Review of Existing Methods in K-means Clustering Algorithm

MS. Kavita Shiudkar[1], Prof. Sachine Takmare[2]

[1] ME CSE, Bharti Vidyapeeth college of Engineering, Kolhapur, Maharashtra, India

[2] Assistant Professor, Dept. of CSE, Bharti Vidyapeeth college of Engineering Kolhapur, Maharashtra, India

---

Abstract – *Data mining is the process of extracting useful information from the large amount of data and converting it into understandable form for further use. Clustering is the process of grouping object attributes and features such that the data objects in one group are more similar than data objects in another group. But it is now very challenging due to the sharply increase in the large volume of data generated by number of applications. Kmeans is a simple and widely used algorithm for clustering data. But, the traditional k-means is computationally expensive; sensitive to outlier's i.e. unnecessary data and produces unstable result hence it becomes inefficient when dealing with very large datasets. Solving these Issues is the subject of many recent research works. In this paper, we will do a review on k-means clustering algorithms.*

Key Words: Initial Centroids, Clustering, Data mining, Data sets, K-means clustering, Map-Reduce.

1. INTRODUCTION

Big Data is evolving term that describes any voluminous amount of structured, semi-structured and unstructured data. It is characterized by "5Vs", volume (size of data set), variety (range of data type and source), velocity (speed of data in and out), value (how useful the data is), and veracity (quality of data). It creates challenges in their collection, processing, management and analysis. As new data and updates are constantly arriving, there is need of data mining to tackle challenges.

The purpose of the data mining technique is to mine information from a bulky data set and make over it into a reasonable form for supplementary purpose. Data mining is also known as the knowledge discovery in databases (KDD). Technically, data mining is the process of finding patterns among number of fields in large relational database. It is the best process to differentiate between data and information. Data mining consists of extract, transform, and load transaction data onto the data warehouse system, Store and manage the data in a multidimensional database system, Provide data access to business analysts and information technology professionals, analyze the data by application software, Present the data in a useful format, such as a graph or table.

2. CLUSTERING

It makes an important role in data analysis and data mining applications. Data divides into similar object groups based on their features, each data group will consist of collection of similar objects in clusters. Clustering is a process of unsupervised learning. Highly superior clusters have high intra-class similarity and low inter-class similarity. Several algorithms have been designed to perform clustering, each one uses different principle. They are divided into hierarchical, partitioning, density-based, model based algorithms and grid-based.

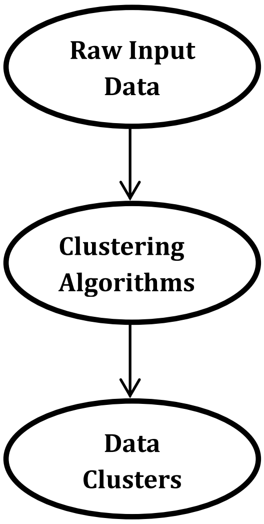

Fig: 1 Clustering stages

There are two types of Clustering Partitioning and Hierarchical Clustering.

1. **Hierarchical Clustering** - A set of nested clusters organized in the form of tree.
2. **Partitioning Clustering** - A division of data objects into subsets (clusters) such that each data object is in exactly one subset.

3. K-MEANS CLUSTERING

K-means clustering technique is widely used clustering algorithm, which is most popular clustering algorithm that is used in scientific and industrial applications. It is a method of cluster analysis which is used to partition N objects into k clusters in such a way that each object belongs to the cluster with the nearest mean [3].

The Traditional KMeans algorithm is very simple [3]:

1. Select the value of K i.e. Initial centroids.
2. Repeat step 3 and 4 for all data points in dataset.
3. Find the nearest point from that centroids in the Dataset.
4. Form K cluster by assigning each point to its closest centroid.
5. Calculate the new global centroid for each cluster.

Properties of k-means algorithm [3]:

1. Efficient while processing large data set.
2. It works only on numeric values.
3. The shapes of clusters are convex.

K-means is the most commonly used partitioning algorithm in cluster analysis because of its simplicity and performance. But it has some restrictions when dealing with very large datasets because of high computational complexity, sensitive to outliers and its results depends on initial centroids, which are selected randomly. Many solutions have been proposed to improve the performance of KMeans. But no one provide a global solution. Some of proposed algorithms are fast but they fail to maintain the quality of clusters. Some generate clusters of good quality but they are very expensive in term of computational complexity. The outliers are major problem that will effect on quality of clusters. Some algorithm only works on only numerical datasets.

4. LITERATURE REVIEW

Amira Boukhdhir, Oussama Lachiheb , Mohamed Salah Gouider [1] proposed algorithm an improved KMeans with Map Reduce design for very large dataset. The algorithm takes less execution time as compared to traditional KMeans, PKMeans and Fast KMeans. It removes the outlier from numerical datasets also Map Reduce technique used to select initial centroids and forming the clusters. But it has limitations like the value of numbers of centroids required as input by user. It works on numerical datasets only. Also numbers of clusters are not determined automatically.

