IJITKM Volume 7 • Number 2 Jan– June 2014 pp. 173-179 (ISSN 0973-4414)

SURVEY OF META CLUSTERING ALGORITHM

M Santhosh kumar[1] and Dr.Supreethi K.P[2]

[1]Student, M.Tech, Jawaharlal Nehru Technological University, Hyderabad, Telangana, India [2]Assistant Professor, Jawaharlal Nehru Technological University, Hyderabad, Telangana, India

santoshm91@gmail.com,supreethi.pujari@gmail.com

ABSTRACT

Meta clustering is a Clustering of clustering's that groups similar base-level clustering's together. Clustering algorithm belonging to the domain of soft computing. Meta clustering algorithm is using applications of fuzzy c-means algorithm to create soft clustering. Fuzzy c-means is a prominent example for such soft computing clustering algorithms with many effective real life applications. Clustering and Meta clustering's are two main extensions of the clustering problems. Given dataset as a input clusterings,in that clustering aim is to find a single final clustering which is better than some existing clustering's and meta clustering aim is creating a new mode of interaction between users, clustering system and the data. Fuzzy members are obtained by the applications of previous clusters. This fuzzy member granular is mainly derived from the events such as transactions, phone calls, user sessions and security breaches. Fuzzy membership extends Meta clustering algorithm based on crisp k-means clustering.

1. INTRODUCTION

Most clustering methods useful for finding optimal or near optimal clustering's which are based on specific clustering criteria.Meta clustering is mainly a new approach to the problem of clustering. Meta clustering main aim is creating a new mode of interaction between users, clustering system and the data [1][13]. one optimal clustering data will finds many alternate good clustering's of data that allows the user to select which clustering is more useful and exploring the space of reasonable clustering.meta clustering is nothing but clustering of clustering's will be groups similar base level clustering which are placed together. It will make it easier for users and to evaluate the clustering's.

Meta clustering can be divided in to 3 fundamental steps:

1. Generate many good, yet qualitatively different, base-level clustering's of the same data.
2. Measure the similarities between the base-level clustering's will generated in the first step. So that similar clustering's can be grouped together.
3. Organize the base-level clustering's at a Meta level and present them to users.

Generally, clustering algorithms can be classified in two categories that are hard clustering and soft (fuzzy) clustering. Hard clustering is data's are divided into distinct clusters, where each data element will be belongs to exactly one cluster. Soft clustering, data elements will be belongs to more than one cluster and which are associated with each element of set of membership levels.

Cluster analysis is a partition of set of objects in to clusters, such objects within cluster should be similar to each other and also different cluster of objects are should be dissimilar with each other[15].Clustering is like mathematical tool that useful to discover structures or certain patterns in a dataset. Cluster analysis is not an automatic process but an iterative process. It is necessarily to modify the preprocessing and parameter until the result achieved [16].

2. CLASSIFICATION OF CLUSTERS

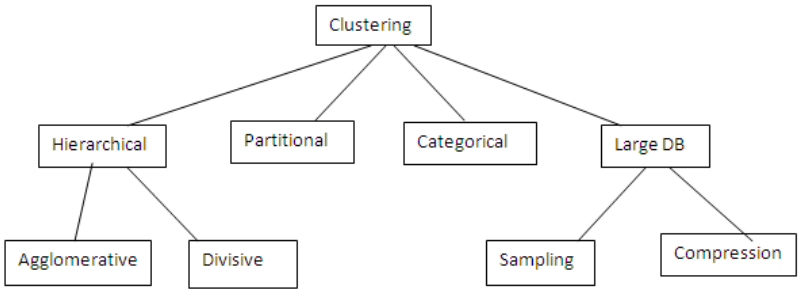

Fig: 1. Classification of clustering Algorithm [8]

2.1 Hierarchical clustering

It works by grouping data objects into a tree of clusters. Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on the decomposition is formed in a bottom-up(merging) or top-down(splitting).i.e. if a particular merge or split decision later turns out to have been a poor choice, the method cannot backtrack and correct it.[11]

2.1.1Agglomerative and divisive hierarchical clustering

This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters. Until all of the objects are in a single cluster or until certain termination conditions are satisfied. Divisive hierarchical clustering is top-down strategy will do the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster. It subdivides the cluster into smaller and smaller pieces, until each object forms a cluster on its own. [11]

