Middle-East Journal of Scientific Research 12 (7): 959-963, 2012
ISSN 1990-9233
© IDOSI Publications, 2012
DOI: 10.5829/idosi.mejsr.2012.12.7.1845

Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity

Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed

Department of Computer Science University of Peshawar Peshawar, Pakistan

Abstract: Clustering technique is used to put similar data items in a same group. K-mean clustering is a common approach, which is based on initial centroids selected randomly. This paper proposes a new method of K-mean clustering in which we calculate initial centroids instead of random selection, due to which the number of iterations is reduced and elapsed time is improved.

Key words: Data mining ・ Clustering ・ K-mean clustering algorithm

INTRODUCTION

Data mining is the analysis of data and the use of software techniques for finding patterns and regularities in the set of data [1]. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid pattern and relationship in large data set [2].

Clustering is one of the broad fields of data mining. In clustering data elements having similarities are placed in respective groups. Clustering algorithms mainly fall into two categories: Hierarchical clustering and partition clustering [3]. The main difference between partitioned and hierarchical clustering is that, in partitioned clustering algorithm data is partitioned into more than two subgroups in one steps and in hierarchical clustering algorithm data is divided into two subgroups in each step. K-mean clustering is a partitioning clustering technique in which clusters are formed with the help of centroids. On the basis of these centroids, clusters can vary from one another in different iterations. Moreover, data elements can vary from one cluster to another, as clusters are based on the random numbers known as initial centroids.

Several attempts have been made by the researchers to improve the efficiency of the basic K-mean algorithm. A new algorithm is introduced and implemented in this research. The rest of the paper is organized in this way. First the basic K-mean clustering algorithm is discussed and then proposed K-mean clustering algorithm is explored. The implementation work and the results of experiments are followed by the comparison of both algorithms.

Related Work: In [2], researchers introduced an enhanced K-means algorithm to improve the time complexity using uniform data. They make clusters in two phases. In phase one, they find initial clusters on similarity basis, while in second phase they finalize clusters. Similarly in [4], all of the requirements, advantages and disadvantages of basic K-mean clustering are discussed.

In [5], the technique of Ant Colony Optimization (ACO) is proposed to improve K-means clustering. The researchers have contributed to improve the cluster quality after grouping. Their proposed method has two phases. In the first phase, on the basis of statistical modes, initial centroids for K-mean clustering are selected. In the second phase, they improve the cluster quality by using ant refinement algorithm.

In [6], an alternate distance measure namely Max-min measure is proposed. By using the Max-min normalization, entire data is adjusted in limits [0,1] and after this normalization clustering is done.

In [7], A data clustering technique by using K-means algorithm is presented, which is based on the initial mean of the cluster, According to this algorithm, whole data space is divided into segments (k*k) and the frequency of data points in each segment is calculated. The segment having the highest frequency will have maximum probability of having centroid. If more than one

Corresponding Author: Azhar Rauf, Department Of Computer Science University Of Peshawar Peshawar, Pakistan.

consecutive segments having the same frequency then that segments are merged. After this, distances of data points and centroids are calculated. In same manner the process is continued.

Basic K-mean Clustering Algorithm: According to the basic K-mean clustering algorithm, clusters are fully dependent on the selection of the initial clusters centroids. K data elements are selected as initial centers; then distances of all data elements are calculated by Euclidean distance formula. Data elements having less distance to centroids are moved to the appropriate cluster. The process is continued until no more changes occur in clusters.

The following figure shows steps of the basic K-mean clustering algorithm [8].

Following are the algorithmic steps for basic K-mean algorithm [9].

INPUT:	Number of desired clusters *K* Data objects D= $\{d_1, d_2 \ldots d_n\}$
OUTPUT: A set of *K* clusters	

Steps:

- Randomly select k data objects from data set D as initial centers.
- Repeat;
- Calculate the distance between each data object d_i (1 <= i<=n) and all *k* clusters C $_j$(1 <= j<=k) and assign data object d_i to the nearest cluster.
- For each cluster j (1 <= j<=k), recalculate the cluster center.
- Until no change in the center of clusters.

Time complexity of K-mean Clustering is represented by O(*nkt*). Where *n* is the number of objects, *k* is the number of clusters and *t* is the number of iterations [10].

Proposed Algorithm: The process and algorithmic steps of proposed algorithm are given.

Clustering Process: In proposed algorithm, the input remains in the same order in which data items are entered. The whole process is divided into two phases.

Phase-I: In phase-I, the cluster size is fixed and the output of the first phase forms initial clusters. Here, the input array of elements is scanned and split up into sub-arrays, which represent the initial clusters.

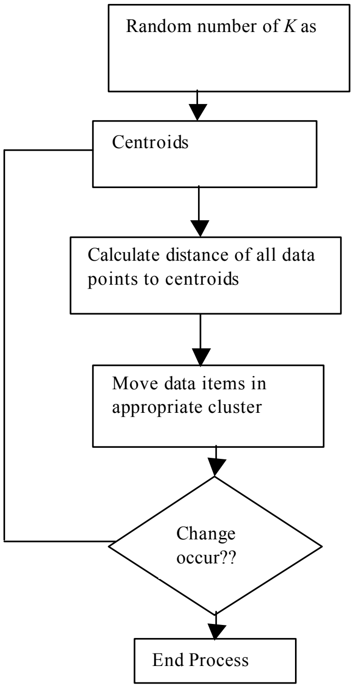
Random number of *K* as
Centroids
Calculate distance of all data points to centroids
Move data items in appropriate cluster
Change occur??
End Process

Fig. 1: K-Mean Clustering process

Phase-II: In phase-II, the cluster sizes vary and the output of this phase are the finalized clusters. Initial clusters are inputs for this phase. The centroids of these initial clusters are computed first, on the basis of which distance from other data elements are calculated. Furthermore the data elements having less or equal distance remains in the same cluster otherwise they are moved to appropriate clusters. The entire process continues until no changes in the clusters are detected.

Steps of Algorithm: Algorithm is divided into two Phases.

In Phase-I, we find the initial clusters, while in Phase-II, data elements are moved in appropriate clusters.

Phase-I:

Algorithm 2.1- *To find the initial clusters*

INPUT:	Array $\{a_1, a_2, a_3, \ldots\ldots a_n\}$ K //Number of Required clusters
OUTPUT: A set of Initial Clusters.	

Steps:

- Find the size of cluster Si (1= i = k) byFloor (n/k). Where n= number of data points D_p $(a_1, a_2, a_3, \ldots\ldots a_n)$ K= number of clusters.
- Create K number of Arrays A_k
- Move data points (D_p) from Input Array to A_k untill S_i =Floor (n/k).

- Continue Step 3 untill all D_p removedfrom input array
- Exit with having k initial clusters.

Phase-II:

Algorithm 2.2- *To find the final clusters*

INPUT:	A set of Initial Clusters.
OUTPUT:	A set of k Clusters.

Steps:

- Compute the Arithmetic Mean M of allinitial clusters C_I
- Set $1 \leq j \leq k$
- Compute the distance D of all D_p to M ofInitial Clusters C_j
- If D of D_p and M is less than or equal toother distances of M_i $(1 \leq i \leq k)$ then D_p stay in same cluster
- Else D_p having less D is assigned to Corresponding C_i
- For each cluster C_j $(1 \leq j \leq k)$, Recomputethe M and move D_p untill no change inclusters.

Experimental Work: Experimental work was designed to compare the performance of proposed K-mean algorithm.

Number of data elements selected was 1000. And for the sake of experiment, 8 numbers of clusters (k) were entered at run time. The process was repeated 10 times for different data sets generated by MATLAB. The proposed K-mean algorithm is efficient because of less number of iterations and improved cluster quality, as well as reduced elapsed time.

In Figure 2, Basic and proposed K-mean clustering algorithms are compared in terms of different data sets. For each run different data sets are generated by MATLAB and entered, to observe the number of iterations.

In Figure 3, Basic and proposed K-mean clustering algorithms are compared in terms of same data set. For each run same data set is entered, to observe that at each time numbers of iterations are different in basic K-mean clustering algorithm. The numbers of iterations are fixed in proposed K-mean clustering algorithm because initial centroids are not selected randomly.

Basic K-mean clustering algorithm gives different clusters, as well as clusters size differs in different runs. Table 1 shows different results for same data set as well as elapsed time.

Proposed K-mean clustering algorithm gives same clusters, as well as clusters size is same in different runs. Table 2 shows same number of iterations and cluster size.

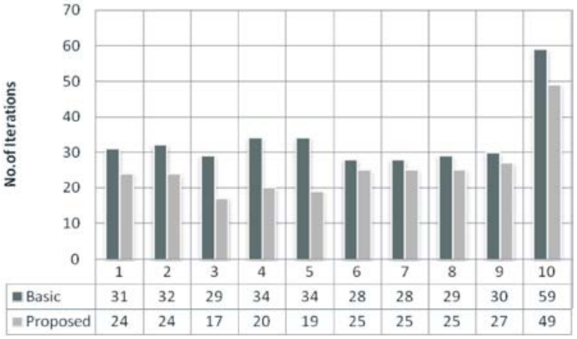

Fig. 2: For different data sets

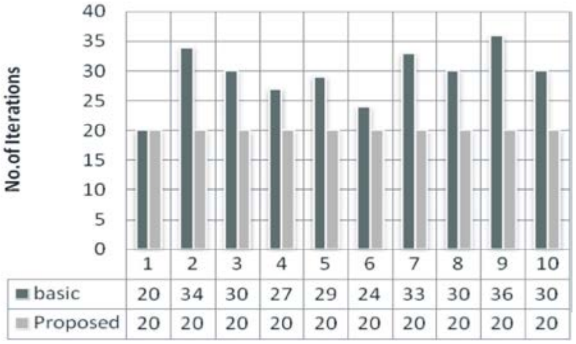

Fig. 3: For same data set

Comparison with Other K-Mean Clustering Algorithms: Comparison of proposed K-mean clustering algorithm with basic and other enhanced algorithm [2] is given.

Comparison of Basic and Proposed K-Mean Clustering Algorithm: Proposed K-mean algorithm is efficient from basic K-mean algorithm in terms of iterations, cluster quality as well as elapsed time.

As in basic K-mean algorithm, initial centroids are selected randomly from the input data, so clusters vary from one another, because of which the number of iterations and total elapsed time also changes in each run of the same data. In proposed K-mean algorithm initial centroids are calculated and as the data is same, it results in same calculations, so the number of iterations remains constant and elapsed time is also improved. This is the reason that proposed K-mean clustering algorithm is efficient from basic K-mean algorithm.

Comparison of Proposed K-Mean Clustering Algorithm with Other Enhanced K-Mean Clustering Algorithm: In [2], initial clusters are based on the searching mechanism. First two smallest elements are searched and those elements are then deleted from the input array and

Table 1: Basic K-mean algorithm

	Number Of Clusters									
Run	1	2	3	4	5	6	7	8	Iterations	Time elapsed in ms
1	55	21	57	21	269	204	343	30	30	71.14
2	659	163	35	21	31	60	17	14	27	71.13
3	39	116	190	567	1	50	13	24	26	67.62
4	39	655	68	36	49	57	48	48	31	94.05
5	155	230	139	81	53	37	30	275	34	74.11
6	106	26	586	90	68	54	31	39	30	70.48
7	655	87	48	57	36	33	35	49	35	121.97
8	57	555	120	36	97	35	81	19	25	120.54

Table 2: Proposed K-mean algorithm

	Number Of Clusters									
Run	1	2	3	4	5	6	7	8	Iterations	Time elapsed in ms
1	18	106	16	83	724	10	16	27	20	6.18
2	18	106	16	83	724	10	16	27	20	3.27
3	18	106	16	83	724	10	16	27	20	5.47
4	18	106	16	83	724	10	16	27	20	5.93
5	18	106	16	83	724	10	16	27	20	6.10
6	18	106	16	83	724	10	16	27	20	6.03
7	18	106	16	83	724	10	16	27	20	6.08
8	18	106	16	83	724	10	16	27	20	4.09

moved to the new sub arrays. The threshold value is set to fix the size of initial clusters and the process is continued to find initial clusters.

In proposed K-mean algorithm, there is no searching mechanism, so the running time of the proposed algorithm is improved as compared to the other techniques.

CONCLUSION

One of the partitional clustering algorithms is K-mean clustering algorithm which depends on initial clusters. In basic K-mean clustering, initial clusters are based on randomly selected centroids.

In this paper, an enhanced K-mean algorithm is introduced and compared with the basic K-mean algorithm. In enhanced K-mean clustering algorithm any type of integer data is used. The performance of basic K-mean clustering algorithm in terms of number of iterations and time complexity is improved. In future, this idea can be tested on text based clustering.

REFERENCES

1. Chen, Z.X., 2009. Shixiong, "K-means Clustering Algorithm with improved Initial Center," in Second International Workshop on Knowledge Discovery and Data Mining, Moscow.
2. Napoleon, D. and P.G. Lakshmi, 2010. "An Efficient K-means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points," in Trendz in Information Sciences and Computing (TISC), Chennai.
3. Jieming Zhou, J.G. and X. Chen, 2009. "An Enhancement of K-means Clustering Algorithm," in Business Intelligence and Financial Engineering, BIFE '09. International Conference on, Beijing.
4. Master, C.P. and X.G. Professor, 2011. "A Brief Study on Clustering Methods Based on the K-means algorithm," in 2011 International Conference on E-Business and E-Government (ICEE), Shanghai, China.
5. Mary, C.I. and S.V.K. Raja, 2009. "Refinement of clusters from k-means with ant colony optimization," Journal of Theoretical and Applied Information Technology, 9(2): 28-32.
6. Visalakshi, N.K. and J. Suguna, 2009. "K-means Clustering using Max-min Distance Measure," in Annual Meeting of the North American Fuzzy Information Processing Society, NAFIPS, Cincinnati, OH.
7. Singh, R.V. and M.P. Bhatia, 2011. "Data Clustering with Modified K-means Algorithm," in International Conference on Recent Trends in Information Technology (ICRTIT), Chennai, Tamil Nadu.

8. Wang, J. and X. Su, 2011. "An improved K-means clustering algorithm," in 3rd International Conference on Communication Software and Networks (ICCSN), Xi'an.
9. Na, S. and L. Xumin, 2010. "Research on K-means Clustering Algorithm An Improved K-means Clustering Algorithm," in Third International Symposium on Intelligent Information Technology and Security Informatics (IITSI), Jinggangshan.
10. Dong, J. and M. Qi, 2009. "K-means Optimization Algorithm for Solving Clustering Problem," in Second International Workshop on Knowledge Discovery and Data Mining (WKDD), Moscow.