International Journal of Scientific & Engineering Research, Volume 4, Issue 10, October-2013
ISSN 2229-5518

A Comparative Study of Hard and Soft Clustering Using Swarm Optimization

Bijayalaxmi Panda, Soumya Sahoo, Sovan Kumar Patnaik

Abstract— Cluster analysis is one of the major techniques in pattern recognition, which is basically considered as one of the unsupervised learning technique. We can apply clustering techniques in various areas like clustering medicine, business, engineering systems and image processing, etc.,The traditional hard clustering methods restrict that each point of the data set belongs to exactly one cluster. But fuzzy clustering proposed that the belongingness of each data points is based on a membership function.Now a days fuzzy clustering has been widely studied and applied in a variety of substantive areas.To find the global optimal solution we have also applied the concept of particle swarm optimization on K-means clusterings and modified particle swarm optimization on Fuzzy– c–means and performed a comparative study on four clustering algorithms on the basis of compactness,separability time complexity.

.**Index Terms**— swarm optimization,fuzzy c-means,k-means,compactness,separability, inertia weight

1 INTRODUCTION

The word cluster is used to define a group of points close to each other. The term "Cluster analysis "is first used by Tryon in 1939.Clustering is an explanatory data analysis technique which is used to find the natural groupings in the data. Classification of data is of 2 types, supervised and unsupervised. Clustering is considered as an unsupervised classification technique. To arrange data into meaningful clusters several algorithms are proposed. Cluster analysis deals with finding similarities in the data and grouping them. Object having similar characteristics belongs to one cluster and objects of one cluster differ from objects of another cluster. There are 2 types of distance measure known as "inter cluster distance" which is the distance between objects of same cluster. "Inter cluster distance" which is the distance between the objects of different clusters. Objects having similar nature are grouped on the basis of similarity measure. For this different types of distances are used, the most popular distance measure is Euclidean distance .

2 RELATED WORKS

2.1 S.Ghosh and S. K. Dubey[1] in 2013 have included two clustering algorithms in their research for comparision i.e
centroid
based K-Means and representative object based FCM (Fuzzy C-Means) cluster. These algorithms are applied and performance is evaluated on the basis of the efficiency of clustering output. The numbers of data points as well as the number of clusters are the factors upon which the behaviour patterns of both the algorithms are analyzed. FCM produces close results to K-Means clustering but it still requires more computation time than K-Means clustering.

- *Bijayalaxmi Panda CV RamanCollege of Engineering ,BPUT, India, PH-9861357434. E-mail:bijayalaxmi.panda81@rediffmail.com*

  Soumya Sahoo CV Raman College of Engineering, BPUT, India,PH-9861775097. E-mail: ritun_08@yahoo.co.in

  Sovan kunmar Patnaik Directorate of Horticulture,India,PH-9338211813, E-Mail-patnaiksovankumar@yahoo.co.in

2.2 Dr.T Velmurugan[2] in 2012 analyzes the performance of two partition based clustering i.e K-means and Fuzzy C Means.He has performed the comparision by clustering of arbitrarily distributed data points.Different shapes of arbitrarily distributed data points are given as input and the number of data points in each cluster and time complexity is the output of this algoritm.According to the result he showed that the performance of K-means is better than that of Fuzzy C Means.

2.3 M.-S. YANG [3] in 1993 has performed a survey of fuzzy set theory applied in cluster analysis. He had given a survey of fuzzy clustering in three categories. The first category is the fuzzy clustering based on fuzzy relation. The second one is the fuzzy clustering based on objective function. Finally, he gave an overview of a nonparametric classifier. That is the fuzzy generalized k-nearest neighbor rule.

2.4 S.Chattopadhyay, D. K. Pratihar,, S. C. D. Sarkar[4] in 2011 has done a performance comparision of Fuzzy C Means by choosing the cluster centres virtually and another clustering called Entropy based fuzzy clustering which works on similarity threshold value .They have compared it on four data sets, such as IRIS, WINES, OLITOS and psychosis (collected with the help of forty doctors), in terms of the quality of the clusters which is, discrepancy factor, compactness, distinctness obtained and the computational time.They have also mapped the best set of clusters into 2-D for visualization using a self-organizing map (SOM).