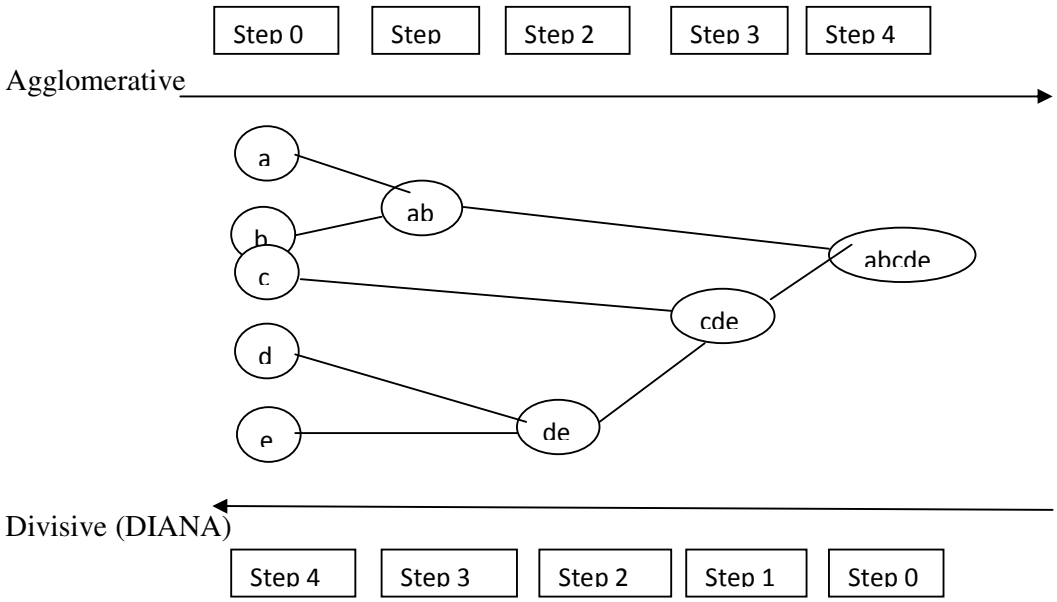

Fig: 2.Agglomerative & divisive hierarchical clustering on data objects (a,b,c,d,e);

3. FUZZY CLUSTERING ALGORITHM

In clustering, fuzzy clustering algorithm is one of the most widely used algorithms. Fuzzy set theory was mainly first proposed by Zadeh in 1965.it was described by a membership function. The fuzzy set use is provides the class membership function in fuzzy clustering, the non-unique partitioning of data into a collection of clusters. The data is assigned membership values for each cluster and fuzzy clustering algorithm is allows the clusters to grow into their natural shapes **[17]**

In fuzzy clustering, objects are not assigned to a specific cluster. They possess a membership function indicates to the strength of membership in all or some of the clusters. but most of clustering techniques described as, 'strength of membership' has been either 0 or 1, with an object would be either in or not in a cluster, except in case of the mixture. In fuzzy clustering jargon, methods where strength of membership is 0 or 1 are known as crisp methods.

Fuzzy clustering has two main advantages over crisp methods.[2]

- Memberships can be combined with other information also. In particular, memberships are probabilities; results can be combined from the different sources by using Bayes' theorem.
- Memberships for any given object indicate whether there is a second best cluster. a fuzzy cluster analysis, the number of subsets assumed is known, and the membership function of each object in every cluster is estimated using a iterative method, but membership functions do not obey the rules of probability theory although, if it once found, memberships can be scaled to lie between 0 and 1, and can then be interpreted as probabilities.

Fuzzy clustering algorithms can be divided in 2 types.

i. Classical fuzzy clustering algorithms
ii. Shaped based fuzzy clustering algorithms.

Again Classical fuzzy clustering algorithm divides in 3 types.
i. The fuzzy c-means algorithm
ii. The Gustafson-kessel algorithm
iii. The Gath-Geva algorithm

And also shape based Fuzzy clustering divides in 3 types
i. Circular shape based clustering algorithm
ii. Elliptical shape base clustering algorithm
iii. Genetic shape base clustering algorithm

3.1. Fuzzy k-means:

K-means or hard c-means clustering is partitioning methods which are applied to analyze data and treats observations of data as objects based on locations and distance between the various input data. Each cluster is specialized by its center point.ie "centroid" means centroid based. This k-means mainly finds the desired number of distinct clusters and their centroids **[14].**k-means clustering is also appropriate for larger tables, up to hundreds of thousands of rows. It makes a fairly good guess at cluster seed points. Then it starts an iteration of alternately assigning the points to clusters and recalculating cluster centers. *K*-Means clustering only supports numeric columns. *K*-Means clustering ignores model types (nominal and ordinal), and treat all numeric columns as continuous columns.

