International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Review on Various Enhancements in K means Clustering Algorithm

Gurpreet Virdi, Neena Madan
CSE, GNDU RC, Jalandhar, Punjab, India

ABSTRACT

Data Mining is the technique used to mine the data that is finding the useful information from the raw data. As day-by-day data is increasing it becomes difficult for us to analyzing such a huge amount of data. For analyzing such data, we have various clustering techniques in data mining. Clustering is the technique used to divide the data into the various clusters. Clustering is done based on similarities within the elements that are to be clustered. K means is one of the clustering algorithm that is widely used because of its efficiency and simplicity. In this paper, we will review various enhancements in k means clustering algorithm.

Keywords: Data Mining, Clustering, K means

I. INTRODUCTION

Data Mining[10] is known as the process of analysing data to extract interesting patterns and knowledge. It turns the large collection of data into knowledge. That is it mines the data in to useful information. This information is currently used for wide range of applications like customer retention, education system, production control, healthcare, market basket analysis, manufacturing engineering, scientific discovery and decision making etc.

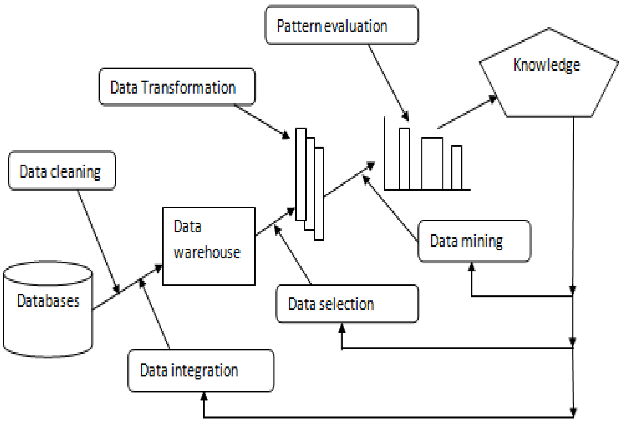

Figure1. Data Mining Process

A. KDD process

The Discovery of knowledge in Databases process includes following steps

1) Data, which is not relevant and contains noise, is cleaned. This step contains cleaning the data.
2) At data integration step, heterogeneous data is combined with the different data sources.
3) In the selection step, applicability of analyzed data is taken into consideration.
4) Under the transform step, with respect to various mining techniques changes in the data are occurs.
5) Data Mining is used to find the required and unique patterns using many available techniques.
6) The pattern evaluation step involves evaluating the required patterns.
7) In the last step the exposed results, which includes knowledge are represented.

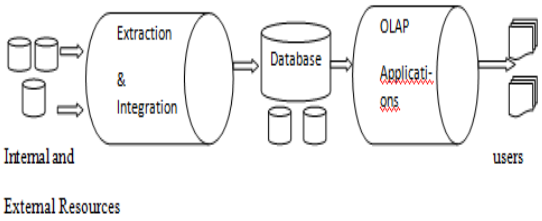

Figure 2. Data Cleaning Process

Components involved in KDD process:

1. External and Internal Resources: First step is to collect the raw data from number of resources either internal resources or external resources.

2. Extraction and Integration: In this step, first it extracts data from different resources and converts it into original format. Data cleaning is a process in which missing values are filled, it smooth noisy data and remove inconsistencies. In data integration, integration of multiple database, data cubes and files takes place.

3. Database: After that data is stored in database and data mart.

4. OLAP Application: OLAP can be used as for discovery in data mining for previously discerned relationship between data items.

II. CLUSTERING

Data clustering[10] is the technique of clustering the data into different groups and these formed groups are known as Clusters[20] in Data mining. Data elements are clustered into different groups based on the similarities and dissimilarities. Basic motive is to keep the similar data elements together in one group. The elements in one cluster have similar properties or similar behavior.

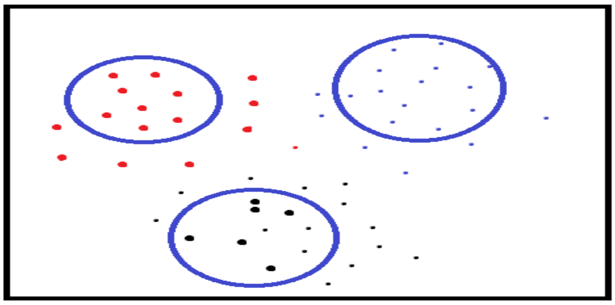

Figure 3. Clustering

There are many clustering algorithms used for clustering. The major fundamental clustering methods can be classified into following categories:

A. Partitioning Methods

The general criterion for partitioning is a combination of high similarity of the samples inside of clusters with high dissimilarity between distinct clusters. Most partitioning methods are distance-based. Given k, the number of partitions to construct, a partitioning method creates an initial partitioning and then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. In this there is partition of a data set containing n objects into a set of k clusters. The goal is, given a k, find a partition of k clusters that optimizes the chosen partitioning criterion. Here k is input parameters. E.g. K-means and K-centriod.

