Improving Performance of Fuzzy C-mean Algorithm using K- D tree Approach

[1]Jitendrasinh Raulji, [2]Vaibhavi Pandya,

[1]Assistant Professor, [2]Assistant Professor,
[1]Computer Engineering,
[1]Parul Polytechnic Institute, Vadodara, India.

Abstract: Large data are any data that you cannot load into your computer's working memory. This is not an objective definition, but a definition that is practical and one that easy to understand. For data mining and pattern recognition communities, to search data and image clustering is used as one of the primary tasks in various applications, and so, clustering algorithms that scale well to data are important and useful [1].

For improved Fuzzy C-Means and to find cluster center first randomly select the observed object and compute the density of the observed object. The density of the observed object is compared with given density parameter. It is found that the density is not less than the given density parameter, the observed object can be seen as the cluster center. For selecting the second cluster center, it is required to satisfying the above constraints. For more specify or easy input to the Fuzzy C-Mean we have tried to apply the decision tree approach.

***Index Terms* – Fuzzy C-Mean, Data mining, Decision tree.**

I. INTRODUCTION

Data mining is the pulling out of hidden, prognostic information patterns from large databases. Now-a-days Data Mining is especially when there is enormous Amount of data and identifying the useful portions of it can be a very tough task itself. Data mining allows us to be proactive about trades rather than retrospective – its means we can now try and guess the future trends before they have already taken place. [13]

Fig_1.Data Mining Process Diagram [7].

The data mining is divided in four group of techniques.

Classification: It creates group of predefined data. For example shopping mall they might attempt to classify items as cloth section, game section, food section etc.

Clustering: It is also create group of data like classification, but the groups are not predefined, some algorithm is using logic to group similar data collected.

Regression: It models the data with the least errors by some specific function.

Association rule learning: It shows the relation between different variables. For an example a supermarket it might collect data of what each buyer buys. Using Association rule they can determine which products are often bought together, which is useful for marketing purposes.

Clustering

Cluster is a set of objects that apply to the same class. In other words the related object are grouped in one cluster and not related are grouped in other cluster[5]. It group of abstract objects into classes of similar objects. When performing the cluster analysis, first we split the set of data into groups based on data likeness and then assign the name to the groups. Clustering analysis is mainly used in such applications like pattern recognition, market research, data analysis, and image processing. It can also help marketers find out distinct groups in their customer basis. They can also describe their customer groups on the base of purchasing patterns. It can help in to identifying documents on the web for fact discovery. It is also used in detection of outlier in such application as detection of credit card fraud.

C-Mean Algorithm

To be a part of cluster, every object has some degree to be part of object in fuzzy clustering, as in fuzzy logic. Rather than belonging completely too only clusters it require degree. So object at the edge of a cluster may be in the cluster to a lesser degree. For being in the kth cluster wk(x), any point x has given the degree. The centroid of a cluster is weighted by their degree of being part to the cluster. Here, we can say that the k-mean and fuzzy c-means algorithm are similar to each other [13]. It assigns coefficients to each point randomly for being in the clusters. Repeat algorithm until the process get cover all. The centroid for each cluster can be find by the formula provided by algorithm. It also computes its coefficients of being in the clusters for each point. The intra-cluster variance also be minimized by algorithm, but k-means the results depend on the initial choice of weights.

Decision Tree Approach

Learning of decision trees from class –labeled training tuples is explored by decision tree approach. Like tree structure the decision tree is flow chart, where test on attribute is denoted by each internal node, an outcome of the test is denoted by each branch, and Class label is hold by a leaf node. The root node is topmost node. Decision trees are commonly involved in the operations like research, analysis of decision, to identify a plan most likely to reach a target. By a probability model a decision tree should be paralleled as online choice model algorithm. To calculating conditional probabilities the decision trees is as a descriptive means for use.

II. PROBLEM STATEMENT

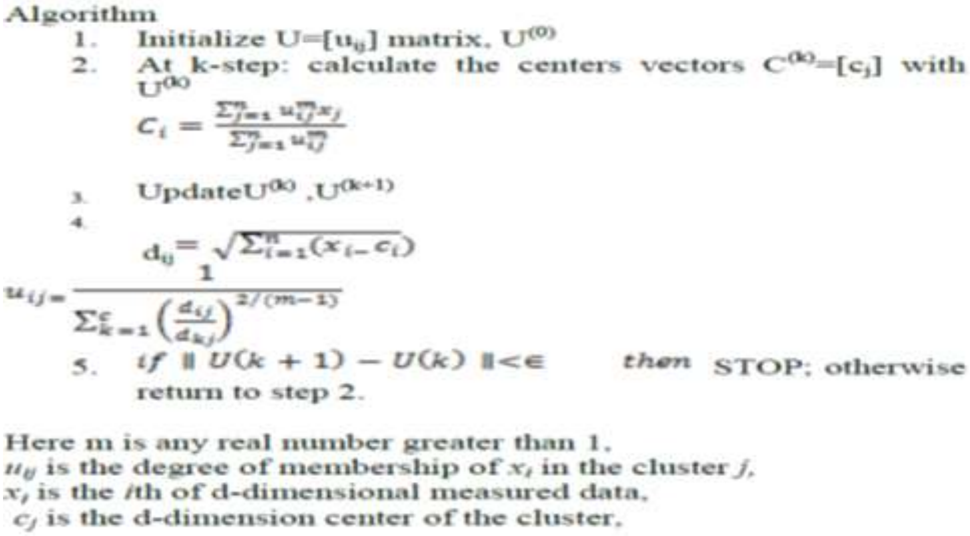
The cluster center in traditional FCM algorithm is determines randomly. This method is so simple and applicable to all data, but sometime it causes high iteration and squared errors. This problem may cause time consuming problem, and in this fast world it may generate business risk for the company's welfare. And to making quick decision at right time it is required to have some after algorithm that make cluster in less time.

III. OBJECTIVES AND GOALS

Initially we did our literature survey on sub areas clustering in date mining. Then we decided the base algorithm we are going to work on some objectives like, to do more analysis on our proposed method of Clustering, to design and develop a concept that minimizes the overall generation time of cluster, To evaluate the performance of the proposed work.

