A Systematic Review on K-Means Clustering Techniques

Swagita Dwivedi swagitadwivedi23@gmail.com
Prof. Lalit Kumar P.Bhaiya Lalit04_bhaiya@yahoo.com
Dept. of Computer Science and Engineering
Bharti College of Engineering ,Durg
Durg, Chhattisgarh, India

Abstract- **number of data points are grouped together to form a cluster. Data of same class are grouped together. K-Means clustering is most essential and fundamental clustering technique through which data centers are analyzed. K-means is most broadly used algorithm for clustering with known arrangements of median points. It is likewise called as nearest neighbor clustering. Beforehand, a different endeavor has been done to improve the performance of k-means algorithm. The result of improved k-means has extraordinary performance improvement for little to medium size of data. Be that as it may, for vast and exceptionally expansive amount of data, k-means fall behind. This paper reviews different methods and techniques used in literature and its advantages and limitations, to analyze the further need of improvement of k-means algorithm.**

Keywords - K-Means, Nearest Neignbour, data points, unsupervised learning, clustering.

I. INTRODUCTION

utilization of internet generates bunches of data. These data are picking up its size as the year passes. The data are generated at record rate each day. To analyze those data and gathering into cluster is tedious task. The problem additionally lies in storing and retrieving of data. The analysis of these data points into different cluster is likewise a difficult task. Researchers have assessed that amount of information in the world doubles for every 20 months. Anyway raw data can't be used legitimately. Its real value is predicted by extracting information helpful for decision support. In many regions, data analysis was traditionally a manual process. At the point when the size of data manipulation and exploration goes past human capacities, individuals look for computing technologies to automate the process [1]. Data mining is process of extraction, transformation and loading of information to/from database or warehouse framework. Storing and overseeing data, give access to data analyst and data scientist to investigations the data for advantage of their business. [2][3]. There are two learning method presents to mine valuable data from raw data.

1. Supervised Learning- In this type of learning Dataset is given as input and get output as desired, however in presence of trainer. Trainer generally prepares the input dataset and classifies it. Case of supervised learning techniques are: Neural network, Multilayer perception, Decision tree.

2. Unsupervised Learning- The ideal result isn't gave to the unsupervised model amid learning procedure. This method can be used to cluster the input data in classes on the basis of their statistical properties as it were. These models are for different type of clustering, k-means, distances and normalization, self-organizing maps.

This paper reviews different methods and techniques used in literature and its advantages and limitations, to analyze the further need of improvement of k-means algorithm.

1. K-Means Clustering- Clustering is imperative and basic idea of data mining field used in different applications. In Clustering, data are partitioned onto different classes. These classes represent some critical features. Means, classes are the container of similar behavior of objects. The objects which carry on or are closest to each other are clustered in one class and who are far or non-similar are clustered in various class. Clustering is technique unsupervised learning. Exceedingly superior clusters have high intra-class similarity and low between class similarities.

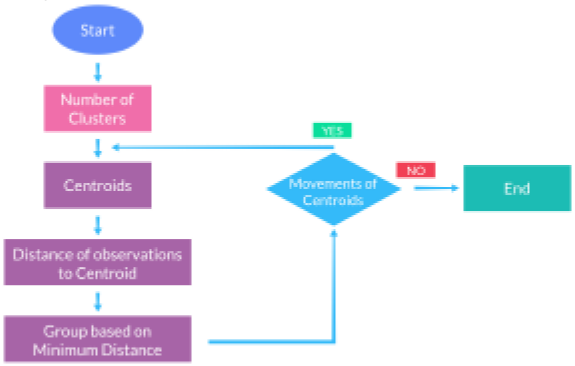
Start
Number of Clusters
Centroids
Distance of observations to Centroid
Group based on Minimum Distance
YES
Movements of Centroids
NO
End

Fig.1. K-Means Generic Algorithm

K-means clustering method is a system of clustering which is extensively utilized. This algorithm is the very mainstream clustering instrument that is utilized in scientific and industrial applications. It is a system of cluster investigation which expects to separate observations into k clusters where each observation has a space with the cluster with the closest mean [4].K-means clustering: K-Means clustering is unsupervised clustering technique where data points are given as input and without and predefined result it generate clustering results. It is intensely used in scientific and industrial applications. For example clustering of similar gene expression, weather data, text classification and so forth [5]. The generic algorithm is very simple as presented in fig.1.

- Choose K points as starting centroids.
- Repeat.
- Form K cluster by allotting every point to its
- closest centroid.
- Recomputed the centroid of every cluster until
- centroid does not modified.

II. LITERATURE SURVEY

K. A. Abdul Nazeer et al. [6] proposes k-means algorithm, for variety of arrangements of values of initial centroids, produces various clusters. Result cluster quality in algorithm relies upon the choosing of initial centroids. Two stages incorporate into actual k means algorithm: first to determining initial centroids and second to allotting data points to the nearest clusters and after that re-calculating the clustering mean. Soumi Ghosh et al. [7] present a relative exchange of two clustering algorithms in specific centroid based K-Means and representative object based FCM (Fuzzy C-Means) clustering algorithms. This discourse is on the basis of evaluation of performance of the efficiency of clustering yield by applying these algorithms.

Shafeeq et al. [8] present an altered K-means algorithm to enhance the cluster quality and to fix the optimal numerous amounts of groups. As input number of clusters (K) given to the K-means algorithm by the client. Be that as it may, in the useful situation, it is uncommonly critical to fix the numerous amounts of clusters ahead of time. The technique proposed in this paper works for both the cases for instance for known number of clusters ahead of time similarly as obscure number of clusters. The client has the adaptability either to fix the number of clusters or input the base number of clusters required. The new cluster centers are figured by the algorithm by expanding the cluster counter by one in each cycle until it fulfills the legitimacy of cluster quality. This algorithm will beat this problem by finding the satisfy less amount of clusters on the run.

