Journal of Engineering and Applied Sciences 14 (18): 6742-6746, 2019
ISSN: 1816-949X

New Modified Dynamic Clustering Algorithm

[1]Mahdi Nsaif Jasim and [2]Mohamed Ibrahim
[1]Department of MIT, Business Informatics College,
University of Information Technology and Communications, Baghdad, Iraq
[2]Department of Information Networks, College of IT, University of Babylon, Babil, Hilla, Iraq
mohamed.ibrahim@uobabylon.edu.iq

Abstract: k-clustering is one of the most common ways to divide the extracted data into clusters which is considered a type of knowledge discovery. While there is a great research effort to determine the key features of mass K, further investigation is needed to determine whether the optimal number of clusters can be found during the process based on the cluster quality scale. This study presents a modified k-means algorithm used to improve cluster quality and optimizing the optimal number of clusters. The k-means algorithm takes the number of clusters (k) as input from the user. But in the practical scenario, it is difficult to determine the number of clusters in advance. The evolution of the proposed method is equivalent to finding the value of the threshold. The suggested threshold value will be used as a distance between the center of each group and other group's centers. Applying the modified algorithm improves the results of enter cluster is 0.111 and entra cluster is 0.0034.

Key words: Dynamic clustering, optimal clustering, k-means algorithm, clustering quality, weather data, optimal number

INTRODUCTION

Data clustering is important and fundamental issue often occurs in a wide range of applications such as knowledge discovery and knowledge extraction. Data partitioning and clustering are fundamental knowledge discovery operations. Data mining has become increasingly important, since, the past decade and in recent times where there is very strong competition in the market where the quality of information and information in a timely manner plays very crucial role in political and marketing decision-making that has attracted much attention in industry information and society as a whole (Li, 2018). There is very high level of data availability in the real world and it is very difficult to extract useful information from this large data and to provide the necessary information in a timely and appropriate manner. Hence, data extraction is a tool for extracting information from a large database and submitting it as required for each specific task. The use of data mining is very broad. It is very useful in application such as knowing markets directions, fraud detection, customer shopping pattern, production control, science exploration and so on. Using data mining, we can predict the nature or behavior of any data pattern. Data cluster analysis is an important and interesting task in discovering, extracting knowledge and exploring data in order to be useful (Aliahmadipour *et al*., 2017). Cluster analysis aims to collect data based on similarities and differences between data elements. The process can be performed in a supervised, semi-controlled or unattended manner (Shmueli *et al*., 2017). Various algorithms have been proposed to deal with the nature of data and input parameters for data collection. Most of the clustering algorithms deals with constant number clusters (k) (Patel and Mehta, 2011a, b).

In real-world application, it is very difficult to predict the number of unknown domain data sets. If the clusters number is very small there is a chance to place different objects in the same group and while when the number of fixed groups is large and then the most similar objects are placed in different groups, both situations are not accurate. In this study, the proposed algorithm which is modified k-means algorithm produces a dynamic number of clusters. The algorithm takes the number of clusters (k) as input from the user and the user must indicate whether the number of groups is fixed. If the number of clusters is fixed, the algorithm works in the same way of k-means algorithm. Assume that the number of clusters is not predefined then the threshold is computed

Corresponding Author: Mohamed Ibrahim, Department of Information Networks, College of IT, University of Babylon, Babil, Hilla, Iraq, mohamed.ibrahim@uobabylon.edu.iq

as in Eq. 2, taking into account the total number of data points in addition to the maximum and minimum values in the data points.

Original k-means clustering: k-means is a cluster analysis methodology designed to split n observations into k groups, so that, each point belongs to the cluster with the nearest seed (Shmueli *et al*., 2017). It is an indicative algorithm that can reduce the sum of the space boxes from all samples originating in the assembly area to the collection of centers to find the minimum of clustering based on the target function (Li, 2018). First, k is accepted as input and the data objects that belong to the assembly domain (including n, n>k) are divided into k types. As a result, the similarities between the same cluster samples are higher but lower among the heterogeneous clusters. K data objects as the original assembly centers are randomly selected from the assembly domain by km algorithm. k-clustering is an algorithm for learning data extraction/machine that is used to collect notes in sets of relevant notes without any prior knowledge of those relationships. The k-means algorithm is one of the simplest assembly techniques and is commonly used in medical imaging, biometrics and related fields. The k-means algorithm is an evolutionary algorithm that acquired its name from its mode of operation. The algorithms gr notes in groups k where k is provided as an input parameter. Then, each note sets for groups based on the proximity note of the average block. The cluster average is then recalculated and the process begins again. Here's how the algorithm works (Shmueli *et al*., 2017) (Algorithm 1).

