Comparative Study On Clustering Algorithms As Fuzzy C-Means And K-Means

N.Savithiri[1], Dr. A. Banumathi[2]
[1]Dept of Computer Science
[2]Assistant Professor, Dept of Computer Science
[1, 2] Government arts college(A)

Abstract- *Software Engineering is a set of problem solving skills, instructions, methods and algorithms are applied on many problem having domain to identify and create useful systems. The created systems are applied on many practical problems to solve it. While creating such kind of systems we need bulk of data managing those data is the most important in Software Engineering. For that many algorithms are there to cluster the data and manage it. Here we have compared two clustering algorithms such as K-Means and Fuzzy-C Means which is used in Software Engineering.*

Keywords- Software Engineering, Clustering, K-Means, Fuzzy-C Means.

I. INTRODUCTION

The combination of computer science and software engineering gave rise to revolutionary developments in both the fields of software engineering and computer science. Computer science deals the problems theoretically and software works with problems in a practical manner. Software engineering is also referred as the art of writing software is always a controversial problem among various experts which covering software design principles, so-called "best practices" for writing code.

The management of team size, work culture, product in-time delivery procedure, out sourcing etc. are some of the key practices included in the software design principles.

Clustering technique defines classes and put objects which are related to them in one class on the other hand in classification objects are placed in predefined classes. Clustering means put the objects which have similar properties into one group and objects which have dissimilar properties into another.

In clustering, above threshold values objects can be placed in one cluster and values below into another cluster. Clustering has alienated the large data set into groups or clusters according to similarity in properties [1]. There are various algorithms which are used to solve this problem. In this research work two important clustering algorithms namely centroid based K-Means and representative object based FCM (Fuzzy C-Means) clustering algorithms are compared. 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 behavior patterns of both the algorithms are analyzed. FCM produces close results to K-Means clustering but it still requiresmore computation time than K-Means clustering[2].

The main advantage of a clustered solution is automatic recovery from failure, that is, recovery without user intervention. Disadvantages of clustering are complexity and inability to recover from database corruption.

In a clustered environment, the cluster uses the same IP address for Directory Server and Directory Proxy Server, regardless of which cluster node is actually running the service. That is, the IP address is transparent to the client application [3]. In a replicated environment, each machine in the topology has its own IP address. In this case, Directory Proxy Server can be used to provide a single point of access to the directory topology. The replication topology is therefore effectively hidden from client applications.

To increase this transparency, Directory Proxy Server can be configured to follow referrals and search references automatically. Directory Proxy Server also provides load balancing and the ability to switch to another machine when one fails.

II. K-MEANS CLUSTERING

K-Means or Hard C-Means clustering is basically a partitioning method applied to analyze data and treats observations of the data as objects based on locations and distance between various input data points.

Partitioning the objects into mutually exclusive clusters (K) is done by it in such a fashion that objects within each cluster remain as close as possible to each other but as far as possible from objects in other clusters.

Each cluster is characterized by its center point i.e. centroid. The distances used in clustering in most of the times do not actually represent the spatial distances. In general, the only solution to the problem of finding global minimum is exhaustive choice of starting points. But use of several replicates with random starting point leads to a solution i.e. a global solution.

In a dataset, a desired number of clusters K and a set of k initial starting points, the K-Means clustering algorithm finds the desired number of distinct clusters and their centroids. A centroid is the point whose co- ordinates are obtained by means of computing the average of each of the co-ordinates of the points of samples assigned to the clusters[2].

Algorithmic steps for K-Means clustering

1) Set K – To choose a number of desired clusters, K.

2) Initialization – To choose k starting points which are used as initial estimates of the cluster centroids. They are taken as the initial starting values.

3) Classification – To examine each point in the dataset and assign it to the cluster whose centroid is nearest to it.

4) Centroid calculation – When each point in the data set is assigned to a cluster, it is needed to recalculate the new k centroids.

5) Convergence criteria – The steps of (iii) and (iv) require to be repeated until no point changes its cluster assignment or until the centroids no longer move[2].

The actual data samples are to be collected before the application of the clustering algorithm. Priority has to be given to the features that describe each data sample in the database. The values of these features make up a feature vector (Fi1, Fi2, Fi3,…….., Fim) where Fim is the value of the M-dimensional space.

As in the other clustering algorithms, k- means requires that a distance metric between points is to be defined. This distance metric is used in the above mentioned step (iii) of the algorithm. A common distance metric is the Euclidean distance. In case, the different features used in the feature vector have different relative values and ranges then the distance computation may be distorted and so may be scaled.