Junatao Wang et al. [9], in this paper author propose an improved k-means algorithm using noise data filter. The deficiencies of the conventional k-means clustering algorithm are overpowered by this proposed algorithm. The algorithm makes density put together detection strategies based with respect to features of noise data where the revelation and processing ventures of the noise data are attached to the original algorithm. By pre-processing the data to dismiss these noisy information before grouping data sets the cluster union of the clustering yield is improved inside and out and the impact of noise data on k-means algorithm is diminished adequately and the clustering yield are progressively accurate.

Shi Na et al. [10] present the analysis of shortcomings of the improved k-means algorithm. As k-means algorithm needs to ascertain the distance between every data object and all cluster focuses in every iteration. This recurrence process impacts the efficiency of clustering algorithm. An enhanced k-means algorithm is proposed in this paper. A basic data structure is needed to store some data in each iteration which is to be used in the next iteration. Computation of distance in each iteration is stayed away from by the proposed technique and saves the running time.

Below shows Comparison Between Various Existing Methods And Its Limitation.

1. Author - K. A. Abdul Nazeer et al.
Method Used - K-Means Algorithm
Dataset - Iris Dataset
Review - An enhanced clustering method propose to find initial centroids efficiently assign data points to cluster. Improve the efficiency and accuracy of k means algorithm.
Limitation - Limitation in this enhanced algorithm that is the value of k, the number of desired clusters, is still required to be given as an input, regardless of the distribution of the data points.

2. Author - Soumi Ghosh et al.
Method Used - K-means algorithm, Fuzzy C-means algorithm,
Dataset - Iris and plant Dataset
Review - Comparative analysis of Fuzzy C-means and K-means on the basis of time complexity. K-means algorithm seems to be superior than Fuzzy C-means
Limitation - Computation time is more than k-means due to involvement of the fuzzy measure calculations.

3. Author - Shafeeq et al.
Method Used - modified K-means algorithm
Dataset - random numbers of 300,500 and 1000 data points
Review - Number of clusters are find in the proposed method on the run based on the cluster quality output. It is work for both known no. of cluster in advance as well as unknown no. of cluster.
Limitation - Proposed approach takes more computational time than the K-means for larger data sets

4. Author - Junatao Wang et al.
Method Used - K-Means algorithm
Dataset - Data set from UCI Repository of Machine Learning Databases
Review - Modified algorithm decrease the impact of noise data on k-means algorithm and clustering results are more accurate
Limitation - Impact of noise are more in forming cluster
5. Author - Shi Na et al.
Method Used - K-Means algorithm
Dataset - Data set from UCI Repository of Machine Learning Databases
Review - Improve the speed and accuracy of clustering, reducing the computational complexity of the k-means
Limitation - Centroid selection algorithm is not effective

III. CONCLUSION

In this paper k-means clustering techniques and method are reviewed. K-means being most renowned among data scientist need further improvement in different section of algorithm. The outliers, empty clusters and selecting centroid for datasets are as yet a difficult task. Subsequently different further research expected to concentrate on these referenced issues. presents different techniques and its limitation are available in proposed k-means algorithm. They need further improvement because of increase of size of data starting at now. This paper has made an endeavor to survey a critical number of papers to deal with the present algorithm of k-means. Present examination illustrate that k-means algorithm can be enhanced by selecting centroid point appropriately.

REFERENCES

[1] E. A. Khadem, E. F. Nezhad, M. Sharifi, “Data Mining: Methods & Utilities”, Researcher2013; 5(12):47-59. (ISSN: 1553-9865).
[2] Namrata S Gupta, Bijendra S Agrawal, Rajkumar M. Chauhan, ìSurvey On Clustering Technique of Data Mining, American International Journal of Research in Science, Technology, Engineering & Mathematics, ISSN:2328-3491
[3] Malwindersingh, Meenakshibansal ,î A Survey on Various KMeans algorithms for Clustering, IJCSNS International Journal of Computer Science and Network Security, VOL.15 No.6, June 2015
[4] A. Saurabh, A. Naik, "Wireless sensor network based adaptive landmine detection algorithm, " 2011 3rd International Conference on Electronics Computer Technology (ICECT), vol.1, no., pp.220, 224, 8-10 April 2011
[5] Amandeep Kaur Mann, Navneet Kaur Mann, ìReview Paper On Clustering Techniquesî ,Global Journal Of Computer Science And Technology Software & Data Engineering, VOL. 13 ,2013.
[6] K. A. Abdul Nazeer, M. P. Sebastian,îImproving the Accuracy and Efficiency of thek-means Clustering Algorithm, Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K.
[7] Soumi Ghosh, Sanjay Kumar Dubey, Comparative Analysis of K-Means and Fuzzy C-Means Algorithmsî, International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013
[8] Shafeeq,A., Hareesha,K.,ìDynamic Clustering of Data with Modified K-Means Algorithm, International Conference on Information and Computer Networks, vol. 27 ,2012
[9] Junatao Wang, XiaolongSu,îAn Improved K-means Clustering Algorithm, Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on 27 may,2011 (pp. 44-46).
[10] Shi Na, Liu Xumin, Guan Yong, ìResearch on K-means Clustering Algorithm: An Improved K-means Clustering Algorithm, Intelligent Information Technology and Security Informatics,2010 IEEE Third International Sy