© 2016 IJSRSET | Volume 2 | Issue 1 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

FCM : Fuzzy C-Means Clustering – A View in Different Aspects

B Kalai Selvi, M Ashwin

Department of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, Tamil Nadu, India

ABSTRACT

Data Mining is the process of obtaining or exploring data from the large amount of raw data. It produces the meaningful information. To obtain the information data mining has multiple techniques such as classification, regression, prediction, clustering, and summarization. There are multiple tasks in data mining to obtain the information such as cleaning, integrating, selection, transformation, pattern evaluation. One of the challenging techniques in the data mining is clustering. Clustering is the process of grouping the data under some condition. The main aim of the paper is to describe about the Fuzzy C-Means Clustering (FCM) and compared with K-Means clustering. The pitfalls overcome by the FCM are also measured theoretically.

Keywords: Clustering, Data Mining, FCM, C-means, K-Means, Fuzzy

I. INTRODUCTION

Data have spread all over the world in multi forms such as text, numbers, sounds, pictures, motion pictures. The data's are cumulative and not in the format to understand. To make the data in the proper form a concept called data mining, which is used to obtain knowledge from the raw data [1]. To obtain knowledge multiple techniques are used such as classification, regression, prediction, clustering, summarization, sequence discovery. Clustering is the process of describing the data's in the form of group according to relations.

Generally clustering is classified into hard cluster and soft cluster [13]. If the data present in only one cluster then it comes under Hard cluster, if the data is allowed to present in more than one set then it comes under Soft cluster. K-means algorithm comes under hard cluster and fuzzy clustering algorithm comes under soft cluster. Sometimes the clustering is classified as Exclusive clustering, overlapping clustering, hierarchical clustering, and probabilistic clustering. Exclusive clustering is also defined as hard clustering and the best example is k-means clustering. Overlapping clustering is soft clustering and the exa example for it is Fuzzy C-Means clustering. Similarly the example for probabilistic clustering is mixture of Gaussians.

Fuzzy Clustering is also called as soft clustering i.e. data elements belong to more than one cluster. Sometimes fuzzy cluster is defined as the soft version of k-means so it is also called as Fuzzy C-Means Clustering (FCM). It incorporates the fuzzy technique with the K-means clustering technique.

The paper is organized as follows: Section II contains the related work and Section III explains about the evaluation methodology of clusters. Section IV indicates the observation and Section V discusses about the sample datasets of the methodologies. Finally, Section VI proposes some suggestions and conclusion.

A. Related Work

The data in the world is growing enormously. To form the group of data that is related to each other data mining uses a new technique called clustering. The use of clustering is applied in many places and its development is growing day by day. Some of the real time examples for clustering are like opening malls, placing telephone towers, opening hospitals etc., thus the clustering can be classifies according to the users. This section is reviewed about Fuzzy C-Means clustering.

Bo Gao and Jun Wang [2] have used the technique called Fast Generalize C-Means (FGFCM) and Xie-Bie (XB) index. FGFCM incoperatates both information of spatial and gray image and produces the advance image. The only difference between FCM and FGFCM is it produces the new image. XB index defines the difference between the mean quadratic error and the minimum of the minimal squared distances between the points in the clustering. B.G.Lee et al. [3] proposed a paper with Kernel based Fuzzy C-Means Classifier which is incorporates kernel method with FCM instead of using distance function in FCM the kernel function is used.

Guoying Liu et al. [4] proposed that unsupervised Fuzzy C-Means based image segmentation method helps to select the local information of the image which reduced the noise when compared to normal segmentation techniques. Li Liu et al. [5] used Neighbor searching and Kernel Fuzzy C-Means to find which algorithm is best when compared to each other. It classifies the complex dataset easily. This technique is mainly used to clarify the dimensional behavior of the mechanical system.

Jonathon K. Parker and Lawrence O. Hall [6] proposed Geometrics Progressive Fuzzy C-Means (GOFCM) and Minimum Sample Estimate Random Fuzzy C-Mean (MSERFCM) to estimate the subsample size. GOFCM offers a short run time as it has faster convergence and MSERFCM is used to find the better set for starting clusters. Pradipta Maji and Sushmita Paul [7] incorporated the Fuzzy C-Means algorithm with the rough set and defined as Rough Fuzzy C-Means Algorithm. This algorithm has the ability to handle the overlapping partitions.

Nikhil R. Pal et al. [8] proposed the Possibility Fuzzy C-Means Clustering Algorithm (PFCM) which is based on the typicality value of data to the clusters. It produces membership and possibilities value simultaneously. Pradipta Maji and Sankar K. Pal [9] proposed Generalized Fuzzy C-Means Algorithm which considers the solution for the minimization problem.

II. METHODS AND MATERIAL

We conducted a survey of research work in Fuzzy C-Means Clustering. In this section the working o FCM, K-Means algorithm and other methods that is incorporated with FCM. Some of the algorithms are Fast Generalized Fuzzy C-Mean Clustering (FGFCM), Kernel Fuzzy C-Means Clustering (KFCM), that are explained below:

A. K-Means Clustering

K-Means Clustering partitions the data into K clusters according to the centers. The data that are near to the centers are grouped [14]. The general equation of K-Means Clustering is defined below:

$$V = \sum_{i=1}^{C} \sum_{j=1}^{C_i} ||x - y||^2 \qquad \text{-------- (1)}$$

Where V is the K-Means Clustering, C is the number of cluster centers, C_i is the data point in the i^{th} cluster, $\|x-y\|^2$ is the squared Euclidean distance. To calculate the distance other functions are also used such as Kullback - Leibler divergence also called as information divergence or KL divergence, cosine distance and L_p distance.

B. Fuzzy C-Means Clustering(FCM)

The fuzzy clustering is classified under Soft Clustering i.e., overlapping clustering. It is also indicated as the soft version of K-Means clustering [11]. The general equation of FCM is defined below

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} U_{ij}^m ||x_i - C_j||^2 \qquad \text{--------- (2)}$$

Where U_{ij} is the membership of xi

$$U_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{||x_i - C_j||}{||x_i - C_k||}\right)^{\frac{2}{m-1}}} \qquad \text{--------- (3)}$$

$$C_j = \frac{\sum_{i=1}^{N} U_{ij}^m . x_i}{\sum_{i=1}^{N} U_{ij}^m} \qquad \text{--------- (4)}$$

Where C-No of Clusters, m-fuzziness exponent.

PROCEDURE:

1. Initialize $U=[U_{ij}]$ matrix, $U^{(0)}$
2. At K-step: Calculate the center vectors $c^{(k)}=[c_j]$ with $U^{(k)}$
3. Update $U^{(k)}, U^{(k+1)}$
4. If $\|U^{(k+1)}-U^{(k)}\| < \varepsilon$ then STOP, otherwise return to step 2

where ε-termination tolerance

Figure 1 : FCM Algorithm

C. Fast Generalized Fuzzy C-Means Clustering (FGFCM)

Fast Generalized FCM is used mainly in images. It is a similarity measure which combines the information of spatial and grey level in the image [11].The general equation of FGFCM is as follows:

$$S_{ij} = \begin{cases} e^{-max(|p_i-p_j|,|q_i-q_j|)/\lambda_s - ||x_i-x_j||^2/\lambda_g \sigma_i^2}, i \neq j \\ 0 \quad , i = j \end{cases} \text{--------} (5)$$

Where S_{ij} is the local similarity measure, i is the center of the local window and j is the neighbor around i window.

D. Kernel Fuzzy C-Means Clustering(KFCM)

Kernel Fuzzy C-Means Clustering is used to find the degree of membership and the kernel weight simultaneously. The weights of the cluster incorporated with the cluster procedure and also have multiple kernels. This kernel FCM is mainly used in the images [10]. This algorithm mainly needs the membership matrix. The procedure of MKFCM (Multiple Kernel Fuzzy C-Means) is explained below:

INPUT: Data point X,
No. of Clusters C,
Kernel Function K_k
PROCEDURE:

1. Initialize membership matrix
2. Calculate Normalized memberships
3. Calculate Coefficient, $B_k = \sum_{i=1}^{N} \sum_{c=1}^{C} (u_{ic}^t)^m$
4. Update Weights, $w_k^t = \frac{\frac{1}{B_k}}{\frac{1}{B_1}+\frac{1}{B_2}+\cdots+\frac{1}{B_M}}$
5. Calculate the distance, $D_{ic} = \sum_{k=1}^{M} (w_k^{(t)})^2$
6. Update Membership, $u^{(t)} = \frac{1}{\sum_{c'}^{C} (\frac{D_{ic}^2}{D_{ic'}^2})^{\frac{1}{m-1}}}$
7. Repeat until $||U^{(t+1)}-U^{(t)}|| < \varepsilon$
8. Return $U^{(k)}$

Figure 2 : KFCM Algorithm

III. RESULTS AND DISCUSSION

A. OBSERVATION

TABLE I
ANALYSED PAPERS AND RESULTS

TITLE	AIM	METHODOLOGY	PROS	CONS
Multi-Objective Fuzzy Clustering for Synthetic Aperture Radar Imagery[2]	To optimize the energy function of fast generalize fuzzy C-means and XB index for SAR images	Fast Generalize Fuzzy C-Means (FGFCM) and Xie-Bie (XB) Index	More robust to noise and outliers	Gives best only for images
Smartwatch-based Driver Vigilance Indicator with Kernal-Fuzzy-C-Means-Wavelet Method[3]	To develop the vigilance monitoring application in the smartwatch which can be able to extract the features and predict the class of vigilance driver based on KSCM model	Kernal based Fuzzy C-means Classifier	Increases the vigilance prediction accuracy rate Classification time is less	This method is not suitable for low processing smart phone
Incorporating Adaptive Local Information into Fuzzy Clustering for Image Segmentation[4]	To select the local information from the images the unsupervised FCM based image segmentation method is used	Unsupervised FCM based image segmentation method	Improves segmentation quality and reduces the influences of image noise	Automatic selection of parameter is still a problem

Robust Dataset Classification Approach Based on Neighbor Searching and Kernel Fuzzy C-Means[5]	To clarify the dimensional behavior of acceleration datasets which is achieved from micro electro mechanical systems (MEMS) and complex image segmentation.	Neighbor Searching and Kernel Fuzzy C-Means	Better adaptiveness and robustness in complex dataset classification	Must focus on high dimensional dataset classification and 3D segmentation
Accelerating Fuzzy-C Means Using an Estimated Subsample Size[6]	The statisitical method to estimate the subsample size using two new accelerated algorithm	Geometric Progressive Fuzzy C-Means (GOFCM) and Minimum Sample Estimate Random Fuzzy C-Means (MSERFCM)	Improvement in terms of speed and quality	It could not reach the quality of MODSPFCM(Modified Single-Pass Fuzzy C-Means)
Rough-Fuzzy Clustering for Grouping Functionally Similar Genes from Microarray Data[7]	To demonstrate the effictiveness of rough fuzzy c-means clustering both in qualitative and quantitative on 14 yeast microarray datasets	Rough Fuzzy C-means clustering,	Performs better in 87.50 % than random initialization 82.14 % for optimum parameter value	Quantitatively it doesn't produce good result
A Possibilistic Fuzzy c-Means Clustering Algorithm[8]	PFCM produces memberships and possibilities simultaneously	Possibilistic Fuzzy c-Means Clustering Algorithm	Behaves like FCM or PCM Can be modified for necessary condition	Similarity of algorithm is a problem
Rough Set Based Generalized Fuzzy *C*-Means Algorithm and Quantitative Indices[9]	The efficiency of the algorithm is compared with other algorithm and demonstrated qualitatively and quantitatively	Generalized Fuzzy C-Means Algorithm	Several new measures are defined to evaluate the performance	Its quality is not defined properly

B. DISCUSSION

The necessary of the clustering is rapidly increasing day by day. It reviews about the working of Fuzzy C-Means clustering and other methods such as generalization, kernel, and geometric progressive are embedded with FCM. The FCM algorithms have several advantages and disadvantages. The advantage of FCM is it provides the best result for overlapped dataset; the data point may belong to more than one cluster. It is also defined that it executes faster than K-Means algorithm [12]. The disadvantages that are defined by the user needs to give the number of clusters. It affords multiple numbers of iterations.

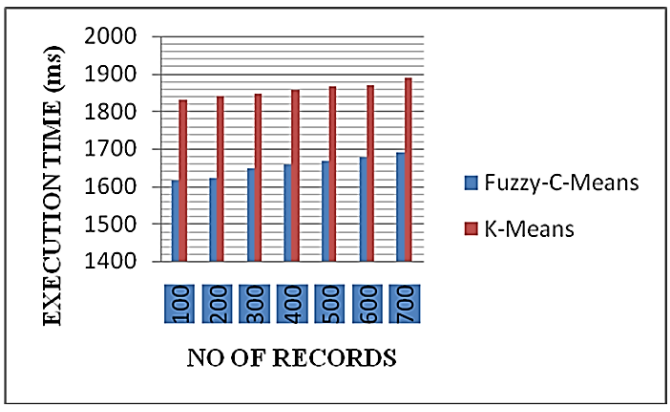

Figure 3 : Comparison of FCM and K-Means Algorithm according to execution time

IV. CONCLUSION

The technology in the storage of data is developing every day. Though the process of storing the data is the challenging process and many new techniques are in the developing stage. The paper focuses about the Fuzzy C-Means Clustering and the different techniques that are incorporated with it. The discussion part gives the comparative result of FCM and K-Means with respect to time. Even if the number of iterations in FCM is more when compared to K-Means the working time is small in FCM.

V. REFERENCES

[1] Jiawei Han, Micheline Kamber and Jian Pei, ”Data Mining Concepts and Techniques”, 3rd ed. Elsevier

[2] Bo Gao and Jun Wang,” Multi-Objective Fuzzy Clustering for Synthetic Aperture Radar Imagery”, IEEE TRANS. ON GEOSCIENCE AND REMOTE SENSING LETTERS .,VOL 12., ISSUE 11.,2015

[3] B.G.Lee., J.H.Park., W.Y.Chung ,” Smartwatch-based Driver Vigilance Indicator with Kernal-Fuzzy-C-Means-Wavelet Method” IEEE TRANSACTION ON SENSORS JOURNAL, VOL 16., ISSUE 1., 2015

[4] Guoying Liu, Yun Zhang, and Aimin Wang,” Incorporating Adaptive Local Information IntoFuzzy Clustering for Image Segmentation “ IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 24, NO. 11, NOVEMBER 2015

[5] Li Liu, Aolei Yang ,Wenju Zhou, Xiaofeng Zhang, Minrui Fei, and Xiaowei Tu,” Robust Dataset Classification Approach Based onNeighbor Searching and Kernel Fuzzy C-Means” IEEE/CAA JOURNAL OF AUTOMATICA SINICA, VOL. 2, NO. 3, JULY 2015

[6] Jonathon K. Parker and Lawrence O. Hall.,” Accelerating Fuzzy-C Means Using an EstimatedSubsample Size”,IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 22, NO. 5, OCTOBER 2014

[7] Pradipta Maji and Sushmita Paul,” Rough-Fuzzy Clustering for GroupingFunctionally Similar Genes fromMicroarray Data”,IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 10, NO. 2, MARCH/APRIL 2013

[8] Nikhil R. Pal, Kuhu Pal, James M. Keller, and James C. Bezdek.” A Possibilistic Fuzzy c-Means Clustering Algorithm”,IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 13, NO. 4, AUGUST 2005

[9] Pradipta Maji and Sankar K. Pal,” Rough Set Based Generalized Fuzzy C-MeansAlgorithm and Quantitative Indices”, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 37, NO. 6, DECEMBER 2007

[10] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1 .300.7332&rep=rep1&type=pdf

[11] http://infoman.teikav.edu.gr/~stkrini/pdfFiles/journals/2010 _TIP.pdf

[12] http://www.indjst.org/index.php/indjst/article/viewFile/ 47757/41449

[13] www.ise.bgu.ac.il/faculty/liorr/hbchap15.pdf

[14] Junjie Wu , ” Advances in k-means Clustering A Data Mining Thinking”, Springer Theses Recognizing Outstanding Ph.D. Research.