shown on fig-4 and fig-5 respectively. This cluster contains 600 data points on 2d space. Fitness of 10 number of run of PSO based clustering program on artificial 2d data is displayed on table-5. Deviation of fitness of each clustering technique on different run can be determined from these data. The best fitness of each clustering technique on artificial 2d dataset is highlighted on the table-5. PSO clustering and K-means has been applied to robot navigation dataset having 5456 number of instances and results are displayed on fig-6 and fig-7 respectively. Performance of PSO clustering and K-mean clustering on haber man dataset is demonstrated on fig-8 and fig-9 respectively.



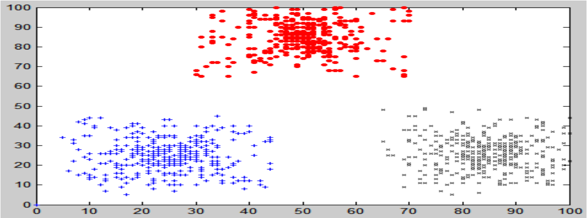
(Fig-4: Cluster generation using PSOC on 2d artificial dataset having 600 instances)



(Fig-5: Cluster generation using K-mean on 2d artificial dataset having 600 instances)

Table-2: cluster centers of datasets (artificial 2d dataset, iris, lense, haberman, balance scale, Wisconsin breast cancer, and contraceptive method choice dataset) generated from simulation

Datasets	Dimension	K-Mean	Fuzzy C-Mean	PSOC
Artificial data (3 class)	**2**	(50.06, 84.37)	(50.21, 84.64)	(49.22, 84.21)
		(83.56, 25.69)	(83.68, 25.47)	(84.86, 26.17)
		(24.80, 24.31)	(25.06, 25.93)	(24.92, 24.92)
Iris (3 cluster)	**4**	(6.31, 2.89, 4.97, 1.70)	(6.77, 3.05, 5.64, 2.05)	(6.83, 3.07,5.71, 2.14)
		(5.20, 3.63, 1.47, 0.27)	(5.88, 2.76, 4.36, 1.39)	(5.87, 2.81, 4.28, 1.39)
		(4.73, 2.93, 1.76, 0.33)	(5.01, 3.40, 1.48, 0.25)	(5.07, 3.4, 1.58, 0.26)
Lense (3 class)	**4**	(2.50, 1.67, 1.33, 1.50)	(2.78, 1.49, 1.49, 1.49)	(2.29, 1.52, 1.74, 0.96)
		(2.50, 1, 2, 1.50)	(1.99, 1.50, 1.50, 1.50)	(2.92, 1.29, 1.42, 2.06)
		(1, 1.50, 1.50, 1.50)	(1.22, 1.49, 1.49, 1.49)	(1.20, 1.57, 1.17, 1.69)
Haber man (2 class)	**3**	(44.54, 62.60, 4.41)	(44.03 ,62.65, 3.67)	(43.15, 63.70, 3.18)
		(62.35, 63.16, 3.54)	(61.80, 63.07, 3.10)	(61.08, 62.22, 4.83)
Balance scale (3 class)	**4**	(3.52, 3, 3.20, 1.52)	(3.01, 2.98, 2.99, 3.01)	(2, 3, 4.4229, 3)
		(4.12, 3, 3, 4.12)	(3.08, 3.03, 3.04, 3.05)	(4.30, 3, 3, 3)
		(1.53, 3, 2.82, 3.31)	(2.90, 2.98, 2.96, 2.93)	(1.89, 3, 1.89, 3)
Wisconsin breast cancer (2 class)	**10**	(616261.11, 4.45, 3.22 ,3.38, 3.23, 3.31, 4.16, 3.63, 3.04, 1.70)	(642100.46, 4.40, 3.06, 3.24, 3.16, 3.24, 4.04, 3.50, 2.92, 1.65)	(616250.96, 4.71, 3.11, 3.67, 3.12, 3.64, 4.56 ,3.80, 4.94, 1.36)
		(1241496.84, 4.43, 3.12, 3.15, 2.68, 3.20, 3.32, 3.37, 2.80, 1.56)	(1216027.38, 4.45, 3.15, 3.18, 2.70, 3.22, 3.35, 3.40, 2.84, 1.57)	(1241526.73, 7.11, 5.46, 5.48, 2.35, 5.51, 6.53, 3.67, 5.30, 5.62)
Contraceptive Method Choice (3 class)	**9**	(43.85, 2.83, 3.32, 4.86,0.81, 0.76, 1.88, 3.33, 0.11)	(44.01, 2.85, 3.35, 4.82, 0.81, 0.76, 1.88, 3.34, 0.11)	(43.22, 2.96, 3.09, 4.12, 0.94, 0.93, 1.95, 3.85, 0.03)
		(33.50, 3.03, 3.47, 3.71,0.80, 0.69, 2.14, 3.22, 0.07)	(33.55, 3.08, 3.51, 3.63, 0.78, 0.69, 2.09, 3.26, 0.067)	(33.17, 3.03, 3.13, 3.99,0.94, 0.17, 2.01, 3.03, 0.01)
		(24.17, 2.97, 3.45, 1.77, 0.91, 0.79, 2.30, 2.91, 0.04)	(24.03, 2.98, 3.46, 1.76, 0.92, 0.79, 2.31, 2.91, 0.04)	(24.11, 2.95, 3.13, 1.96, 0.97, 0.96, 2.84, 2.97, 0.015)