3.1.1Steps for k-means clustering algorithm:

1. Set k: choose a number of desired clusters, K.

2. **Initialization:** choose k-starting points for initial estimates of the cluster centorids.

3. Classification: To examine each point of dataset then assign it to cluster which centroid is nearer to it.

4. Calculation of centroid: Each point of dataset is assigned to a cluster. Its need to recalculate the new k-centroids.

5. Criteria of convergence: step 3,4 requires repeated until no point will changes its cluster assignment or until centroids no longer move.

3.2 Fuzzy c-means:

Fuzzy c-means is very natural than hard clustering. This fuzzy c-means algorithm mostly used in widely. It was developed by james Bezdek [17].PFCM is a hybridization of possibilistic c-mean(PCM) and FCM that avoids the various problems of PCM,FCM and FPCM.FCM employees the fuzzy partitioning such that a data point can be belongs to all the groups with different membership grades between 0 to 1.**[16]**This algorithm works with assigning membership of each data point corresponding to each cluster center and data point. Mostly this data cluster center is near to its membership towards of particular cluster center.

Advantages:

1. Unsupervised
2. Converges

Limitations:

1. Long computation time occurs.
2. Sensitivity to initial guess

PCM is a new clustering model, which is useful to overcome the difficulties of FCM.

3.3 Rough Sets to Clustering:

Traditionally, clustering partitions will be in group of objects into a number of non-overlapping sets based on similarities. Sometimes the boundaries of these sets or clusters may not be clearly defined. Some objects may be almost equidistant from center of multiple clusters. Traditional set theory is explains that these objects be assigned to a single cluster. This rough set theory is used to represent the overlapping of clusters. Rough sets are more flexible representation than conventional sets and also at this same it is less descriptive than the fuzzy sets. Mainly rough clustering based on k-means, genetic algorithms, kohonen self-organizing maps **[18].**It is includes a review of rough cluster validity measures, applications of rough clustering to forestry,medicine,web mining, super markets and traffic engineering**[19][20].**

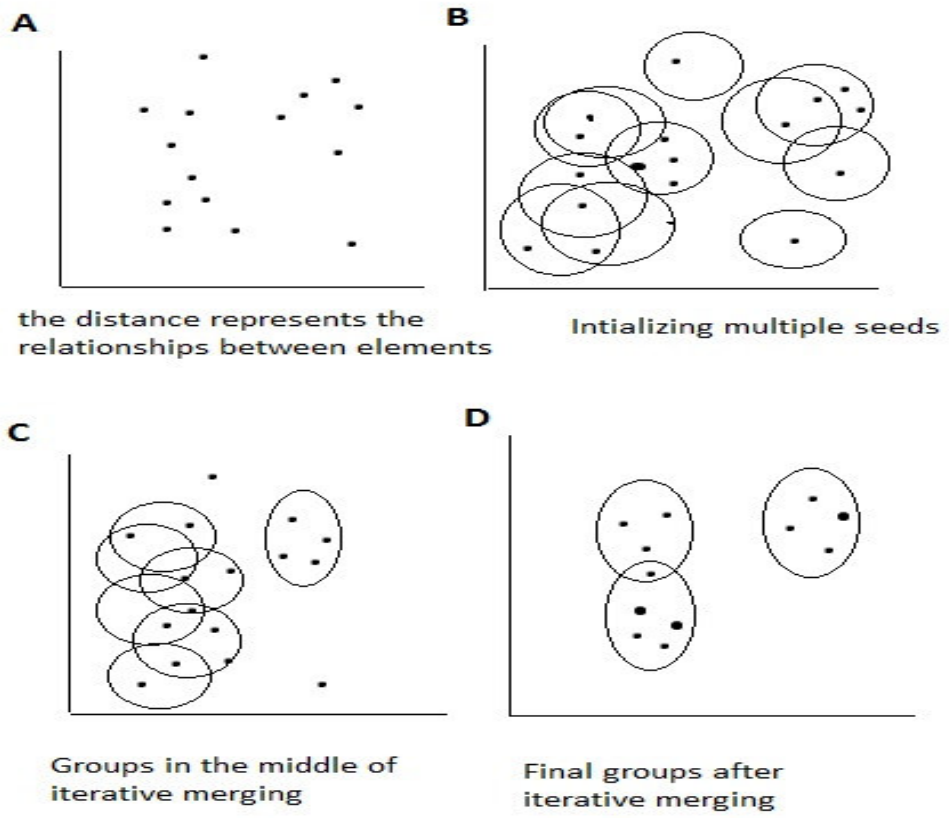

Fig:3 group of clusters

3.4Comparing of k means and fuzzy c means:

