International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 03, Issue: 05 (SEPTEMBER -OCTOBER, 2017)

A Survey of Data Mining Clustering Algorithms Based on Partitioning Clustering

Omlata Dohare
M. Tech. Scholar Department of CSE
TIT., RGPV, Bhopal,
M.P., India
omlata72@gmail.com

Prof. Aishwaryai Vishwakarma
A.P, Department of CSE
TIT., RGPV, Bhopal
M.P., India
Aishwarya_vishvakarma@gmail.com

Abstract--*In clustering is used to explain methods for grouping of unlabeled information. Clustering is a very important task in data processing to cluster information into significant subsets to retrieve data from a given dataset. Clustering is also called unsupervised learning since the information objects are pointed to a group of clusters which might be taken as categories additionally. The chief objective of the clustering is to present a group of comparable records. The clustering drawback has been targeted by several researchers. Knowledge bunch could be a technique within which logically similar info is physically stored together. The numbers of disk accesses are to be reduced so as to extend the efficiency within the information systems. 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. 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 (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 objects 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, than 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 algorithms 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].

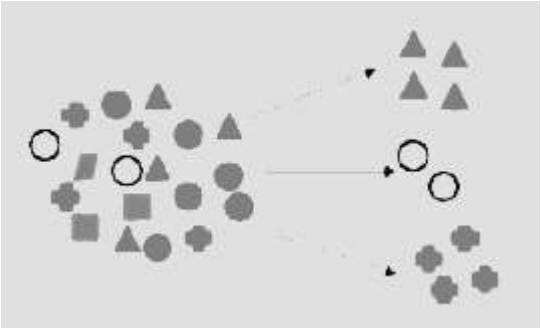

Fig1 Partitioning clustering technique

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].

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 fundamental case developed by Jim Bezdek in his PhD thesis at Cornell University in 1973. It is improved by Bezdek in 1981. The FCM indicates 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

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

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]..

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.

Anil K. Jain [7] provided a brief overview of clustering, summarize well known clustering methods, discuss the major challenges and key issues in designing clustering algorithms, and point out some of the emerging and useful research directions, including semi-supervised clustering, ensemble clustering, simultaneous feature 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 portioning 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

Ornella Cominetti 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 Clustering Algorithm the FRBC employs a supervised classification approach to do the unsupervised cluster analysis. It tries to automatically explore the potential clusters in the data patterns and identify them with some interpretable fuzzy rules. Simultaneous classification of data patterns with these fuzzy rules can reveal the actual boundaries of the clusters. To illustrate the capability of FRBC to explore the clusters in data, the experimental results on some benchmark datasets are obtained and compared with other fuzzy clustering algorithms. The clusters specified by fuzzy rules are human understand able 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 results 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 Two Levels of Prediction Model framework is explained in this paper which works better for general cases. However, Two Levels of Prediction Model suffer from the heterogeneity user's behavior. To overcome this difficulty, this paper uses Fuzzy Possibility algorithm for clustering. The experimental result shows that the proposed technique results in higher hit rate.

Zhanlong Chen et al., [12] is facing to the better performance parallel GIS operation requirements and developed a spatial data partitioning approach depending on the minimum distance clustering, understanding load balance when partitioning spatial data. Developing a new approach to fix the clustering centers depending on K-Means approach, the centers arranged based on the ascending coordinate sort order and distributed smoothly in the space

Thirumurugan et al., [13] discussed spatial clustering based on statistical method of analysis for determining the knowledge which is encapsulated in the spatial database. This study shows the importance of the spatial clustering approach accomplished through approaches for instance, PAM and CLARA and permits to come across the restrictions of PAM technique.

Pham et al. [14] modified the FCM objective function by including a spatial penalty on the membership functions. The penalty term leads to an iterative algorithm, which is very similar to the original FCM and allows the estimation of spatially smooth membership functions.

Ahmed et al. [15] proposed FCM_S where the objective function of the classical FCM is modified in order to compensate the intensity in homogeneity and allow the labelling of a pixel to be influenced by the labels in its immediate neighborhood. One disadvantage of FCM_S is that the neighborhood labelling is computed in each iteration step, something that is very time-consuming

In Wei Du et al. [16] as a partition based clustering algorithm, K-Means is widely used in many areas for the features of its efficiency and easily understood. However, it is well known that the K-Means algorithm may get suboptimal solutions, depending on the choice of the initial cluster centers. In this paper, we propose a projection-based K-Means initialization algorithm. The proposed algorithm first employ conventional Gaussian kernel density estimation method to find the highly density data areas in one dimension. Then the projection step is to iteratively use density estimation from the lower variance dimensions to the higher variance ones until all the dimensions are computed. Experiments on actual datasets show that our method can get similar results compared with other conventional methods with fewer computation tasks.

III. EXPECT OUTCOME

Research in the field of data mining clustering based on fuzzy techniques and identifies various challenges in the field data mining. Dataset analysis bases on clustering and divided different cluster with Error free data and useful information increase accuracy and minimize error based on our proposed clustering technique. A efficiently cluster according to their importance and best possible answer.

IV. CONCLUSION

In this paper has presented a survey of most recent research work done in this area. 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. Market basket analysis is one of the main applications of clustering in supermarkets. In marketing, clustering finds groups of customers with similar behavior given a large database of customer data containing their properties and past buying records. 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 vital techniques in the field of data mining. Recently, clustering is applied in various applications. This survey concentrates on efficiency of the clustering approaches. This survey utilized a clustering algorithm to provide the best clustering results with greater clustering accuracy and reduced mean squared error and execution time, respectively with quick convergence. Survey is done on the data mining clustering based on fuzzy techniques hence it is the most efficient technique when compared with the clustering techniques.

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 Algorithms", 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.