Journal of Theoretical and Applied Information Technology
15 August 2012. Vol. 42 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

COOPERATIVE SWARM INTELLIGENCE BASED EVOLUTIONARY APPROACH TO FIND OPTIMAL CLUSTER CENTER IN CLUSTER ANALYSIS

[1]BIGHNARAJ NAIK, [2]SARITA MAHAPATRA
[1]Asstt Prof., Department of Information Technology, Siksha 'O' Anusandhan University, India
[2]Lecturer., Department of Information Technology, Siksha 'O' Anusandhan University, India

Email: [1]bighnaraj_naik@yahoo.co.in, [2]saritamahapatra4@gmail.com

ABSTRACT

The Centroid-based clustering is an NP-hard optimization problem and the common approach is to search for cluster centers only for approximate solutions. In this paper we have proposed swarm intelligence based nature-inspired center-based clustering method using PSO optimization. Proposed PSO clustering method is capable to search best cluster with maximum fitness using social-only and cognition-only model, such that the square distances from the cluster are minimized. In this article, how PSO based clustering can be used to get N number of cluster specified by the user in a dataset is demonstrated. Our suggested method has been tested with artificial dataset and several datasets from UCI Machine learning repository. Effectiveness and usefulness of the proposed method is shown by comparing fitness of this method with K-means and Fuzzy c-means technique. For better comparative result, we have ended up with comparison of proposed clustering model with subtractive clustering (extension of the mountain clustering) to ensure proposed method computes optimal number of cluster in a dataset. Results shows that, this method is quite simple, effective and has much potential to search best cluster centers in multidimensional search space.

Keywords :*Swarm intelligence (SI); Particle swarm optimization (PSO); Centroid-based clustering; Cluster Analysis; Fuzzy C-means clustering (FCM); K-Means clustering; Subtractive Clustering; Euclidean distance;*

1. INRODUCTION

1.1 Swarm intelligence:

Swarm Intelligence (SI) [12]-[13] is an artificial intelligence technique inspired by nature, based on the study of collective behavior in centralized and self-organized systems. SI was introduced by Beni & Wang in 1989, in the context of cellular robotic system. 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. Two 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 basic 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.

1.1.2 Particle swarm optimization(PSO):

PSO is originally attributed to Kennedy, Eberhart and Shi[12] and was first intended for simulating social behavior, as a 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.

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

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

corresponding center it is closer to. Center is the representative of a cluster. The most popular 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.

1.2.2 Fuzzy c-means clustering (FCM):

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 belongingness to that cluster and another data point that lies far away from the center of a cluster will have low degree of belongingness [1]-[19]-[20] to that cluster.

1.2.3 Subtractive clustering:

If you do not have a clear idea how many clusters there should be for a given set of data, Subtractive clustering, [20]-[19]-[21], is a fast, one-pass algorithm for estimating the number of clusters and the cluster centers in a set of data. The algorithm does the following:

- Selects the data point with the highest potential to be the first cluster center
- Removes all data points in the vicinity of the first cluster center (as determined by radii), in order to determine the next data cluster and its center location
- Iterates on this process until all of the data is within radii of a cluster center

The subtractive clustering method is an extension of the mountain clustering method proposed by R. Yager[20].

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 other 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.

$$V_{id}(t+1)= V_{id}(t) + c_1 * rand() * (lbest_{id} - X_{id}) + c_2 * rand()*(gbest_{id} - X_{id}) \quad (1)$$

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

Here c1 and c2 are the constants and rand() is random function which generates random number in between 0 to 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 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. Basis concept of PSO can be used for clustering [2]-[5]-[3] and classification[6].

3. CLUSTER ANALYSIS USING PSO

Cluster analysis is a collection of methods, which identifies groups of instances that have similar characteristics. Clustering can be achieved by various algorithms [7]-[8] that differ significantly in their methods of generation of cluster. Typically a cluster includes groups with low distances among the cluster members, dense areas of the data space [4] or particular distributions [10]. The appropriate clustering algorithm and parameter settings depend on the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

individual data set being used for clustering and intended use of the results [11]. In this paper, a cluster analysis model is proposed which is based on most popular nature inspired technique known as PSO. It has following steps (Algorithm-1). 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 dataset X, V=<V_1,V_2,….V_n> represents 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. The Algorithm computes n number of clusters in a given dataset. Here n (n>1) is the number of cluster provided by the user. Initial cluster centers will be selected randomly. Euclidian distances are computed from randomly selected cluster center. Fitness of all instances of generated clusters has been calculated as it is used as lbest.

