A dynamic K-means clustering for data mining

Md. Zakir Hossain[1], Md. Nasim Akhtar[2], R.B. Ahmad[3], Mostafijur Rahman[4]
[1,2]Department of Computer Science and Engineering, Dhaka University of Engineering and Technology, Bangladesh
[3]Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin (UniSZA), Malaysia
[4]Department of Software Engineering, Daffodil International University (DIU), Bangladesh

Article Info

Article history:

Received Sep 25, 2018
Revised Nov 24, 2018
Accepted Dec 8, 2018

Keywords:

Centroid
Clustering
Data mining
Euclidean distance
K-Means
Threshold value

ABSTRACT

Data mining is the process of finding structure of data from large data sets. With this process, the decision makers can make a particular decision for further development of the real-world problems. Several data clusteringtechniques are used in data mining for finding a specific pattern of data. The K-means method isone of the familiar clustering techniques for clustering large data sets. The K-means clustering method partitions the data set based on the assumption that the number of clusters are fixed. The main problem of this method is that if the number of clusters is to be chosen small then there is a higher probability of adding dissimilar items into the same group. On the other hand, if the number of clusters is chosen to be high, then there is a higher chance of adding similar items in the different groups. In this paper, we address this issue by proposing a new K-Means clustering algorithm. The proposed method performs data clustering dynamically. The proposed method initially calculates a threshold value as a centroid of K-Means and based on this value the number of clusters are formed. At each iteration of K-Means, if the Euclidian distance between two points is less than or equal to the threshold value, then these two data points will be in the same group. Otherwise, the proposed method will create a new cluster with the dissimilar data point. The results show that the proposed method outperforms the original K-Means method.

Corresponding Author:

Md. Zakir Hossain,
Department of Computer Science and Engineering,
Dhaka University of Engineering and Technology, Gazipur, Bangladesh.
Email: zakircse11.duet@gmail.com

1. INTRODUCTION

The new interdisciplinary field of computer science is data mining. This is the process of finding data pattern automatically from the large database [1]. The necessity of data mining is increasing day by day since previous ten or fifteen years and so now in this time on the marketplace is very challenging competition to efficiency of information and information rapidly performed an important role to find out a decision of plan and provided a great offer of information in industry, society and all together. In real-world, a large number of data is available in which it is difficult to retrieve the useful information. Due to the practical importance, it is important to retrieve the structure of data within the given time budget. The data mining provides the way of eliminatingunnecessary noises from data. It helpstoprovide necessary information from the large dataset and present it in the proper form when it is necessary for a specific task. It's very helpful to analyze the market trend, search the new technology, production control based on the demand of customer and so on. In a word, the data mining is harvesting of knowledge from a large amount of data. We can predict the type or behavior of any pattern using data mining.

Cluster evaluation of data is an important task in knowledge finding and data mining. Cluster formation is the process of creating data group based on the data similarities from large dataset. The clustering process is done by supervised, semi-supervised or unsupervised manner [2].

The clustering algorithms are powerful meta-learning tools for analyzing the data produced by modern applications. The purpose of clustering is to classify the data into groups according to similarities, traits, and behavior of data [3].

Many clustering algorithms have been proposed for classification of data. Most of these algorithms are based on the assumption that the number of clusters in a large data is fixed.The problem with this assumption is that if the assumed number of cluster is small then there is a higher chance of adding dissimilar items into the same group. On the other hand, if the number of cluster is large, then there is a higher chance of adding similar data placed into different groups [4]. In addition, in the real situation, it is difficult to know the number of clusters in advance.

In this paper, we develop a dynamic K-Means clustering algorithm. This algorithm firstly calculates a threshold value based on the data set and then groups the data set without fixing the number of clusters (K). In the proposed algorithm analyze the data set based on the threshold value and finally the data set is clusters. The threshold value is the key to this proposed method. The threshold value determines the data are same group or create a new group.

2. THE K-MEANS CLUSTERING ALGORITHM

In this section, we describe the K-Means algorithm first then the detail of the proposed algorithm will be provided in the following section. The K-Means clustering algorithm is a popular algorithm which works for various types of data namely medical image, text and so on. The performance of clustering algorithms depends on the initial centroid of K-Means. If the selection of centroid is wrong, then clustering result is volatile and the number of iterations will be increased. Therefore, both the time and space complexity will be increased proportionally [5].

