International Journal For Technological Research In Engineering
Volume 1, Issue 9, May-2014 *ISSN (Online): 2347 - 4718*

PERFORMANCE ENHANCEMENT OF K-MEANS CLUSTERING ALGORITHM

Ms. Neha Shrimali[1], Prof. Avani Jadeja[2]
[1]M.E. Department of Computer Engineering
[2]Assistant Professor Department of Computer Engineering
[1,2]Hasmukh Goswami College of Engineering, Vahelal, Ahmedabad, Gujarat, India.

Abstract: Data clustering is one of the important unsupervised learning methods of data miming for grouping the data. In data clustering a set of objects are classified so that one member is similar to the other one. Data clustering techniques used in many fields like pattern recognition, machine learning, image analysis, information retrieval and in many more fields. The main important factors of clustering are to create good quality cluster and doing efficient task. K-means clustering is simple and most commonly used clustering method and produce clusters for many practical applications. But k-means clustering algorithm works efficiently only for small amount of data and becomes computationally complex for large amount of data. One more disadvantage of k-means clustering is that there is no mechanism for selecting initial cluster center which is selected randomly and result is depended on initial center points. It can contain dead unit problem. In this thesis, with the help of proposed method the problems of finding initial center points and assigning data items to appropriate clusters are solved. The proposed algorithm improves and enhances the execution speed of clustering the data set and also solves the dead unit problem. With the help of mathematical calculations the proposed algorithm decreases the complexity which we face in k-means clustering algorithm.

Index Terms: K-means clustering, K value, clustering analysis

I. INTRODUCTION

Clustering is the most important unsupervised-learning problem as every problem is of this type. The main purpose is finding a structure in a collection of unlabeled data. Totally, the clustering involves partitioning a given dataset into some groups of data whose members are similar in some way. The usability of cluster analysis has been used widely in data recovery, text and web mining, pattern recognition, image segmentation and software reverse engineering [1]. Clustering is one of the broad fields of data mining. In clustering data elements having which are having similarities are placed in respective groups. Clustering algorithms have main into two categories: Hierarchical clustering and partition clustering [3]. The main difference between partitioned and hierarchical clustering is that, in the first category, partitioned clustering algorithm data is partitioned into more than two subgroups in one step 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, 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 data 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.

II. K-MEANS CLUSTERING ALGORITHM

The simple definition of k-means clustering, as that is mentioned earlier, this is to classify data to groups of objects based on attributes/features into K number of groups. K is positive integer number. K-means is Prototype-based (center-based) clustering technique which is one of the algorithms that solve the well-known clustering problem. This creates a single one-level partitioning of the data objects.

K-means (KM) define a prototype in terms of a centroid, which is the mean of a group of points and is applied to dimensional continuous space. Another technique as prominent as K-means is K-medoid, which defines a prototype that is the most representative point for a group and can be applied to a wide range of data since it needs a proximity measure for a pair of objects. The difference with centroid is the medoid correspond to an actual data point. [3]

K-means algorithm is given below (Algorithm 1).

Algorithm: K-means Clustering Algorithm

Input: D = {d1, d2, d3,…,dn} //set of n data items
K // number of desired clusters

Output: A set of k clusters

Steps:

1. Choose k data items from D randomly as initial centroids;
2. Repeat Assign each item di to the cluster which has closest centroid; Calculate new mean for each cluster;

Advantages:

- K-mean value algorithm is a classic algorithm to resolve cluster problems; this algorithm is relatively simple and fast.
- For large data collection, this algorithm is relatively flexible and high efficient, because the Complexity is O (ntk). Among which, n is the times of iteration,

k is the number of cluster, t is the times of iteration. Usually, k"n and t"n. The algorithm usually ends with local optimum.

- It provides relatively good result for convex cluster.
- Because the limitation of the Euclidean distance. It can only process the numerical value, with good geometrical and statistic meaning.