3 K-MEANS CLUSTERING ALGORITHM

In general creates K partitions of the datasets with n objects, each partition represent a cluster, where k<=n. It tries to divide the data into subset or partition based on some evaluation criteria[5].K-means is the simplest and most widely used algorithm in many areas like image segmentation, object recognisation, etc. K-means algorithm similarly measure is based on Euclidean distance's-means is the most popular unsupervised

International Journal of Scientific & Engineering Research Volume 4, Issue 10, October-2013
ISSN 2229-5518

algorithm's-means algorithm is the most commonly used partitioning method, which uses centroid-based technique.The K-means algorithm works only for datasets that consist of numerical attributes. It takes number of desired clusters, take data points as input and produce k-clusters as output[6].In K-means algorithm n set of objects are grouped into K-clusters. Similarly measure of the cluster defined by mean value of the object in a cluster which is regarded as centroid.

1. Consider a data set having no. of objects, let k is the no. of clusters form λ set of k clusters on the data set.

2. Randomly choose k object from the data set as the initial cluster centers.The data points which are of min distance to a particular counter are assigned to that counter.

3. Update the old centers with the mean of data points assigned to that center. process is repeated until the convergence is achieved i.e. the cluster centers do not change.There are some advantages and drawbacks of k-means clustering. This algorithm is easy to implement and an handle large data set very efficiently. It can produce spherical clusters. But some of the main disadvantages of K-means algorithms are no. of cluster should be specified in advance. Algorithm is very sensitive to initial centers. There are also chances of occurrence of empty cluster. It does not converge to global optimum.

4 Fuzzy C-Means Clustering

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method developed by Dunn and improved by Bezdek is frequently used in pattern recognition.In fuzzy clustering the elements are assigned not only to one cluster but to all the clusters with certain degree of membership. This membership to groups is not hard/crisp, rather soft and and is represented by a numeric value between 0 to 1 (called, "membership grade"). Amongst various fuzzy clustering algorithms, Fuzzy C-Means (FCM) is the basic one. As it has some limitations, several algorithms have been developed further to improve its performance[7]This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis

of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. So summation of membership of each data point should be equal to one. After each iteration membership and cluster centers are updated according to the formula:It is based on minimization of the following objective function:

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

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

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

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i of cluster j,is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center

5 Particle Swarm Optimization

Particle swarm optimization is proposed by American social psychology James Kennedy and Russell Eberhart in 1995.It follows the simple basic idea that biotic population share information. The algorithm is easy to implement and converge rapidly. It can be applied when there is large number of samples. Each particle is a point of N-dimensional solution space and has a speed which is also a N-dimensional vector. Each particle has a fitness function(value) associated with it. Each particle adjusts its position and move closer to optimal point[8]PSO- is one of the swarm intelligence methods that explore global optimal solution.It is based on social behavior of birds flocking and fish schooling. It uses swarm of particles as the individuals in the population for searching through solution space.Each candidate solution is called PARTICLE and represents one individual of a population.The population is a set of vectors and is called swarm. The particle changes their components and more (fly) in a space R^2. .They can evaluate their actual position using function to be optimized.This function is called fitness function.Particles also compare themselves to their neighbors and imitate the best of that neighbor.Flexibility,Robust,self-organised having no clear leader can use post memory,swarmming nature, colonial life are benefits of PSO.

5.1 Search space D- dimensional

$X_i=[X_{i1},\ldots\ldots\ldots X_{iD}]^T$=ith particle of swarm

$V_I=[V_{i1},\ldots\ldots\ldots V_{iD}]^T$=velocity of ith particle

$P_i=[P_{i1}, \ldots\ldots\ldots P_{iD}]T$=Best previous position of ith particle.

5.2 PSO algorithm

Swarm of particles is flying through the parameter space and searching for optimum.Each particle is characterized by

Position vector…………………..$X_I(t)$

Velocity vector……………………$V_i(t)$

Each particle has individual knowledge pbest,its own as well as

-Social knowledge gbest

International Journal of Scientific & Engineering Research Volume 4, Issue 10, October-2013
ISSN 2229-5518

-Pbest of its best neighbor.