The K-Means algorithm is widely used technique which is a simple clustering technique in data mining. It isa non-supervised learning algorithm which is used to solve well-known cluster problem [6]. Partition based clustering is a way to cluster large data sets in which a number of objects are given first, then theseobjects are partitioned into a number of groups and each group contains similar data points [7].

The K-means algorithm classifies the data into K different cluster through the iterative, converging process. The generated clusters of K-Means are independent. The K-Means clustering algorithm works in two different parts. Firstly, it selects a K-value, where K is the number of clusters. Another part is to consider each data point to the nearest center [8]. After completing the first step then calculate the Euclidean distance between the data point to K centroids. Then all the data points are used to create some group. This process will be continuing until minimum. The K-Means algorithm given below.

Here, K is the number of clusters and D is the data set which contains n data objects.

Step-1: Select k data objects from D as an initial cluster centers.

Step-2: Repeat Step 3 and Step 4, if the center of clusters remains unchanged.

Step-3: Calculate the distance between each data object di, where i=0,1,2,…K-1 and all k cluster centers cj, where j=0,1,2…K-1. Assign data object di to the nearest cluster.

Step-4: For each cluster j, recalculate the cluster center.

The K-Means clustering algorithm result, it is so close to each data points in each data group. In K-Means algorithm, the data groups are created before calculating the distance between centroid to each data point and this process continues a number of times until each data points are purely group [9]. So the time complexity of the K-Means clustering algorithm is O(mkt). Where ‘m’ is the data points, ‘k’ is the initial centroids,‘t’ is the number of iterations [10].

3. RELATED WORK

In this section, we will give a brief discussion of the existing K-means algorithms. In [2], a modified K-Means algorithm is proposed to select the initial center of cluster based on the improvement of the sensitivity. This algorithm divides the whole space in segment and calculates the frequency between the segment and each data point. The maximum frequency of data point selects the centroid. In this method, the number K is defined by user as defined by the traditional K-mean algorithm. For this algorithm, the number of divisions will be k*k, where ‘k’ vertically as well as ‘k’ horizontally.

In [10], an improved k-means algorithm is proposed. In this algorithm, the information of data structure needs to store in each iteration. This information used in next iteration. This proposed method without calculating the distance between each data points and cluster centers repeatedly, so saving the running time.

In [11], an optimized k-means clustering method is proposed based on three optimization principles named k*-means. Firstly, a hierarchical optimization principle initialized by k* cluster centers (k*> k) to reduce the risk of randomly seeds selection. Secondly, a cluster pruning strategy is proposed for improving the efficiency of k-means. Finally, it implements an optimized update theory to optimize the k- means iteration updating.

4. PROPOSED METHOD

Our proposed method clusters dynamically all data from a large data set without specifying (K) value, where (K) is the number of clusters. In K-Means firstly select the (K) value then start clustering based on the value of (K). But, at first, it is the difficult task to select. For this reason K-means clustering result quality becomes poor. In our proposed method to cluster large data set based on the threshold value and the result of clustering quality is improved.

$$\sum_{i=0}^{N-1} \frac{\sum_{j=0}^{N-1} \frac{dist(x_i, x_j)}{N}}{N} \tag{1}$$

$$Min\left(\sum_{i=0, j=0}^{N-1} dist(x_i, x_j) \right) \tag{2}$$

Our proposed algorithm given below.