IV. EXISTING TECHNIQUE

The current research shows the comparison between two important clustering algorithms namely centroid based K-Means and representative object based FCM (Fuzzy C-Means). And the results represents that the FCM generates results close to K-Means clustering. It is also observed that FCM takes more computation time than K-Means clustering [1].



```
Algorithm
    1.  Initialize U=[$u_{ij}$] matrix, $U^{(0)}$
    2.  At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$
        $C_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}}$
    3.  Update$U^{(k)}$ ,$U^{(k+1)}$
    4.
        $d_{ij} = \sqrt{\sum_{i=1}^{n}(x_i - c_i)}$
$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}}$
    5.  $if \parallel U(k+1) - U(k) \parallel < \epsilon$   $then$ STOP; otherwise
        return to step 2.
```

Here m is any real number greater than 1.
u_{ij} 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.

Fig_2.Fuzzy C-Mean Algorithm [1].

V. PRAPOSED TECHNIQUE

In proposed work the Fuzzy Means algorithm should be improved in terms of its accuracy and efficiency. It may be enhance the output of fuzzy c means clustering algorithm by using decision tree approach with it which mine the data in accurate and sequential manner. Here some steps that shows the overall flow of the work.

Step_1. We shall be taking the dataset on which the clustering is will be apply.

Step_2. Applying same dataset for the other clustering algorithm.

Step_3. Generate the Results.

Step_4. Applying the decision tree Approach in same dataset.

Step_5. Process the result of decision tree by the Fuzzy C Mean algorithm.

Step_6. Generate the Results.

Step_7. Comparing the both result of Step_3 and Step_7.

Step_8. Showing the analysis.

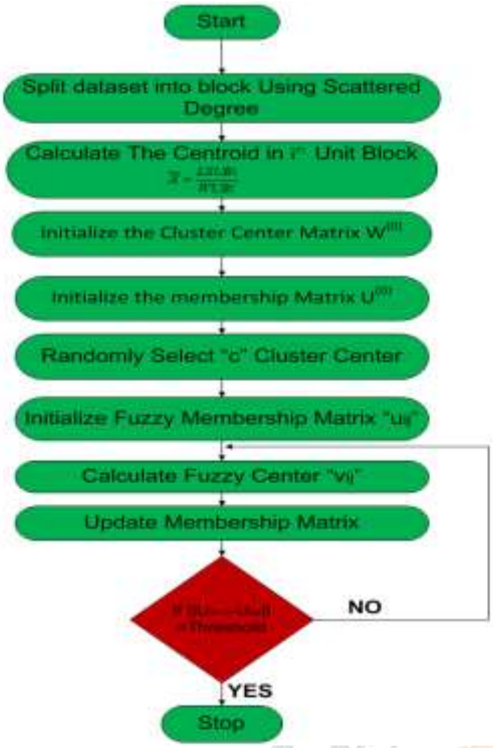

Start
Split dataset into block Using Scattered Degree
Calculate The Centroid in i^{th} Unit Block
Initialize the Cluster Center Matrix W
Initialize the membership Matrix U
Randomly Select "c" Cluster Center
Initialize Fuzzy Membership Matrix "uij"
Calculate Fuzzy Center "vij"
Update Membership Matrix
NO
YES
Stop

Fig_3. A flow chart of proposed system

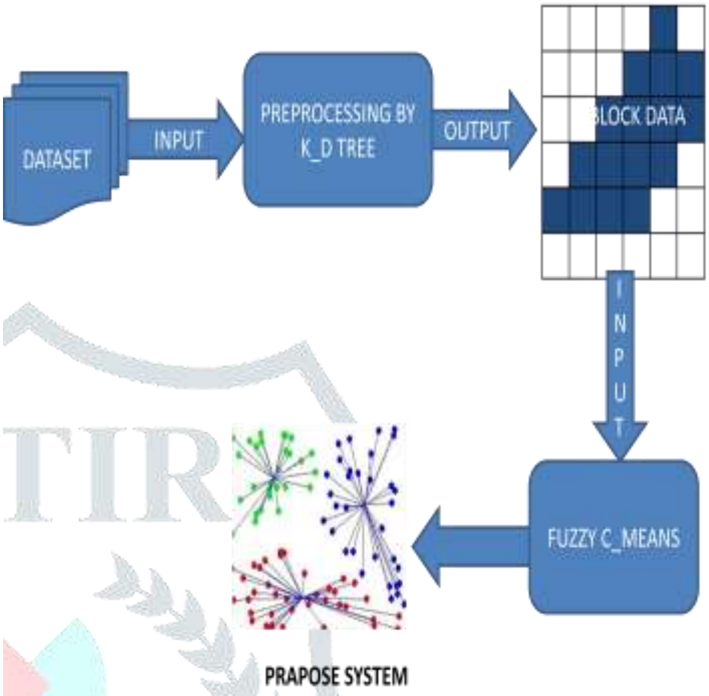


Fig_4. Proposed System

VI. IMPLEMENTATION ON MAT LAB

MATLAB is multi-paradigm numerical computing environment. It is a 4G language of programming which is developed by Math Works. Implementation of algorithm and plotting of function, development of user interfaces is carry out by MATLAB. Matrix manipulation is also done by MATLAB [15]. The numerical computing can be done by MATLAB.

