Volume 4, Issue 5, May 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Performance Comparison of Fuzzy C Means with Respect to Other Clustering Algorithm

Tejwant Singh[1], Mr. Manish Mahajan[2]
[1] Research Scholar, [2] Associate Professor,
Dept of IT, Chandigarh Engineering College, Mohali, India

Abstract: Fuzzy C-Mean (FCM) is an unsupervised clustering algorithm based on fuzzy set theory that allows an element to belong to more than one cluster. Where fuzzy means "unclear" or "not defined" and c denotes "clustering". In FCM the number of cluster are randomly selected. [15] FCM is the advanced version of K-means clustering algorithm and doing more work than K-means. K-Means just needs to do a distance calculation, whereas fuzzy c means needs to do a full inverse-distance weighting. This, plus the overhead needed for computing and managing, explains why FCM is quite slower than K-Means.[4]

Keywords:-Clustering,Time-complexity, Accuracy, FCM, K-Means and Y-means.

I. Introduction:-

Data Mining is the process of extracting hidden, previously unknown and useful information from large databases and data warehouses. Data mining process involve steps like data cleaning, integration, selection, transformation, data mining technique, pattern evaluation and knowledge representation. Various data mining techniques are used like classification, clustering, association rules, Sequential patterns, Prediction, Decision trees, etc. are used in various applications.
Here we discuss about clustering algorithms like fuzzy c-means, k-means, etc.

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 of analysis, clustering and classifier design. FCM is widely applied in agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis. [2]

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. The degree of membership of each data item to the cluster is calculated which decides the cluster to which that data item is supposed to belong.[1]
For each item, we have a coefficient that specifies the membership degree (uij) of being in the k^{th} cluster as follows:---

$$u_{ij} = \sum_{k=1}^{n} (d_{ij} / d_{ik})^{(2/m-1)}$$

where,
dij – distance of i^{th} item from j^{th} cluster,
dik –distance of i^{th} item from k^{th} cluster,
and m – fuzzification factor.

The existence of a data item in more than one cluster depends on the fuzzification value (m) defined by the user in the range of [0, 1] which determines the degree of fuzziness in the cluster.[1]
Thus, the items on the edge of a cluster may be in the cluster to a lesser degree than the items in the center of the cluster. When m reaches the value of 1 the algorithm works like a crisp partitioning algorithm and for larger values of m the overlapping of clusters tends to be more.
The main objective of fuzzy clustering algorithm is to partition the data into clusters so that the similarity of data items within each cluster is maximized and the similarity of data items in different clusters is minimized. Moreover, it measures the quality of partitioning that divides a dataset into C clusters.[1]

© 2014, IJARCSSE All Rights Reserved

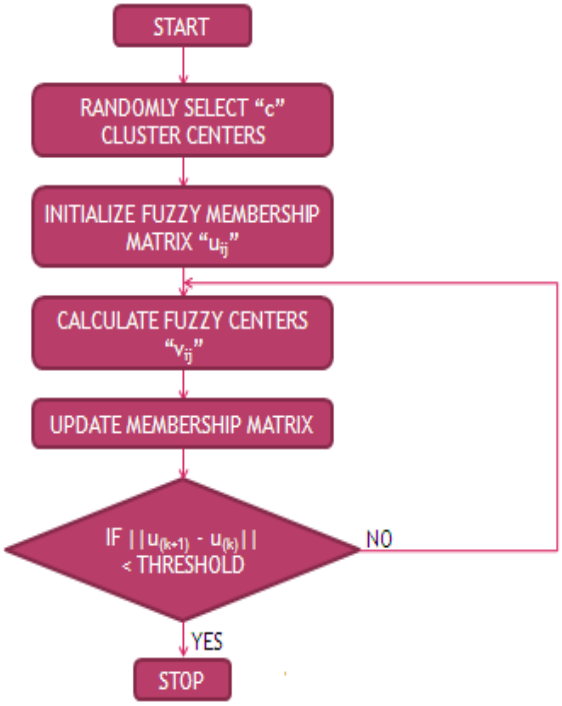

Fig 1:- Fuzzy c means algorithm

II. Comparisons:-

1.

Data mining techniques play a vital role in intrusion detection by analyzing the large volumes of network data and classifying it as normal or anomalous. Firewalls that are used for intrusion detection possess certain drawbacks which are overcome by various data mining approaches.

Several data mining techniques like Classification, Clustering and Association rules are widely used to enhance intrusion detection. Among them clustering is preferred over classification since it does not require manual labeling of the training data and the system need not be aware of the new attacks.
Three different clustering algorithms namely K-Means Clustering, Y-Means Clustering and Fuzzy C-Means Clustering. K-Means clustering results in degeneracy and is not suitable for large databases. Y-Means is an improvement over K-means that eliminates empty clusters. Fuzzy C-Means is based on the fuzzy logic that allows an item to belong to more than one cluster and concentrates on the minimization of the objective function that examines the quality of partitioning.
These methods extract previous unknown significant relationships and patterns from large databases. The extracted patterns are then used as a basis to identify new attacks.