Figure4. Partitioning Clustering

B. Hierarchical Methods

In this method hierarchical decomposition of the given set of data objects is created. It can be classified as being either agglomerative or divisive based on how hierarchical decomposition is formed. Agglomerative approach is the bottom up approach starts with each object forming a separate group. It then merges groups close to one another until all the groups are merged into one. In this type of clustering it is possible to view partitions at different level of granularities using different types of K. E.g. Flat Clustering

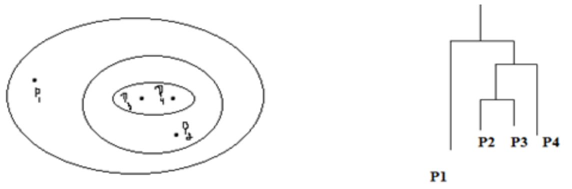

Figure5. Hierarchical Clustering

C. Density Based Methods

Most partitioning methods cluster objects based on distance between objects. Spherical shaped clusters can be discovered by these methods and encounter difficulty in discovering clusters of arbitrary shapes. So for arbitrary shapes new methods are used known as density-based methods which are based on the notion of density. It helps to discover arbitrary shape clusters. It also handles noise in the data. It is one time scan. It requires density parameters also.

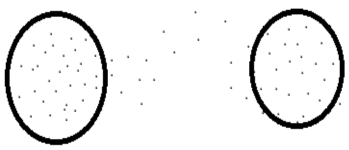

Figure 6. Density based clustering

D. Grid Based Methods

Grid based methods quantize the object space into a finite number of cells that form a grid structure. It is a fast method and is independent of the number of data objects and depends only on the number of cells in each dimension in the quantized space. In this objects are together to form grid. Grid-based algorithms quantize the space into a finite number of grids and perform all operations on this quantized space. These approaches have the advantage of fast processing time independent of the data set size and are dependent only on the number of segments in each dimension in the quantized space.

III. K-MEANS CLUSTERING ALGORITHM

The k-means clustering algorithm is the basic algorithm which is based on partitioning method which is used for many clustering tasks especially with low dimension datasets. It uses k as a parameter, divide n objects into k clusters so that the objects in the same cluster are similar to each other but dissimilar to other objects in other clusters. The algorithm attempts to find the cluster centres, (*C1 …… Ck*), such that the sum of the squared distances of each data point, *xi*, $1 \leq i \leq n$, *to* its nearest cluster centre *Cj*, $1 \leq j \leq k$, is minimized. First, the algorithm randomly selects the k objects, each of which initially represents a cluster mean or centre. Then, each object *xi* in the data set is assigned to the nearest cluster centre i.e. to the most similar centre. The algorithm then computes the new mean for each cluster and reassigns each object to the nearest new centre. This process iterates until no changes occur to the assignment of objects. The convergence results in minimizing the sum-of-squares error that is defined as the summation of the squared distances from each object to its cluster centre. The following procedure summarizes the k-means algorithms:

Algorithm: k-means:-The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

k: the number of clusters,

D: a data set containing n objects.

Output:

A set of k clusters.

Method:

1) randomly choose k objects from D as the initial cluster centers;
2) repeat
3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;
5) until no change;

IV. LITERATURE REVIEW

Xiaoyan Wang and Yanping Bai[1] explained that global k-means algorithm results in singleton clusters and sometimes had bad initialization positions. They modified global k-means to eliminate singleton clusters at first and then apply Minmax k-means cluster error method to global k-means to eliminate the effect of bad initialization. The proposed global Minmax k-means implemented on some popular data sets and compared with k-means, global k-means, Minmax k-means. They concluded that proposed global MinMax k-means algorithm shows better results than others.

Wei Du, Hu Lin, Jianwei Sun, Bo Yu and Haibo Yang[2] explained that k means algorithm is widely used in many areas because of its simplicity and efficiency. They propose a projection based k means initialization algorithm. The proposed algorithm first use conventional Gaussian kernel density estimation method to find highly density areas in one dimension. That in this they get good initial center for k means clustering algorithm. Then they compared their results with conventional method's results and conclude that their method gives similar results with same accuracy but has fewer computation tasks as compared to those conventional methods.