VII. DATA SET DESCRIPTION

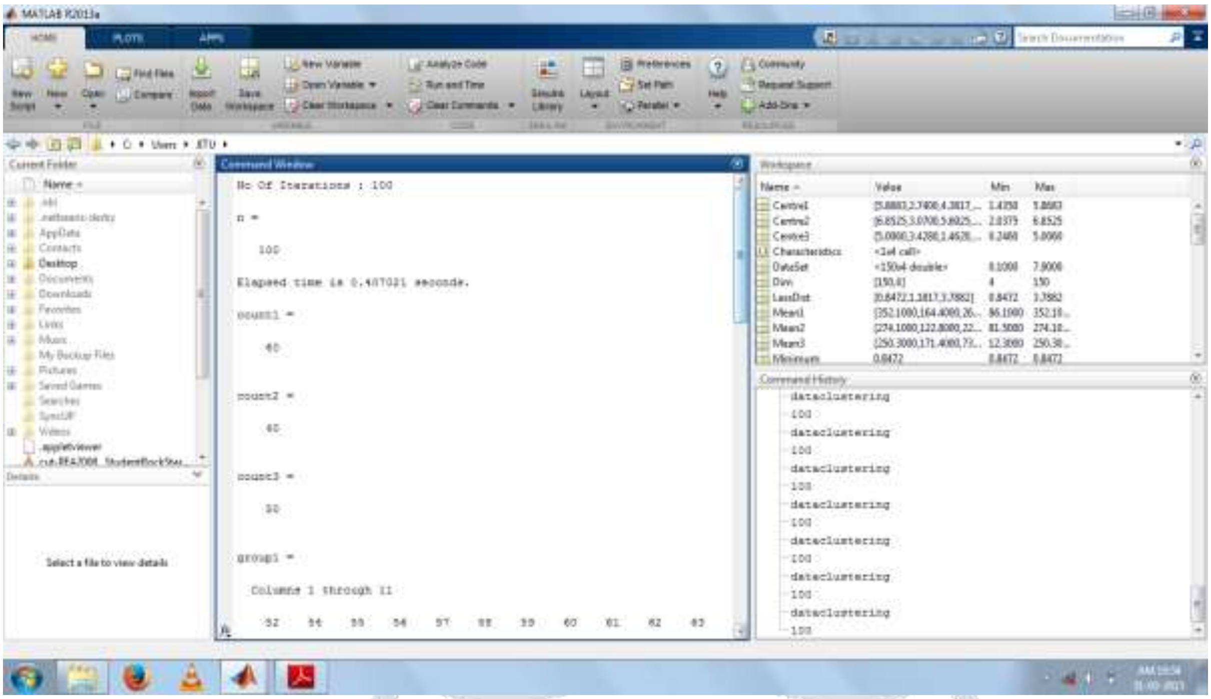
Data Input: - Iris Dataset This dataset is based on three flower named Virginia, setose, versicolor. Here, they have taken the data of sepal length, sepal width in numerical value. Output:-The output will generates the cluster of given data when any of the existing or proposed algorithm will apply.

VIII. PERFORMANCE EVOLUTION

The fuzzy c_means algorithm will generates the cluster and will measure the time required for generating the cluster and we also see the performance of the algorithm. We also apply our proposed method on same dataset and measure the time for generating the cluster and compare both of them on the base of time.

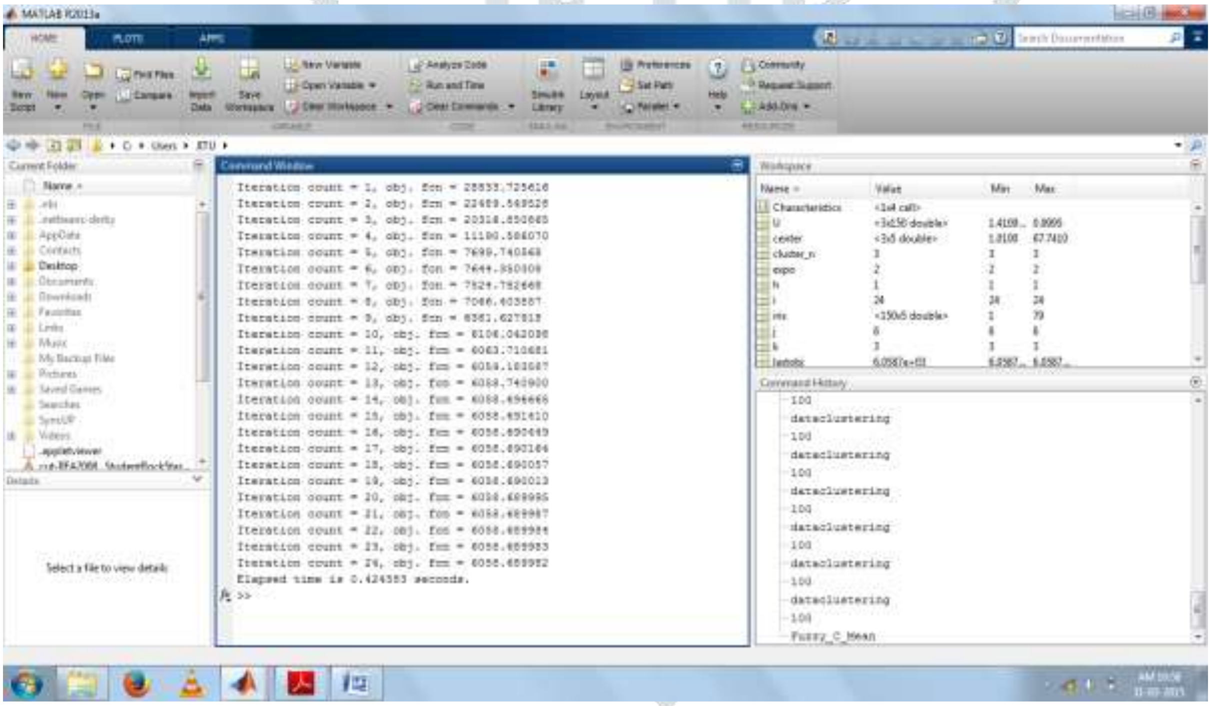
IX. IMPLEMENTATION

Here, we have implemented algorithms in MATLAB. A K-means Clustering algorithm to compare with FCM clustering algorithm and tried to generate results to compare.



```
No Of Iterations : 100

n =

   100

Elapsed time is 0.487021 seconds.

count1 =

    40


count2 =

    60


count3 =

    50


group1 =

  Columns 1 through 11

    53    54    55    56    57    58    59    60    61    62    63
```

