Survey on Different Enhanced K-Means Clustering Algorithm

Kapil Joshi[#1], Himanshu Gupta[#2], Prashant Chaudhary[#3] Punit Sharma[#4]

[#1234] *Assistant Professor, Dept. Of Computer Science, Uttaranchal University, Dehradun, India*

Abstract *— Data Mining is justify technique used to extract, meaning ful information from mountain of data and Clustering is an important task in Data Mining process which can be used for the purpose to make groups or clusters of the particular given data set which is based on the similarity between them. K-Means clustering is a clustering procedure in which the given data set is divided into K i.e number of clusters. The impact factor of k-means is its simplicity, high efficiency and scalability. However, is also comprises of number of limitations: random selection of initial centroids, number of cluster K need to be initialized and influence by outliers. In view of these deficiencies, this paper represents a survey of improvements done to traditional k-means to handle such limitations and we will compare K-means clustering algorithm with various clustering algorithm.*

Keywords *— Data Mining, Clustering, K-means algorithm, Fuzzy C-means algorithm, Genetic algorithm, Genetic algorithm-K-Means (GAKM).*

I. INTRODUCTION

The field of data mining and knowledge discovery is emerging as a new, fundamental research area with important applications to science, engineering, medicine, business, and education. Data mining attempts to formulate analyse and implement basic induction processes that facilitate the extraction of meaningful information and knowledge from unstructured data [1]. Size of databases in scientific and commercial application is huge where the number of records in a dataset can vary from some thousand to thousand of millions [2].

Clustering may be defined as a data reduction tool i.e. used to create subgroups that are more and more manageable than individual datum. Basically, clustering is justify as a process used for grouping a large number of data into meaningful groups or clusters based on some similarity between data. Clusters are the groups that have data similar on basis of common features and dissimilar to data in other clusters.

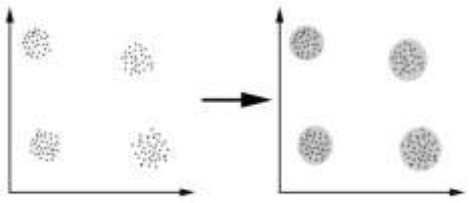

Fig. 1 Group Clustering

II. TYPES OF CLUSTERS

A. *Well-separated clusters:*

A cluster is a collection of points such that any other point in a cluster closer or more similar to each and every other point in the cluster than to any point not in the cluster.

Fig. 1 Well-separated Clusters

B. *Centre based clusters:*

A cluster is a group of objects so that an object in a cluster is more closer to the "centre" of a cluster, than to the centre of other cluster – The centre of a cluster is called a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster**.**

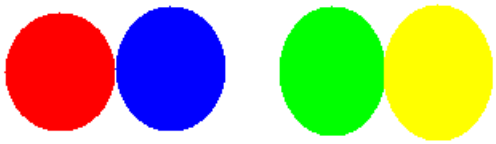

Fig. 3 Centre Based Clusters

C. *Contiguous clusters:(Nearest neighbour or transitive)*

A cluster is a group of points such that a single point in a cluster is closer to one or more other points in the cluster than to any other point not in the cluster.

Fig. 4 Contiguous clusters

D. *Density- based clusters:*

A cluster is dense region of points, which is individual separated by low-density regions, from the other regions of high density regions. It used when the clusters are very irregular, and when noise and outliers are available.

Fig. 5 Density-based clusters

III.LITERATURE REVIEW

Wang Shunye et al[3] Motivated by the problem of random selection of initial centroid and similarity measures ,the researcher presented a new K-means clustering algorithm based on dissimilarity. This improved k-means clustering algorithm basically consists of 3steps.The first step discussed is the construction of the dissimilarity matrix i.e dm.Secondly ,Huffman tree based on the Huffman algorithm is created according to dissimilarity matrix. The output of Huffman tree gives the initial centroid. Lastly the k-means algorithm is applies to initial centroids to get k cluster as output. Iris, Wine and Balance Scale datasets are selected from UIC machine learning repository to test the proposed algorithm. Compared to traditional k-means the proposed algorithm gives better accuracy rates and results.

Navjot Kaur, Navneet Kaur[4]enhanced the traditional k-means by introducing Ranking method. Author introduces Ranking Method to overcome the deficiency of more execution time taken by traditional k-means. The Ranking Method is a way to find the occurrence of similar data and to improve search effectiveness. The tool used to implemented the improved algorithm is Visual Studio 2008 using C#. The advantages of k-means are also analysed in this paper. The author finds k-means as fast ,robust and easy understandable algorithm. He also discuss that the clusters are non-heirarchical in nature and are not overlapping in nature. The process used in the algorithm takes student marks as data set and then initial centroid is selected.Euclidean distance is then calculated from centroid for each data object. Then the threshold value is set for each data set. Ranking Method is applied next and finally the clusters are created based on minimum distance between the data point and the centroid. The future scope of this paper is use of Query Redirection can be used to cluster huge amount of data from various databases.

