Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 11, November 2015, pg.72 – 76

RESEARCH ARTICLE

Implementing and Improvisation of K-means Clustering

Unnati R. Raval, Chaita Jani

Computer Science & KIRC (Gujarat Technological University), India

Computer Science & KIRC (Gujarat Technological University), India

Unnatiraval31@gmail.com

Jani.chaita@gmail.com

Abstract--- The clustering techniques are the most important part of the data analysis and k-means is the oldest and popular clustering technique used. The paper discusses the traditional K-means algorithm with advantages and disadvantages of it. It also includes researched on enhanced k-means proposed by various authors and it also includes the techniques to improve traditional K-means for better accuracy and efficiency. There are two area of concern for improving K-means; 1) is to select initial centroids and 2) by assigning data points to nearest cluster by using equations for calculating mean and distance between two data points. The time complexity of the proposed K-means technique will be lesser that then the traditional one with increase in accuracy and efficiency.

Keywords--- K-means, Centroid based Clustering, enhanced K-means, Clustering, Pro and Cons of K-means

I. INTRODUCTION

At present the applications of computer technology in increasing rapidly which created high volume and high dimensional data sets [2]. These data is stored digitally in electronic media, thus providing potential for the development of automatic data analysis, classification and data retrieval [2]. The clustering is important part of the data analysis which partitioned given dataset in to subset of similar data points in each subset and dissimilar to data from other clusters [1]. The clustering analysis is very useful with increasing in digital data to draw meaningful information or drawing interesting patters from the data sets hence it finds applications in many fields like bioinformatics, pattern recognition, image processing, data mining, marketing and economics etc. [3].

There have been many clustering techniques proposed but K-means is one of the oldest and most popular clustering techniques. In this method the number of cluster (k) is predefined prior to analysis and then the selection of the initial centroids will be made randomly and it followed by iterative process of assigning each data point to its nearest centroid. This process will keep repeating until convergence criteria met. However, there are shortcomings of K-means, it is important to proposed techniques that enhance the final result of analysis. This article includes researched on papers [3],[6],[7],[8] and [9] which made some very important improvements towards the accuracy and efficiency of the clustering technique.

The main purpose of the article is to proposed techniques to enhance the techniques for deriving initial centroids and the assigning of the data points to its nearest clusters. The clustering technique proposed in this paper is enhancing the accuracy and time complexity but it still needs some further improvements and in future it is also viable to include efficient techniques for selecting value for initial clusters (k).

II. K-MEANS: A CENTROID-BASED CLUSTERING TECHNIQUE

This partitioning clustering is most popular and fundamental technique [1]. It is vastly used clustering technique which requires user specified parameters like number of clusters k, cluster initialisation and cluster metric [4]. First it needs to define initial clusters which makes subsets (or groups) of nearest points (from centroid) inside the data set and these subsets (or groups) called clusters [1]. Secondly, it finds means value for each cluster and define new centroid to allocate data points to this new centroid and this iterative process will goes on until centroid [5] does not changes. The simplest algorithm for the traditional K-means [4] is as follows;

Input: D = {d1, d2, d3,.....dn} // set of n numbers of data points

K // The number of desire Clusters

Output: A set of k clusters [4]

1. *Select k points as initial centroids.*
2. *Repeat*
3. *From K clusters by assigning each data point to its nearest centroid.*
4. *Recompute the centroid for each cluster until centroid does not change.*

However, the algorithm has its own pros and cons, which is as follows;

PROs:

1. It is relatively faster clustering technique [1].
2. It works fast with the Large data set since the time complexity is O(nkl) where n is the number of patterns, k is the number of clusters and l is the number of the iterations [5, 6].
3. It relies on Euclidian distance which makes it works well with numeric values with interesting geometrical and statistic meaning [6].

CONs:

1. The initial assumption of value for K is very important but there isn't any proper description on assuming value for K and hence for different values for K will generate the different numbers of clusters [6].
2. The initial cluster centroids are very important but if the centroid is far from the cluster center of the data itself then it results into infinite iterations which sometimes lead to incorrect clustering [6].
3. The K-means clustering is not good enough with clustering data set with noise [6].