Caiquan Xiong, Zhen Hua, Ke Lv, Xuan Li[3] explained that K-means algorithm is widely used algorithm in many fields. But k-means algorithm has sensitivity to the initial centers, which leads to the final outcome that depends upon the initial centers. For this they proposed improved k-means text clustering algorithm by optimizing initial centers. The key idea of this algorithm is to determine the initial cluster centers with the density parameter of the data objects. Through this they show that improved k means algorithm can improve the stability and accuracy of text clustering. By the proposed algorithm they are able to eliminate the sensitivity of k means algorithm to the initial centers.

Akanksha Choudhary, Mr. Prashant Sharma, Mr. Manoj Singh[4][13] focused on improving the efficiency and effectiveness of the k means clustering algorithm. For this they proposed three score based initialization methods. They implement these methods on normalized data to prove the superiority of the proposed work. They use two normalization techniques that are MinMax normalization and Z-score normalization to normalize the data. The efficiency and accuracy of the proposed methods are demonstrated through several experiments. Also the comparison was made with other works. This was done to show the impact of improvement.

Anshul Yadav, Sakshi Dhingra's paper [5][14] explained clustering is the technique of grouping the similar objects in to the groups known as clusters. Then they discussed the k-means algorithm. This algorithm has many pitfalls. They focused on one of the limitation of generating empty clusters. For eliminating this limitation they proposed the enhanced k-means algorithm. They implement this algorithm on two databases from real life projects that are from sugar mill and MLM Company. And concluded that the empty clusters that are produced by old k-means algorithm are removed when enhanced k-means algorithm is implemented on the data.

Shruti Kapil and Meenu Chawla's paper [6][16]explained that Data Mining plays very important role in extracting useful information from huge amount of data. They studied one of the Clustering algorithm that is k-means algorithm. K-means algorithm is widely used algorithm because of its efficiency. They make use of two distance functions that were Euclidean distance function and Manhattan distance function. Then they evaluate the performance of k-means clustering algorithm using these two distance functions. They concluded that the performance of algorithm with Euclidean distance is better than the algorithm with Manhattan distance.

K.Rajalakshmi et.al,[7][17] represent that the medical field is growing extremely fast. A huge amount data is generated by this field every day. It becomes very difficult to handle this data, so there is a need of a technique to handle this data. To turn these data into useful pattern, there is a need to mine the data. The medical data mining are useful to produce optimum results on prediction based system of medical line. This paper analyzes various disease predictions techniques using K-means algorithm. This data mining based prediction system reduces the human effects and is cost effective.

Rishikesh Suryawanshi, Shubha Puthran's paper [8][18] explained that K means algorithm is widely used algorithm in clustering. But there are many pitfalls regarding this algorithm. K means algorithm is inefficient for big data, also its final result depends upon the initial selected centroids. Its computation is also expensive. For overcoming its pitfalls they discussed various enhancement techniques. These enhancement techniques are Refined initial cluster centers method, A parallel K-means algorithm, A parallel k-means clustering algorithm based on Map Reduce technique, Determine the initial centroids of the clusters and Assign each data point to the appropriate clusters, An efficient enhanced k-means clustering algorithm, Variation in K-means algorithm and proposed parallel K-means clustering algorithm, A New Initialization Method to Originate Initial Cluster Centers for K-Mean Algorithm, Dynamic Clustering of Data with Modified K-Means Algorithm. In this paper

the limitations of clustering technique are mapped with the all these mentioned modified k-means approaches.

Akhilesh Kumar Yadav, Divya Tomar, Sonali Agarwal[9][19] explained Lung Cancer is the most crucial and serious problem noticed now a days in all over the world. This paper explained that as there is huge amount of data in the medical field and it needs to be mined. It analyze the real data set from SGPGI (Sanjay Gandhi Post Graduate Institute of Medical Sciences) Lucknow. As the real data set contains problems such as missing values, highly dimensional, noise, and outlier etc., so it needs some clustering technique to make clusters. In this paper author applies the proposed foggy k-means algorithm instead of traditional k-means. Because foggy k means clusters the data of patient most accurately and efficiently. That is foggy k means give better results instead of simple k means.

Arpita Nagpal, Aman Jatain, Deepti Gaur's paper [10][21] discussed various clustering techniques like partition, hierarchical, density based, grid based, categorical data clustering etc.Clustering is used to discover similar data and to put similar data into one group that is called cluster. Clustering is used in many fields to discover similar data. This paper forms the hypothesis that most widely used clustering algorithm is the k-means algorithm. Most of the algorithms are its variants. A lot of research in this clustering field has focused on numerical data sets and there are only a small number of techniques that are available for categorical and other databases.