Fig_5. Elapsed Time for Cluster Generation of K-mean Algorithm



```
Iteration count = 1, obj. fcn = 28833.725616
Iteration count = 2, obj. fcn = 22489.549528
Iteration count = 3, obj. fcn = 20318.850685
Iteration count = 4, obj. fcn = 11180.586070
Iteration count = 5, obj. fcn = 7699.740568
Iteration count = 6, obj. fcn = 7644.850308
Iteration count = 7, obj. fcn = 7524.752668
Iteration count = 8, obj. fcn = 7066.403587
Iteration count = 9, obj. fcn = 6561.627818
Iteration count = 10, obj. fcn = 6106.042086
Iteration count = 11, obj. fcn = 6063.710681
Iteration count = 12, obj. fcn = 6058.183687
Iteration count = 13, obj. fcn = 6058.740900
Iteration count = 14, obj. fcn = 6058.696666
Iteration count = 15, obj. fcn = 6058.691610
Iteration count = 16, obj. fcn = 6058.690469
Iteration count = 17, obj. fcn = 6058.690164
Iteration count = 18, obj. fcn = 6058.690057
Iteration count = 19, obj. fcn = 6058.690013
Iteration count = 20, obj. fcn = 6058.689996
Iteration count = 21, obj. fcn = 6058.689987
Iteration count = 22, obj. fcn = 6058.689984
Iteration count = 23, obj. fcn = 6058.689983
Iteration count = 24, obj. fcn = 6058.689982
Elapsed time is 0.424383 seconds.
>>
```

