A Systematic Review on K-Means Clustering Techniques

Ankita Dubey
Shri Shankaracharya group of institutions
Dept. of Computer Science and Engineering
Durg, Chhattisgarh, India
ankitadubey2410@gmail.com

Asso.Prof. Dr. Abha Choubey
Shri Shankaracharya group of institutions
Dept. of Computer Science and Engineering
Durg, Chhattisgarh, India
abha.is.shukla@gmail.com

Abstract

In the field of data mining, clustering is a technique where millions of data points are grouped together to form a cluster. Data of same class are grouped together. K-Means clustering is most important and basic clustering technique through which data points are analyzed. K-means is most widely used algorithm for clustering with known sets of median points. It is also called as nearest neighbor clustering. Previously, various efforts has been done to improve the performance of k-means algorithm. The outcome of improved k-means has great performance improvement for small to medium size of data. But for large and very large amount of data, k-means fall behind. This paper reviews various 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

Recent years, there are tremendous increase in the usage of internet. The usage of internet generates lots of data. These data are gaining its size as the year passes. The data are generated at record rate every day. To analyze those data and group into cluster is tedious task. The problem also lies in storing and retrieving of data. The analysis of these data points into different cluster is also a challenging task. Researchers have estimated that amount of information in the world doubles for every 20 months.

However raw data cannot be used directly. Its real value is predicted by extracting information useful for decision support. In most areas, data analysis was traditionally a manual process. When the size of data manipulation and exploration goes beyond human capabilities, people 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 system. Storing and managing data, provide access to data analyst and data scientist to analyses the data for benefit of their business. [2][3].

There are two learning method presents to mine useful data from raw data.

1. Supervised Learning: In this type of learning, dataset is given as input and get output as desired, but in presence of trainer. Trainer generally trains the input dataset and classify it. Example of supervised learning techniques are: Neural network, Multilayer perception, Decision tree.

2. Unsupervised Learning: The desired result is not provided to the unsupervised model during learning procedure. This method can be used to cluster the input data in classes on the basis of their statistical properties only. These models are for various type of clustering, k-means, distances and normalization, self-organizing maps.

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

A. K-Means Clustering

Clustering is important and essential concept of data mining field used in various applications. In Clustering, data are divided onto various classes. These classes represents some important features. Means, classes are the container of similar behavior of objects.

The objects which behave or are closer to each other are grouped in one class and who are far or non-similar are grouped in different class. Clustering is a process of unsupervised learning. Highly superior clusters have high intra-class similarity and low inter-class similarity.

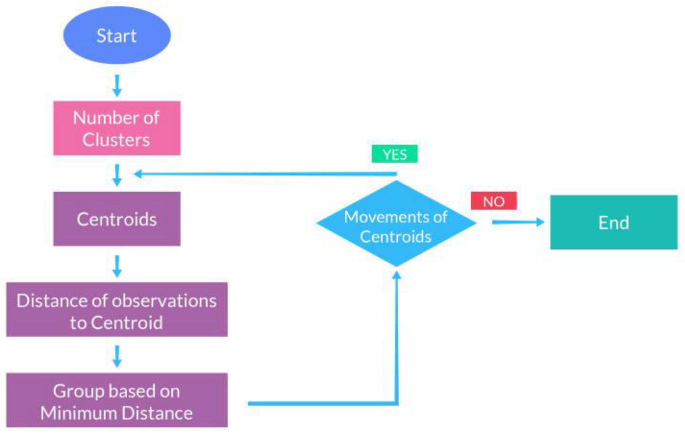
Start
Number of Clusters
Centroids
Distance of observations to Centroid
Group based on Minimum Distance
YES
Movements of Centroids
NO
End

Fig.1. K-Means Generic Algorithm

K-means clustering technique is a technique of clustering which is widely used. This algorithm is the most popular clustering tool that is used in scientific and industrial applications. It is a method of cluster analysis which aims to partition observations into k clusters in which each observation belongs to the cluster with the nearest mean [4].

K-means clustering: K-Means clustering is unsupervised clustering technique in which data points are given as input and without and predefined result it generate clustering results. It is heavily used in scientific and industrial applications. For. E.g. clustering of similar gene expression, weather data, text classification etc. [5].