The input parameters of the clustering algorithm are the number of clusters that are to be found along with the initial starting point values. When the initial starting values are given, the distance from each sample data point to each initial starting value is found using equation. Then each data point is placed in the cluster associated with the nearest starting point. After all the data points are assigned to a cluster, the new cluster centroids are calculated. For each factor in each cluster, the new centroid value is then calculated. The new centroids are then considered as the new initial starting values and steps (iii) and (iv) of the algorithm are repeated. This process continues until no more data point changes or until the centroids no longer move.

III. FUZZY C-MEANS CLUSTERING

Bezdek introduced Fuzzy C-Means clustering method in 1981, extend from Hard C-Mean clustering method. FCM is an unsupervised clustering algorithm that is applied to wide range of problems connected with feature analysis, clustering and classifier design.

FCM is widely applied in agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, shape analysis and target recognition. With the development of the fuzzy theory, the FCM clustering algorithm which is actually based on Ruspini Fuzzy clustering theory was proposed in 1980's. This algorithm is used for analysis based on distance between various input data points. The clusters are formed according to the distance between data points and the cluster centers are formed for each cluster [2]. Infact, FCM is a data clustering technique in which a data set is grouped into n clusters with every data point in the dataset related to every cluster and it will have a high degree of belonging (connection) to that cluster and another data point that lies far away from the center of a cluster which will have a low degree of belonging to that cluster.

If $||U(k+1)-U(k)||<\delta$ then we are to stop otherwise we have to return to step 2, updating the cluster centers iteratively and also the membership grades for data point. FCM iteratively moves the cluster centers to the right location within a dataset. To be specific introducing the fuzzy logic in K-Means clustering algorithm is the Fuzzy C-Means algorithm in general. Infact, FCM clustering techniques are based on fuzzy behavior and they provide a technique which is natural for producing a clustering where membership weights have a natural interpretation but not probabilistic at all. This algorithm is basically similar in structure to K-Means algorithm and it also behaves in a similar fashion [2].

IV. COMPARATIVE ANALYSIS

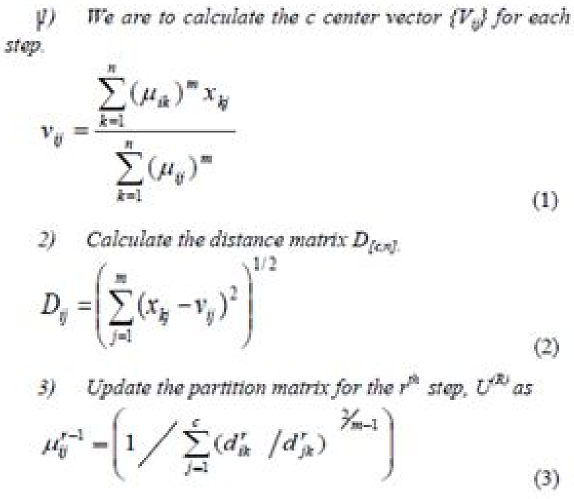
K-Means clustering and Fuzzy-C Means Clustering are very similar in approaches. The main difference is that, in Fuzzy-C Means clustering, each point has a weighting associated with a particular cluster, so a point doesn't sit "in a cluster" as much as has a weak or strong association to the cluster, which is determined by the inverse distance to the center of the cluster.

Fuzzy-C means will tend to run slower than K means, since it's actually doing more work.

Algorithmic steps for Fuzzy C-Means clustering

We are to fix c where c is (2<=c<n) and then select a value for parameter ‘m’ and there after initialize the partition matrix U(0).

Each step in this algorithm will be labelled as ‘r’ where r = 0, 1, 2 …



1) We are to calculate the c center vector {V_{ij}} for each step.

$$v_{ij} = \frac{\sum_{k=1}^{n} (\mu_{ik})^m x_{kj}}{\sum_{k=1}^{n} (\mu_{ij})^m} \qquad (1)$$

2) Calculate the distance matrix $D_{[c,n]}$.