Fig_6. Elapsed Time for Cluster Generation of FCM Algorithm

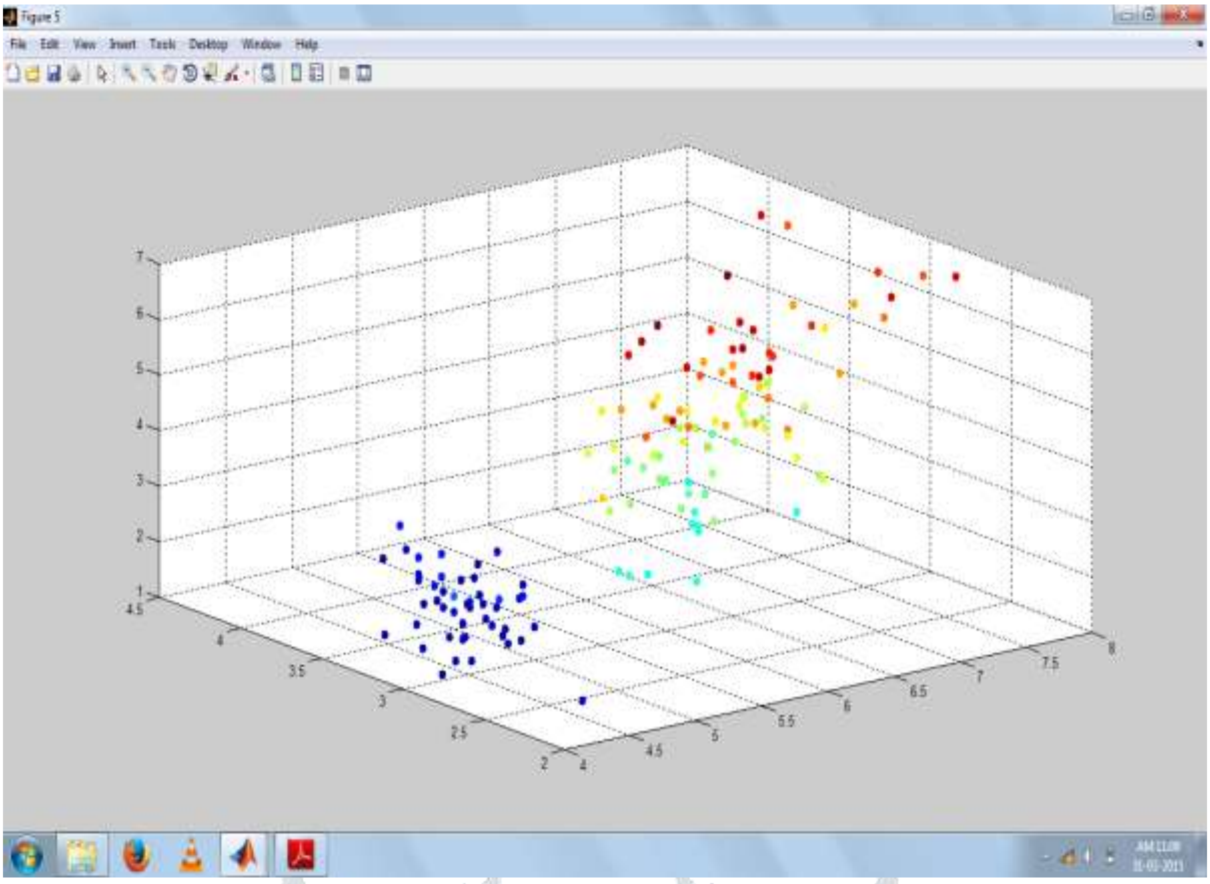


Fig_7. Graph of K-Means with iris dataset for clusters A

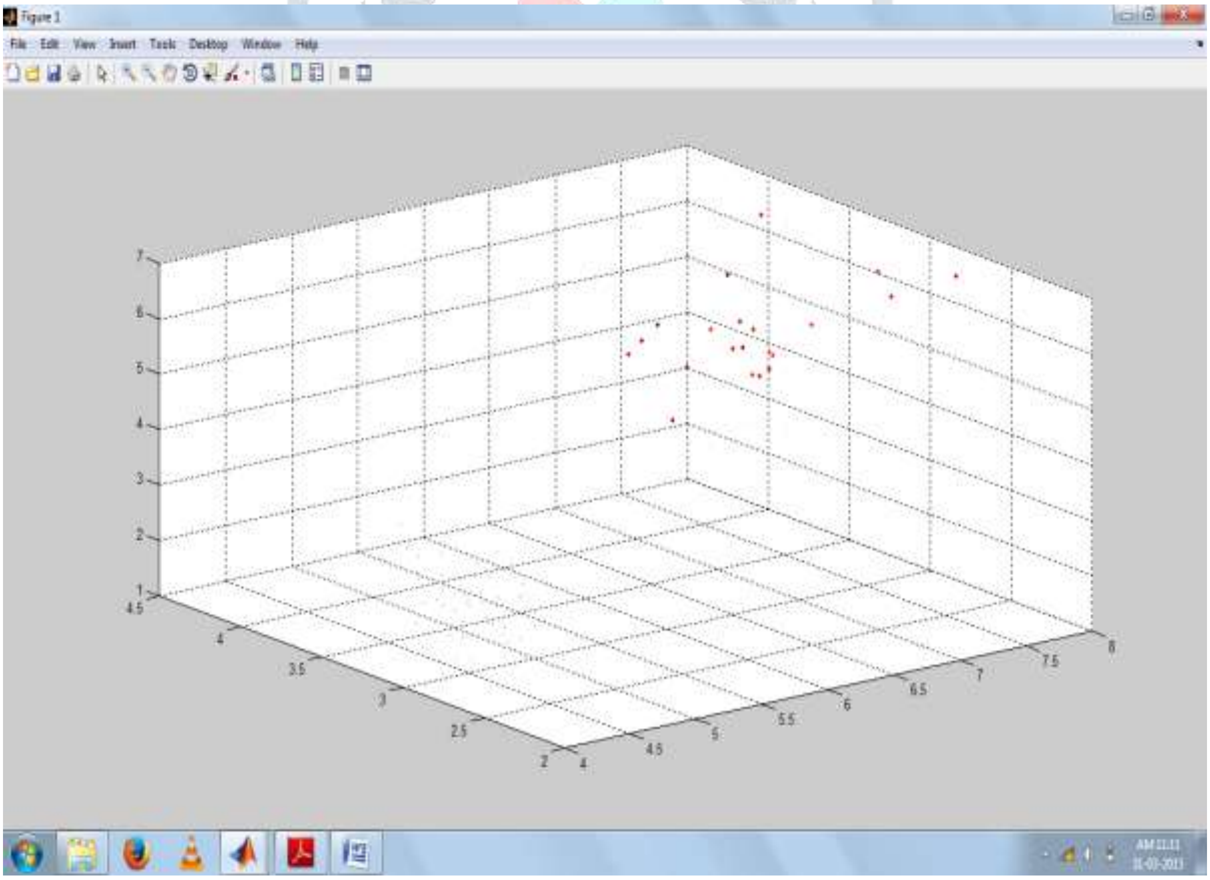


Fig_8. Graph of K-Means with iris dataset for clusters B

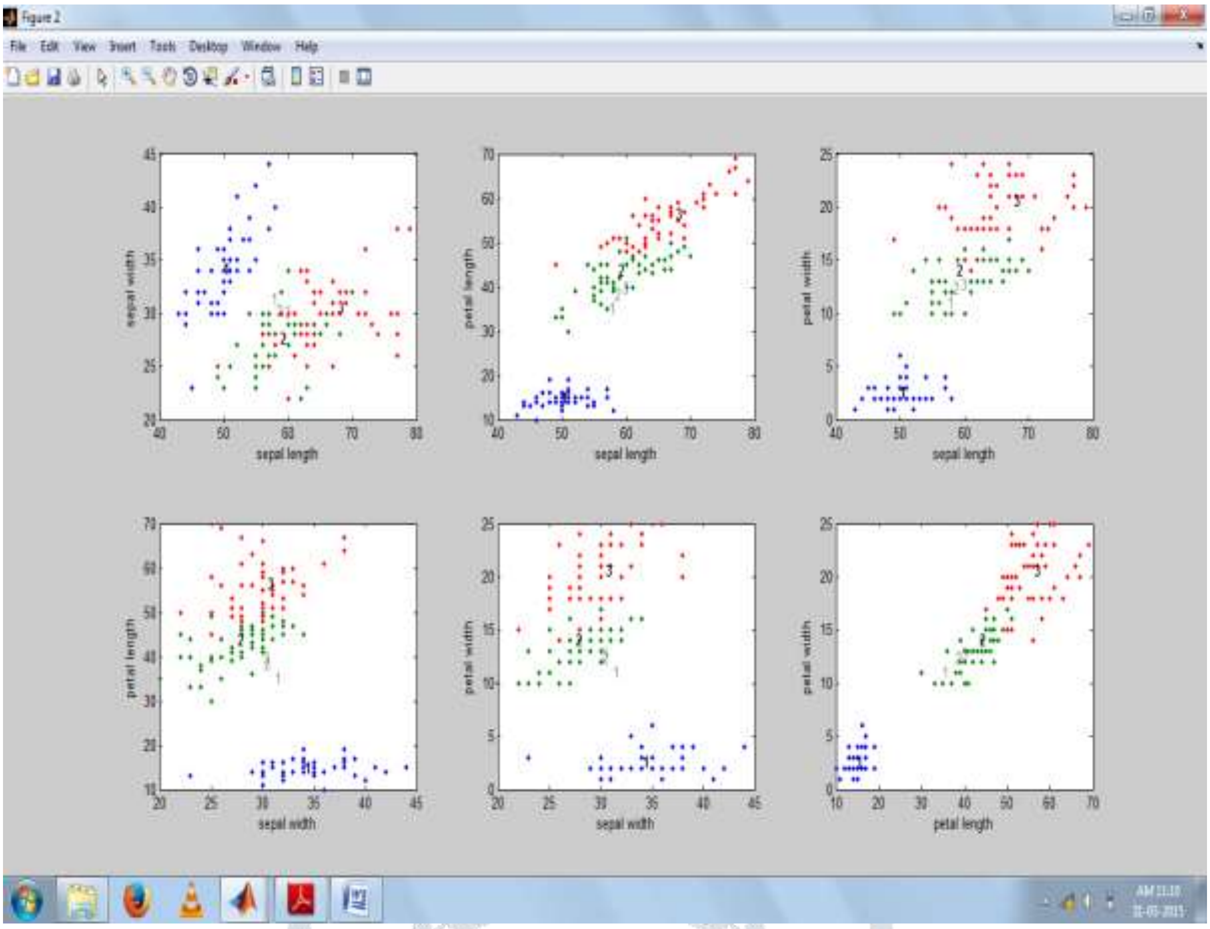
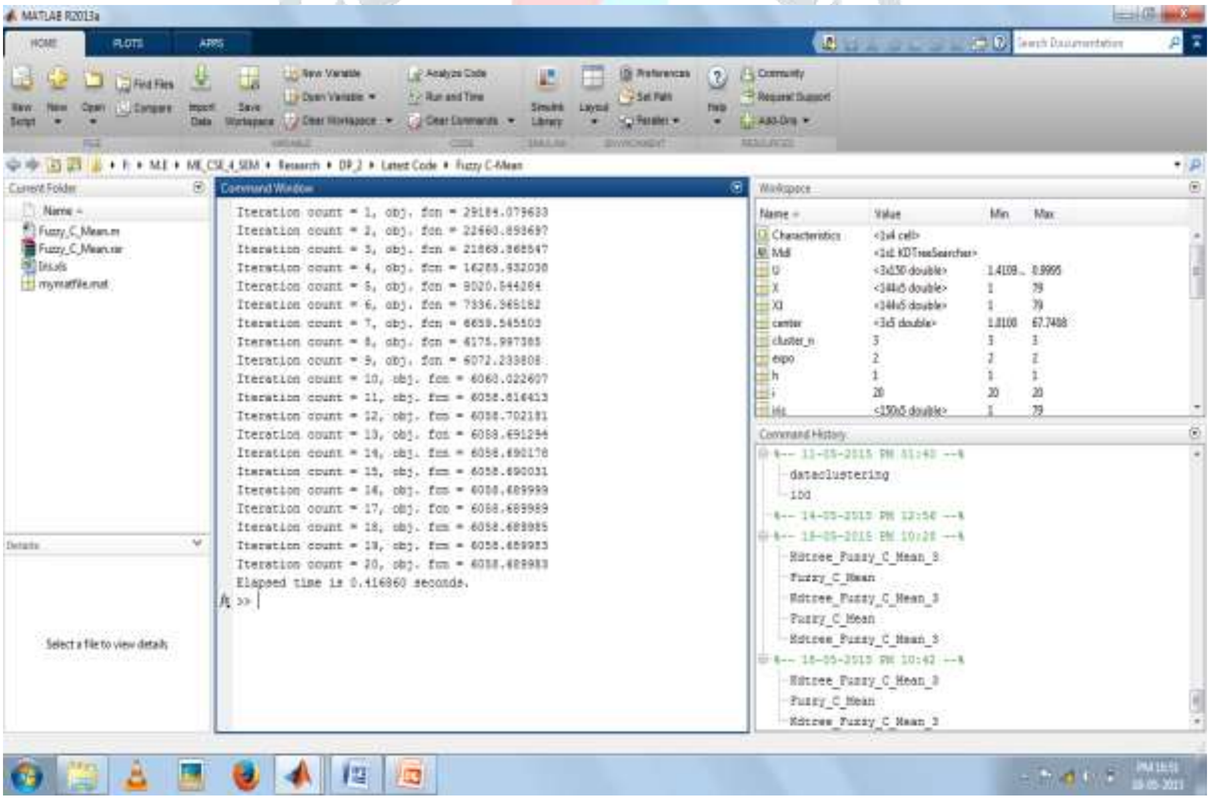