(Fig-4: Cluster generation using PSOC on 2d artificial dataset having 600 instances)

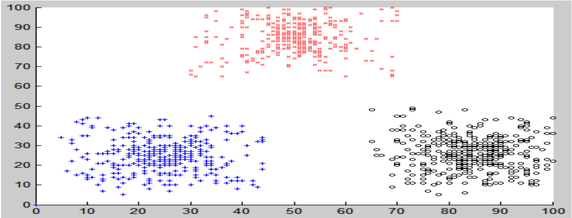
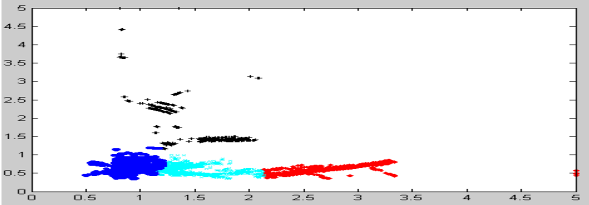
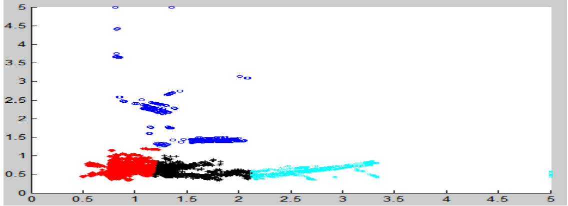


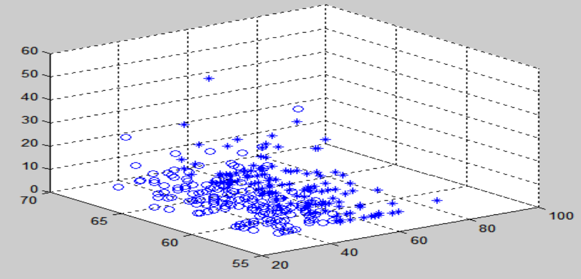
Fig-5: Cluster generation using K-mean on 2d artificial dataset having 600 instances)



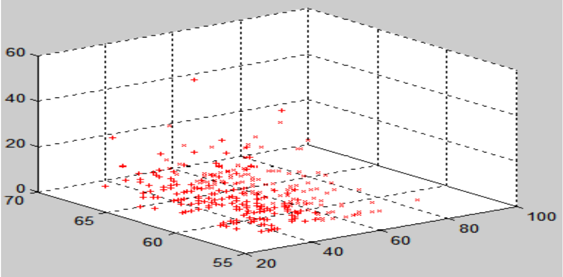
(Fig-6: Cluster generation using PSOC on robot navigation dataset (2d) having 5456 instances)



(Fig-7: Cluster generation using k-mean on robot navigation dataset (2d) having 5456 instances)



(Fig-8: Cluster generation using PSOC on haber man dataset (3d) having 306 instances)



(Fig-9: Cluster generation using K-mean on haber man dataset (3d) having 306 instances)

Table-3: cluster centers of datasets (hayes-roth, robot navigation, spect heart and wine dataset) generated from simulation