Three clustering algorithms namely K-means, Y-means and Fuzzy C-means have been discussed. Result obtained by discussion state that FCM is better than among.
The main advantage of Fuzzy C-Means clustering for intrusion detection is the high detection rate and lower false positive rate that it offers. Although Fuzzy C-Means is an efficient technique but it is time consuming.

The performance of intrusion detection systems can be still improved by combining the features of Fuzzy C- Means clustering technique with some other technique so that it reduces the time required by Fuzzy C-Means for the clustering process and also increases the detection rate and decreases the false positive rate thereby making the intrusion detection system more accurate and effective.[1]

2.

In the arena of software, data mining technology has been considered as useful means for identifying patterns and trends of large volume of data. This approach is basically used to extract the unknown pattern from the large set of data for business as well as real time applications.
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. FCM produces close results to K-Means clustering but it still requires more computation time than K-Means clustering.

This experiment reveals the fact that K-Means clustering algorithm consumes less elapsed time i.e. 0.443755 seconds than FCM clustering algorithm which takes 0.781679 seconds.
On the basis of the result drawn by this experiment it may be safely stated that K-Means clustering algorithm less time consuming than FCM algorithm

© 2014, IJARCSSE All Rights Reserved

Algorithm	Time Complexity	Elapsed Time (Seconds)
K-Means	O(ncdi)	0.448755
FCM	$O(ndc^2i)$	0.781679

Fig 2:-comparative analysis of k-means and FCM

The time complexity of the K-Means algorithm is O(ncdi) and the time complexity of FCM algorithm is $O(ndc^2i)$. where n = number of data points, c = number of cluster, d = number of dimension and i = number of iterations.

Keeping the number of data points constant. we may assume that n = 100, d = 3, i = 20 and varying number of clusters. The following graph represents the comparison in details.

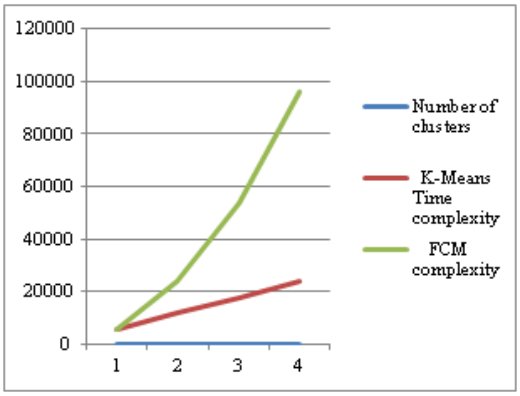

Fig.3. Time complexity of K-Means and FCM by varying number of clusters

Now keeping no. of cluster constant, lets assume n=150, d=2, c=2 and varying no. of iteration, we obtain the following graph.

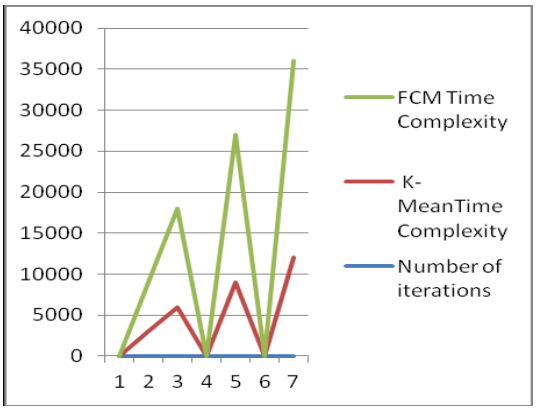

Fig 4:-. Time complexity of K-Means and FCM by varying number of iterations

From the obtained results we may conclude that K-Means algorithm is better than FCM algorithm.

FCM produces close results to K-Means clustering but it still requires more computation time than K-Means because of the fuzzy measures calculations involvement in the algorithm.

In fact, FCM clustering which constitute the oldest component of software computing are really suitable for handling the issues related to understand ability of patterns, incomplete/noisy data, mixed media information, human interaction

© 2014, IJARCSSE All Rights Reserved

and it can provide approximate solutions faster. They have been mainly used for discovering association rules and functional dependencies as well as image retrieval. So, overall conclusion is that K-Means algorithm seems to be superior than Fuzzy C-Means algorithm.[2]

3.

Web usage mining has become very critical for effective Web site management, creating adaptive Web sites, business and support services, personalization, network traffic flow analysis etc., It was crucial to understand who the users were, what they looked at, and how their interests changed with time. To achieve this, one of the promising approaches is web usage mining, which mines web logs for user models and recommendations. Web usage mining algorithms have been widely utilized for modeling user web navigation behavior.[6]

The web-usage mining framework processed with fuzzy C means clustering algorithm (to discover web data clusters) and compare with Expected Maximization cluster system to analyze the Web site visitor trends.
The clustered data is then used to analyze the trends using inference system. By linking the Web logs with cookies and forms, it is further possible to analyze the visitor behavior and profiles.[7]
Experimentation conducted with C Fuzzy means and Expected Maximization clusters in Syskill Webert data set from UCI, shows that EM shows 5% to 8% better performance than Fuzzy c means in terms of cluster number.