Fig_9. Graph of Fuzzy C-Means with iris dataset for three clusters


MATLAB R2013a
HOME PLOTS APPS
New Script New Open Find Files Compare Import Data Save Workspace New Variable Open Variable Clear Workspace Analyze Code Run and Time Clear Commands Simulink Library Layout Preferences Set Path Parallel Help Community Request Support Add-Ons
F: ▸ M.E ▸ ME_CSE_4_SEM ▸ Research ▸ DP_2 ▸ Latest Code ▸ Fuzzy C-Mean
Current Folder
Name
Fuzzy_C_Mean.m
Fuzzy_C_Mean.asv
Iris.xls
mymatfile.mat
Details
Select a file to view details
Command Window
Workspace
Command History

```
Iteration count = 1, obj. fcn = 29184.079633
Iteration count = 2, obj. fcn = 22660.893697
Iteration count = 3, obj. fcn = 21868.868547
Iteration count = 4, obj. fcn = 16285.932030
Iteration count = 5, obj. fcn = 9020.844284
Iteration count = 6, obj. fcn = 7336.966182
Iteration count = 7, obj. fcn = 6638.545503
Iteration count = 8, obj. fcn = 6175.997385
Iteration count = 9, obj. fcn = 6072.233808
Iteration count = 10, obj. fcn = 6060.022607
Iteration count = 11, obj. fcn = 6058.816413
Iteration count = 12, obj. fcn = 6058.702181
Iteration count = 13, obj. fcn = 6058.691294
Iteration count = 14, obj. fcn = 6058.690178
Iteration count = 15, obj. fcn = 6058.690031
Iteration count = 16, obj. fcn = 6058.689999
Iteration count = 17, obj. fcn = 6058.689989
Iteration count = 18, obj. fcn = 6058.689985
Iteration count = 19, obj. fcn = 6058.689983
Iteration count = 20, obj. fcn = 6058.689983
Elapsed time is 0.416860 seconds.
>>
```

Name	Value	Min	Max
Characteristics	<1x4 cell>		
Mdl	<1x1 KDTreeSearcher>		
U	<3x150 double>	1.4108...	0.9995
X	<144x5 double>	1	79
X1	<144x5 double>	1	79
center	<3x5 double>	1.8100	67.7408
cluster_n	3	3	3
expo	2	2	2
h	1	1	1
i	20	20	20
iris	<150x5 double>	1	79

```
%-- 11-05-2015 PM 11:40 --%
datadlustering
idd
%-- 14-05-2015 PM 12:50 --%
%-- 18-05-2015 PM 10:28 --%
Kdtree_Fuzzy_C_Mean_3
Fuzzy_C_Mean
Kdtree_Fuzzy_C_Mean_3
Fuzzy_C_Mean
Kdtree_Fuzzy_C_Mean_3
%-- 18-05-2015 PM 10:42 --%
Kdtree_Fuzzy_C_Mean_3
Fuzzy_C_Mean
Kdtree_Fuzzy_C_Mean_3
```