Efficiency:

Efficiency of k-means is fairer; fuzzy c-means is slower. Because of k-means just needs a distance calculation. Whereas fuzzy c-means needs to full inverse distance weighting.

Performance:

Performance of k-means is traditional and limited use, while the performance of c-means is that it can be used only in variety of clusters and this can be handle uncertainty

Application:

Application of k-means is applicable for image retrieval algorithms, c-means can be applied to segmentation of magnetic response imaging (MRI) and also for analysis of network traffic.

Fuzzy c means algorithm is slower than k means algorithm in efficiency but gives better results in cases where data is incomplete or uncertain and has a wider applicability.

In this two clustering algorithms namely centroid based k-means and representative object based fuzzy c-means (FCM) clustering algorithms are compared. Fuzzy c-means produces close results to compare with k-means clustering but still requires more computational time than k-means.[14]

4. FUZZY META CLUSTERING ALGORITHM

We can describe any phone numbers as our objects. The algorithm can applicable to any network of objects. Now let's take $\mathbf{Pn_j}$ is the $\mathbf{j^{th}}$ phone number and also represent $\mathbf{Pn_j}$ by a static part of data S_j. In this same way dynamic part of data $\mathbf{d_j}$ is exists.i.e $\mathbf{Pn_j=(s_j,d_j)}$.K is taken as number of clusters, then

$$\mathbf{d_j=(m_{j1},m_{j2},\ldots,m_{jk}).}$$

Where $\mathbf{m_{jk}}$=average to membership of phone numbers. That $\mathbf{P_{nj}}$ is calling that falls in $\mathbf{k_{th}}$ cluster, which are taken from the previous iteration. Let **call** $(\mathbf{P_{nj}})$ be the set of phone numbers which are called by $\mathbf{P_{nj}}$.

$$\frac{\mathbf{mj,k=\sum pnjcall(pnj)\mu ik}}{\mathbf{Call(pnj)}} \quad \text{---------- (1)}$$

Where,$\mu\mathbf{ik}$= membership of $\mathbf{P_{ni}}$ to cluster K. $\mathbf{Call(pnj)}$= cardinality of **call** $(\mathbf{p_{nj}})$.

Clustering scheme $\mathbf{d_i}$ keeps on changing through every iteration.i^{th} iteration, represents the corresponding quantities by a superscript **i** .

$$S_j^i = S_j^{i-1} = S_j^0 \quad \text{---------- (2)}$$

$$D_j^i = (m_{j1}^{i-1}, m_{j2}^{i-1}, \ldots\ldots, m_{jk}^{i-1}) \text{----------(3)}$$

Where m_{jk}^{i-1} = average membership of destination numbers belongings to k^{th} cluster. Which is called by P_{nj} in $(i-1)^{th}$ iteration.

$$P_{nj}^i = (S_j^i, d_j^i) = (s_j^0, d_j^i) \quad \text{----------(4)}$$

1. Apply fuzzy c-means clustering to the phone numbers by using their static representations, i.e. a phone number $\mathbf{p_{nj} = s_j}$,Where Sj is a vector of attribute values retrieved from the data set for the phone number.

2. Calculate the dynamic representation d_j of a phone number Pnj:

$$\mathbf{dj = [m^1_i | m^2_i | \ldots\ldots | m^k_i]}$$

Where m_{ji} is the average membership of all the phone numbers called by p_{nj} to cluster i in the previous clustering.

3. Re-cluster the phone numbers with concatenation of static and dynamic representations, i.e. a phone number

$$\mathbf{p_{nj} = (s_j , d_j)}$$

4. If the values of dj for a phone P_{nj} have changed, go back to step 2.

The application of algorithm to the mobile phone call dataset can be explained in the following way.