Yang [5] 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.

Md.SohrabMahmud et al [7] gave an algorithm to compute better initial centroids based on heuristic method. The newly presented algorithm results in highly accurate clusters with decrease in computational time. In this algorithm author firstly compute the average score of each data points that consists of multiple attributes and weight factor. Merge sort is applied to sort the output that was previously generated. The data points are then divided into k cluster i.e number of desired cluster. Finally the nearest possible data point of the mean is taken as initial centroid. Experimental results shows that the algorithm reduces the number of iterations to assign data into a cluster. But the algorithm still deals with the problem of assigning number of desired cluster as input.

Pallavi Purohit and Ritesh Joshi et al[8] proposed an improved approach for original K-means clustering algorithm due to its certain limitations. The main reason for poor performance of K-means algorithm is selection of initial centroids randomly. The proposed algorithm deals with this problem and improves the performance and cluster quality of original k-means algorithm. The new algorithm selects the initial centroid in a systematic manner rather than randomly selecting. It first find out the

closest data points by calculating Euclidian distance between each data point and then these points are deleted from population and forms a new set. This step is repeated on new set by finding data points that are closest to each other. Performance comparison is done using MATLAB tool. The proposed algorithm gives more accurate results and also decreases the mean square distance. But the proposed algorithm works better for dense dataset rather than sparse.

Juntao Wang et al [9] discuss an improved k-means clustering algorithm to deal with the problem of oulier detection of existing k-means algorithm. The proposed algorithm uses noise data filter to deal with this problem. Density based outlier detection method is applied on the data to be clustered so as to remove the outliers. The motive of this method is that the outliers may not be engaged in computation of initial cluster centres. In the next step fast global k-means algorithm proposed by Aristidis Likas is applied to the output generated previously. The results between k-means and improved k-means are compared using Iris,Wine,Abolone datasets. The Factors used to test are clustering accuracy and clustering time. The disadvantage of the improved k-means is that while dealing with large data sets ,it will cost more time.

Raju G, et al [10] gave a comparative analysis between k-means clustering algorithm and fuzzy clustering algorithm. In this paper the researcher also discuss the advantages and limitations of fuzzy c-means algorithm. K-means is a partional based clustering algorithm whereas Fuzzy c-means is non partional based clustering algorithm.Fuzzy c-means

mainly works in two process. In the first process cluster centers are calculated and in second the data points are assigned to calculated cluster centre with the help of Euclidean distance. This process is almost similar to conventional k-means with a little difference. In fuzzy c-means algorithm membership value ranging from 0 to 1 is assigned to data item in cluster.0 membership indicates that the data point is not a member of cluster whereas 1 indicates the degree to which data point represents a cluster. The problem faced by fuzzy c-means algorithm is that the sum of membership value of data points in each cluster is restricted to 1.Algorithm also face problem in dealing with outliers. On the other hand comparison with k-means shows that the fuzzy algorithm is efficient in obtaining hidden patterns and information from natural data with outlier points

IV. CLUSTERING ALGORITHM

Clustering algorithms are basically used in an unsupervised manner. They are presented with a set of data instances that should be grouped according to some standard manner of similarity. The basic algorithm has finally access only to the particularly group of features describing each and every object; This is not given any information (e.g., labels) as to where each of the instances must be placed within the partucular partition .Data clustering is a data exploration technique that allows objects with similar characteristics to be grouped together in order to facilitate their further processing. Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. Clustering problem is about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters.

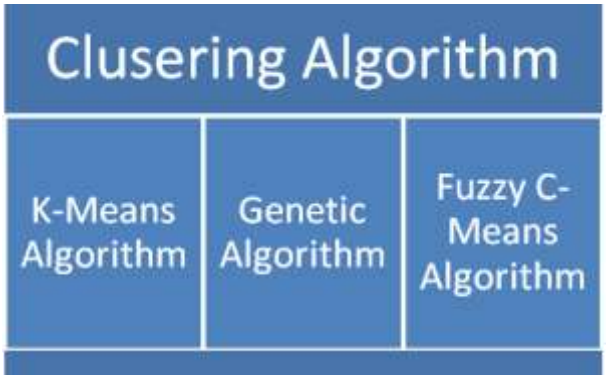

Fig. 6 Clustering Algorithm

A. *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.

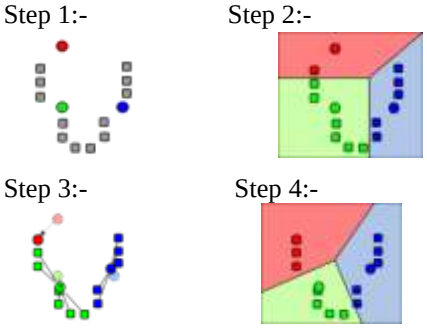

Fig 7 Demonstration of the Standard K-Means Algorithm

B. *Fuzzy c--means Algorithm:*

Fuzzy clustering is really a powerful unsupervised method for the analysis of data and construction of models. Fuzzy clustering is more and more natural than other hard clustering. Objects on the boundaries between multiple classes are not forced to totally relations to classes, but rather are to be assigned membership degrees between zero and one indicating their partial membership. Fuzzy c-means algorithm is widely used. Fuzzy c-means clustering reported in the literature for a unique case (m=2) by Joe Dunn in 1974. The basic case developed by Jim Bezdek in his PhD thesis at Cornell University in 1973. It can be improved by Bezdek in 1981. The FCM indicates fuzzy partitioning like that a data point can be a part of all groups with various membership grades between zero and one.