III. APPROACH FOR ENHANCING K-MEANS

Survey on Enhanced K-means algorithms

The K-means is very old and most used clustering algorithm hence many experiments and techniques have been proposed to enhance the efficiency accuracy for clustering. Let's discuss some of the improved K-means clustering proposed by different authors.

The first paper [3] proposing algorithm by improving the methods of finding of initial centroids and assigning data points to appropriate clusters [3]. Initially, it starts with the checking for negative value attributes in given data set and if there is some, then it a transom all negative value attributes to positive by subtracting it for the minimum positive value of the data set itself [3]. Next, calculate the distance from center to each data point then the original data points are sorted in accordance with the sorted distance and partitioned it with K equal cluster to find better initial centers [3]. In next step, all data points will be assigned to new centroids by using heuristic approach to reduce computational time. The time complexity of the proposed algorithm in this case will be O(nk) [3] instead of O(nkl) [2] for traditional K-means, is much faster.

In this proposal [7], instead of calculating distance from centroids to data points in all iterations it goes for, in traditional technique, it calculates the distance from new centroid once and if the distance is less than or equal to the previously calculated distance then it stays in cluster itself and no further calculation will be carry on for this particular data point otherwise it goes for same procedure again until all data points are assigned to its closest pre tagged centroids [7]. The time complexity of the proposed algorithm will be O(nk) [7] which will be faster than the traditional K-means.

The enhanced K-means [8] divides the algorithm in two phases and uses different algorithms to make proposed algorithm more efficient and accurate. In the first phase the initial centroids are determined systematically to produce the clusters with better accuracy and in second phase it uses various methods described in the algorithm to assigned data points to the appropriate clusters. There have been discussed algorithms for both phases in the paper. The time complexity of the algorithm, as claimed in paper, is O(n) where n is number of data point[8] with makes the algorithm much faster than the others.

In the other paper [6], initial centroids are optimised to find a set of data to reflect the characteristics of data distribution as the initial cluster centers and then optimising the calculation of distance between data points to the cluster centers and make it more match with the goal of clustering [6]. The test has been carried-out using IRIS and WINE data set and results are clearly showing improvements in accuracy for proposed techniques.

Finally, this technique proposed in [9], is base on the assumptions that the value of k (clusters) known in priori and then divide all data set into k clusters to calculate the new centroids. After that decide the membership of the patterns according to minimum distance from cluster center criterion and then do iterations using formula discussed in the paper [9]. Then the iteration will start with calculating distance from cluster center criterion and assigning of data points to cluster until no changes found in the cluster centers [9].

IV. POINTS CONSIDER FOR ENHANCING K-MEANS CLUSTERING

On the based on survey that have been carried –out on some proven enhanced K-means algorithms, there have been some areas which could be improved to get better accuracy and efficiency from altering traditional K-means. These areas have been discussed in this section with reference to proven theorems and methods.

As per research it is clear that we needs to make better assumption or find method for determining initial centroids with assigning data points to closed centroid clusters after iteration to enhance the results of traditional K-means.

Important Equations

For given dataset $D = \{x_1, x_2, x_3, ..., x_n\}$

Equation 1: The mean value is **Dmean=** $(x_1 + x_2 + x_3 + ... + x_n)/n$ **.....(Equ1)**

Where n // total number of data points in data set d

Equation 2: The distance between to data points could be calculate

$$Dist_{x_1 x_2} = \left| x_1 - x_2 \right| \quad \text{.....(Equ2)}$$

Equation3: The Centroid of dataset

$$Ci = Nearest(POINT) toD_{mean} \quad \text{.....(Equ3)}$$

Enhanced K-Means algorithm

This algorithm will be divided in two parts,

Input: $D = \{x_1, x_2, x_3, ..., x_n\}$ // set of n numbers of data points

K // The number of desire Clusters

Output: A set of k clusters

V. IMPROVED TECHNIQUE

Part1: Determine initial centroids [3]

Step1.1: Find the mean value for the given Dataset using Equation (Equ1).

Step1.2: Find the distance for each data point from mean value.

Step1.3: Sort data points according to their distance from the mean value calculated in step2.1.

Step1.4: Derive K number of equal subsets from data set.

Step1.5: Calculate middle point for each subset which will be initial centroids using Equ3.

Step1.6: Compute distance from each data point to initial centroids.