Datasets	Dimension	K-Mean	Fuzzy C-Mean	PSOC
Hayes-roth (3 class)	5	(22.50, 1.88, 2.04, 2.04, 2.15)	(20.77, 1.89, 2.10, 2.06, 2.10)	(21.94, 2.23, 2.09, 1.57, 2.34)
		(66.50, 2.11, 2.04, 1.79, 1.77)	(66.50, 2.15, 2.02, 1.83, 1.80)	(62.85, 1.31, 1.37, 1.31, 2.20)
		(110.50, 2, 1.77, 2.02, 1.93)	(112.24, 1.99, 1.77, 2.03, 1.97)	(107.04, 1.95, 1.67, 1.35, 2.82)
Robot navigation (4 class)	2	(0.86837 , 0.64510)	(0.84114, 0.65298)	(0.85425, 0.64654)
		(1.50637, 1.83635)	(1.81802, 0.60456)	(1.45730, 1.79110)
		(1.49497, 0.58152)	(1.33356, 0.65478)	(1.45431, 0.58266)
		(2.74827, 0.59520)	(2.83389, 0.65440)	(2.71498, 0.58820)
Spect heart (2class)	22	(0.301, 0.245, 0.0566, 0.075, 0.075, 0.226, 0.094, 0.113, 0.132, 0.056, 0.169, 0.132, 0.094, 0.132, 0.037 , 0.0377, 0.037, 0.018, 0, 0.132, 0.169, 0.037)	(0.312, 0.198, 0.085, 0.133, 0.120, 0.171, 0.080, 0.139, 0.145, 0.096, 0.142, 0.116, 0.125, 0.189, 0.096, 0.0394, 0.079, 0.044, 0.031, 0.102, 0.145, 0.110)	(0.362, 0.241, 0.080, 0.120, 0.120, 0.214, 0.094, 0.174, 0.161, 0.094, 0.188, 0.134, 0.147, 0.214, 0.094, 0.040, 0.053, 0.026, 0.013, 0.120, 0.134, 0.120)
		(0.888, 0.592, 0.370, 0.629, 0.481, 0.444, 0.185, 0.555, 0.555, 0.444, 0.518, 0.333, 0.555, 0.777, 0.518, 0.148, 0.444, 0.259, 0.222, 0.296, 0.333, 0.629)	(0.720, 0.562, 0.260, 0.434, 0.326, 0.452, 0.173, 0.411, 0.436, 0.302, 0.470, 0.293, 0.405, 0.552, 0.336, 0.119, 0.303, 0.176, 0.134, 0.288, 0.319, 0.404)	(0.936, 0.810, 0.556, 0.873, 0.506, 0.620, 0.189, 0.620, 0.746, 0.620, 0.683, 0.379, 0.683, 0.873, 0.683, 0.189, 0.746, 0.379, 0.316, 0.379, 0.620, 0.746)
Wine (3class)	13	(12.929, 2.504, 2.408,19.890, 103.596, 2.111, 1.584, 0.388, 1.503, 5.650, 0.883, 2.365, 728.338)	(12.991, 2.563, 2.390, 19.635, 104.027, 2.140, 1.635, 0.387, 1.529, 5.646, 0.891, 2.408, 742.707)	(13.036, 3.696, 2.368, 20.826, 98.893, 1.946, 1.160, 0.486, 1.524, 6.648, 0.757, 1.972, 726.856)
		(12.516, 2.494, 2.288,20.823, 92.347, 2.070, 1.758, 0.390, 1.451, 4.086, 0.941, 2.490 , 458.231)	(12.515, 2.425, 2.295, 20.777, 92.423, 2.075, 1.788, 0.387, 1.453, 4.135, 0.945, 2.490, 459.580)	(12.598, 3.104, 2.338, 21.789, 96.188, 2.407, 2.109, 0.407, 1.672, 3.413, 1.051, 2.774, 460.348)
		(13.804, 1.883, 2.426,17.023, 105.510, 2.867, 3.014, 0.285, 1.910, 5.702, 1.078, 3.114, 1195.148)	(13.803, 1.867, 2.456, 16.966, 105.354, 2.866, 3.026, 0.291, 1.921, 5.825, 1.080, 3.071, 1221.035)	(13.811, 1.824, 2.423, 15.681, 107.947, 3.316, 3.336, 0.300, 1.892, 6.300, 1.047, 3.178, 1192.863)

Clusters formed after applying PSO based clustering and K-Means clustering on artificial 2d dataset is shown on fig-4 and fig-5 respectively. This cluster contains 600 data points on 2d space. Fitness of 10 number of run of PSO based clustering program on artificial 2d data is displayed on table-5. Deviation of fitness of each clustering technique on different run can be determined from these data. The best fitness of each clustering technique on artificial 2d dataset is highlighted on the table-5. PSO clustering and K-means has been applied to robot navigation dataset having 5456 number of instances and results are displayed on fig-6 and fig-7 respectively. Performance of PSO clustering and K-mean clustering on haber man dataset is demonstrated on fig-8 and fig-9 respectively.

