IJCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Volume-4, Issue-3 E-ISSN: 2347-2693

A Study of Various Methods to find K for K-Means Clustering

Hitesh Chandra Mahawari[1*] and Mahesh Pawar [2]

[1*]*School of Information Technology, RGPV, India*
[2] *Department of IT, UIT, RGPV, India*

www.ijcseonline.org

Received: Feb/22/2016 Revised: Feb/28/2016 Accepted: Mar/17/2016 Published: Mar/31/ 2016

Abstract— Clustering is the technique which used to group data from a set of unlabeled data, in a way that data containing similar properties contains in a same group. There are many cluster techniques are used to cluster data thus there is no suitable definition for cluster is available. Techniques like link based clustering, centroid based clustering, distribution based clustering, density based clustering are used. A survey over centroid based K-mean clustering techniques is presented which is widely used for clustering purpose. K-mean clustering technique suffers drawbacks like sensitive to initialization centroid, sensitive to noise, and there is no. of clusters also not defined. Thus an enhanced k-mean technique is presented to reduce such drawbacks and provide an enhanced functionality for clustering.

Keywords— K-Means, Clustering, Centroid, Centroid based clustering, partition based clustering, point based, convex, euclidian

I. Introduction

In recent decade a huge amount of data flooded over the internet, which data presents in form of unorganized or heterogeneous format. Thus a technique is required to organize that data. Clustering is the technique which is generally used for that purpose. There are various clustering techniques like centroid based clustering, connectivity based clustering, distribution based clustering, and density based clustering. Connectivity based clustering is the technique which based on the concept that nearby objects are more related as compare to the farther objects. In density based clustering, clusters are defined on the basis of area of density. In distribution based clustering, clusters form by the objects which contain same distributions.

Centroid based techniques are used a point vector to form cluster either a member of cluster or not. Here a centroid based technique called K-mean clustering technique is presented. K-mean is widely used technique for clustering purpose, that technique uses properties of voronoi cells to form cluster. K-mean form clusters in way that each data point having same distance or less from the other clusters centroid.

But this technique suffers defects like depends on initial centroid and also sensitive to noise and density.
Thus an enhanced technique is required to reduce such defects like in [1] a technique to select initial centroid is presented, in that technique distance for each data point from origin is calculated and arranged in sorting order. Then data point containing mean distance is selected as a centroid for first cluster, and iterated in that way to form other clusters.

In [2] an anonymous cluster based technique is used to decide the no. of cluster to be formed. In [3] a technique to select k in k-mean clustering is presented. In [4] a brief review over the techniques which used select initial cluster in k-mean clustering, is presented. [5] Presented a technique which reduces the mean square value among various data points, that way it increases the efficiency of the K-mean clustering technique. A parameter learning technique for K-mean clustering, is presented in [6] . Some other techniques are also discussed.

K-mean provides various applications like clustering analysis, classification, and some others. In [7]a clustering analysis and classification techniques are presented for various graphical data, in [8]a cluster analysis for student's marks is also presented. Clustering provides wide range of applications. An efficient clustering technique can improve the performance of whole task. Thus a survey over various techniques which used for enhancement in K-mean clustering is presented in Literature Review section.
Further this paper organizes as follows:-
II Literature review, III Conclusion, IV References.

In II^nd section of this paper, we presented a brief information regarding the relevant past work on finding K for K Means. The III^rd part of this paper contains derived conclusions out of our study. Finally the IV^th part is the citations or references.

Corresponding Author: *Hitesh Chandra Mahawari,*
MahawariHitesh@gmail.com, School of Information Technology, RGPV, India

II. Literature Review

In [1] an enhanced K-mean clustering technique for clustering is presented. K-mean is widely used for clustering. In K-mean clustering performance of the clustering depends on the selection of initial centroid. K-mean also subjected to local minima, and not provides global solution. In the proposed technique an enhanced initial centroid selection mechanism and an efficient way to assign data point for each cluster is presented. In that technique check for negative value attribute within the dataset is conducted. If any negative value attribute is found then transform it in to the positive space. Then distance for each data point from origin is calculated and arrange these data points in sorting order with reference to the distance to origin. In that sorted order middle point of the data points selects as the initial centroid for the cluster. In that way reduces the defect of the existing technique which selects initial centroid on random basis.

In [9] a less similarity based clustering technique to provide better centroid initialization and better time complexity, is presented. There are mainly three type of clusters can be formed, text based, partition based, hybrid.
In text based clustering, cluster can be formed on the basis of text content, like if to document contains same clusters then these document can be put into same cluster. Two type of clustering techniques can be used called partition based clustering, hierarchical clustering.

In graph based clustering, graph structures are used to form clusters. In that edge cut used to form clusters, each node denotes the documents. Link based clustering technique is used to form clusters in that links are contains all the information about the content of the node. But in both these technique there are disadvantages like if content based technique using for different languages it cannot provide better results, in link based technique a dense structure of links can be used which generate complexity. A hybrid technique, which takes advantages of both these techniques and restricts demerits of these techniques are used to form cluster the web document.

In [10] an optimized K-mean clustering technique is presented, which optimized the running time of the process. In that technique an assumption will be made which shows that a small part of the component of the cluster is changed after the iteration thus there no need to recalculate whole data, like after the iteration a centroid can be move but small part of the cluster can be affected by this movement these component cannot change their cluster thus there is no need to recalculate the whole data. An optimized technique is presented which provides better efficiency and running time complexity for the data.