Algorithm 1; k-means:

The partitioning algorithm where each center is represented in the cluster
Through the average value of objects in the group
Inputs: k: number of clusters, D: dataset containing n objects
Output: A set of k groups
Methods:
1. Select the k objects from D as the centers of the initial assemblies
2. Repeat
3. (Re) assign each object to the cluster that has the most similar object by using the equalizer. 1, based on the average value of the objects in the group
4. Update the cluster seeds, that is, calculate the mean value of objects per cluster
5. So do not change

$$\text{Distance} = \sum_{i=1}^{n} |X_i - Y_i| \quad (1)$$

Literature review: Many researchers have attempted to enhance the efficiency of the km algorithm (Fahim *et al*., 2006; Huang, 1998; Yuan *et al*., 2004). The most compatible alternative of the km algorithm is the k-modes (Huang, 1998; Chaturvedi *et al*., 2001) method which replaces the modes of the groups with situations. Like the k-means method, the k-modes algorithm also produces ideal solutions locally based on the choice of initial patterns. The k-prototypes (Huang, 1998) integrate k-modes and k-modes to aggregate the data. In this way, the different scale is determined by considering both numeric and class characteristics. The original km algorithm consists of two phases: one to determine the initial and other central points to assign data points to the nearest groups and then to recalculate the cluster. The second phase is performed repeatedly until the groups are settled, meaning that the data points stop when crossing the group boundaries. Yuan *et al*. (2004) A systematic method was suggested to find primary centroids. The central pulses obtained in this way are consistent with the data distribution. Thus, groups were produced more accurately, compared with the original k-means algorithm. However, the RMB method of doing does not indicate any improvement in the complexity of the time of the k-means algorithm. Singh and Bhatia (2011) proposed a data clustering approach which works by partitioning the space into different segments and calculating the frequency of data point in each segment and the segment which shows maximum frequency of data point has maximum chances to contain the centroid of the cluster.

The researchers have introduced the concept of threshold distance for each cluster's centroid for comparing the distance between a data point and cluster's centroid and using this method, efforts to calculate the distance between data point and cluster's centroid is minimized. This algorithm effectively decreases the complexity and makes calculations easier. Fahim *et al*. (2006) an effective way to assign data points to groups was suggested. The original k-means algorithm is very expensive computationally because each iteration calculates the distances between data points and all centroids. Fahim's approach benefits from two distance functions-one like the k-means algorithm and another based on inference to reduce the number of distance calculations. But this method assumes that the initial costs are randomly selected as in the original k-means algorithm. Thus, the accuracy of the final groups is not guaranteed.

Contributions:

- The proposed formula calculates the threshold for initial centroids before applying the proposed algorithm

- The proposed algorithm produces optimal number of clusters
- Time complexity of the proposed algorithm is less than from classic k-means (O(2kn))

Original k-means algorithm has a predetermined number of clusters. In the actual action, it is very necessary to find the number of groups for unknown datasets at runtime. Installing several clusters may result in poor quality assembly. We apply the proposed algorithm with the calculation of the initial centroids based on the weighted average score of network devices features. Next, we perform preprocess operations and normalize the data set before applying the modified method algorithm. This proposed method works in three phases. During the first phase a pre-processing technique is adopted for data that transforms the raw data into an understandable form. During the second phase, normalization is done to standardize data objects in a given range. During the third phase, the modified algorithm is applied to create clusters. The proposed method finds the number of playback groups based on a dynamic threshold. This method works in an unknown number of groups (clusters). In the proposed method, the user does not need to specify the number of clusters but a dynamic configuration is specified. The proposed method works based on developed equation that plays important role by determining the optimal number of clusters dynamically.

MATERIALS AND METHODS

Data preprocessing: Preprocessing is very important step to produce acceptable clustering results. This step uses parameters such as constant, mean, minimum, maximum and standard deviation to calculate missing values in data records (Patel and Mehta, 2011a, b). Missing values should be reduced to the minimum acceptable number to get accurate results. Initial processing includes steps such as data cleaning, data integrity, data conversion, data reduction and data reduction.

Normalization: Extracting data yields effective results if normalization is applied to the data set. Normalization is a process used to standardize all data set attributes and assign them the suitable weight, so that, redundant or disturbing elements can be eliminated and valid and reliable data improves the accuracy of the result (Qin *et al*., 2017). K-mean algorithm that uses Euclidean space which is highly prone to irregularities in the size of different features (Patel and Mehta, 2011a, b). There are many ways to normalize data such as Min-Max, Z-score and decimal scaling. The best way of normalization depends on the data to be normalized. Here, we used the Min-Max technique to normalize our data because our data set is limited and does not have significant variability between minimum and maximum. The Min-Max adjustment technique performs a linear conversion of the data (Singh and Bhatia, 2011). In this way, we fit the data within predefined limits or at a predetermined interval.

Threshold based initial centroids calculation: Equation 2 takes into consideration the total number data points and maximum, minimum of a weighted average score of network devices properties. In the proposed method, initialize centroids not randomly but using a suitable threshold computed using Eq. 2:

$$\text{Threshold} = \frac{1}{1+\sqrt[3]{\log_2 N^{*}(\text{Max-Min})}} \quad (2)$$

Where:
N = Total Number of data points
Max = Maximum value in features average
Min = Minimum value in features average

To illustrate the method and how threshold determines the clusters centroids take this example: If the value of the threshold is 0.3 and the values of data point between 1 and 0, so, the value of the second centroid will be 1 the threshold value and equal to 0.7, to find third centroid by the value of the second centroid minus the value of the threshold and thus will be (0.4). It is necessary to apply normalization process on values of the data points to ensure that their values are between 0 and 1 in order to apply the threshold correctly because the threshold value is <1. The modified algorithm can be presented (Algorithm 2):

Algorithm 2; Threshold-based clustering:
Input: D: a data set containing n objects
Output: A set of k clusters
1. Using (Eq. 2) to find the threshold, (the threshold value is in range 1-0
2. While value of last centroid more than zero
3.
   a- Find the first centroid by subtracting 1 from the value of threshold in 1
   b- find the next centroid by subtracting the value of the previous centroid by value threshold in 1
4. End While
5. Re) Assign each object to the cluster that has the most similar object by using (Eq. 1) based on the average value of the objects in the cluster
6. Update the cluster seeds that is calculate the mean value of objects for each cluster

RESULTS AND DISSCUSSION

In this study, the effectiveness of the proposed algorithm is analyzed, 1400 data records are taken in account. Experimental data is defined as network devices

Table 1: Experimental results

No. test	Data points	k-by user	k by (proposed algorithm)	Inter-distance (k-means)	Inter-distance (proposed algorithm)	Intra-distance (k-means)	Inter-distance (proposed algorithm)
1 (Fig. 1)	63	2	4	0.2274	0.27956	0.02189	0.01650
2 (Fig. 2)	170	4	6	0.10379	0.21489	0.01697	0.01357
3 (Fig. 3)	517	8	10	0.10331	0.1327	0.01907	0.00836

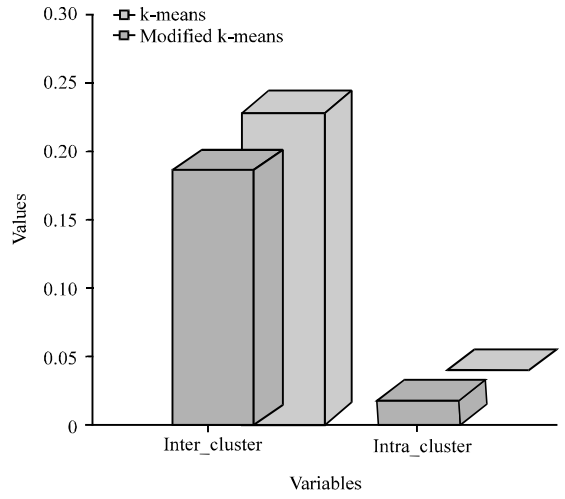

Fig. 1: First test

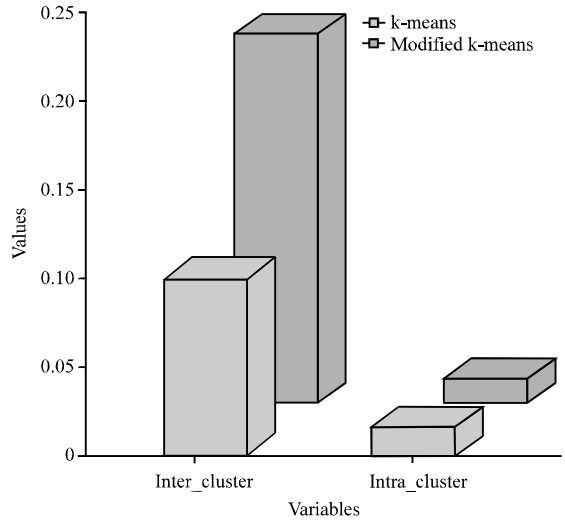

Fig. 2: Second test

with their specifications where the specifications are temperature, humidity and product brand. The test results explained that the proposed method performs better than the k-means algorithm in the quality of clustering and the ability to deal with incomplete data. The experiment was done on a real data set Table 1. The proposed algorithm works for an unknown number of clusters and gives the optimal number of clusters. As shown in the above Table 1 and Fig. 1-3 the test results is done on real data,

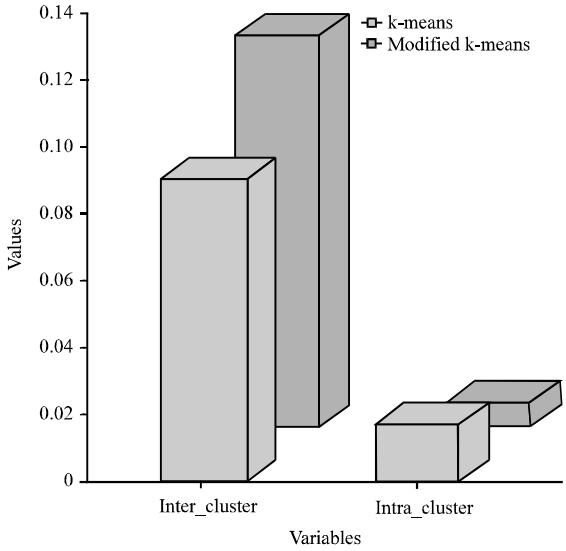

Fig. 3: Third test

the proposed algorithm gives an ideal number of clusters for the data set. It is also noted that the time spent in the proposed method is roughly the same as the k-means algorithm for the smaller data set. The algorithm was developed and tested for the efficiency of various data points in C # language. The algorithm takes a less computational time than the k-means algorithm for the large data set in some cases. The clustering results including measuring the distance between each object with all seeds to do the best clustering results.

CONCLUSION

The proposed algorithm provides an ideal number of clusters for the tested data sets. It is also noted that the time taken in the proposed method is almost less than the k-mean algorithm method for the same data size. This study presents an enhanced modified k-means algorithm by prepare formula to calculate threshold-based initials seeds. When we compare the results, the higher the value of the threshold, the fewer the clusters and vice versa. The time complexity of the proposed algorithm is (2kn) where, n is the number records and k is the number of clusters. Finally, the number of not clustered records is approaches to zero.

REFERENCES

Aliahmadipour, L., V. Torra and E. Eslami, 2017. On Hesitant Fuzzy Clustering and Clustering of Hesitant Fuzzy Data. In: Fuzzy Sets, Rough Sets, Multisets and Clustering, Torra, V., A. Dahlbom and Y. Narukawa (Eds.). Springer, Cham, Switzerland, ISBN:978-3-319-47556-1, pp: 157-168.

Chaturvedi, A.D., P.E. Green and J.D. Caroll, 2001. K-modes clustering. J. Classification, 18: 35-56.

Fahim, A.M., A.M. Salem, F.A. Torkey and M.A. Ramadan, 2006. An efficient enhanced k-means clustering algorithm. J. Zhejiang Univ. Sci. A, 7: 1626-1633.

Huang, Z., 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining Knowledge Discovery, 2: 283-304.

Li, W., 2018. Applications of clustering algorithms to entity resolution and human genome variation. Ph.D Thesis, Penn State University, Pennsylvania, USA.

Patel, V.R. and R.G. Mehta, 2011a. Impact of outlier removal and normalization approach in modified K-means clustering algorithm. Intl. J. Comput. Sci. Issues, 8: 354-359.

Patel, V.R. and R.G. Mehta, 2011b. Performance analysis of MK-means clustering algorithm with normalization approach. Proceedings of the 2011 World Congress on Information and Communication Technologies, December 11-14, 2011, IEEE, Mumbai, India, ISBN: 978-1-4673-0127-5, pp: 974-979.

Qin, J., W. Fu, H. Gao and W.X. Zheng, 2017. Distributed k-Means Algorithm and Fuzzy c-Means Algorithm for sensor networks based on multiagent consensus theory. IEEE. Trans. Cybern., 47: 772-783.

Shmueli, G., P.C. Bruce, I. Yahav, N.R. Patel and K.C. Lichtendahl Jr, 2017. Data Mining for Business Analytics: Concepts, Techniques and Applications in R. John Wiley & Sons, Hoboken, New Jersey, USA., ISBN:9781118879368, Pages: 574.

Singh, R.V. and M.S. Bhatia, 2011. Data clustering with modified K-means algorithm. Proceedings of the 2011 International Conference on Recent Trends in Information Technology (ICRTIT), June 3-5, 2011, IEEE, Chennai, Tamil Nadu, India, ISBN: 978-1-4577-0588-5, pp: 717-721.

Yuan, F., Z.H. Meng, H.X. Zhang and C.R. Dong, 2004. A new algorithm to get the initial centroids. Proceedings of the 3rd International Conference on Machine Learning and Cybernetics, August 26-29, 2004, Shanghai, China, pp: 1191-1193.