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

Here n number of lbest are generated and next velocity vectors have been computed by using initial velocity, lbest and gbest, Next positions of cluster centers are generated by using new velocity. These steps will be repeatedly executed until and unless the target clusters are found. The positions and velocities of the particles initially in search space denoted by V and X. The new velocities and positions of the particles for next iterations [5] can be evaluated by using the equations 1 and 2. Fig-1 and fig-2 demonstrates generation of initial cluster and target cluster respectively. Flowchart of this procedure is shown on fig-3.

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

ALGORITHM-1 PSOClUSTERING (X, n, S)

X – Dataset to be clustered, n – Number of cluster.

S – Small positive valued constant

V=<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

.

1. load dataset X and set number of cluster 'n' to be found.

2. Set initial random cluster center vector <C_1,C_2,….C_n> and random velocity V=<V_1,V_2,….V_n>.

3. Compute Euclidian distance from all clusters<C_1,C_2,….C_n> to all the instances of X and

4. Create clusters based on Euclidian distances.

5. Calculate fitness of all instances (F_{xi}) of clusters by using the equation-3 and generate lbest.

6. Choose the instance having highest fitness in each cluster is chosen as gbest of that cluster. Generate n number of gbest.

7. Compute new velocity V_{NEW} out of initial velocity, lbest and gbest by use of equation-1.

8. Update the position of all cluster centers (centroid) with new velocity V_{NEW} and generate C_{NEW} by using equation-2.

9. if (Euclidian distance(C,C_{NEW}) <= S)

10. goto step-3

11. else display final clusters

12. goto step-14

13. end if

14. Calculate the performance of PSO (F_{CT}) using equation-4.

15. end

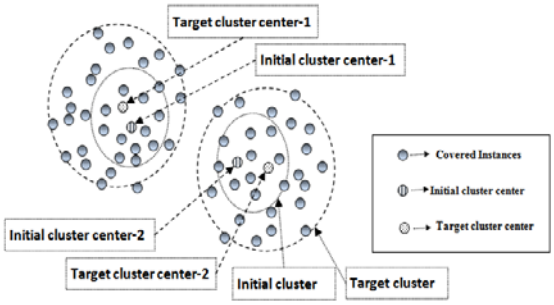
$$Fc = \frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{n}|C_j - X_i|^2} \quad (5)$$

F_{Xi} represents fitness of an instance, where X is the dataset, N is the number of instances in X, X_i is the i^{th} 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 i^{th} 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 i^{th} instance of X, k is a positive constant and d is a small-valued constant. Most of time the PSO algorithm stops based on two parameter.1- exceed maximum velocity range and 2- maximum number of iterations. Our proposed model will stop in neither of these conditions. It will stop when it reaches a value (S) . S is the difference between old cluster center and new cluster center. Here s is small valued constant. Value of S depends upon dataset being used. During experiments, values of S has been chosen for different datasets and listed at table-2 and table-3.