Duong Van Hieu, Phayung Meesad [2] proposed algorithm for reducing executing time of the k-means .They implemented this by cutting off a number of last iterations. In this experiment method 30% of iterations are reduced, so 30% of executing time is reduced, and accuracy is high. However, the choosing randomly the initial centroids produces the instable clusters. Clustering result may be affected by noise points, so it produces inaccurate result.

Li Ma and al [3] developed a solution for improving the quality of traditional k-means clusters. They used the technique of selecting systematically the value of k i.e number of clusters as well as the initial centroids. Also they reduced the number of noise points so the outlier's problem solved. This algorithm produces good quality clusters but it takes more computation time.

Xiaoli Cui and al [4] proposed an algorithm i. e. an improved k-means. This algorithm works on only representative points instead of the whole dataset, using a sampling technique. The result of this the I/O cost and the network cost reduced because of Parallel K-means. Experimental results shows that the algorithm is efficient and it has better performance as compared with k-means but, there is no high accuracy.

Yugal Kumar, G. Sahoo [5] focused on K-Means initialization problems. The K-Means initialization problem of algorithm is formulated by two ways; first, how many numbers of clusters required for clustering and second, how to initialize initial centers for clusters of K-Means algorithm. This paper covers the solution for of the initialization problem of initial cluster centers. For that, a binary search initialization method is used to initialize the initial cluster points i.e. initial centroid for K-Means algorithm Performance of algorithm evaluated using UCI repository datasets.

Huang Xiuchang, SU Wei [6] focused on problem of user behavior pattern analysis, which has the insensitivity of numerical value, uneven spatial and temporal distribution characteristics strong noise. The traditional clustering algorithm not works properly. This paper analyses the existing clustering methods, trajectory analysis methods, and behavior pattern analysis methods, and combines clustering algorithm into the trajectory analysis. After modifying the traditional K-MEANS clustering algorithm, the new improved algorithm designed which is suitable to solve the problem of user behavior pattern analysis compared with traditional clustering methods on the basis of the test of the simulation data and actual data, the results shows that the improved algorithm more suitable for solving the trajectory pattern of user behavior problems.

Nidhi Singh, Divakar Singh [7] K-means is widely used for clustering algorithm. This paper proves that the accuracy of k-means for iris dataset is much than the hierarchical clustering and for diabetes dataset accuracy of hierarchal clustering is more than the k-means algorithm. The time taken to cluster the data sets is less in case of k-means. A good clustering method produces high-quality clusters to ensure that objects of a same cluster are more similar than members of different cluster. Kmeans algorithm in this paper works well for large datasets.

Kedar B. Sawan [8] existing K-means clustering algorithm has a number of drawbacks. The selection of initial starting point will have effect on the results of number of clusters formed and their new centroids. Overview of the existing methods of choosing the value of K i.e. the number of clusters along with new method to select the initial centroid points for the K-means algorithm has been proposed in the paper along with the modified K-Means algorithm to overcome the deficiency of the classical K-means clustering algorithm. The new method is closely related to the approach of K-means clustering because it takes into account information reflecting the performance of the algorithm. The improved version of the algorithm uses a systematic way to find initial centroid points which reduces the number of dataset scans and will produce better accuracy in less number of iteration with the traditional algorithm. The method could be computationally expensive if used with large data sets because it requires calculating the distance of every point with the first point of the given dataset as a very first step of the algorithm and sort it based on this distance. However this drawback could be taken care by using multi-threading technique while implementing it within the program. However further research is required to verify the capability of this method when applied to data sets with more complex object distributions.

Bapusaheb B. Bhusare, S. M. Bansode [9] the K means clustering algorithm which mainly based on initial cluster centers. In this paper K means clustering algorithm by designed in such way that the initial centroids selected using Pillar algorithm. Pillar algorithm effectively chooses the initial centroids and improves accuracy of clusters. However, proposed algorithm has outlier problem leads to reduced performance. So there is need to choose the appropriate parameter in data set for outlier detection mechanism. An improvement in pillar algorithm is done and the number of distance calculation reduced for the previous initial centroids neighbors and used for next step of iterations which causes to increase in the computational time. The experimental results show that the use of pillar algorithm with change improved solution.

Kamaljit Kaur, Dr. Dalvinder Singh Dhaliwal, Dr. Ravinder Kumar Vohra [10] found that the K-Means algorithm has two major limitations 1. Several distance calculations of each data point from all the centroids in each iteration. 2. The final clusters depend upon the selection of initial centroids. This work improves k-Means clustering algorithm designed in MATLAB and the datasets from UCI machine learning repository used. The initial centroids initial centroids not selected randomly. By using new approach good clustering results obtained. The new method of selection of initial centroid is better than selecting the initial centroids randomly.

Abhijit Kane's [11] paper includes the automatically find the number of clusters in a dataset. Here every step requires re-clustering of the dataset, total O (n) operations computed. This method works well for clusters that are distinctly separated. This method is also density-independent, making it useful for clustering algorithms like the Expectation-maximization algorithm.