REPEAT

Part2: Assigning data points to nearest centroids [7]

Step2.1: Calculate Distance from each data point to centroids and assign data points to its nearest centroid to form clusters and stored values for each data.

Step2.2: Calculate new centroids for these clusters.

Step2.3: Calculate distance from all centroids to each data point for all data points.

IF

The Distance stored previously is equal to or less then Distance stored in Step3.1

Then Those Data points don't needs to move to other clusters.

ESLE

From the distance calculated assign data point to its nearest centroid by comparing distance from different centroids.

Step2.5: Calculate centroids for these new clusters again.

Until

The convergence criterion met.

OUTPUT

A Set of K clusters.

VI. TIME COMPLEXITY

As the concept is drawn from the [3] for deriving initial centroids the time it will take for first phase will be O(nlogn) [3] where n is the number of data points . However these technique needs to go through lot of sorting hence the overall rime complexity becomes O(nlongn) in both and worst case [3]. In the second phase of clustering, if the data point remains in the clusters itself then the time complexity becomes the O(1) and for others it else O(K)[7]. If half of the data points retains its clusters then time complexity will become O(nK/2) hence the total time complexity becomes O(nk). Hence the total time complexity fir the enhance K-means clustering proposed will become O(nlogn) which has less time complexity than the traditional k-means which runs with time complexity of O(nKl)[7].

VII. FUTURE WORK

In future there are still some areas which need to be improved to produce more accurate and faster clustering results. As research made on many proposed methods and texts, there is space for improvements and one of them explained below.

Selecting Value for K (Number of clusters)

For some of the popular datasets, the values for K have been set and it cloud be used for accurate results [10]. However, in reported papers [3],[6],[7],[8] and [9], there isn't any explanation on selecting values for K or about any particular method or techniques to assume values in case of other less popular datasets. Moreover, K-means clustering on any data-mining or data processing software, the values for K (Clusters) are always needs to be specified by users [10]. Since, performance of the clustering may be affected by the selected values for K (clusters) [2], we needs to find proper techniques for selecting K (number of clusters).

VIII. CONCLUSION

The traditional K-means clustering is most used technique but it depends on selecting initial centroids and assigning of data points to nearest clusters. There are more advantages than disadvantages of the k-means clustering but it still need some improvements. This paper explains the techniques that improves the techniques for determining initial centroids and assigning data points to its nearest clusters with more accuracy with time complexity of O(nlogn) which is faster than the traditional k-means. The initial value for the K (number of clusters) is still area of concern because it can improve accuracy of the clustering, which will be improved by enhancing the traditional way in future.

REFERENCES

[1] H. Jiawei,M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, San Francisco California, Morgan Kaufmann Publishers, 2012.

[2] A.K. Jain, Data clustering: "50 years beyond K-means, *Pattern Recognition Letters", Elsevier,* vol.31, pp.651-666, 2010.

[3] M. Yedla, S.R. Pathakota, and T.M. Srinivasa, "Enhancing K-means Clustering algorithm with Improved Initial Center", *International Journal of Computer Science and Information Technologies*, vol.1 (2), pp.121-125, 2010.

[4] P. Rai, and S. Sing, "A survey of clustering techniques", *International Journal of computer Applications,* vol. 7(12), pp.1-5, 2010.

[5] T. Soni Madhulatha, "An overview on clustering methods", *IOSR Journal of engineering*, vol. 2(4), pp.719-725, 2012.

[6] C. Zhang, and Z. Fang, "An improved k-means clustering algorithm", *Journal of Information & Computational Science*, vol. 10(1), pp.193-199, 2013.

[7] S. Na, G. Yong, and L. Xumin, Research on k-means clustering algorithms, *IEEE Computer society*, vol.74, pp.63-67, 2010.

[8] K.A.Abdul Nazeer and M.P. Sebastian, "Improving the accuracy and efficiency of the K-means clustering algorithm", *The World Congress on Engineering*, vol. 1, 2009.

[9] M.A. Dalal, N.D. Harale, and U.L. Kulkarni, "An iterative improved k-means clustering,", *ACEEE International Journal on Network Security*, vol. 2(3), pp.45-48, 2011.

[10] D.T. Pham, S.S. Dimov, and C.D. Nguyen, "Selection of K in K-means clustering", *IMechE 2005*, vol.219, pp.103-119, 2014.