Where ‘D’ (d_1, d_2, ………d_n) is data sets. ‘n’ is the data points. ‘K’ is the clusters. ‘X’ (x_1, x_2, x_3, ………x_n) is the data point. The Th is the threshold. ‘c’ is cluster center.

```
Step-1: Calculate distance matrix distance di($x_i$,$x_j$), where i=0, 1, 2, …….N-1 and j=0, 1, 2, …N-1.
Step-2: Calculate the threshold value Th using (1).
Step-3: Find the minimum Mean from $x_i$ to $x_j$ using (2).
Step-4: Find the minimum mean value index $x_i$. Select $x_i$ th data point as a first centroid.
Step-5: Repeat Step-6 and Step-7 until data points changes group otherwise Step-8.
Step-6: Calculate the distance between each data point $x_i$ and all K cluster centers $c_j$.
            if (Th>=$d_i$)
            Assign data point xi to the nearest cluster.
            else
            K=K+1;
Step-7: recalculate the each cluster center.
Step-8: End.
```

5. EXPERIMENTAL RESULT AND ANALYSIS

5.1. Experimental Setup

We have simulated our proposed method using MATLAB, Java, MapReduce, and C++ in a personal computer. The personal computer specification is 4GB RAM 2.4 GHz Corei5 processor. At first, our proposed method developed in C++ and then it is converted into java MapReduce. The result of our proposed method is applied in the MATLAB to see the cluster and data point position. Then we have developed general K-Means algorithm to compare with our proposed method.

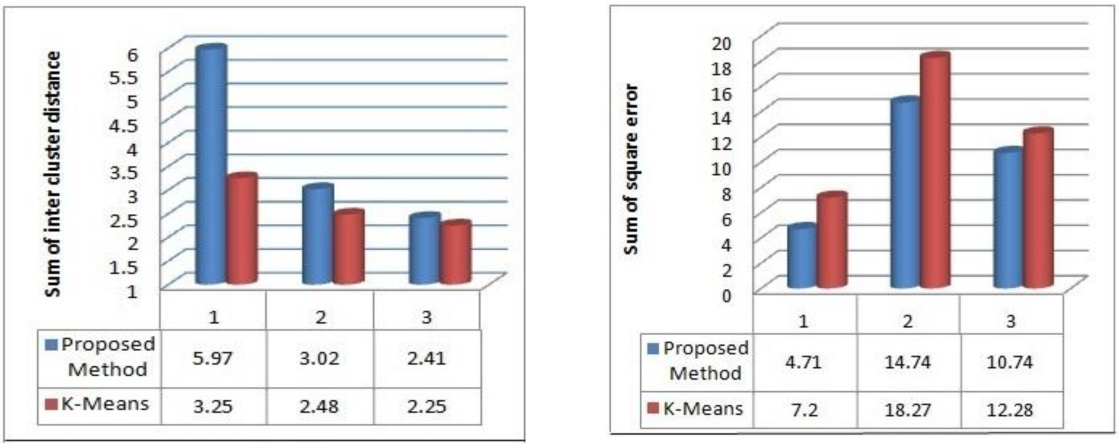
5.2. Result Analysis

In the result analysis, compare between proposed method and general K-Means clustering based on various parameters such as inter cluster distance, intra cluster distance and sum of square error (SSE). If SSE and intra cluster distance are minimum, then the quality of cluster is good. If inter cluster distance is maximum, then the quality of cluster is good. For result analysis, we are generated some data set using java. The range of data sets between 0 to 100 and number of data point is 100, 200,300, 400, 500, 1000 and also use iris data set. Table 1 shown the iris data set result and compare between proposed method and K-Means clustering based on sum of inter cluster distance and sum of square error. In iris setosa and Iris versicolour each has 50 instances.

Table 1. Result for Iris Data Set

Data point group	The proposed Method Algorithm			General K-Means Algorithm		
	# of cluster(K)	Sum of inter cluster distance	Sum of square error	# of cluster(K)	Sum of inter cluster distance	Sum of square error
Iris setosa (petal lenght) (petal width)	6	5.97	4.71	6	3.25	7.20
Iris versicolour (sepal lenght) (sepal width)	3	3.02	14.74	3	2.48	18.27
Iris versicolour (petal lenght) (petal width)	3	2.41	10.74	3	2.25	12.28

Figure 1(a) showing the comparison between K-Means algorithm and propose algorithm based on sum of inter-cluster distance. Our proposed algorithms apply in iris setosa. It creates six data group dynamically based on similarity. So the sum of inter-cluster distance is increased. In K-Means algorithm sum of inter-cluster distance is decreased shown in Figure 1(a). Figure 1(b) showing the comparison between K-Means algorithm and propose algorithm based on sum of square error. Our proposed algorithms apply in iris data sets. Then sum of square error is decrease. In K-Means algorithm sum of square error is increased shown in Figure 1(b). For Our generated data sets result given in Table 2.