Omar Kettani, Faical Ramdani, Benaissa Tadili [12] work covers an algorithm designed for automatic clustering. This method computes the correct number of clusters on tested data sets. This method was compared with G-means. The comparison of algorithm shows that the proposed approach much better than G-means in terms of clustering accuracy.

Avni Godara, Varun Sharm [13] covers the prime algorithm. The KMeans clustering is a powerful algorithm used most of the application in daily life dataset, but problem of initial centroid selection. In past years number of papers presented to improve classical k means algorithm. To remove problem of initial centroid selection need to define data points for centroid before next iteration. The use of prim's algorithm gives better results for selection of initial centroid and choose easily data points for future iterations. Experimental result also shows that the prime algorithm gives better and optimal performance for initial centroids, accuracy of result not adjusted.

D. Sharmila Rani, V. T. Shenbagamuthu [14] K-means is a typical clustering algorithm and it is used for clustering large sets of data. This work includes K-means algorithm and analyses the standard K-means clustering algorithm. The standard K-means algorithm is computationally complex and need to reassign the data points, a number of times during every iteration, which makes effect on the efficiency of standard K-means clustering. This paper work covers a simple and efficient way for assigning data points to clusters. This work ensures that the entire process of clustering in O (nk) time without sacrificing the accuracy of clusters.

Effat Naaz, Divya Sharma, D Sirisha, Venkatesan M. [15] Paper build a system to know the accuracy of medication associated with each symptom. To do this K-means Clustering on the clinical note corpus applied. The document clustering results in improving the medication recommendation. An experimental result shows that pre-processing before clustering results in efficient process of clustering. For experimental work different tools used like, section annotator, symptom annotator, negation annotator and medication annotator to get different views of clinical notes which improves the visibility of clinical note. The result of this is increase of the accuracy of medications associated with the symptoms.

5. CONCLUSION

In this review work most widely used k-means clustering techniques of data mining is analyzed. This work shows that there are several methods to improve the clustering with different approaches. Various clustering techniques are reviewed which improve the existing algorithm with different perspective. Some limitations of existing algorithm will be eliminated in future work. This technique will be useful in extraction of useful information using cluster from huge database. It removes the limitation of K-means clustering algorithm and gives accurate result in less time so we can say it's very efficient than standard K-means clustering algorithm and quality of cluster also improved. From Our analysis of different K-means approaches, we conclude that it's better than traditional K-means clustering algorithm.

6. REFERENCES

[1] Amira Boukhdhir Oussama Lachiheb, Mohamed Sala Gouider. "An improved Map Reduce Design of Kmeans for clustering very large datasets", IEEE transaction.

[2] V. Duon, M. Phayung. "Fast K-Means Clustering for very large datasets based on Map Reduce Combined with New Cutting Method (FMR KMeans)", Springer International Publishing Switzerland, 2015.

[3] M. Li and al. "An improved k-means algorithm based on Map reduce and Grid", International Journal of Grid Distribution Computing, (2015)

[4] C. Xiaoli and al. "Optimized big data K-means clustering using Map Reduce", Springer Science + Business Media

New York (2014).

[5] Yugal Kumar and G. Sahoo, "A New Initialization Method to Originate Initial Cluster Centers for K-Means Algorithm", International Journal of Advanced Science and Technology Vol.62, (2014).

[6] Huang Xiuchang , SU Wei ,"An Improved K-means Clustering Algorithm" ,JOURNAL OF NETWORKS, VOL. 9, NO. 1, JANUARY 2014

[7] Nidhi Singh, Divakar Singh," Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (3) , 2012.

[8] Kedar B. Sawant, "Efficient Determination of Clusters in K-Mean Algorithm Using Neighborhood Distance "International Journal of Emerging Engineering Research and Technology Volume 3, Issue 1, January 2015.

[9] Bapusaheb B. Bhusare, S. M. Bansode, "Centroids Initialization for K-Means Clustering using Improved Pillar Algorithm" , International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 4, April 2014 .

[10] Kamaljit Kaur, Dr. Dalvinder Singh Dhaliwal, Dr. Ravinder Kumar Vohra ,"Statistically Refining the Initial Points for K-Means Clustering Algorithm ",International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 11, November 2013.

[11] Abhijit Kane," Determining the number of clusters for a Kmeans clustering algorithm", Indian Journal of Computer Science and Engineering (IJCSE) Vol. 3 No.5 Oct-Nov 2012

[12] Omar Kettani, Faical Ramdani, Benaissa Tadili, "AK-means: An Automatic Clustering Algorithm based on Kmeans ", Journal of Advanced Computer Science & Technology, 4 (2) (2015) .

[13] Avni Godara, Varun Sharma," Improvement of Initial Centroids in Kmeans clustering Algorithm", Vol-2 Issue-2 2016 IJARIIE

[14] D. Sharmila Rani, V.T. Shenbagamuthu,"Modified K-Means Algorithm for Initial Centroid Detection", International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

[15] Effat Naaz, Divya Sharma, D Sirisha, Venkatesan M," Enhanced Kmeans clustering approach for healthcare analysis using clinical documents", International Journal of Pharmaceutical and Clinical Research 2016.