First, Normal data can be dividing the values of each seven attributes by the mean value for attribute. it means all the 7 attributes are more or less equally weighted. Determining the appropriate numbers of clusters is the major challenge. This is resolved by Lingras and Rathinavel [7] by applying k-means clustering to phone numbers data. They are mainly used minimization of scatter process in cluster and maximizing the separation between clusters using Davis-Bouldin(DB)index[12].the final number of phone number clusters were found to be five.

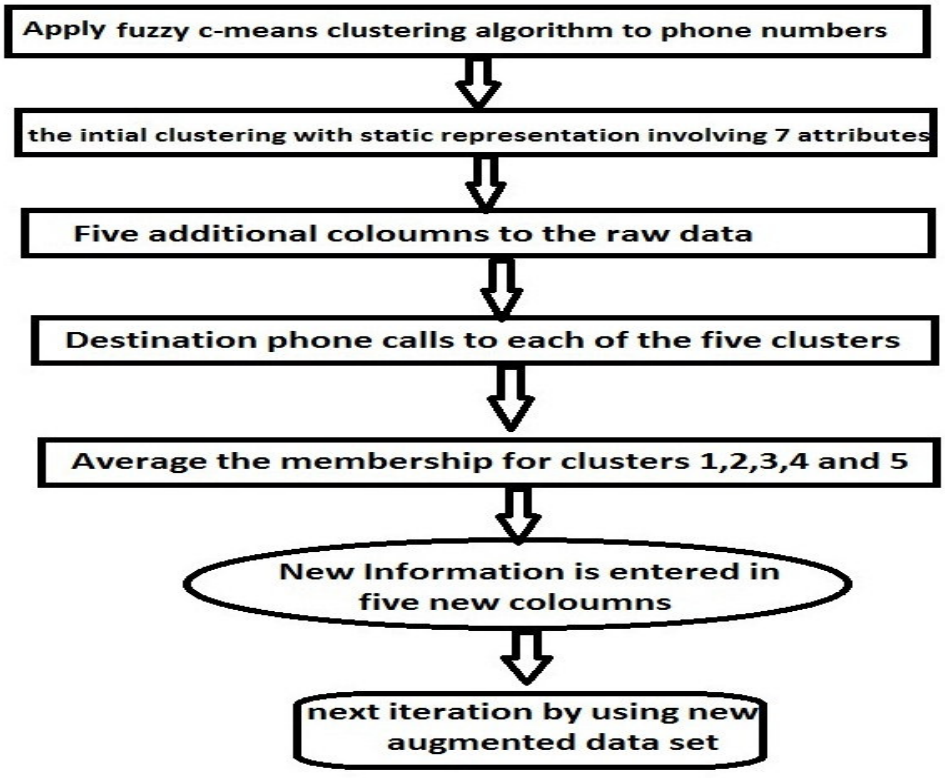

Fig.4.1.Steps for Fuzzy Meta clustering algorithm.

5. DATA SETS

Data preparation applied to the original data set which is provided by Eagle [8] followed by the design of the experiment. This data set is used and experimentally done by Lingras and Rathinavel [7] in the implementation of their crisp recursive meta-clustering algorithm.

The data set comprised of 182,208 phone calls data was collected from about 102 users which is over a period of nine months. Here some variables which are chosen to represent a phone call:

1) Weekend/Weekday (1/0)
2) Daytime/night-time (1/0)
3) Duration of phone call (normalized using different weighting schemes)
4) Voice call (1/0)
5) Direction (outgoing/incoming = 1/0)
6) SMS (1/0)
7) Packet data (1/0)
8) MMS (1/0)
9) Data Call (1/0)
10) Long Duration call (1/0).

The first clustering analysis of the phone calls will mainly exists after the fields were added. The need for this field will be justified later. If the duration is above 1512s, it is taken a long duration call. In the representation of phone calls, it mainly takes the binary values of data. In original data set, the type of call was classified as SMS, data call, voice and MMS calls. The MMS and data calls amount to twenty-five and sixteen calls respectively. Hence these calls are excluded from any further of analysis. The data set for the phone numbers will be represented by the following variables:

1) Average duration of Phone calls.
2) Average number of Weekend/Weekday [0, 1]
3) Average number of daytime/night-time [0, 1]
4) Average number of outgoing/incoming [0, 1].
5) Average number of SMS [0, 1]
6) Average number of Voice calls [0, 1]
7) Average number of long duration calls [0, 1].