Table-4: Effectiveness (fitness) of KMean , FCM and PSOC

S.No.	Dataset	No. of instances	No. of classes	Dim	PSOC	KMean Clustering	FCM Clustering
1	art-2d-data	600	3	2	4.94309E-06 (s = 0.02)	4.94137E-06	4.91855E-06
2	Iris	150	3	4	0.014432895 (s = 1)	0.012395396	0.012738542
3	Lenses	24	3	4	0.354960239 (s = 1.5)	0.339904827	0.381339952
4	haber man	306	2	3	0.00034265 (s = 0.03)	0.000317745	0.000316547
5	balance scale	625	3	4	0.002742756 (s = 0.002)	0.002573387	0.003332606
6	Wisconsin breast cancer	699	2	10	7.25929E-14 (s = 1)	7.25935E-14	7.48861E-14
7	Contraceptive Method Choice	1473	3	9	8.19498E-05 (s = 0.5)	7.80139E-05	7.69432E-05
8	hayes roth	132	3	5	4.71204E-05 (s = 3)	4.59807E-05	4.43056E-05
9	Robot Navigation	5456	4	2	0.001896439 (s = 0.1)	0.001583094	0.002000381
10	spect heart	80	2	22	0.076041565 (s = 0.03)	0.069341756	0.077804472
11	Wine	178	3	13	4.86902E-07 (s = 0.02)	4.83293E-07	4.6507E-07

Table-5: Fitness of K-Mean , FCM and PSOC on 2d artificial dataset

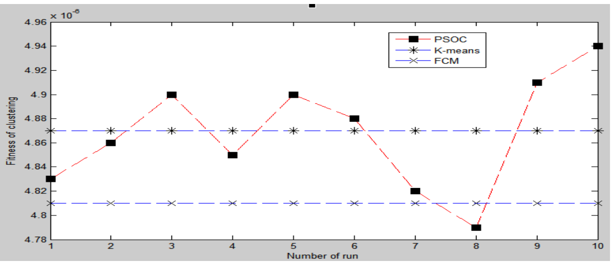
No. of run	PSOC	KMean	FCM
1	4.82761E-06	4.86614E-06	4.80668E-06
2	4.85616E-06	4.86416E-06	4.91855E-06
3	4.8951E-06	4.94015E-06	4.87772E-06
4	4.85183E-06	4.76226E-06	4.86309E-06
5	4.89927E-06	4.91396E-06	4.82431E-06
6	4.88002E-06	4.83975E-06	4.80431E-06
7	4.82072E-06	4.89246E-06	4.80044E-06
8	4.78785E-06	4.83876E-06	4.83746E-06
9	4.90774E-06	4.94137E-06	4.80802E-06
10	4.94309E-06	4.85081E-06	4.82524E-06

Table-6: Fitness of K-Mean , FCM and PSOC on 4d iris dataset

No. of run	PSOC	K-Mean	FCM
1	0.0108764	0.0123954	0.0127382
2	0.0144329	0.0123954	0.0127384
3	0.0110948	0.0123954	0.0127384
4	0.012737	0.0117522	0.0127385
5	0.0108248	0.0114606	0.0127382
6	0.0112495	0.0123954	0.0127384
7	0.0026524	0.0123954	0.0127383
8	0.0133923	0.0123954	0.0127383
9	0.0143069	0.0117522	0.0127384
10	0.0123246	0.0123954	0.0127382

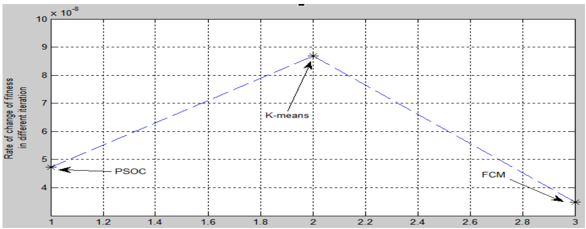
5. Result Analysis

Fitness which is generated from 10 run of PSO-based clustering (PSOC) method, k-means and fuzzy c-mean on 2d artificial datasets and iris datasets are collected in teble-1 and table-2 respectively. Fig-10 and fig-11 shows the change in fitness of all clustering (K-means, FCM and PSOC) on 2d artificial dataset and iris dataset respectively.