The generic algorithm is very simple as presented in fig.1.

- Select K points as initial centroids.
- Repeat
- Form K cluster by assigning each point to its closest centroid.
- Recomputed the centroid of each cluster until centroid does not change

II. LITERATURE SURVEY

K. A. Abdul Nazeer et al. [6] proposes k-means algorithm, for different sets of values of initial centroids, produces different clusters. Final cluster quality in algorithm depends on the selection of initial centroids. Two phases includes in original k means algorithm: first for determining initial centroids and second for assigning data points to the nearest clusters and then recalculating the clustering mean.

Soumi Ghosh et al. [7] present a comparative discussion of two clustering algorithms namely centroid based K-Means and representative object based FCM (Fuzzy C-Means) clustering algorithms. This discussion is on the basis of performance evaluation of the efficiency of clustering output by applying these algorithms.

Shafeeq et al. [8] present a modified K-means algorithm to improve the cluster quality and to fix the optimal number of cluster. As input number of clusters (K) given to the K-means algorithm by the user. But in the practical scenario, it is very difficult to fix the number of clusters in advance. The method proposed in this paper works for both the cases i.e. for known number of clusters in advance as well as unknown number of clusters. The user has the flexibility either to fix the number of clusters or input the minimum number of clusters required. Thenew cluster centres are computed by the algorithm by incrementing the cluster counter by one in each iteration until it satisfies the validity of cluster quality. This algorithm will overcome this problem by finding the optimal number of clusters on the run.

Junatao Wang et al. [9] propose an improved k-means algorithm using noise data filter in this paper. The shortcomings of the traditional k-means clustering algorithm are overcome by this proposed algorithm. The algorithm develops density based detection methods based on characteristics of noise data where the discovery and processing steps of the noise data are added to the original algorithm. By pre-processing the data to exclude these noise data before clustering data sets the cluster cohesion of the clustering results is improved significantly and the impact of noise data on k-means algorithm is decreased effectively and the clustering results are more accurate.

Shi Na et al. [10] present the analysis of shortcomings of the standard k-means algorithm. As k-means algorithm has to calculate the distance between each data object and all cluster centers in each iteration. This repetitive process effects the efficiency of clustering algorithm. An improved k-means algorithm is proposed in this paper. A simple data structure is required to store some information in every iteration which is to be used in the next iteration. Computation of distance in each iteration is avoided by the proposed method and saves the running time.

***TABLE I**. Shows comparison between various existing approaches and its limitation*

S. No.	Author	Method Used	Data source	Review	Limitation
1	K. A. Abdul Nazeer et al.	K-Means Algorithm	Iris Dataset	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 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	Soumi Ghosh et al.	K-means algorithm, Fuzzy C-means algorithm,	Iris and plant Dataset	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	computation time is more than k-means due to involvement of the fuzzy measure calculations
3	Shafeeq et al.	modified K-means algorithm	random numbers of 300,500 and 1000 data points	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.	proposed approach takes more computational time than the K-means for larger data sets
4	Junatao Wang et al.	K-Means algorithm	Data set from UCI Repository of Machine Learning Databases	Modified algorithm decrease the impact of noise data on k-means algorithm and clustering results are more accurate	Impact of noise are more in forming cluster
5	Shi Na et al.	K-Means algorithm	Data set from UCI Repository of Machine Learning Databases	Improve the speed and accuracy of clustering, reducing the computational complexity of the k-means	Centroid selection algorithm is not effective

III. CONCLUSION

In this paper k-means clustering techniques and method are reviewed. K-means being most famous among data scientist need further improvement in various section of algorithm. The outliers, empty clusters and selecting centroid for datasets are still a challenging task. Hence various further research needed to focus on these mentioned issues. Table I. presents various techniques and its limitation are present in proposed k-means algorithm. They need further enhancement due to increase of size of data as of now.

This paper has make an attempt to review a significant number of papers to deal with the present algorithm of k-means. Present study 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 ,201

[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 Symposium on 2-4 April, 2010(pp. 63-67)