6. RESULTS

The clustering results can be analyzed in two parts one is static and another is dynamic. The static results are mainly derived from the matrix D^0 and the dynamic results are derived from the information added with each iteration. The dynamic results are derived from

$[m^1_i|m^2_i|.....|m^k_i]$

7. COMPARISON OF CRISP AND FUZZY META-CLUSTERING

When the results of fuzzy meta-clustering and crisp meta-clustering approaches are compared, we noticed, the clusters tend to have the same labels (low, moderate and high) across all the variables in their static fuzzy clustering. For example, cluster 5, which has a moderate label across all variables or cluster 2 which has a moderate label across most of its variables except 3 which are both high. This kind of flexibility and a general trend of cluster centers are not noticed in the crisp recursive meta-clustering [7].

8. CONCLUSION

Each granule is represented by static attributes which are derived from the raw data set, and dynamic attributes represents the average fuzzy membership of connected granules to different clusters. This meta-clustering algorithm is applied to a network of phone numbers is resulting clustering shows meaningful profiles for each cluster in terms of static attributes and also dynamic attributes. These profiles are making easy and it possible for us to analyze sociability and popularity of different clusters. The use of fuzzy clustering not only allows us to describe partial membership of a phone number to a given cluster, but it leads more moderate values of the dynamic attributes based on graded fuzzy memberships as opposed to the binary memberships in crisp clustering.

9. REFERENCES

[1]RuiXu,"survey of clustering Algorithm", IEEE transactions on neural networks, vol.16, 2005.

[2]Brian S.Everitt, Sabine Landau, MorvenLeese, Daniel Stahl," Cluster analysis", Wiley, 2011.

[3] Kishore Rathinavel, Pawan Lingras,"A Granular Recursive Fuzzy Meta-clustering Algorithm for Social Networks",proc.IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), Edmonton AB,24-28 June 2013,pp 567-572.

[4] P. Lingras and K. Rathinavel, "Recursive Meta-clustering in a Granular Network," in Plenary talk at the Fourth International Conference of Soft Computing and Pattern Recognition, Brunei, 2012.

[5]Pawan Lingras, Parag Bhalchandra, Satish Mekewad, Ravindra Rathod, Santosh Khamitkar: "Comparing Clustering Schemes at Two Levels of Granularity for Mobile Call Mining**".** RSKT 2011: 696-705, 2011.
[6] Pawan Lingras, Sarjerao Nimse, N. Darkunde, A. Muley: "Soft clustering from crisp clustering using granulation for mobile call mining". GrC 2011: 410-416, 2011
[7]P. Lingras and K. Rathinavel, "Recursive Meta-clustering in a Granular Network," in Plenary talks at the Fourth International Conference of Soft Computing and Pattern Recognition, Brunei, 2012.
[8] N. Eagle, "The Reality Mining Data", 2010.
[9] P.Lingras and C.J.Butz, "Conservative and Aggressive Rough SVR modeling, "theoretical Computer Science journal section on Theory of Natural Computing, vol. to appear, 2011.
[10]Margaret H.Dunham,"Data mining-Introductory & Advanced Topics",2003.
[11] Jiawei Han and Micheline Kamber, "Data Mining: concepts and Techniques", 2006.
[12]D. Davies and D. Bouldin, "A cluster separation measure," Pattern Analysis and Machine Intelligence, IEEE Transactions on, no. 2, pp.224–227, 1979.
[13]Rich Caruana, Mohamed Elhawary, Nam Nguyen,Casey smith,"Meta Clustering", proc ICDM'06, 2006.
[14] Soumi Ghosh, Sanjay Kumar Dubey, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms", IJACSA, 2013
[15]M.s.yang,"A survey of Fuzzy clustering", mathematical computing modeling, vol 18, no.11, pp 1-16, 1993.
[16]R.suganya,R shanthi,"Fuzzy c-means Algorithm-A Review", International Journal of Scientific and Research Publications,Vol 2,2012.
[17]Pal N.R,Pal k,keller J.M. and bezdek J.c,"A possibilistic fuzzy c-means clustering Algorithm",IEEE transactions on fuzzy systems,vol.13,N0.4,517-530,2005.
[18]Pawan Lingras,Georg peters, rough clustering, WIREs data mining knowledge discovery, Wiley publications,1:64-72,2011.
[19]Hirano S, Tsumoto,"rough clustering and its application to medicine", journal of information sciences124, 125-137, 2000.
[20]Lingras.P, West C,"Interval set clustering of web users with rough k-means", journal of Intelligent Information Systerms, 23, 5-16, 2003.