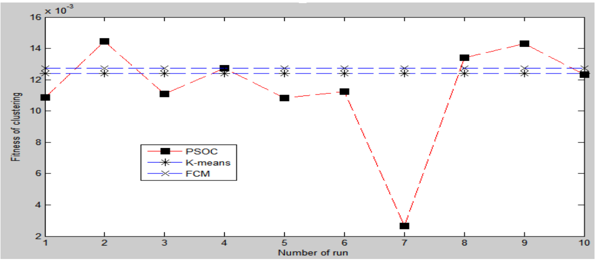
(Fig-10: Comparison of fitness of K-Mean, FCM and PSOC on 2d artificial dataset)

The standard deviation in fitness of all clustering (K-means, FCM and PSOC) on 2d artificial dataset has been demonstrated on fig-12. Out of this simulation, we conclude that standard deviation of K-means is larger than PSOC and FCM has least standard deviation in fitness on 2d artificial dataset. Fig-13 describes standard deviation in fitness of K-means, FCM and PSOC on iris dataset. Highest standard deviation in fitness is noted on PSOC and FCM has least deviation.



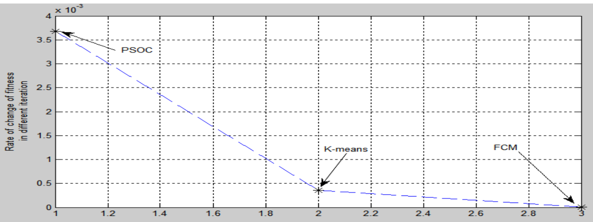
(Fig-12 : Rate of change of fitness of K-mean, FCM and PSOC in 10 numbers of run on 2d artificial dataset)

Performance of PSOC is compared with performance of k-mean and fuzzy c-mean on various datasets (artificial 2d dataset, irs, lense, habaer man, balancescale, breast cancer ,contraceptive method choice, hayes-roth, robot navigation, spect heart and wine). Fig-14 shows the comparison of performance of PSOC with K-means and FCM.

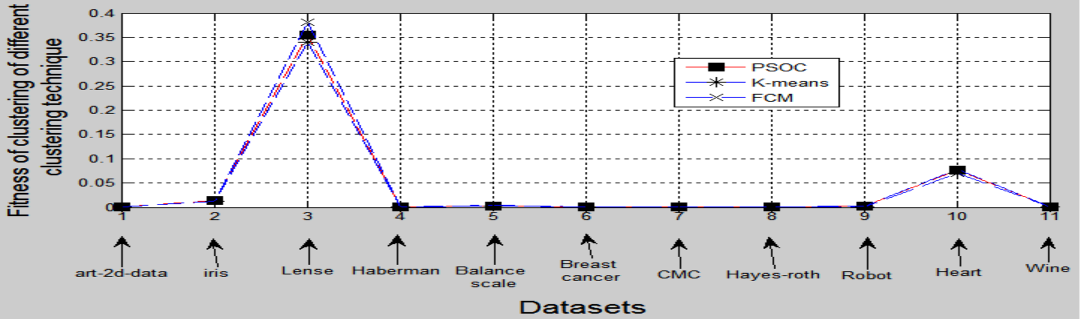


(Fig-11 : Comparison of fitness of K-Mean, FCM and PSOC on iris dataset)

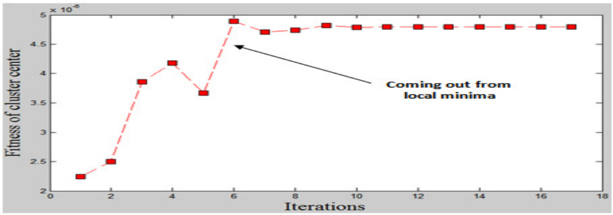
PSO based clustering is providing good cluster center vector of a cluster but the time of convergence is uncertain. As the number of iteration increases, the initial random cluster center moves toward center of respective cluster which results the final cluster center with best fitness. The lbest and gbest are calculated in each iteration. Based on lbest and gbest, new velocity has been calculated and new position of cluster centers during each iteration is computed. Change of gbest in every iteration on 2d artificial dataset has been noted and demonstrated on fig-15. Fig-16 shows the change of gbest in different iteration on iris dataset. The fluctuation curve in fig-15 and fig-16 describes how PSO clustering avoids local minima. This helps the PSO clustering not to fall in local minima.