In [11] a k-mean clustering technique to analyze student's data is presented. In existing techniques fuzzy theory, rough set theory used to analyze students data, but these techniques are not able to provide better results. Thus, a k-mean clustering technique which combines with deterministic model to analyze the student's results is used. In that distance based method is used to categorize that data.

In [12] an improved k-mean clustering technique is presented. Existing K-mean technique is sensitive to initialization thus to reduce this drawback an improved technique is presented. In that technique Neyman-Pearson based technique is used to estimate the threshold value of the end-members in the clusters and a spectral signature based technique is used to extract data from the image to initialize cluster. MVES (minimum value enclosing algorithm) technique is used extract spectral signature value from the end members. In that way enhanced technique is developed to cluster hyper-spectral data.

In [13] a new improved k-mean clustering technique is presented. In that technique an improved focal point and K value is used to overcome the drawback of the k-mean clustering technique. In that first a data point is selected from the dataset and calculate the distance its distance from other data points and compute the density of that data point, if its distance with in threshold value and density is also above the threshold value then put it into the cluster otherwise put it into high density area. Calculate distance for each data-point now calculates the k value for cluster. In that way an improved initialization and K-value calculation framework is provided, which reduce the drawbacks of the existing technique.

In [2] an enhanced K-mean technique, which uses an enhanced mechanism to choose no. of clusters, is presented. In that technique anomalous pattern based technique is used to decide the no. of clusters in that first an origin is selected from the data points and then an anomalous pattern based technique is applied and iterated to form clusters. In anomalous pattern detection iteration starts at the farthest point from the origin, then iterated to form cluster. There are various techniques which are used to find the value of K like structural approach, variance based approach, and consensus based approach, resampling approach also presented.

III. Conclusion

There are many clustering techniques are used to form clusters unlabeled datasets. K-mean is widely used technique which used for clustering purpose but it suffers some drawbacks which are discussed in above section. A survey over various clustering technique which used to enhance the functionality of the k-mean clustering, is presented. In that techniques like improving initialization of the initial centroid, provide a way to decide no. of cluster to be formed.

IV. References

[1] Madhu Yedla, Srinivasa Rao Pathakota, and T M Srinivasa, "Enhancing K-means Clustering Algorithm with Improved Initial Center," in IJCSIT, **2010**.

[2] Boris Mirkin Mark Ming-Tso Chiang, "Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads," journal of classification, **2009**.

[3] S S Dimov, and C D Nguyen D T Pham, "Selection of K in K-means clustering ," IMechE **2005**.

[4] Naveen D Chandavarkar Uday Kumar S, "A Survey on Several Technical Methods for Selecting Initial Cluster Centers in K-Means Clustering Algorithm," IJARCSSE, Dec **2014**.

[5] David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, Angela Y. Wu Tapas Kanungo, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation," IEEE.

[6] Rebecca J. Passonneau, Austin Lee, Axinia Radeva, Boyi Xie And David Waltz Haimonti Dutta, "Learning Parameters Of The K-Means Algorithm, From Subjective Human Annotation," in Association for the Advancement of Artificial Intelligence, **2011**.

[7] H.J. Mucha, "Adaptive cluster analysis, classification and multivariate graphics," Weirstrass Institute for Applied Analysis and Stochastics, **1992**.

[8] N.V. Anand Kumar and G. V. Uma, "Improving Academic Performance of Students by Applying Data Mining Technique," European Journal of Scientific Research, vol. 34, **2009**.

[9] Navjot Kaur Manjot Kaur, "Web Document Clustering Approaches Using K-Means Algorithm," IJARCSSE, **2013**.

[10] Marian Cristian Mihaescu, Mihai Mocanu Cosmin Marian Poteras, "An Optimized Version of the K-Means Clustering Algorithm," in IEEE, **2014**.

[11] O.O. Oladipupo, I.C Obagbuwa O.J. Oyelade, "Application of k-Means Clustering algorithm for prediction of Students’ Academic Performance," IJCSIS, **2010**.

[12] A. Jamshidzadeh , M. Saadatseresht , S. Homayouni A. Alizade Naeini, "An Efficient Initialization Method For K-Means Clustering Of Hyper spectral Data," ISPRS, Nov **2014**.

[13] Zhiyi Fang Chunfei Zhang, "An Improved K-means Clustering Algorithm," in JICS, **2013**.

[14] Wenbin, Yang,Yan &Qu Wu, "Interactive visual summary of major communities in a large network," in Pacific Visualization Symposium, Hangzhou,China, **2015**, pp. 47-54.

AUTHORS PROFILE

Hitesh Chandra Mahawari received his Engineering degree in Computer Science Engineering from RGPV University, India. He is Currently pursuing Masters in Technology in Computer Technology and Applications from SOIT, RGPV, India. His research interests are in the area of Big Data, Data Mining, & Internet of Things.

Dr. Mahesh K Pawar is Sr. faculty in Department of IT, UIT , RGPV, India . With 15 years of Academic Experience & three years of IT Industry Experience as a Software Engineer. His research interests are Software Engineering, Big Data, DBMS and Hadoop.