Madhu Yedla, Srinivasa Rao Pathakota, T M Srinivasa's paper explained [11] that k-means is one of the most popular clustering algorithm. Many improvements were already done to improve the performance of the k-means, but most of these require additional inputs like threshold values for the number of data points in a set. The result of k-means algorithm depends on the initial centroids which are chosen randomly. So, they used the enhanced k-means algorithm which does not require any kind of input like threshold value. They concluded that enhanced k-means algorithm results are more accurate and take less computational time

Oyelade et.al paper [12] defined the ability of the student performance of high learning. To analyze student result based on cluster analysis use standard statistical algorithm to arrange their score according to the level of their performance. In this paper K-mean clustering is implemented to analyze student result. The model was combined with deterministic model to analyze student's performance in the private institutes of Nigeria to monitor the academic performance of the students so that academic planners can make effective decisions.

V. CONCLUSION

As we know that kmeans algorithm is widely used algorithm for clustering. In this paper we had reviewed various improvements that had done in k means algorithm. In future work we can improve the existed kmeans algorithms in terms of accuracy and execution time.

VI. REFERENCES

[1]. Xiaoyan Wang and Yanping Bai, "The global Minmax k - means algorithm", SpringerPlus,2016.

[2]. 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, China,2016.

[3]. Caiquan Xiong, Zhen Hua, Ke Lv, Xuan Li," An Improved K-means text clustering algorithm By Optimizing initial cluster centers", IEEE International Conference on Cloud Computing and Big Data ,2016.

[4]. Akanksha Choudhary , Mr. Prashant Sharma, Mr. Manoj Singh," Improving K-Means Through Better Initialization And Normalization", IEEE Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI),Jaipur, India, Sept. 21-24, 2016.

[5]. Anshul Yadav, Sakshi Dhingra, "An Enhanced K-Means Clustering Algorithm to Remove Empty Clusters", International Journal of Engineering Development and Research, 2016.

[6]. Shruti Kapil and Meenu Chawla, "Performance Evaluation of K-means Clustering Algorithm with Various Distance Metrics", 1st IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems, 2016.

[7]. K.Rajalakshmi,, Dr.S.S.Dhenakaran,N.Roobin "Comparative Analysis of K-Means Algorithm in Disease Prediction", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 7, July 2015.

[8]. Rishikesh Suryawanshi, Shubha Puthran, "Review of Various Enhancement for clustering Algorithms in Big Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering,2015.

[9]. Akhilesh Kumar Yadav, Divya Tomar, Sonali Agarwal, "Clustering of Lung Cancer Data Using Foggy K-Means", International Conference on Recent Trends in Information Technology (ICRTIT), 2013.

[10]. Arpita Nagpal, Aman Jatain, Deepti Gaur, "Review based on Data Clustering Algorithms", Proceedings of 2013 IEEE Conference on Information and Communication Technologies,2013.

[11]. Madhu Yedla, Srinivasa Rao Pathakota, T M Srinivasa, "Enhancing K-means Clustering Algorithm with Improved Initial Center",International Journal of Computer Science and Information Technologies,2010.

[12]. Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", International Journal of Computer Science and Information Security, Vol. 7, o. 1, 2010.

[13]. Ismail Bin Mohamad and Dauda Usman, "Standardization and Its Effects on K-Means Clustering Algorithm", Research Journal of Applied Sciences,Engineering and Technology, Vol. 6(17), pp. 3299-3303, 2013.

[14]. M. Emre Celebi, H. A. Kingravi, P. A. Vela, "A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm",,Expert Systems with Applications, pp. 200-210, vol.40, 2013.

[15]. Vrinda Khairnar, Sonal Patil, "Effcient clustering of data using improved k-means

algorithm: A Review", Imperial Journal of Interdisciplinary Research, 2016.

[16]. Deepak Sinwar and Rahul Kaushik " Study of Euclidean and Manhattan Distance Metrics using Simple K-Means Clustering", IJRASET-9653 , May 2014.

[17]. VelidePhani Kumar and Lakshmi Velide, "Data Mining Approach for Prediction and Treatment Of diabetes Disease", IJSIT, 2014.

[18]. Dr.Urmila R. Pol, "Enhancing K-means Clustering Algorithm and Proposed Parallel K-means clustering for Large Data Sets." International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.

[19]. M.Mahajan,P.Nimbhorkar,K. Varadarajan, "The Planar k-Means Problem isNP-Hard", Lecture Notes in Computer Science 5431: 274–285. doi:10.1007/978-3-642-00202-1_24, 2009.

[20]. Sukhvir Kaur, "SURVEY OF DIFFERENT DATA CLUSTERING ALGORITHMS", International Journal of Computer Science and Mobile Computing,2016.

[21]. S. J. Nanda, G. Panda, "Accurate Partitional Clustering Algorithm Based on Immunized PSO",IEEE, March 30, 31, 2012.