$$D_{ij} = \left(\sum_{j=1}^{m} (x_{kj} - v_{ij})^2 \right)^{1/2} \qquad (2)$$

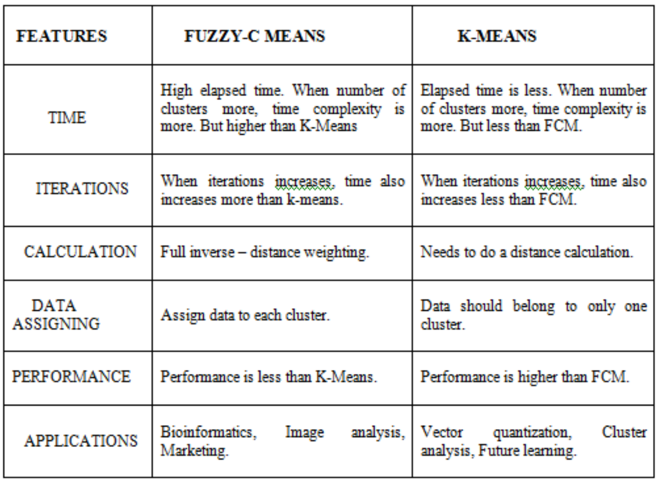
3) Update the partition matrix for the r^{th} step, $U^{(R)}$ as

$$\mu_{ij}^{r-1} = \left(1 \Big/ \sum_{j=1}^{c} (d_{ik}^r \;/ d_{jk}^r)^{2/m-1} \right) \qquad (3)$$

Each point is evaluated with each cluster, and more operations are involved in each evaluation. K-Means just needs to do a distance calculation, whereas fuzzy- c means needs to do a full inverse-distance weighting. K means clustering cluster the entire dataset into K number of cluster where a data should belong to only one cluster. Fuzzy c-means create k numbers of clusters and then assign each data to each cluster, but there will be a factor which will define how strongly the data belongs to that cluster [3].

The following table shows the features of K-Means and Fuzzy- C Means as time, iterations, calculation, data assigning, performance, applications.

Table 1. Table comparison between Fuzzy-C Means and K-Means algorithm



FEATURES	FUZZY-C MEANS	K-MEANS
TIME	High elapsed time. When number of clusters more, time complexity is more. But higher than K-Means	Elapsed time is less. When number of clusters more, time complexity is more. But less than FCM.
ITERATIONS	When iterations increases, time also increases more than k-means.	When iterations increases, time also increases less than FCM.
CALCULATION	Full inverse – distance weighting.	Needs to do a distance calculation.
DATA ASSIGNING	Assign data to each cluster.	Data should belong to only one cluster.
PERFORMANCE	Performance is less than K-Means.	Performance is higher than FCM.
APPLICATIONS	Bioinformatics, Image analysis, Marketing.	Vector quantization, Cluster analysis, Future learning.

V. CONCLUSION

In Software Engineering, data managing is the most important parameter. Because a software consumes more data. This problem can be manageable. when we use clustering technique. Here the two existing clustering algorithms K-Means and Fuzzy-C Means are compared by various parameters like time, number of iterations, data assigning and applications. During this K-Means produces best results in all parameters.

REFERENCES

[1] Shalini Verma and Abinav Mishra “Various Clustering Techniques in Software Engineering-A Review”, International Journal of Computer Science and Mobile Computing, Volume 4, Issue 7, July 2015, pg.no 453-458

[2] Soumi Ghosh and Sanjay Kumar Dubey “Comparative Analysis of K-Means AND Fuzzy-C Means Algorithms” ,International Journal of Advanced Computer Science and Applications ,Volume 4, No.4, 2013.

[3] Velmurugan.T “Performance Comparison between k-Means and Fuzzy C-Means Algorithms using Arbitrary Data Points”, WULFENIA Journal, Klagenfurt, Austria, Volume 19, No. 8; Aug 2012.

[4] Nidhi Grover “A study of various Fuzzy Clustering Algorithms”, International Journal of Engineering Research, Volume 3, Issue No.3, PP: 177-181.

[5] Jogannagari Malla Reddy, S.V.A.V. Prasad and Samabasiva Rao Baragada “A Comprehensive Analysis of Literature A ComprehensiveAnalysis of Literature Reported Software Engineering Advancements Using AHP” International Journal on Recent and Innovation Trends in Computing and
Communication ,Volume: 4 Issue:1, PP 238 – 249.

[6] Namratha .M, and Prajwala .T. R.”A Comprehensive Overview of Clustering Algorithms in Pattern Recognition” IOSR Journal of Computer Engineering (IOSRJCE) Volume 4, Issue 6 (Sep-Oct. 2012), PP: 23-30.

[7] Banumathi.A and Pethalakshmi.A “Increasing Cluster Uniqueness in Fuzzy C-Means through Affinity Measure”, International Conference on Pattern Recognition, Informatics and Medical Engineering, March 21-23, 2012.