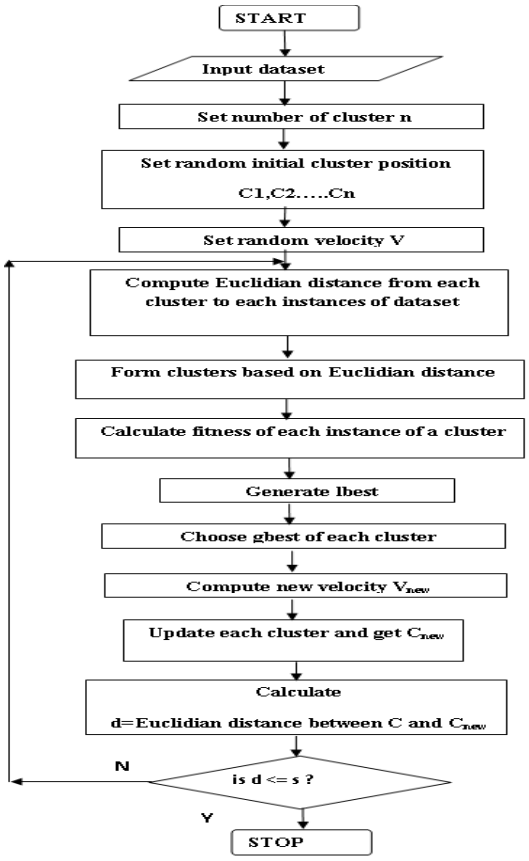
Journal of Theoretical and Applied Information Technology
15 August 2012. Vol. 42 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(Fig-1: Initial cluster centers, bests and gbest)



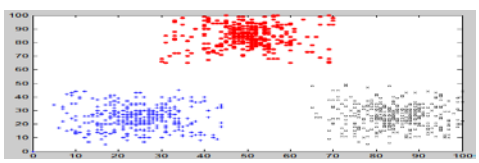
(Fig-2: Formation of Target cluster)



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

4. SIMULATION RESULT & ANALYSIS

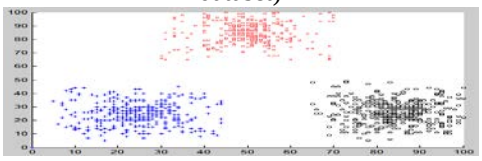
The proposed clustering method using PSO has been tested with datasets from Machine learning repository and has been implemented on a system with MATLAB with the configuration specified (*Platform- MATLAB-10, OS- Window-7, Processor- Intel(R) Core(TM) i3 CPU M380 @ 2.53 GHz , RAM-3.00 GB*). Better configurations will help the program to faster. Dataset used has multiple clusters and multiple dimensions. The resultant cluster centers of different datasets are listed on table-1. A little change in cluster center vector has significant effect of total fitness of cluster. Fitness comparison of proposed PSO based clustering with K-Mean, Fuzzy C-Mean has been done and simulation result is demonstrated on table-2. 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 dataset and iris 4d dataset is displayed on table-4 and table-5 respectively. 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 and iris dataset is highlighted on the table-4 and table-5 respectively. 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. The standard deviation in fitness of K-means, FCM and PSOC on 2d artificial dataset and iris dataset has been demonstrated on fig-12 and fig-13 respectively. From fig-12, we conclude that standard deviation of K-means is larger than PSOC and FCM has least standard deviation in fitness on 2d artificial dataset. As the number of iteration increases, fitness of cluster centers increases. Change of gbest in different iterations on artificial and iris dataset has been noted and demonstrated on fig-15 and fig-16 respectively. The curve in fig-15 and fig-16 describes how PSO clustering avoids local minima and this helps the PSO clustering not to fall in local minima.



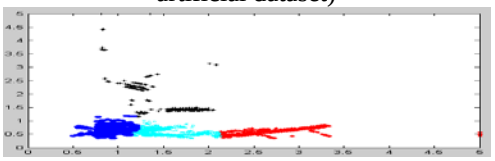
Journal of Theoretical and Applied Information Technology
15 August 2012. Vol. 42 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

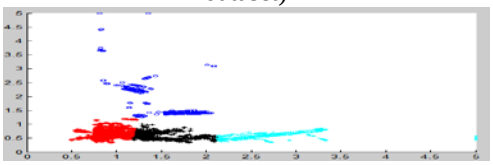
(Fig-4: Cluster generation using PSOC in artificial dataset)



(Fig-5: Cluster generation using K-mean in artificial dataset)



(Fig-6: Cluster generation using PSOC in robot dataset)



(Fig-7: Cluster generation using k-mean in robot dataset)

Table-1: Cluster centers of datasets

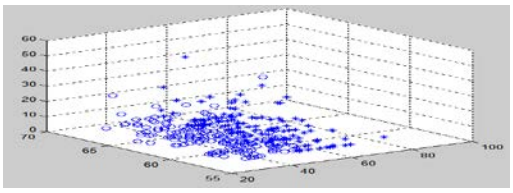
Datasets	No of Class	Dim	K-Mean	Fuzzy C-Mean	PSOC
Iris	3	4	(6.31, 2.89, 4.97, 1.70)	(6.77, 3.05, 5.64, 2.05)	(6.83, 3.07,5.71, 2.14)
			(5.20, 3.630, 1.47, 0.27)	(5.88, 2.76, 4.36, 1.39)	(5.87, 2.81, 4.28, 1.39)
			(4.73, 2.93, 1.76, 0.33)	(5.01, 3.40, 1.48, 0.25)	(5.07, 3.4, 1.58, 0.26)
Balance scale	3	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, 4.120)	(3.08, 3.03, 3.040, 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	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)
Lense	3	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)
CMC	3	9	(43.85, 2.83, 3.320,4.86,0.81, 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)
Haber man	2	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)
Artificial data	3	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)
Hayes-roth	3	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	4	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	2	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	3	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)

Journal of Theoretical and Applied Information Technology
15 August 2012. Vol. 42 No.1

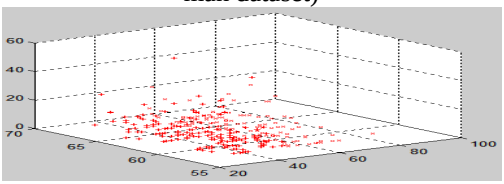
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

1	Iris	150	0.014432895 (s = 1)	0.012395396	0.012738542
2	balance scale	625	0.002742756 (s = 0.002)	0.002573387	0.003332606
3	Wisconsin breast cancer	699	7.25929E-14 (s = 1)	7.25935E-14	7.48861E-14
4	Lenses	24	0.354960239 (s = 1.5)	0.339904827	0.381339952
5	CMC	1473	8.19498E-05 (s = 0.5)	7.80139E-05	7.69432E-05
6	Haberman	306	0.00034265 (s = 0.03)	0.000317745	0.000316547
7	Artificial data	600	4.94309E-06 (s = 0.02)	4.94137E-06	4.91855E-06
8	Hayes-roth	132	4.71204E-05 (s = 3)	4.59807E-05	4.43056E-05
9	Robot	5456	0.001896439 (s = 0.1)	0.001583094	0.002000381
10	Spect heart	80	0.076041565 (s = 0.03)	0.069341756	0.077804472
11	Wine	178	4.86902E-07 (s = 0.02)	4.83293E-07	4.6507E-07

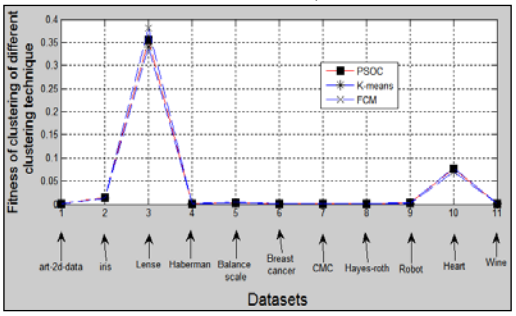
Table-2: Comparison of fitness of clustering methods



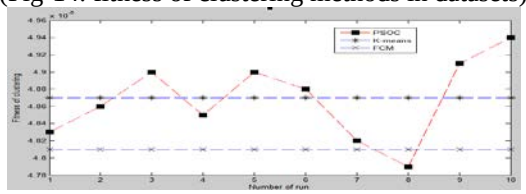
(Fig-8: Cluster generation using PSOC in haber man dataset)



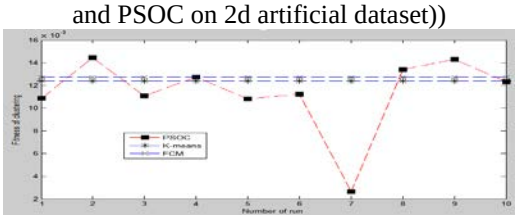
(Fig-9: Cluster generation using K-mean in haber man dataset)



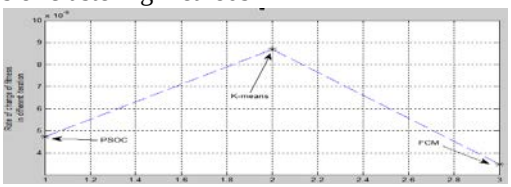
(Fig-14: fitness of clustering methods in datasets)



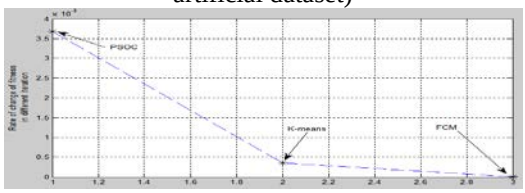
(Fig-10: Comparison of fitness of K-Mean, FCM and PSOC on 2d artificial dataset))



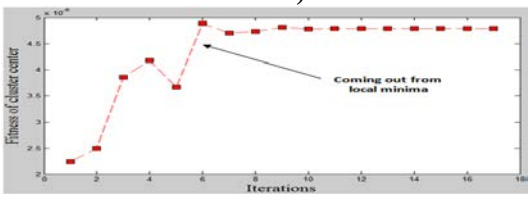
(Fig-11: Comparison of fitness of K-Mean, FCM and PSOC on iris dataset)



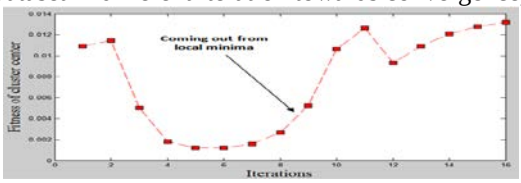
(Fig-12: Rate of change of fitness of K-mean, FCM and PSOC in 10 numbers of run on 2d artificial dataset)



(Fig-13: Rate of change of fitness of K-mean, FCM and PSOC in 10 numbers of run on 4d iris dataset)



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



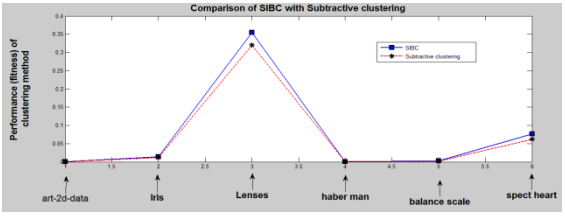
(Fig-16: Change of gbest of PSOC on iris dataset in different iteration towards convergence)

5. COMPARISON OF PSOC WITH SUBTRACTIVE CLUSTERING

In this section, the proposed method has been compared with a advanced and extended version of mountain clustering known as subtractive clustering (*extension of the mountain clustering*) [21]-[10] to ensure proposed method computes optimal number of cluster in a dataset. Subtractive clustering computes clusters in a dataset without prior information on number of cluster to be

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

generated. During subtractive clustering, radii value is set to different value (table-3) to generate cluster. Fitness/performance comparison of PSOC with subtractive clustering is demonstrated in table-3 and fig-17. We have observed that during subtractive clustering the range of radii is [0.009 to 1.7] where as during PSOC the range of S is [0.002 to 1.5].



(Fig-17: Performance comparison of PSOC with subtractive clustering)

Table-3: Performance Comparison of PSOC with Subtractive clustering (SC)

S.No.	Dataset	Instances	PSOC	SC
1	Artificial data	600	4.94309E-06 (s = 0.02)	4.84362E-06 (Radii=0.2)
2	Iris	150	0.014432895 (s = 1)	0.012185374 (Radii=0.7)
3	Lenses	24	0.354960239 (s = 1.5)	0.320307495 (Radii=1.7)
4	Haberman	306	0.00034265 (s = 0.03)	0.000417292 (Radii=0.65)
5	Balance Scale	625	0.002742756 (s = 0.002)	0.001599995 (Radii=1.05)
6	Spect Heart	80	0.076041565 (s = 0.03)	0.062336367 (Radii=0.009)

Table-4: 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-5: 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