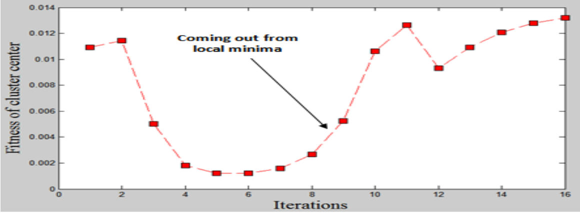
(Fig-13 : Rate of change of fitness of K-mean, FCM and PSOC in 10 numbers of run on iris dataset)



(Fig-14 : Comparison of performance of K-Mean, FCM and PSOC on one artificial dataset and ten real datasets from UCI repository)



(Fig-15 : Change of gbest of PSOC on 2d artificial dataset in different iteration towards convergence)



(Fig-16 : Change of gbest of PSOC on iris dataset in different iteration towards convergence)

6. Parameter Setting

c1 and c2 are the parameters of cognition and social model of PSO. k and d are the parameters of the equation-4, which is used to calculate the fitness clustering technique. Above simulation has been carried out with k=50, d=0.1 to find out the efficiency (fitness) of proposed model and all existing model on different datasets. During PSO based clustering the PSO parameter is set to c1=1 and c2=1 for early convergence. In the algorithm psoBasedClustering (X, n), the parameter s must be set during clustering. Here s is small valued constant. Value of s depends open dataset being used because it depends upon the degree of interference and overlapping among clusters in a particular dataset. Values of s has been chosen for different datasets and listed at table-4.

7. Computational Complexity

Time complexity of proposed PSO based clustering is calculated and has been compared with time complexity of existing algorithm. We conclude that time complexity of proposed algorithm is bounded with O $(m*n*d*t_{max})$. Table-7 shows the comparison of time complexity among K-means, Fuzzy C-means and PSO based clustering. K-means algorithm takes $O(m*n*d*t_{max})$ [10] , Fuzzy C-mean takes $O(m*t_{max})$ [10] and our proposed PSO based clustering takes $O(m*n*d*t_{max})$.

Algorithm psoBasedClustering (X, n)

```
1. load dataset X and set number of clutser 'n' to be found.-------------------- ◯ (c)
2. Set initial random clutser center vector <$C_1$,$C_2$,….$C_n$>.------------------- ◯(c)
3. Set random velocity V=<$V_1$,$V_2$,….$V_n$>, where each V1=<v1,v2,….vk>. here n is the number of cluster and k is dimension of
   dataset.$V_1$,$V_2$,$V_3$ ,…$V_n$ are initial random velocity vector for $C_1$, $C_2$, $C_3$,…$C_n$ respectively.
4. Compute Euclidian distance from all clusters<$C_1$,$C_2$,….$C_n$> to all the instances of X.------$t_{max}$ * ◯(n*m*d)
5. Create clusters based on Euclidian distances computed at step-4.---------------------------- $t_{max}$ * ◯(m*d)
6. Calculate fitness of all instances ($F_{xi}$) of clusters by using the equation-3 and generate lbest.--$t_{max}$ *◯(n*m*d)
7. The instance having highest fitness in each cluster is chosen as gbest of that cluster. Generate n number of gbest, where 'n' is
   the number of cluster.------------------------------------------- $t_{max}$ *◯(m)
8. Compute new velocity $V_{NEW}$ out of initial velocity, lbest and gbest by use of equation-1.--- $t_{max}$ *◯(n*d*c)
9. Update the position of all cluster centers (centroid) with new velocity $V_{NEW}$ and generate $C_{NEW}$ by using equation-2.------------
   ---------------------------------- $t_{max}$ * ◯ (n*c)
10. if(Euclidian distance(C,$C_{NEW}$) <= s)---------------------- $t_{max}$ * ◯(c)
11.         goto step-4-------------------------------------- $t_{max}$ * ◯(c)
12. elsedisplay final clusters------------------------------- ◯(1)
13.         goto step-14---------------------------------------------------- ◯(1)
14. Compute the performance of PSO ($F_{CT}$) using equation-2.------------------------ ◯(n*m)
15. stop ------------------------ ◯(1)
```

$$T(m) = c+c+ n*m*d*t_{max}+ m*d* t_{max}+ n*m*d* t_{max}+m* t_{max}+n*c*d* t_{max} + n*c* t_{max} + c* t_{max} + c* t_{max} +1 +n*m+1$$
$$= ◯(n*m*d*t_{max})$$

Where T(m) is the total number of steps (time), m is the size of dataset being used, n is the number of cluster to be formed, c is a +ve constant and t_{max} (t_{max} >=1) is the maximum number of iteration of PSO.