-Velocity update:

$V_i(t+1)=W*V_i(t)+C_1*rand*(pbest(t)-X_i(t))+ \quad C_2*rand*(gbest(t)-X_i(t))$

5.3 Position update

$X_i(t+1)=Xi(t)+Vi(t+1)$
Where $W>(1/2)(C1+C2)-1$
$0<W<1$

5.4 Maximal velocity

Velocity must be limited

Prevention of swarm explosion

V_{max}-If velocity of particle is greater than V or less than $-v_{max}$ it is set to V

V_{max} is saturation point of velocity.

5.5 Comments on inertial weight factor:

A large inertia weight(w) facilitates a global search while a small inertia weight facilitates a local search.By linearly decreasing the inertia weight from a relatively large value to a small value through the course of pso run gives the pso performance compared with fixed inertia weight settings0.9 to 0.4.Larger weight-greater global search ability.Smaller weight-greater local search.

6 EXPERIMENTAL RESULTS

In the experiment we have taken 3 different data sets(iris,wine and glass) for observation. By considering the algorithms for K-means and KPSO[9] as well as Fuzzy c-Means and MFPSO[10] we have taken the following observations.The details and abstract of all datasets have been shown below.

TABLE 1
DESCRIPTION OF DATASETS

Datasets	Instances	Features	No.of classes
Iris	150	4	3
Wine	178	13	3
Glass	214	10	6

The details of parameters used in PSO for KPSO and MFPSO is given in Table-2

TABLE 2
DESCRIPTION OF PARAMETERS

Type of Approach	Swarm Size	Max iteration	Acceleration(c1)	Acceleration(c2)	Inertia of weight
KPSO	20	100	1.5	1.5	0.7
MFPSO	20	100	1.49	1.49	0.7

Clustering methods can be considered as either hard or soft depending on whether a pattern belongs to exactly one cluster or to many clusters with different degrees. In hard clustering each point of the dataset belongs to exactly one cluster, a membership value of zero or one is assigned to each pattern, whereas in fuzzy clustering, a value between zero and one is assigned to each pattern by a membership function.When the number of clusters is fixed to K,then K-means clustering gives the definition of the optimization problem by finding k cluster centers and assign the objects to the nearest cluster center.It measurses the distance by means of Euclidean distance method.But K-means usually gets trapped in a local optimum,where we are perfoming each run with random initializations.In fuzzy C-means clustering, each point is associated with a weight for particular cluster.So each point has a degree of belongingness to clusters, as in fuzzy logic, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster may be in the cluster to a smaller degree than points in the center of cluster.. Like the k-means algorithm, Fuzzy c-means aims to minimize an objective function.which differs from the k-means objective function by the addition of the membership values and the fuzzifier.The fuzzifier decides the level of cluster fuzziness. A large fuzzifier results in smaller memberships.The function fuzzy c-means accepts the data set and a required number of clusters as input and returns an optimal cluster centers and membership grades for each data point. It assumes an initial guess , which marks the mean location of each cluster. These guesses are basically incorrect.Then it 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.We have also taken into account the hybridized approaches using PSO that is KPSO and modified fuzzy PSO(MFPSO)and compared them with the basic K-means and Fuzzy c- means on the basis of compactness,separability,and time complexity.KPSO algorithm is a combination of two modules i.e PSO module and K-means module.Initiallythe PSO module finds the cluster's centroid locations.These locations are used by the K-means module for finding the optimal clustering solution.Each particle in the swarm represents the data centres for the standard cluster-

International Journal of Scientific & Engineering Research Volume 4, Issue 10, October-2013
ISSN 2229-5518

ing solution.After very iteration the particle adjusts itself with the corresponding vector's position of its own experience and neighbor[9] A swarm represents a number of candidate clustering solutions for the data centroids.Each particle maintains a matrix $X_i = (C_1, C_2, ---C_i------C_k)$ where

C_i represents the ith cluster centroid vector and k is the cluster number ..The average distance between the cluster centroid

TABLE 3
COMPARISON OF COMPACTNSS(CMP)