Fig_10. Time complexity of proposed algorithm at first time

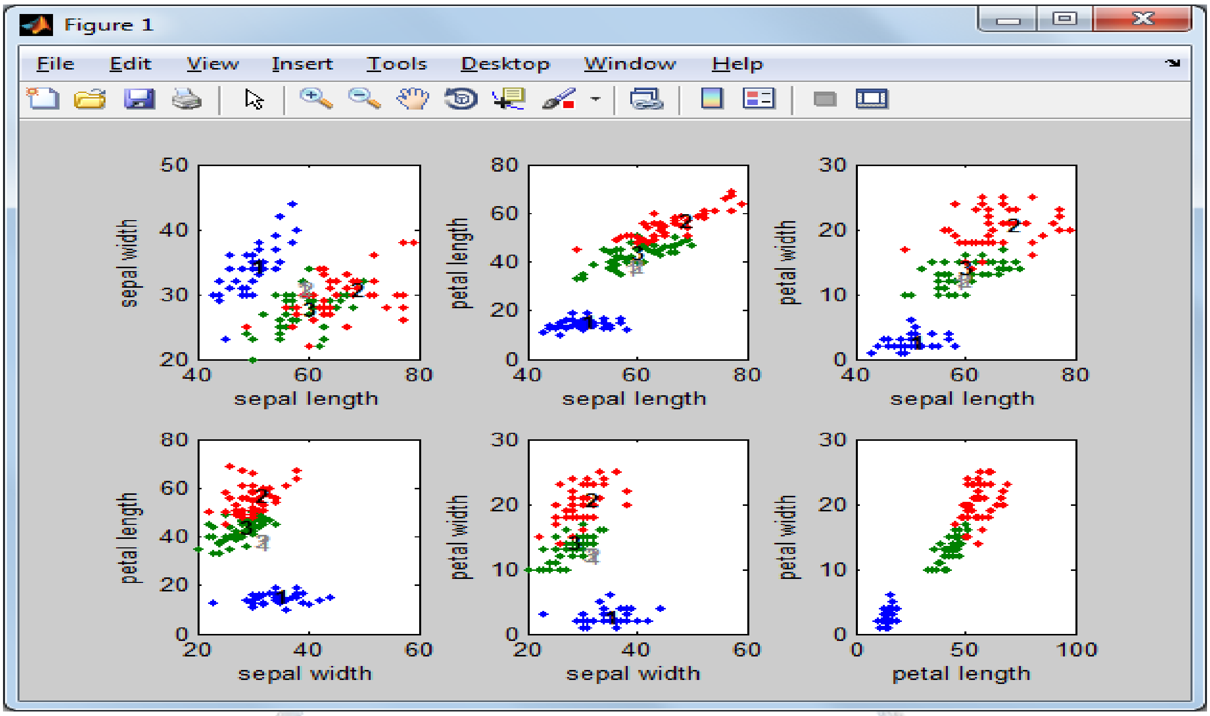


Fig_11. Cluster generated proposed algorithm

X. RESULTS AND DISCUSSION

Table 1: Performance comparison of K-Means and FCM Algorithms [1].

Algorithm	Time Complexity	Elapsed Time(Seconds)
K-Means	O(ncdi)	0.443755
FCM	O (ncd^{2}i)	0.781679

Table 2: Performance comparison of K-Means, FCM Algorithm and Proposed Algorithm.

Algorithm	Time Complexity	Elapsed Time(Seconds)
K-Means	O(ncdi)	0.443755
FCM	O (ncd^{2}i)	0.781679
Pro-FCM	O (ncd^{2}i)	0.416860

Table 3: Performance comparison of FCM Algorithm and Proposed Algorithm with variation in cluster number.

Sr. No.	Number of cluster	FCM Time Complexity	Proposed FCM time Complexity
1	1	6000	5800
2	2	24000	22000
3	3	54000	51000
4	4	96000	94000
:	:	:	:

Table 4: Performance comparison of FCM Algorithm and Proposed Algorithm with variation in number of iteration.

Sr. No.	Number of Iteration	FCM Time Complexity	Proposed FCM time Complexity
1	1	6000	5000
2	2	12000	12000
3	3	18000	17000
4	4	24000	22000
:	:	:	:

Graphical representation for performance of FCM and PFCM

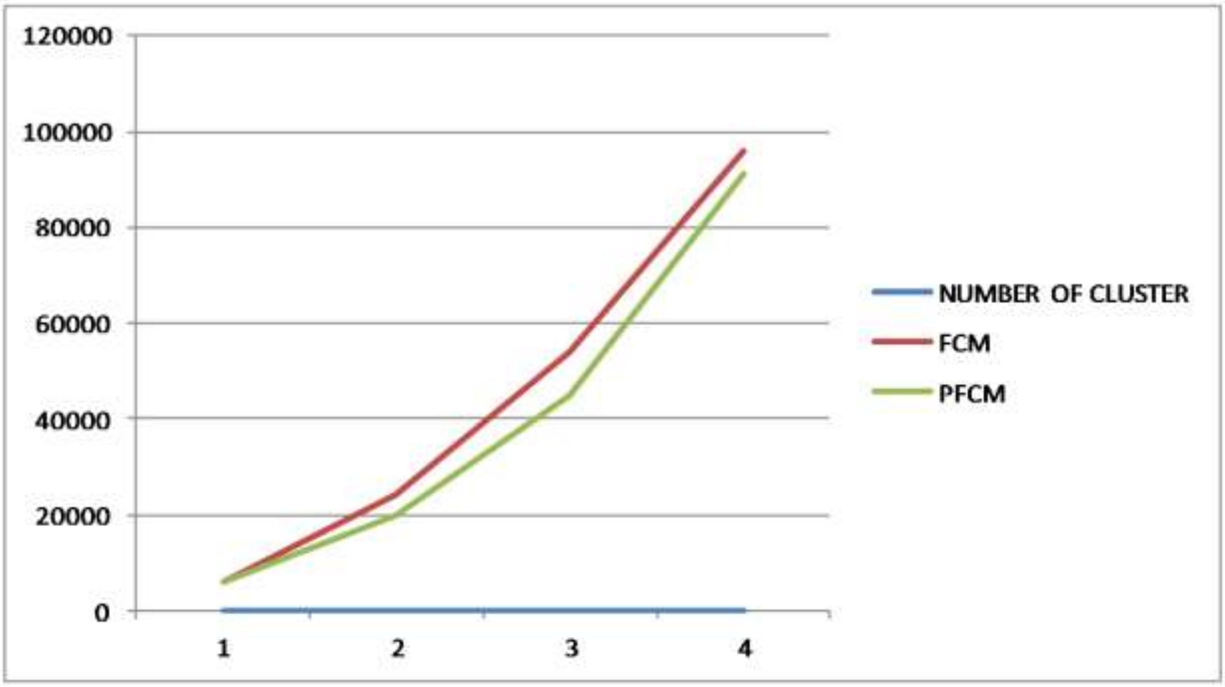


Fig_12. Graphical comparison for number of Cluster

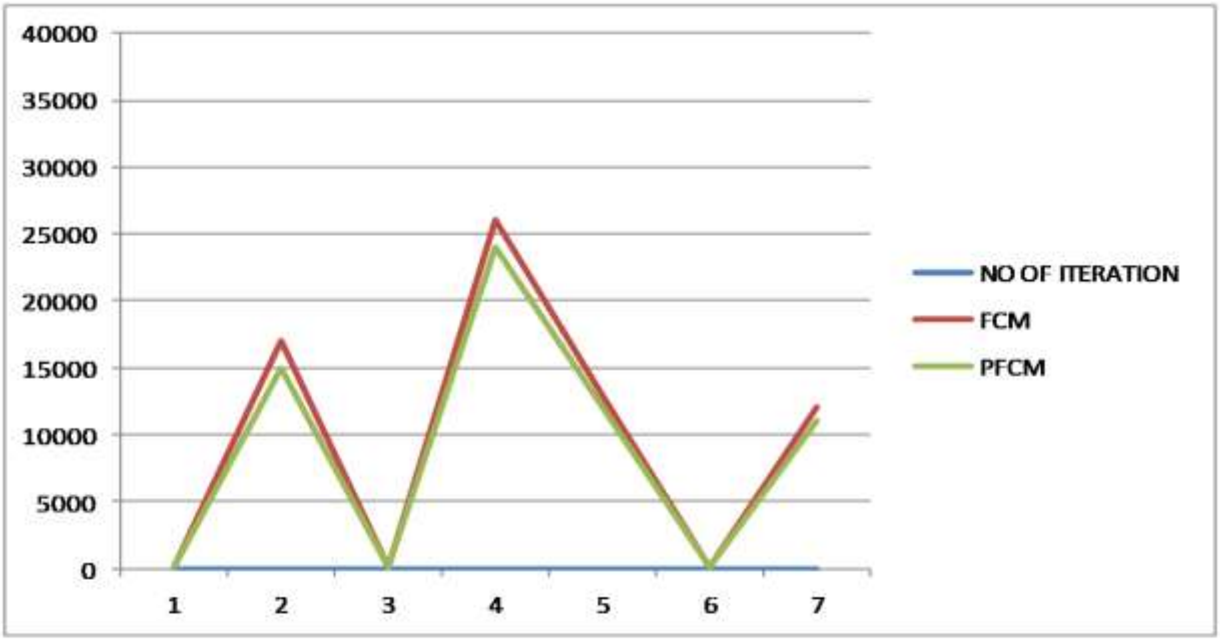


Fig_13. Graphical comparison for number of Iteration

XI. Conclusion

As conclusion fuzzy c means consumes more computation time than other clustering techniques and it may be more faster by pre-processing the input using K_D Tree approach where it provides more specified data in the manner that our fuzzy c means algorithm works more accurate and efficient.

References

[1] Tejwant Singh, Mr. Manish Mahajan,IJARCSSE ,VOLUME 4,2014," Performance Comparison of Fuzzy C Means with Respect to Other Clustering Algorithm".

[2] Timothy C. Havens, Senior Member, IEEE, James C. Bezdek, Life Fellow,IEEE, Christopher Leckie," Fuzzy c-Means Algorithms for Very Large Data", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 20, NO. 6, DECEMBER 2012

[3] Lawrence O. Hall, Fellow, IEEE, and Marimuthu Palaniswami, Fellow, IEEE Ranshul Chaudhary, Prabhdeep Singh, Rajiv Mahajan" A SURVEY ON DATA MINING TECHNIQUES",IJARCCE Vol. 3, Issue 1, January 2014.

[4] Rui Xu, Student Member, IEEE and Donald Wunsch II, Fellow, IEEE" Survey of Clustering Algorithms", IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2010.

[5] Bassam M. El-Zaghmouri ,Marwan A. Abu-Zanona " Fuzzy C-Mean Clustering Algorithm Modification and Adaptation for Applications ",World of Computer Science and Information Technology Journal (WCSIT), 2012.

[6] Prof. Neha Soni , Prof. Amit Ganatra ,"Categorization of Several Clustering Algorithms from Different Perspective: A Review ", International Journal of Advanced Research in Computer Science and Software Engineering ,Volume2, Issue 8, August 2012.

[7] Soumi Ghosh and Sanjay Kumar Dubey,amity university „"Comparative Analysis of K-Means and Fuzzy C-Means Algorithms" ((IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013.

[8] Mrs. Bharati R.Jipkate and Dr.Mrs.V.V.Gohokar "A Comparative Analysis of Fuzzy C-Means Clustering and K Means Clustering Algorithms" International Journal Of Computational Engineering Research / ISSN: 2250–3005

[9] R.Suganya,R.Shanthi "Fuzzy c-Means Algorithm-A Review",(IJSRP) International Journal of Computer Science and Research Publication,Vol. 2,Issue 11.Nov 2012.

[10] B. Liu, W. Hsu, and Y. Ma. Mining association rules with multipleMinimumm Supports. Pages 337–341. ACM Special Interest Group on Knowledge Discovery and Data Mining Explorations, 1999.

Web Links

[11] Http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace

/datamining.htm.

[12] http://www.bindichen.co.uk/post/AI/fuzzy-c-means.html.

[13] http://www.techonthenet.com/information-mining/weka

[14] http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm.