Current Trends in Technology and Science
ISSN: 2279-0535. Volume: VII, Issue: V

Comparative Data Analysis based on Fuzzy Clustering Algorithm and FGA

Omlata Dohare

Department of CSE, TIT., RGPV, Bhopal, M.P., India; omlata72@gmail.com

Prof. Aishwarya Vishwakarma

Department of CSE, TIT., RGPV, Bhopal, M.P., India; aishwarya_vishvakrma@gmail.com

Abstract-- The fundamental knowledge clustering drawback may be defined as searching for groups in data or grouping connected objects together. Many alternative clustering techniques are proposed over the years like Partitioning strategies, Density-based strategies and Grid-based methods. During this analysis work vital clustering algorithms particularly representative object based mostly FCM (Fuzzy C-Means) clustering algorithms are compared our proposed algorithms. The quality analysis of the prediction provided by the proposed algorithm was measured by means of statistical tests. These algorithms are applied and performance is evaluated on the idea of the efficiency of clustering output. During this analysis the information bunch algorithms supported fuzzy techniques. These fuzzy bunch algorithms are wide studied and applied in a type of substantive areas. Our proposed Fuzzy clustering with genetic algorithmic program (FGA). These fuzzy bunch algorithms are wide studied and applied in a type of substantive areas. Our proposed Fuzzy clustering with genetic algorithmic program (FCGA)

Keywords- Data Mining, Clustering, Data clustering, Genetic algorithm, Partition clustering, Fuzzy clustering.

I. Introduction

In usually data processing deals with the issue of extracting patterns from the data by paying suspicious attention to computing, communication and human-computer interface problems. clustering is one in all the main data mining tasks to cluster the similar data or information. All clustering algorithms aim of dividing the gathering all information objects into subsets or similar clusters. A cluster could be a collection of objects that are 'similar' between them and are 'dissimilar' to the object's happiness to alternative clusters and a clustering algorithmic program aims to search out a natural structure or relationship in an unlabeled information set. In data processing clustering bound information are well studied within the numerous areas like data mining, machine learning, Bioinformatics, and pattern recognition. However, there's solely preliminary analysis on clustering unsure information. Cluster analysis is additionally recognized as a vital technique for classifying information, finding clusters of a data set supported similarities within the same cluster and dissimilarities between completely different clusters [1].

Clustering: it is the process of assembling the data records into significant subclasses (clusters) in a way that increases the relationship within clusters and reduces the similarity among two different clusters. The main purpose of clustering is to divide a set of objects into significant Groups. The clustering of objects is based on measuring of correspondence between the pair of objects using distance function. Thus, result of clustering is a set of clusters, where object within one cluster is further similar to each other, then to object in another cluster. The Cluster analysis has been broadly used in numerous applications, including segmentation of medical images, pattern recognition, data analysis, and image processing. Clustering is also called data segmentation in some applications because clustering partitions huge data sets into groups according to their resemblance Other names for clustering are unsupervised learning (machine learning) and segmentation [2]. Clustering is used to get an overview over a given data set. A set of clusters is often enough to get insight into the data distribution within a data set. Another important use of clustering algorithm is the preprocessing for some other data mining algorithm. Cluster analysis could be a most vital technique for categorizing a 'mountain' of data into controllable meaningful piles. Cluster analysis could be an information reduction tool that generates subgroups that are any controllable than individual information. like factor analysis, it observes the whole complement of inter-associations among variables. In cluster analysis there's no previous information regarding that components corresponding to every cluster. The grouping or clusters are defined through an analysis of the information. Subsequent multivariate analyses are performed on the clusters as teams. Clustering drawback is regarding partitioning a given information set into teams (clusters) such the information points in an exceedingly cluster are additional like one another than points in several clusters. [3].

Partitioning technique A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset. This partitioning method consists of a set of M clusters and each object belongs to one individual cluster. A partitional Clustering algorithm divides the objects into number of clusters. This method creates various partitions and then evaluate then by using some criterion. There are various types of partitioning methods are [4].

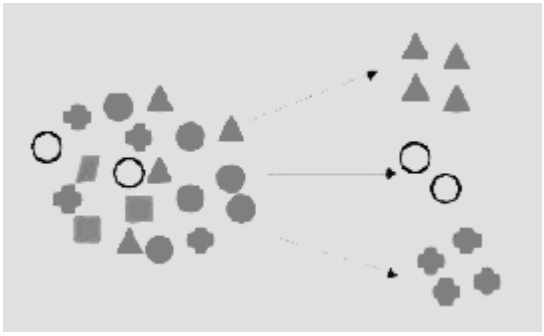

Fig1 Partitioning clustering technique

K-means Algorithm: K-means clustering is a method of vector quantization from signal processing, that is very popular for cluster analysis in data mining. *k*-means clustering defines to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as prototype of the cluster.

FCM Algorithm: FCM clustering algorithms, allocation of information points to clusters is "Fuzzy" instead of being hard. thus, the fuzzy clustering is additionally termed as "Soft clustering". Fuzzy c-means (FCM) may be a typical clustering algorithmic program that permits certain information points to reside in one or a lot of clusters. FCM clustering g algorithmic program is being effectively utilized in pattern detection. The clustering technique depends on minimization of objective function. Fuzzy clustering is basically a strong unsupervised technique for the analysis of information and construction of models. Fuzzy clustering is a lot of and a lot of natural than alternative hard clustering. Objects on the boundaries between multiple categories don't seem to be forced to completely relations to categories, however rather are to be assigned membership degrees between zero and one indicating their partial membership. Fuzzy c-means algorithmic program is wide used. Fuzzy c-means clustering according within the literature for a novel case (m=2) by Joe Dunn in 1974. The FCM point out fuzzy partitioning like that a knowledge purpose is a section of all teams with varied membership grades between zero and one. FCM may be a vital technique of clustering that allows one a part of information to go to extra than two clusters. This technique developed in 1973 and improved in 1981. it's frequently utilized in pattern recognition technique. It depends on minimization of the subsequent objective function, given below figure. where m is any complex quantity larger than one, uij is that the degree of membership of xi within the cluster j, xi is that the ith of d-dimensional measured information, cj is that the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured information and therefore the center [5].

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2$$

II. Literature Survey

Philip K. Maini et al. [6] described a useful survey of fuzzy clustering in main three categories. The first category is basically the fuzzy clustering depends on exact fuzzy relation. The second one is the fuzzy clustering based on single objective function. Finally, it is given an overview of a nonparametric classifier. That is the fuzzy generalized k nearest neighbor rule. The fuzzy clustering algorithms have obtained great success in a variety of substantive areas.

A. K. Jain et al. [7] provided a brief overview of clustering, summarize well known clustering methods, discuss the major challenges and key issues in designing clustering algorithms, and position out some of clustering, immediate attribute selection during data clustering, and large-scale data clustering. Clustering is in the eye of the beholder, so indeed data clustering must involve the user or application needs.

Ngai et al. [8]. K-means & K-medoids are two partioning methods. K-means algorithm in order to cluster the data. This method is referred to as the UK-means algorithm. Proposed the UK-means method extends the k-means method. The UK-means technique measures the distance between an uncertain object and the cluster center (which is a certain point) by the expected distance

OrnellaCominetti et al. [9] showed that the fuzzy spectral clustering method DifFUZZY performs well in a number of data sets, with sizes ranging from tens to hundreds of data points of dimensions as high as hundreds. This includes microarray data, where a typical size of a data set is dozens or hundreds (number of samples, conditions, or patients in medical applications) and dimension is hundreds or thousands (number of genes on the chip). The clustering methodology used in their approach is specifically designed to handle non-Euclidean data sets associated with a manifold structure, as it seamlessly integrates spectral clustering approaches with the evaluation of cluster membership functions in a fuzzy clustering context.

In [10] A Fuzzy Rule-Based cluster algorithmic rule the FRBCemploys a supervised classification approach to try and do the unattended cluster analysis. It tries to mechanically exp

Pham et al. [14] modified lore the potential clusters within the knowledge patterns and establish them with some explicable fuzzy rules. Coincidental classification of knowledge patterns with these fuzzy rules will reveal the particular boundaries of the clusters. for example, the aptitude of FRBC to explore the clusters in knowledge, the experimental results on some benchmark datasets are obtained and compared with different fuzzy cluster algorithms. The clusters such by fuzzy rules are human perceive ready with acceptable accuracy.

M. Punithavalli et al. [11] they enhance the two levels of Prediction Model to achieve higher hit ratio. The uses Fuzzy Possibilistic algorithm for clustering. The experimental result shows that the proposed techniques result in better hit ratio. Forecasting the user's browsing pattern is a significant technique for many applications. The Forecasting results can be utilized for personalization, building proper web site, enhancing marketing strategy, promotion, product supply, getting marketing data, forecasting market trends, and enhancing the competitive strength of enterprises etc. The uses web usage mining technique for predicting the user's browsing behavior. One of the effective existing techniques for web usage mining is the usage of hierarchical agglomerative clustering to cluster users' browsing behaviors. The usage of 2 Levels of Prediction Model framework is explained during this paper that works higher for general cases. However, 2 Levels of Prediction Model suffer from the no uniformity user's behavior. to beat this problem, this paper uses Fuzzy risk formula for bunch. The experimental result shows that the projected technique leads to higher hit rate.

Zhanlong Chen et al., [12] is facing to the higher performance parallel GIS operation necessities and developed a spatial knowledge partitioning approach betting on the minimum distance bunch, understanding load balance once partitioning spatial knowledge. Developing a replacement approach to mend the bunch centers betting on K-Means approach, the centers organized supported the ascending coordinate kind order and distributed smoothly within the area.

Thirumurugan et al., [13] discussed spatial cluster supported statistical procedure of research for determinant the data that is encapsulated within the spatial information. This study shows the importance of the spatial cluster approach accomplished through approaches for example, PAM and CLARA and permits to come back across the restrictions of PAM technique. the FCM objective operate by including a spatial penalty on the membership functions. The penalty term results in an unvaried formula, that is incredibly just like the first FCM and permits the estimation of spatially smooth membership functions.

Ahmed et al. [15] proposed FCM_S wherever the target operate of the classical FCM is changed so as to compensate the intensity in homogeneity and permit the labelling of a picture element to be influenced by the labels in its immediate neighborhood. One disadvantage of FCM_S is that the neighborhood labelling is computed in every iteration step, one thing that's terribly time-consuming.

In Wei Du et al. [16] As a partition primarily based cluster formula, K-Means is wide utilized in several areas for the options of its potency and simply understood. However, it's standard that the K-Means formula might get suboptimal solutions, counting on the selection of the initial cluster centers. during this paper, we propose a projection-based K-Means initialization formula. The planned formula initial use standard Gaussian kernel density estimation methodology to search out the extremely density knowledge areas in one dimension. Then the projection step is to iteratively use density estimation from the lower variance dimensions to the upper variance ones till all the size square measure computed. Experiments on actual datasets show that our methodology will get similar results compared with alternative standard strategies with fewer computation tasks.

III. Simulation Tool and Results Graph

(a) Simulation Tool: The Performance analysis of MATLAB version (R2013a) i.e. used for this thesis Implementation of data mining provides processor optimized libraries for fast execution and computation and performed on input cancer dataset. It uses its JIT (just in time) compilation technology to provide execution speeds that rival traditional programming languages. It can also further advantage of multi core and multiprocessor computers, MATLAB provide many multi-threaded linear algebra and numerical function. These functions automatically execute on multiple computational thread in a single MATLAB, to execute faster on multicore computers. In this thesis, all enhanced efficient data retrieve results were performed in MATLAB (R2013b) to get an enhanced result using fuzzy clustering. MATLAB is the high-level language and interactive environment used by millions of engineers and scientists worldwide. It lets the explore and visualize ideas and collaborate across different disciplines with signal and image processing, communication and computation of results. MATLAB provides tools to acquire, analyze, and visualize data, enable you to get insight into your data in a division of the time it would take using spreadsheets or traditional programming languages. It can also document and share the results through plots and reports or as published MATLAB code.

(b) Results Analysis

In this graph shows the error eliminate analysis in Hepatitis, error eliminate indicate the given condition fulfilled for data clusters, but actually data condition is not fulfilled. In this graph shows error eliminate at time apply one method FGA and FCM with FGA both cases set random values and finding percentage of error eliminate and time. At the time of FGA

proceed of error free data is greater than the previous approach FCM, graph shows minimum error dataset clusters as compare previous approach but time is more as compare previous approach. Below show in figure

Table 1 Error eliminates and time analysis on Hepatitis database

Method	Random Values	Error Rate	Time
FCM	0.45754	3.60362	4.99891
FGA	0.45754	2.13683	6.14644

5.3.2 Results Graph

Hepatitis dataset-based result graph show below. FCM algorithm is more error in dataset clustering as compare to FGA. FGA is better as compare to FCMs because FCM dataset cluster error rate is more but FGA is minim dataset cluster error. FCM algorithm is time take minim as compare to FGA.

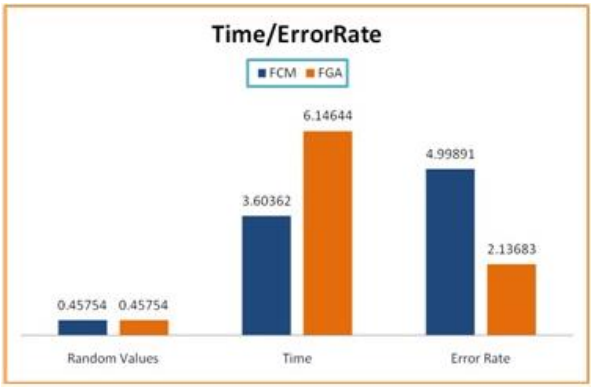

Fig2 Error rate analysis propose algorithm and pervious algorithm

IV. Conclusion

Our recent research work done in this area proposed an improved clustering algorithm based on FGA. Clustering techniques play a key role in many applications. Many researches are being done in this area for the betterment of the overall performance of the clustering techniques. Clustering is a potential technique in many data mining applications. Result analysis processes used two database gravesthairaid and BCWD Dataset. Classification of plants and animals given their features is also a major application area in bioinformatics. In World Wide Web, Document classification and clustering weblog data to discover groups of similar access patterns is an active area of research. Clustering has been one of the most very important techniques in the field of data mining. Recently, clustering is applied in various applications. Data mining clustering based on fuzzy techniques hence it is the most efficient technique when compared with the clustering techniques. Close-cluster groups can find the corresponding solution to update both cluster number and cluster centers. Proposed algorithm is better as compare to FCM because dataset cluster error rate is more but proposed algorithm is minim dataset cluster error. FCM algorithm is time take minim as compare to proposed algorithm. The graph shows above with the proposed algorithm high accuracy as compared to the normal pervious algorithm FCM has because the FGA data set has less error in dataset also called filtered data set.

References

[1]. V. S. Rao and Dr. S. Vidyavathi, "Comparative Investigations and Performance Analysis of FCM and MFPCM Algorithms on Iris data", Indian Journal of Computer Science and Engineering, vol.1, no.2, 2010 pp. 145-151.

[2]. D.H. Fisher. "Conceptual clustering, learning from examples, and inference", Proc. 4th Int. Workshop on Machine Learning, Irvine, CA, Pp. 38–50, 1987.

[3]. S. AnithaElavarasi and Dr. J. Akilandeswari and Dr. B. Sathiyabhama, A Survey on Partition Clustering Algorithms, January 2011.

[4]. Pradeep Rai Shubha Singh, "A Survey of Clustering Techniques", International Journal of Computer Applications (0975 – 8887) Volume 7– No.12, October 2010.

[5]. J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algoritms", Plenum Press, New York,1981.

[6]. Philip K. Maini, Don Kulasiri, Sijia Liu and Radek Erban,, "Diffuzzy: A fuzzy clustering algorithm for complex data sets" , International Journal of Computational Intelligence in Bioinformatics and Systems Biology vol.1, no.4,pp. 402-417, 2010.

[7]. Anil K. Jain, "Data clustering: 50 years beyond Kmeans", Pattern Recognition Letters, no.31, pp. 651– 666, 2010.

[8]. W.K. Ngai, B. Kao, C.K. Chui, R. Cheng, M. Chau, and K.Y. Yip, "Efficient Clustering of Uncertain Data," Proc. Sixth Int'l Conf. Data Mining (ICDM), 2006.

[9]. Jian Yu and Miin-Shen Yang, "A Generalized Fuzzy Clustering Regularization Model with Optimality Tests and Model Complexity Analysis", IEEE Transactions on Fuzzy Systems, Vol. 15, No. 5, Pp. 904–915, 2007.

[10]. Eghbal G. Mansoori, "FRBC: A Fuzzy Rule-Based Clustering Algorithm", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 19, NO. 5, OCTOBER 2011

[11]. R. Khanchana and M. Punithavalli., "Web Usage Mining for Predicting Users' Browsing Behaviors by using FPCM Clustering", IACSIT International Journal of Engineering and Technology, Vol. 3, No. 5, October 2011.

[12]. Zhanlong Chen; Liang Wu; Dingwen Zhang, "Spatial data partitioning based on the clustering of minimum distance criterion", International Conference on Computer Science and Service System (CSSS), Pp. 2583 – 2586, 2011.

[13]. Thirumurugan, S.; Suresh, L., "Statistical spatial clustering using spatial data mining", IET International Conference on Wireless, Mobile and Multimedia Networks, 26 – 29, 2008.

[14]. D. Pham, "Fuzzy clustering with spatial constraints," in Proc. Int. Conf. Image Processing, New York, 2002, vol. II, pp. 65–68.
[15]. M. Ahmed, S. Yamany, N. Mohamed, A. Farag, and T. Moriarty, "A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data," IEEE Trans. Med. Imag., vol. 21, pp. 193–199, 2002.
[16]. Wei Du, Hu Lin, Jianwei Sun, Bo Yu and Haibo Yang, "A New Projection-based K-Means Initialization Algorithm", Proceedings of 2016 IEEE Chinese Guidance, Navigation and Control Conference August 12-14, Nanjing, China, 2016.
[17]. M. Ahmed, S. Yamany, N. Mohamed, A. Farag, and T. Moriarty, "A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data," IEEE Trans. Med. Imag., vol. 21, pp. 193–199, 2002.
[18]. Wei Du, Hu Lin, Jianwei Sun, Bo Yu and Haibo Yang, "A New Projection-based K-Means Initialization Algorithm", Proceedings of 2016 IEEE Chinese Guidance, Navigation and Control Conference August 12-14, Nanjing, China, 2016.
[19]. D. Pham, "Fuzzy clustering with spatial constraints," in Proc. Int. Conf. Image Processing, New York, 2002, vol. II, pp. 65–68.
[20]. Altug Akay, Andrei Dragomir, Bjorn Erik Erlandsson, "A Novel Data-Mining Approach Leveraging Social Media to Monitor Consumer Opinion of Sitagliptin", IEEE Journal of Biomedical and Health Informatics, Vol. 19, No. 1, Pp. 2168-2194, January 2015.
[21]. E. Chandra, V.P. Anuradha, "Survey on Clustering Algorithms for Data in Spatial Database Management Systems", International Journal of Computer Applications vol. 24, no.9, pp.975 – 8887, June 2011.
[22]. Lin Xiao-ping; Mao Zheng-yuan; Liu Jian-hua, "A Spatial Clustering Method by Means of Field Model to Organize Data", Second WRI Global Congress on Intelligent Systems (GCIS), Pp. 129 – 131, 2010.