(a) (b)

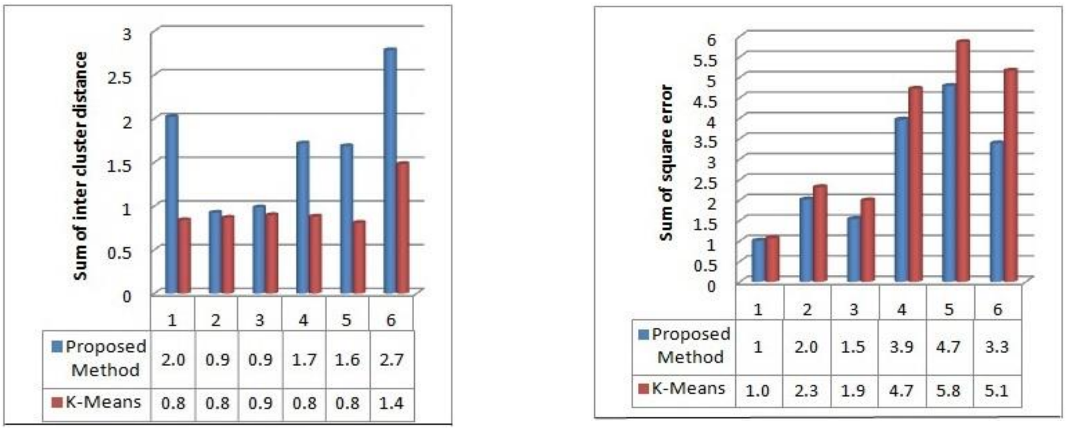
Figure 1(a). Sum of inter cluster distance for iris data set, (b). Sum of square error for iris data set

Table 2. Result for Generated Data Set

Number of Data points	Our proposed Method Algorithm			General K-Means Algorithm		
	# of cluster (K)	Sum of inter cluster distance	Sum of square error	# of cluster (K)	Sum of inter cluster distance	Sum of square error
100	4	2.02	1.003	3	0.84	1.066
200	3	0.9243	2.01	3	0.8673	2.312
300	6	0.9856	1.534	6	0.8975	1.987
400	4	1.7177	3.964	4	0.8813	4.716
500	4	1.686	4.784	4	0.8084	5.858
1000	8	2.78	3.38	6	1.48	5.1

Table 2 shows our generated data set result and compare between proposed method and K-Means clustering based on sum of inter cluster distance and sum of square error. We are generated some data sets. The range of data sets between 0 to 100 and each data set has 100, 200, 300, 400, 500, and 1000 instance.

Figure 2(a) showing the comparison between K-Means algorithm and propose algorithm based on sum of inter-cluster distance using our generated data sets. Figure 2(a) show when number of data points increase then sum of inter cluster distance increase for our proposed method. So, data points are group efficiently. In K-Means algorithm sum of inter-cluster distance is decrease. Figure 2(b) showing the comparison between K-Means algorithm and propose algorithm based on sum of square error. Figure 2(b) show sum of square error is decrease for our proposed method. In K-means sum of square error is increased. So, Cluster quality is poor.



(a)

(b)

Figure 2(a). Sum of inter cluster distance for our generated data set, (b). Sum of square error for our generated data set

6. CONCLUSION

In this paper, we propose a new K-Means algorithm to remove the difficulties of the existing K-Means algorithm. The proposed method dynamically forms the clusters for a given data set. We compare our proposed method with the existing K-Means algorithm. The results show that the proposed method outperforms the existing method for the well-known iris data set.

REFERENCES

[1] S. Sharma, J. Agrawal, S. Agarwal, S. Sharma, *"Machine Learning Techniques for Data Mining: A Survey"*, in IEEE International Conference on Computational Intelligence and Computing Research, 2013.

[2] R. V. Singh, M.P.S. Bhatia, *"Data Clustering with Modified K-means Algorithm"*, in IEEE-International Conference on Recent Trends in Information Technology (ICRTIT 2011), June 2011.