Disadvantages:

The inherent prosperities of the K-means clustering algorithm to determine its limitations, specific performance is as follows:

- The K value is most important for K-means clustering algorithm. There is no applicable evidence for the decision of the value of K (number of cluster to generate), and sensitive to initial value, for different initial value, there may be different clusters generated.
- K-means clustering algorithm has a higher dependence of the initial cluster centers. If the initial cluster center is completely away from the cluster center of the data itself, the of iterations tends to infinity, but also makes it easier for the final clustering results into local optimization, resulting in incorrect clustering results.
- K-means clustering algorithm has a strong sensitivity to the noise data objects. If there is a certain amount of noise data in dataset, it will affect the final clustering results, leading to its error.
- K-means clustering algorithm for the discovery of clusters of arbitrary shape is most difficult.
- K-means clustering algorithm has main limitation on amount of data. In the iterative process, every time you need to adjust the cluster to which data object belongs.

III. RELATED WORK

Lloyd et al. [6], to minimize the mean squared distance from each data point to its nearest center. A popular heuristic for k-means clustering is Lloyd's algorithm. In this paper, presented a simple and efficient implementation of Lloyd's k-means clustering algorithm, which is called the filtering algorithm. This algorithm is easy to implement, requiring a kd-tree as the only major data structure. Efficiency is achieved because the data points do not vary throughout the computation hence, this data structure does not need to be recomputed at each stage. Kiri Wagstaff and Claire cardie et al. [7], first, develop a k-means variant that can incorporate background knowledge in the form of instance-level constraints, thus demonstrating that this approach is not limited to a single clustering algorithm. In particular, we present our modifications to the k-means algorithm and demonstrate its performance on six data sets. Second, while previous work with COBWEB was restricted to testing with random constraints, Demonstrate the power of this method applied to a significant real-world problem. The major modification is that, when updating cluster assignments, we ensure that none of the specified constraints are violated. They attempt to assign each point di to its closest cluster Cj. This will succeed unless a constraint would be violated. If there is another point d= that must be assigned to the same cluster as d, but that is already in some other cluster, or there is another point d=/ that cannot be grouped with d but is already in C, then d cannot be placed in C. We continue down the sorted list of clusters until find one that can be host d. Constraints are never broken; if a legal cluster cannot be found for d, the empty partition ({}) is returned. Researcher et al. [8], in proposed K-mean algorithm initial centroids are calculated and the data is same, which is results in same calculations, so the number of iterations remains constant and elapsed time that is also improved. This will be the reason that proposed K-mean clustering algorithm is efficient from basic K-mean algorithm. Now initial clusters which are based on the searching mechanism. First two smallest elements are searched and those elements are then deleted from the input array and moved to the new sub arrays. Then 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 given no searching mechanism, so the running time of the proposed algorithm is improved as compared to the other techniques. Chunfei zhang and zhiyi fang et al. [9], improvement for K-means cluster algorithm, offer the improved algorithm. In the last, the simulation results show the improved clustering algorithm is not only the clustering process is more stable, at the same time, improved clustering algorithm to reduce or even avoid the impact of the noise data in the data set object to ensure that the final clustering result is more accurate and effective.

The following factors improve the algorithm:

The distance between data points and the cluster center. The distance formula of data point xi and cluster center kj. The density parameter T. The number of data points which is contained by a scope defined as density parameter. The scope is a round which takes space point of not statistics xi as the center, as the radius. The greater the density of xi, the greater the value of the density parameter are. The core data points. If the -neighborhood of a data point contains at least PTS min number of data points, and then the data point called the core data point. The cluster center. Differences from the traditional clustering adjustment, the improved clustering algorithm add the weight of data point to the cluster center. Data points near the center of the cluster weights, on the contrary, the value of data points away from the cluster center is less weight. The Euclidean distance between data points and the cluster center. The distance between data point and the cluster center determine the cluster which data point belongs to. K.A. Abdul Nazir and M.P. Sabestian et al. [11], in this algorithm, here it will start to reduce the number of data points. Generate new data points by selecting randomly, each two data point and calculating their mean, this reduction continues until the number of data points is equal or greater than ten percentage of the total number of data points. The information of each child (generated point) and its parent (the two points whose the new point is generated by) which is kept for use later to

expand the group. After reaching the minimum possible reduction of original data points, run the basic K-means algorithm and group data points to N cluster. Then, the smallest group of data points after clustering, running the loop to optimizing K-means algorithm with this method in amount of ten thousand times. As we know, each points is the result of two data points (parent) in next step in each cluster we expand each point to its parents, which means , each cluster`s size become 2 times. Then again run the step of algorithm for generating new data points. Then continue, until reaching the original number of data points in the first level.

IV. PROPOSED K-MEANS CLUSTERING ALGORITHM

The traditional K-means and simple algorithm is a widely used clustering algorithm, with a wide range of applications. Clustering items from database has become an important in business analysis. In which K-means clustering algorithm is basic one. For getting advance algorithm, many enhancements have been done so far. This will include improving time efficiency; memory utilization and some other propose d methods which have guaranteed better performance than basic k-means clustering. Our focus will be in this thesis is to think about those attributes or factors which are related to data sets and have importance in finding initial clusters. These factors can give some appealing association clusters which may be useful for organization. Here in simple k-means clustering algorithm there are several disadvantages like dead unit problem, how to find proper initial center points etc are there. Because of these problems we need to improve the algorithm to overcome the problem. The new improved proposed algorithm helps to solve the dead unit and optimizes the proper selection of initial centroids of the clusters. The proposed algorithm minimizes the execution time and works efficiently for large data sets.

We divide our proposed algorithm into two parts:
Phase 1: How to select initial centroids?
Clusters (K) and data set is provided by the user. Value of y is calculated by providing value of K. Whole data space is divided into y*y blocks (y horizontally and y vertically). Frequency of data items in each block is calculated. Then K blocks are selected with the highest frequency. To find out the initial centroid, the mean of each selected block is calculated.

Phase 2: How to allocate data points to respective clusters?
For assigning the data items to the appropriate clusters, find distance matrix with the help of calculating distance between each and every cluster's centroid. For each centroid, take the minimum distance from the remaining centroid and make it half and denote it by Ti (threshold value). Then data items have been taken one by one and find the difference from ith centroid. If the generated difference is less than or equal to the threshold value Ti, then the data item will be assigned to the ith cluster, otherwise, then find out the distance from the other centroid. If data item which is not assigned to any of the cluster, then assign data item to the cluster which has nearest centroid. Repeat this process for each data item. Calculate the mean of the clusters and check the termination condition. If the condition is not satisfied then update and modify the centroid of clusters and repeat the Phase 2.

Proposed K-means Clustering Algorithm
Input: D = {d1, d2, d3,…,dn} //set of n data items
K // number of desired clusters
Output: A set of K clusters

Steps:
1. Enter input as the value of K and data set D.
2. Compute the distance between the data points of given data set.
3. Find closest pair from data set and add to new array Am.
4 Repeat step 3 until the number of data items in Am reaches to (n/k).
4. Calculate the value of y as integer y= (K*2.5)1/2 +1 // divide the data space into y*y means y horizontally and y vertically
3. For each dimension
Find the minimum and maximum value of data items from given data set.
Then find range of group (G) using equation G= (max-min)/x
Divide the data space in y group with width G.
5. Calculate data items frequency in each block.
6. Select K groups which is having highest frequency.
7. Calculate mean of selected group and find out initial cluster value.
8. Calculate distance between centroids and find out distance matrix |Ci, Cj|= {d (mi,mj)}// mi,mj denotes the means of i, j clusters respectively.
9. For each cluster take the minimum distance and make it half
Ti=0.5(min {|Ci,Cj|}) 1<=i<=K, 1<=j<=K
10. For each data item p1=1 to N
For each cluster q1=1 to K
Calculate distance between data item and centroid (dp,mq)
If distance (dp,mq)<=Tq then
Assign data itemdp to cluster Cq.
Break;
11. For each data item p2=1 to N
If data item dp will not assign to any cluster then
Assign data item dp to the cluster which has nearest centroid.
12. Check the termination condition of algorithm and take mean of each cluster seperately.
13. If satisfied then exit.
Else update the centroid of cluster.
Go to Step 7.

V. EXPERIMENTAL RESULT

In this experiment section, here we evaluate the performance of the proposed algorithm. I have implemented proposed algorithm in Visual Basic.NET. Mentioned experiments were run on 2.30 GHz Intel® core™ i3-2350 machine with 2 GB of RAM using Visual Studio 2008 on Windows xp. Data set Flame (shape) real life is taken for experiments. There is two

dimensional data which is having 2 clusters. The data set has 240 data points. The data set is used for testing the efficiency and accuracy of the proposed algorithm [15]. We have examined the time needed for execution by the traditional K-means algorithm and the new improved proposed K-means algorithm. Here we have also calculated the sum of squared error. The number of clusters is entered by the user. The data points in each cluster are showed by using different colors and the calculation of execution time is in milliseconds. Both of the algorithms are executed for the different number of clusters for five times and the average is taken. Different factors like Execution time, sum of squared error and number of iterations for clustering are shown in Table. From the Table, here it is seen that the new improved proposed algorithm gives better result than previous algorithm. Execution time and sum of squared error required by proposed K-means is less as compared to that of original k-means algorithm. In proposed algorithm we get better accuracy than older algorithm. The proposed algorithm gives efficient outputs than traditional algorithm. Dead unit problem has been solved. Quality, accuracy of clusters depends on how we choose initial center points. Here figure 1 and 2 shows the experimental result of the traditional k-means clustering algorithm and proposed k-means clustering algorithm which also explain the differences in accuracy, execution time in ms and sum of squared error.

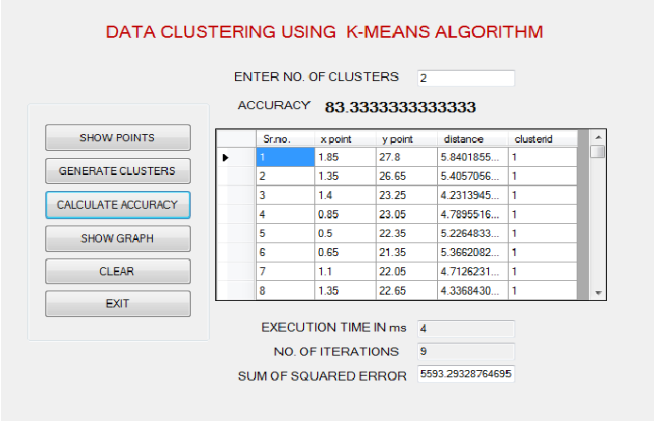

Figure 1: Experimental result of Traditional K-Means clustering algorithm

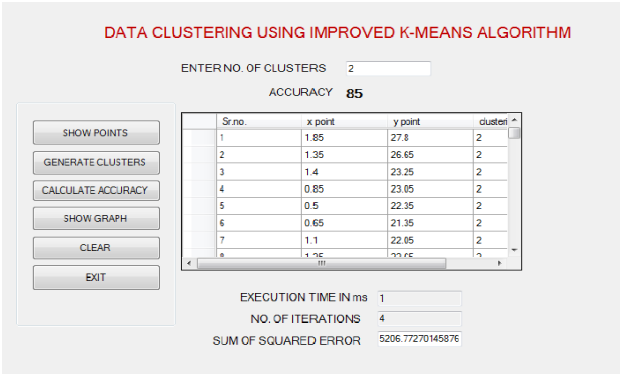

Figure 2: Experimental result of proposed K-Means clustering algorithm

TABLE 1COMPARISON OF K-MEANS ALGORITHM AND PROPOSED K-MEANS ALGORITHM

No. of clust-ers	K-means algorithm			Proposed K-means Algorithm		
	SSE	Exec. time in ms	No. of iterations	SSE	Exec. time in ms	No. of iterations
2	5574	4	8	5202	1.0	4
4	2089	11.4	14	2082	2.9	7
5	1820	9.4	9	1770	2.4	5
7	1295	19.2	13	1217	7.0	12
10	898	28.8	14	879	8.2	10
12	749	30	13	742	6.2	6
17	573	46.4	14	512	13.2	10