The web usage mining framework presented in this work evaluates the performance of expectation-maximization (EM) and Fuzzy C means cluster algorithms.[9]

4.

Segmentation of an image entails the division or separation of the image into regions of similar attribute. Clustering is one of the methods used for segmentation. In this section we compare the performance of various segmentation techniques for color images. K-means clustering and Fuzzy C-Means clustering techniques are compared for their performance in segmentation of color images.

Segmentation is the process of partitioning a digital image into multiple segments. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful easier to analyze.
Clustering algorithms may be classified as listed below:

- **Exclusive Clustering**
- **Overlapping Clustering**

In the first case data are grouped in an exclusive way, so that if a certain data belongs to a definite cluster then it could not be included in another cluster.
On the contrary the second type, the overlapping clustering, uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership.
K-means is an exclusive clustering algorithm, Fuzzy C-means is an overlapping clustering algorithm.
This section compares k-means and fuzzy c-means clustering image segmentation algorithms .The algorithms are developed in MATLAB for analysis and comparison.[8]
K-means clustering produces fairly higher accuracy and requires less computation. Fuzzy C means clustering produces close results to K-means clustering, yet it requires more computation time than K-means because of the fuzzy measures calculations involved in the algorithm.[14]

III. Conclusion:

We have make the internal review of above said papers in which data mining techniques are used for various application like image segmentation, web usage mining, intrusion detection system, software arena, bioinformatics, stock market, web analysis etc. Fuzzy c means is one of the algorithm which is used in data mining for clustering. As compare to other clustering algorithms fuzzy c means is more efficient, reliable and robust than others in certain cases or applications by its performance.[3] But after the internal review of above papers we conclude that fuzzy c means take/have more computation time than other clustering techniques.

Proposed Work:-

Finally, we conclude that fuzzy c means algorithm should be improved in terms of its computation time. We can improve the performance of fuzzy c means clustering algorithm by using two steps first is by decision tree approach with it which mine the data in accurate and sequential manner and second by creating the noise-free log file. If we process the log file which are free from noisy data and have sequence order data then we can deduct the compilation time of fuzzy c means algorithm.

References:-

[1] Mitchell D'silva and Deepali Vora, "Comparative Study of Data Mining Techniques to Enhance Intrusion Detection", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622

© 2014, IJARCSSE All Rights Reserved

www.ijera.com. Vol. 3, Issue 1, January -February 2013, pp.1267-1275

[2] 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.*

[3] O. A. Mohamed Jafar and R. Sivakumar, "A Comparative Study of Hard and Fuzzy Data Clustering Algorithms with Cluster Validity Indices",emerging research in computing, information, communication and applications ERCICA 2013.

[4] Subhagata chattopadhyay,dilip kumar pratihar,sanjib chandra de sarkar,"A comparative study of fuzzy c-means algorithm and entropy-based fuzzy clustering algorithms", computing and informatics, vol. 30, 2011, 701–720.

[5] Eugene Levner, David Pinto, Paolo Rosso, David Alcaide, R.R.K. Sharma levner@hit.ac.il, "Fuzzifying clustering algorithms:- The case study of MajorClust", {dpinto, prosso}@dsic.upv.es, dalcaide@ull.es, rrks@iitk.ac.in(2008)

[6] Marathe Dagadu Mitharam "Preprocessing in Web Usage mining", International Journal of Scientific & Engineering Research, February -2012.

[7] Aditi Shrivastava, Nitin Shukla, "Extracting Knowledge from User Access Logs", International Journal of Scientific and Research Publications. April-2012.

[8] Soniya P. Chaudhari , Prof. Hitesh Gupta , Prof. S. J. Patil(2013), "Web Log Clustering using FCM and Swarm Intelligence Based Algorithms", International Journal of Innovative Research in Science, Engineering and Technology.janurary-2013.

[9] K.Poongothai ,M.Parimala and Dr. S.Sathiyabama"Efficient Web Usage Mining with Clustering", IJCSI International Journal of Computer Science Issues, November 2011.

[10] Kuo-lung wu, "Parameter Selections of Fuzzy C-Means Based on Robust Analysis", World Academy of Science, Engineering and Technology 41 2010.

[11] Deepak Kumar Niware Dr. Setu Kumar Chaturvedi, "Web Usage Mining through Efficient Genetic FuzzyC-Means" ,(IJCSIS) International Journal of Computer Science and Information Security,Vol. 11, No. 9,September 2013.

[12] K. Suresh ,R .Madana Mohana, A. Rama Mohan Reddy, A. Subrmanayam, "Improved FCM algorithm for Clustering on Web usage Mining"published by IEEE 2011.

[13] N. Sujatha and Dr. K. Iyakutti "Improved Fuzzy C-Means Clustering of Web Usage Data with Genetic Algorithm" CiiT International Journal of Data Mining and Knowledge Engineering, Vol 1, No 7, October 2009.

[14] 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

[15] www.google.com

[16] https://en.wikipedia.org/

© 2014, IJARCSSE All Rights Reserved