Datasets	Clusters	K-Mean	KPSO	Fuzzy-C-means	MFPSO
iris	Cluster-1	3.635	2.915	17.0396	18.5895
	Cluster-2	4.605	3.445	18.5895	21.464
	Cluster-3	4.075	3.6	24.8017	28.5092
wine	Cluster-1	407.0233	567.4769	47.9155	52.8768
	Cluster-2	544.3163	844.7049	39.8128	46.9354
	Cluster-3	401.3728	601.0592	37.6207	40.3589
glass	Cluster-1	34.265	41.923	16.4457	18.3546
	Cluster-2	13.3399	16.7966	7.1491	9.3654
	Cluster-3	45.044	52.562	18.3736	21.3569
	Cluster-4	7.6869	6.2231	21.9647	23.1456
	Cluster-5	6.1191	4.3425	9.9829	11.2457
	Cluster-6	17.1565	14.0911	20.3145	20.9875

TABLE 4
COMPARISON OF SEPARABILITY(SEP)

Datasets	K-mean	KPSO	Fuzzy C Means	MFPSO
Iris	5.8196	10.7192	0.4821	0.5421
Wine	748.16	810.2	0.3129	0.5841
Glass	5.2713	9.256	0.4319	0.4963

and a data is used as fitness value to evaluate the solution represented by each particle .The fitness value is measured as :

$$f = \frac{1}{N}\sum_{i=1}^{N_c}\left(\sum_{k=1}^{p} d(C_k, X_{ki})/P_i\right) \qquad (1)$$

Where X_i denotes the kth data vector, which belongs to cluster I. C_{ki} is the centroid vector of ith cluster, $d(C_k, X_{ki})$ is the distance between document X_{ki} and the cluster centroid C_k, P_i stands for data number ,which belongs to cluster C_i, N_c stands for cluster number.

We find that theKPSO can performed a globalized search but requires more iterations and computations.In case of MFPSO [10] the position and velocity of particles are redefined to represent the fuzzy relation between the variables.The corresponding position matrix after updations are normalized and all the negative elements are made as zero which are re-evaluated again using the random numbers between 0 &1.We have observed that the MFPSO requires more iterations than

KPSO rather produces more optimal results and also don't get trapped in localoptima.Its performance is also increasing in case of large data sets.The optimization of the objective function of each dataset is performed using K-means,KPSO,Fuzzy C-means,Modified fuzzy PSO and the compactness and separability have been calculated..We can say that overally we have tested the performance of a soft clustering and a hard clustering to compute the differencesbetween the datasets.

The compactness of the algorithms has been calculated using the following equation(2) and(3)

$$cmp = \frac{1}{c}\sum_{1}^{c} v(c_i)/v(X) \qquad (2)$$

$$v(X) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d^2(x_i - \bar{x})} \qquad (3)$$

Where Cmp is compactness, C is number of clusters, $v(Ci)$ is the variance of clusters, $v(X)$ is the variance of datasets & x is the mean of X.

$$sep = \max(D_c) \qquad (4)$$

Where Dc is the intercluster distances.

The separability for K-means and KPSO has been calculated by the equation(5).

$$Sep = \sum_{i=1}^{k}(C_i - C_j)/k \qquad (5)$$

Where k is the number of cluster & C_i is the cluster center.

International Journal of Scientific & Engineering Research Volume 4, Issue 10, October-2013
ISSN 2229-5518

TABLE 5
CALCULATION OF TIME COMPLEXITY

Sl.no.	Number of Iterations	K-Means Time Complexity	FCM Time Complexity	KPSO Time complexity	MFPSO time complexity
1	5	3000	6000	120000	1200000
2	10	6000	12000	240000	9600000
3	15	9000	18000	360000	14400000
4	20	12000	24000	480000	76800000

TABLE 6
COMPARISON OF TIME COMPLEXITY BY VARYING NO.OF CLUSTERS

Algorithm	Time Complexity
K-Means	O(ncdi)
Fuzzy c-Means	O(ncd²i)
KPSO	O(P(ncdi)²)
MFPSO	O(P(ncd³i))

TABLE 7
COMPARISON OF TIME COMPLEXITY BY VARYING NO.OF ITERATIONS

Sl.no	Number of clusters	Time complexity of K-means	Time complexity of KPSO	Time complexity of Fuzzy-c-means	Time complexity of MFPSO
1	1	6000	240000	6000	4800000
2	2	12000	480000	24000	19200000
3	3	18000	720000	54000	43200000
4	4	24000	960000	96000	76800000

We have analyzed that the result of all the four approaches varies as per the size and number of predefined clusters of a data set. In case of iris dataset the MFPSO performs better in comparision to K-means,Fuzzy -c-means and KPSO in terms of compactness.Higher compactness results good clusters. In case of wine dataset KPSO gives better compactness in comparison to other approaches.Similarly in case of glass data set for cluster 1,2,& 3 KPSO gives better compactness and for cluster 4,5 &6 MFPSO gives better compactness respectively in comparison to other approaches.If we are going to analyze the separability,KPSO gives better separability incomparison to others.We can say that K-means converges faster in comparison to Fuzzy c-Means but here every data point in the dataset related to every cluster with a high degree of belongingness. Similarly KPSO converges faster but MFPSO gives a global optimal clustering solution. Keeping the number of data points constant we may assume that n = number of data points, c = number of cluster, d = number of dimension ,i = number of iterations and P=population size of Swarm. where n = 100,d = 4, i = 20,P=40 and varying number of clusters The following table and graph represents the comparison in details. The K-Means clustering takes as input the matrix of the corresponding data sets as well as the number of clusters.It iteratively minimises the overall sum within a cluster using squared Euclidean distances.With regards to performance, FCM may converge faster than K-Means, but needs more computational requirement as it needs to perform k multiplications for each point, for each dimension where as K-Means just needs to do a distance calculation. Fuzzy-c-means clustering is also an iterative process. The process stops when the maximum number of iterations is or when the objective function improvement between two consecutive iterations is less than the minimum amount of improvement specified.We have tested it for Iris dataset.

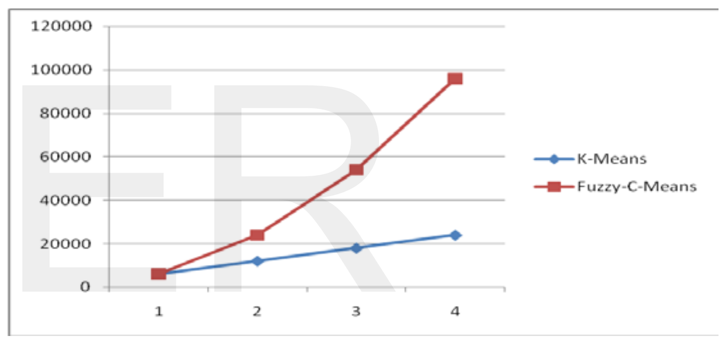

Fig 1.Comparison of time complexity of K-means and Fuzzy-C-Means by varying number of clusters

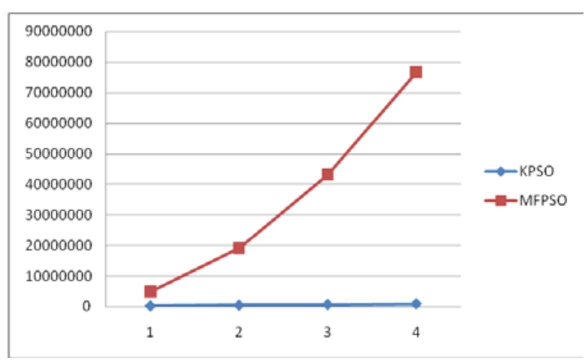

Fig 2.Comparison of Time Complexity of KPSO & MFPSO by varying number of clusters

International Journal of Scientific & Engineering Research Volume 4, Issue 10, October-2013
ISSN 2229-5518

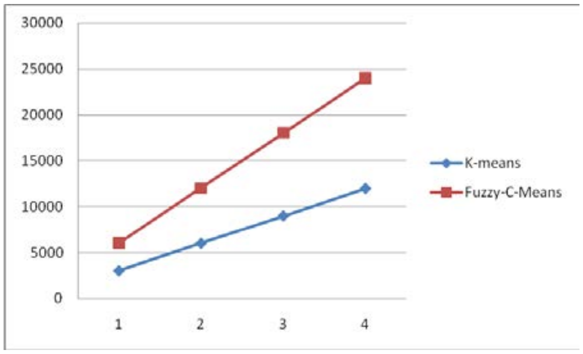

Fig 3.Comparison of Time complexity of K-means and Fuzzy-C-Means by varying number of iterations