[3] V. W. Ajin, L.D. Kumar, *"Big data and clustering algorithms"*, in International Conference on Research Advances in Integrated Navigation Systems (RAINS), May 2016.

[4] A. Shafeeq. B. M, Hareesha. K. S, *"Dynamic Clustering of Data with Modified K-Means Algorithm"*, in International Conference on Information and Computer Networks (ICICN 2012), IPCSIT vol. 27 IACSIT Press, Singapore, 2012.

[5] L. Guoli, W. Tingting, Y.Limei, *"The improved research on k-means clustering algorithm in initial values"*, in International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC), Shengyang, China, 2013.

[6] S. Jigui, L. Jie, Z. Lianyu, "Clustering algorithms Research", *in Journal of Software*, 2008; 19(1): 48-61.

[7] D.Neha, B.M. Vidyavathi, "A Survey on Applications of Data Mining using Clustering Techniques", *in International Journal of Computer Applications* (0975–8887) Volume 126 – No.2, 2015.

[8] M.Fahim, A. M. Salem, F. A. Torkey, "An efficient enhanced k-means clustering algorithm", in *Journal of Zhejiang University Science A*, 2006; 10: 1626-1633.

[9] K.A. Abdul Nazeer, M.P. Sebastian, *"Improving the Accuracy and Efficiency of the k-means Clustering Algorithm"*, in Proceeding of the World Congress on Engineering, vol 1, London, July 2009.

[10] L. ShiNa, G. Xumin, *"Research on k-means Clustering Algorithm an Improved k-means Clustering Algorithm"*, in Third International Symposium on Intelligent Information Technology and Security Informatics.

[11] J. Qi, Y. Yu, L. Wang, J. Liu, *"K*-Means: An Effective and Efficient K-means Clustering Algorithm"*, in IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom), 2016.

BIOGRAPHIES OF AUTHORS

	Md. Zakir Hossain received the B.Sc Engineering degree in Computer Science and Engineering Department from Dhaka University of Engineering and Technology (DUET), Gazipur, Bangladesh, in 2015 and he is currently pursuing the M.Sc Engineering degree in Computer Science and Engineering Department in Dhaka University of Engineering and Technology (DUET), Gazipur. His research interest includes Data Mining, Big Data, AI, Machine Learning, Cloud Computing, Software Engineering, Computer Network, IoT. He has presented papers at conferences both at home and abroad.
	Md. Nasim Akhtar received the M.Eng and Ph.D degrees from National Technical University of Ukraine, Kiev, Ukraine and Moscow State Academy of Fine Chemical Technology, Russia, in 1998 and 2010, respectively. Currently, he is a Professor in the Department of Computer Science and Engineering, Dhaka University of Engineering and Technology (DUET), Gazipur, Bangladesh. His research interests includes Distributed Data Warehouse System On Large Clusters, Digital Image Processing and Water Marking, Peer to Peer Networking, Cloud Computing, Operating System. He has presented papers at conferences both at home and abroad, published articles and papers in various journals.
	R. Badlishah Ahmad obtained Bachelor of Engineering with Honors (B.Eng. (Hons)) in Electrical & Electronic Engineering from Glasgow University, UK in 1994. Continued Master of Sciences (M.Sc.) in Optical Electronic Engineering at University of Strathclyde, UK and graduated in 1995 and in 2000 completed PhD. Research interests are in Computer and Telecommunication Network Modelling include WSN and Optical Network using discrete event simulators (OMNeT++), Optical Networking and Embedded System based on GNU/Linux.
	Mostafijur Rahman completed his BSc in Computer Science from National University of Bangladesh (2003). He Pursued his MSc (2009) and PhD (2017) in Computer Engineering, from UNIMAP, Malaysia. He worked as Lecturer since 2009 to September, 2017 for School of Computer and Communication Engineering in UNIMAP. Currently he is serving as Assistant Professor in the Department of Software Engineering at Daffodil International University (DIU), Bangladesh. His research interest in Software Testing, Multimedia and Creativity in Medical Science, Computer Security, Cloud Computing, Algorithm Optimization, Parallel and Distributed System, Device Driver for GNU/Linux based embedded OS.