6. PARAMETER SETTING & COMPUTATIONAL COMPLEXITY

k and d are the parameters of the objective function (equation-4). Simulation has been carried out with k=50, d=0.1. c1 and c2 are the parameters of cognition and social model of PSO and is set to c1=1 and c2=1 for early convergence. In the algorithm PSOClUSTERING (X, n, S), the parameter S must be set during clustering. Here S is small valued constant and it depends upon 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-3. As an observation based on following datasets, a values of S is chosen within the range [0.002 to 3]. Time complexity of proposed PSO based clustering is calculated and has been compared with time complexity of existing algorithms. Time taken by each step of PSO based clustering has been calculated and based on that total time complexity T(m) computed. Time complexity is found to be bounded with $O(m*n*d*t_{max})$. Table-6 shows the comparison of time complexity among PSO based clustering and exiting clustering algorithm. K-means algorithm takes $O(m*n*d*t_{max})$ [10] , Fuzzy C-mean takes $O(m*t_{max})$ [10], Subtractive clustering takes $O(m^2*d*t_{max})$ [21]-[10] and our proposed PSO based clustering takes $O(m*n*d*t_{max})$.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ALGORITHM-1 PSOClUSTERING (X, n, S)

X – Dataset to be clustered, n – Number of cluster to be generated.

S – Small positive valued constant

V=<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.

1. load dataset X and set the value of n .---------- **◯(c)**
2. Set initial random cluster center vector <C_1,C_2,….C_n>.---------- **◯(c)**
3. Set random velocity V=<V_1,V_2,….V_n>---------- **◯(c)**
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.----------**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. else display final clusters---------- **◯(1)**
13. goto step-14---------- **◯(1)**
14. Compute the performance of PSO (F_{CT}) using equation-4.---------- **◯(n*m)**
15. stop ---------- **◯(1)**

So T(m) = c+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 **(6)**

Here 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, d is the dimension of dataset to be clustered, c is a +ve constant and t_{max} (t_{max} >=1) is the maximum number of iteration of PSO. The growth above equation-6 is dominated by n*m*d*t_{max} .Among all the function that are involves in this equation n*m*d*t_{max} has highest growth. So total time complexity T(m) can be described as T(m) = ◯(n*m*d*t_{max}).

Table-6: Time Complexity of clustering methods

Clustering Algorithm	Time Complexity	Capability of handling high dimensional data
K-means	◯(m*n*d*t_{max})	No
Fuzzy C-means	◯(m*t_{max})	No
Subtractive Clustering	◯(m^2*d*t_{max})	No
PSOC	◯(m*n*d*t_{max})	Yes

7. CONCLUSION

In this paper, a clustering analysis algorithm based on PSO has been proposed, called PSO-clustering. PSO-based clustering is based on the objective function F_C and Fx_i to search automatically the data cluster centers of n-dimension. Depending on the choice of the initial random cluster centers, traditional clustering algorithm such as K-means may falls at local optimal solution. It can't make sure to solve the global optimal solution every time. Like other evolutionary algorithm, PSO can avoid entering into the local optimal solution (shown at fig-15 and fig-16). The experimental result shows that the PSO clustering has better performance over the traditional clustering methods. In this paper, an efficient implementation of PSO clustering algorithm has been demonstrated. This suggested method differs from existing approaches only the way the optimal cluster centers are computed. Analysis shows that, if the dataset contains well-separated clusters, the algorithm will run faster. The fitness of cluster centers is better for distinctly separated clusters. 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

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

performance of clustering, a appropriate value of S has to be set. PSO has obtained competitive results on the data sets used and can be used for other several data sets. Future work includes application of this tool to more data sets with more complex data and different degree of interferences. Simulated results show that, PSO is an effective and competitive technique in DM.

REFERENCES

[1]. Alireza Ahmadyfard, Hamidreza Modares "Combining PSO and k-means to Enhance Data Clustering", Internatioal 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", Internatioal 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.

[21]. Miin-Shen Yang and Kuo-Lung Wu, "A modified mountain clustering algorithm"

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Published online: 24 June 2005, Springer-Verlag, Pattern Anal Applic : pp 125–138, 2005