Table-7: Comparison of Time Complexity

Clustering Algorithm	Time Complexity	Capability of handling high dimensional data
K-means	◯$(m*n*d*t_{max})$	No
Fuzzy C-means	◯(m*tmax)	No
PSOC	◯$(m*n*d*t_{max})$	Yes

8. Conclusion

This paper provides a clustering analysis algorithm based on PSO, called PSO-clustering. PSO-based clustering is based on the object function F_C and Fx_i to search automatically the data cluster centers of n-dimension. Traditional cluster algorithm such as K-means may falls at local optimal solution, depending on the choice of the initial random cluster centers. It can't make sure to solve the global optimal solution every time. Related to the other evolution algorithm, PSO can avoid entering into the local optimal solution (shown at fig-15 and fig-16). The experimental result on real datasets shows that the PSO clustering has better performance than the traditional clustering analysis methods. We have presented an efficient implementation of PSO clustering algorithm, which is easy to implement and only requires that a relative advantage provided by preprocessing in the above manner is greater. Our algorithm differs from existing approaches only the way how optimal cluster centers are computed. The algorithm has been implemented and the source code will be available on demand. We have demonstrated the efficiency of algorithm through experiments on both synthetically generated dataset and 10 numbers of real data sets from UCI repository. Analysis shows that, the algorithm runs faster if dataset contains well-separated clusters. In case of distinctly separated clusters, the fitness of cluster centers is better. Later stages of PSO clustering algorithm, as the centers are converging to their final positions, one would expect that the majority of the data points have the same closest cluster center from one stage to the next. A good algorithm must exploit this coherence to improve the running time. In a dataset, if the degree of interference and overlapping increases, the performance of traditional PSO based clustering decreases. In this suggested PSO clustering, to increase the performance of clustering, a appropriate value of ‘s’ is to be set. Values of s has been chosen for different datasets and listed at table-4.We implemented and compared proposed method with K-means and FCM with some benchmark datasets (table-2 and table-3). From the results, we can conclude that PSO can obtain competitive results on the data sets used and other several real data sets, although there is some increase in the computational effort is needed. We observed that, a little change in cluster centers results cluster with best fitness. Future work includes application of this tool to more demanding data sets with more complex data and different degree of interferences. Simulated results show that, PSO is a real effective and competitive technique in DM. This method can be applied to pattern recognition, classification and various field of data mining.

References

[1]. Alireza Ahmadyfard, Hamidreza Modares “Combining PSO and k-means to Enhance Data Clustering”, International Symposium on Telecommunications,IEEE, 2008, pp 688-691

[2]. A.A.A. Esmin, D.L. Pereira and F.P.A De Araujo “Study of different approach to clustering data by using the Particle Swarm Optimization algorithm” , IEEE Congress on Evolutionary Computation (CEC 2008), pp 1817-1822

[3]. Milad Azarbad , AtaoUah Ebrahimzadeh and Abbas Babajani-Feremi “Brain Tissue Segmentation Using an Unsupervised Clustering Technique Based on PSO Algorithm”, Proceedings of the 17th Iranian Conference of Biomedical Engineering (ICBME2010), 3-4 November 2010, IEEE

[4]. Shi M. SHAN, Gui S. DENG, Ying H. HE “Data Clustering using Hybridization of Clustering Based on Grid and Density with PSO” 2006, IEEE, pp 868-872

[5]. Alireza Ahmadyfard, Hamidreza Modares “Combining PSO and k-means to Enhance Data Clustering”, International Symposium on Telecommunications,2008, IEEE, pp 688-691

[6]. Chih-Cheng Hung and Li Wan “Hybridization of Particle Swarm Optimization with the K-Means Algorithm for Image Classification”, IEEE, 2009

[7]. Abbas Ahmadi, Fakhri Karray and Mohamed Kamel ”MULTIPLE COOPERATING SWARMS FOR DATA CLUSTERING” Proceedings of the 2007 IEEE Swarm Intelligence Symposium (SIS 2007), IEEE

[8]. DW van der Merwe and AP Engelhrecht “Data Clustering using Particle Swarm Optimization”, IEEE, 2003, pp 215-220

[9]. GAO Lei-fu, Qi Wei , LIU Xu-wang “Particle Swarm Optimization algorithm Based on Variable Metric Method and its application of non-linear equations” , IEEE, 2010, pp 514-518

[10]. Rui Xu and Donald Wunsch II “Survey of Clustering Algorithms”, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005, pp 645-678

[11]. Ching-Yi Cheo and Fun Ye “Particle Swarm Optimization Algorithm and Its Application to Clustering Analysis”, lntemationai Conference on Networking, Sensing Control, 2004, IEEE, pp 789-794

[12]. J. Kennedy and R. Eberhart, “Particle swarm optimization,” Proc.IEEE Int. Conf. Neural Networks, 1995, pp. 1942– 1948.