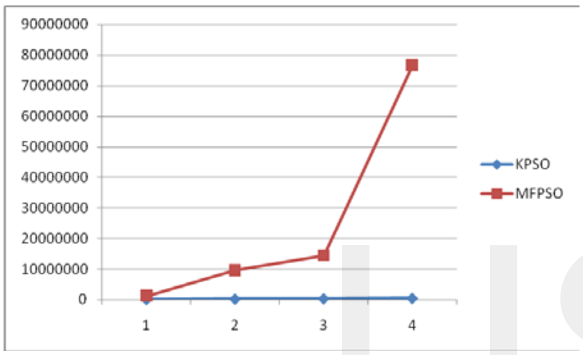

Fig 4.Comparison of Time complexity of KPSO and MFPSO by varying number of iterations

9 Conclusion

K-Means and KPSO clustering algorithm needs to define the number of cluster beforehand. K-Means algorithm is also having problems like getting trapped to local optima, sensitivity to outliers, and unknown number of iteration steps that are required to cluster. To avoid this the PSO hybridization was also considered,The time complexity of all the four algorithms was calculated. From the obtained results we may conclude that K-Means algorithm is performing better than Fuzzy-c-means algorithm and other hybrid approaches in terms of time complexity ,whereas modified fuzzy PSO is performing better in terms of optimal clustering solution.We have also analyzed that KPSO and MFPSO gives better performance in terms of compactness which is purely dependent on the dataset.Infact, Fuzzy c-means clustering is the oldest approach of software computing, which are suitable for pattern recognition, incomplete/noisy data,and media information, and can provide approximate better solutions faster,similarly MFPSO also but in a more slower manner. We have arrived at a conclusion that all the algorithms have performed well.But K-Means performs better than other algorithms in terms of speed,whereas MFPSO performs better in terms of optimal clustering solution.

References

[1] S. Ghosh,S. K. Dubey,"Comparative analysis of K-Means and Fuzzy C-Means" ((IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013.

[2] T. Volmurgan, " Austria performance comparision between K-means and Fuzzy C means" Wulfenia Journal using Arbitary data points vol-19 n0-8 aug 2012.

[3] M.-S. YANG,"A Survey of Fuzzy Clustering" Mathematical and Computer Modelling Volume 18, Issue 11, December 1993, Pages 1–16

[4]S.Chattopadhyay, S.C. D. Sarkar" A Comparative Study of Fuzzy C-means Algorithm and Entropy-Based Fuzzy Clustering Algorithms"Computing and Informatics, Vol. 30, 2011, 701–720.

[5]N. Soni1, Dr. A. Ganatra "Comparative study of several Clustering Algorithms"International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)Volume-2 Number-4 Issue-6 December-2012.

[6]M. Gupta, V. Shrivastava"Review of various Techniques in Clustering "International Journal of Advanced Computer Research (ISSN (print):2249-7277 ISSN (online):2277-7970) Volume-3 Number-2 Issue-10 June-2013.

[7] L. P. Maguluri, S. Begum, T .V .M Rao," A Novel Based Fuzzy Clustering Algorithms for Classification Remote Sensing Images "International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June 2013.

[8] C. Sethi, G. Mishra,"A Linear PCA based hybrid K-Means PSO algorithm for clustering large dataset" International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June-2013 .

[9].B.B Mishra,S.Sahoo,S.K Patnaik,Ch.D.P.Dash,"KPSO:An Evolutionary Approach for Data clustering",Proceedings of National Conference on "Future trends in information and communication Technology & Applications".

[10].S.C.Satapathy,S.K.Patnaik,Ch.D .P.Dash,S.Sahoo,"Data clustering using modified fuzzy-PSO(MFPSO)",Proceedings of 5^{th} International Workshop,MIWAI 2011,LNAI 7080 Springer.

.