Fuzzy C-means (FCM) is a ultimate method of clustering that permits one part of data to belong to more than two clusters. This method developed in 1973 and improved in 1981. It is frequently used in pattern recognition method. It depends on minimization of the following 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 real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center.

C. *Genetic Algorithm K-means (GAKM):*

Jenn-Long Liu, Yu-Tzu Hsu and Chih-Lung Hung [6] proposed GAKM a hybrid method that combines a genetic algorithm (GA) and K-means algorithm. The

function of GAKM is to determine the optimal weights of the attributes and cluster centres of clusters that are needed to classify the dataset. Genetic algorithm is a stochastic search algorithm which is based on the Darwinian principal of natural selection and natural genetics.

The Genetic algorithm automatically obtain the knowledge about the search space that is the space for possible and feasible solutions. GAs are inspired by Darwinian theory of the survival of the fittest.

V. TABLE I

COMPARISION OF VARIOUS CLUSERING ALGORITHMS

Algorithm Name / Parameter	K-Means Algorithm	Fuzzy C-Means Algorithm	Genetic Algorithm
Method	Partitioning	Non Partitioning	metaheuristic
Proposed by	James MacQueen	Tamura	John Henry Holland
Speed	Fast	Fast	Fast
Dataset	Yes	Yes	Yes
Perfoamance	Simple and Robust	Simple and Robust	Simple to understand
Approach	Based on vector quantization it use Cluster centers to model the data.	This algorithm attempts to a finite collection of n no. of elements.	A genetic algorithm requires a genetic representation of the solution domain.
Learning Process	Unsupervised	Unsupervised	Optimization Method
Used For	To partition n observation into a clusters	Data elements relates to more than one clusters	optimization and search problems
Time Complexity	O(ncdi)	$O(ndc^2i)$	problem dependent

VI. CONCLUSIONS

Clustering has a crucial role in different applications. The commonly used efficient clustering algorithm is k-means clustering. K-means clustering is an important topic of research now a days in data mining. This paper have presented a survey of most recent research work done in this area. However k-means is still at the stage of exploration and development. The survey concludes that many improvements are basically required on k-means to improve problem of cluster initialization, cluster quality and efficiency of algorithm. On the other hand In this paper we study the three clustering algorithms one is Fuzzy C-Means algorithm , simple K-Means partitioning algorithm and the GAKM an hybrid algorithm which is the combination of simple K-Means and Genetic Algorithm. K-means is combine with GA to get the optimize no. of clusters from the result of simple K-Means algorithm .Both algorithm are simple to understand and can be applicable for various type of data like genomic data set, numerical data set.

REFERENCES

[1] M H Dunham, "Data Mining: Introductory and Advanced Topics," Prentice Hall, 2002.

[2] Birdsall, C.K., Landon, "A.B.: Plasma Physics via Computer Simulation," Adam Hilger, Bristol

(1991).

[3] Wang Shunye "An Improved K-means Clustering Algorithm Based on Dissimilarity" 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC)Dec 20-22, 2013, Shenyang, China IEEE

[4] Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur "EFFICIENT KMEANSCLUSTERING ALGORITHM USING RANKING METHOD IN DATA MINING" ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May2012

[5] Don Kulasiri, Sijia Liu, Philip K. Maini and RadekErban, "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.

[6] Jenn-Long Liu, Yu-Tzu Hsu, Chih-Lung Hung , " Development of Evolutionary Data Mining Algorithms and their Applications to Cardiac Disease Diagnosis", WCCI 2012 IEEE World Congress on Computational Intelligence, June 2012.

[7] Md. Sohrab Mahmud, Md. Mostafizer Rahman, and Md.Nasim Akhtar "Improvement of K-means Clustering algorithm with better initial centroids based on weighted average" 2012 7th International Conference on Electrical and Computer Engineering 20-22 December, 2012, Dhaka, Bangladesh, 2012 IEEE

[8] Pallavi Purohit "A new Efficient Approach towards k-means Clustering Algorithm" ,International journal of Computer Applications,Vol 65-no 11,march 2013

[9] Juntao Wang & Xiaolong Su "An improved K-Means clustering algorithm" 2011 IEEE

[10] Raju G, Binu Thomas, Sonam Tobgay and Th. Shanta Kumar"Fuzzy Clustering Methods in Data Mining:A comparative Case Analysis" 2008 International Conference on advanced computer theory and engineering,2008 IEEE