[13]. D.E Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Reading, MA: Addison-Wesley, 1989.

[14]. M. Clerc and J. Kennedy, “The Particle Swarm-Explosion, Stability, and Convergence in a multi-dimensional complex space”, IEEE Trans.Evol.Comput, Vol.6, pp.58-73, Feb.2002.

[15]. W.F Abd-EL-Wahed, A.A Mousa, M.A.EL-Shorbagy,”Integrating Particle Swarm Optimization With Genetic Algorithms For Solving Nonlinear Optimization problems”, Journal Of Computational and Applied mathematics,2010.

[16]. Alejandro Cervantes,In˝es Galv˝an, and Pedro Isasi, “ An Adaptive Michigan Approach PSO for Nearest Prototype Classification”, Spanish founded research MEC project PLINK::UC3M,Ref: TIN2005-08818-C04-02 and CAM project UC3M-TEC-05-029.

[17]. Stefan Janson and Martin Middendorf, Member,IEEE, ”A Hierarchical Particle Swarm Optimizer and Its Adaptive Variant”, IEEE Transactions on Systems, Man and Cybernetic-PartB: Cybernetics, Vol.35, No.6, DEC2005.

[18]. Dorigo, M.; Birattari, M.; Stutzle, T .“Ant colony optimization”, Computational Intelligence Magazine, IEEE, Nov. 2006, pp 28 – 39

[19]. Chiu, S., "Fuzzy Model Identification Based on Cluster Estimation," Journal of Intelligent & Fuzzy Systems, Vol. 2, No. 3, Sept. 1994.

[20]. Yager, R. and D. Filev, "Generation of Fuzzy Rules by Mountain Clustering," Journal of Intelligent & Fuzzy Systems, Vol. 2, No. 3, pp. 209-219, 1994.

Authors

Author’s Name: BIGHNARAJ NAIK
Author’s Profile: He has received B.E degree in Information Technology from National Institute of Science and Technology, India, and the MTech degree in computer science from the Institute of Technical Education and Research, Siksha ‘O’ Anushandhan University, India. He is currently working as an Assistant Professor in the department of Information Technology at Institute of Technical Education and Research, Siksha ‘O’ Anushandhan University, India. His main research interests are in the areas of design and analysis of algorithm, data mining, pattern recognition and natured inspired computing. He is a member of IAENG Society (Member Number: 115149).

Author’s Name: Sarita Mahapatra
Author’s Profile: She received the B.Tech degree in information technology from the Synergy Institute of Technology, under Uttkal University and M.Tech from Institute of Technical Education and Research, Siksha ‘O’ Anushandhan University, india. She is a lecturer in the Department of Information Technology, ITER, Siksha ‘O’ Anushandhan University, where she is teaching since last 5 year. Her research interests are in the areas of data mining and pattern recognition.

Author’s Name: Subhra Swetanisha
Author’s Profile: She has been an Assistant Professor with the department of Computer Science and Engineering in Trident Academy of Technology Bhubaneswar, Odisha, India.She has received MTech degree in Computer Science and Engineering from KIIT University, India. Her current research interest includes Genetic-Fuzzy-Neural systems, Data Mining, and Image Processing.She is a member of Indian Society for Technical Education(ISTE) with member id LM78439.

Author’s Name: Swadhin Kumar Barisal
Author’s Profile: He has received M.TECH. degree in Computer science and engineering from Indian Institute of Technology , Kharagpur, India, and the BTech degree in computer science from Synergy Institute of Engineering and Technology, Odisha,India. He is currently working as an Assistant Professor in the department of Computer science and engineering at Institute of Technical Education and Research, Siksha ‘O’ Anushandhan University, India. His main research interests are in the areas of design and analysis of algorithm, data mining, and pattern recognition and also object oriented technology. He is a member of IAENG Society (Member Number: 119595).