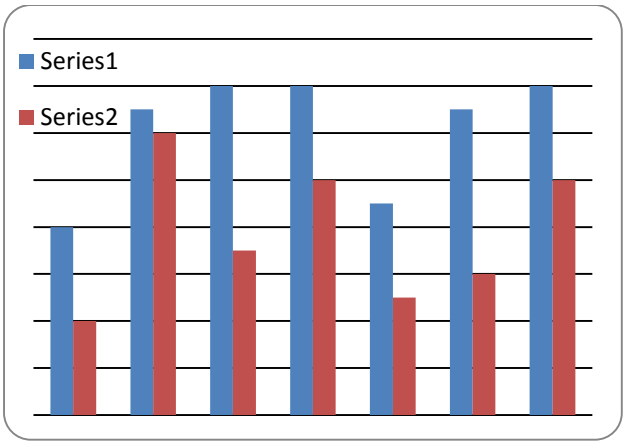

Figure 3: Comparison between K-means Clustering Algorithm and Proposed K-means Clustering Algorithm on based of no of iterations
Series1 k-means clustering value of no. of iterations
Series 2 proposed k-means clustering value of no. of iterations

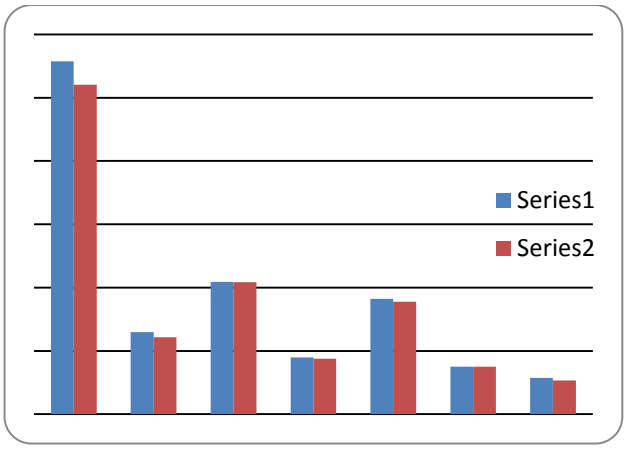

Figure 2: Comparison between K-means Clustering Algorithm and Proposed K-means Clustering Algorithm on based of sum of squared error (SSE)
Series1 k-means clustering value of SSE
Series 2 proposed k-means clustering value of SSE

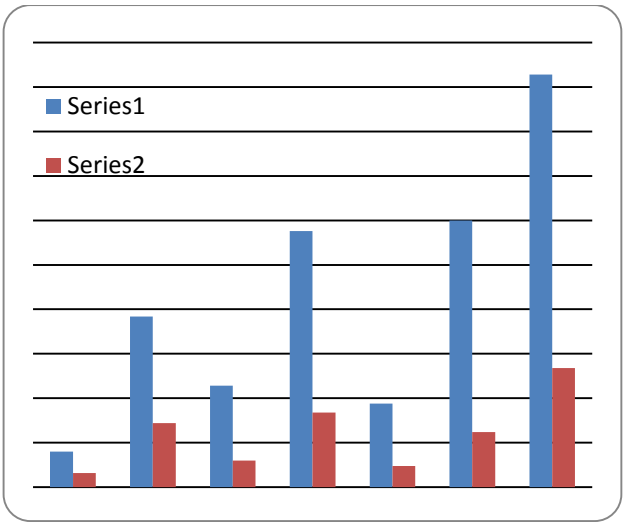

Figure 3: Comparison between K-means Clustering Algorithm and Proposed K-means Clustering Algorithm on based of execution time in ms
Series1: k-means clustering value of exec time in ms
Series 2: proposed k-means clustering value of exec time in ms

VI. CONCLUSION

The k-means algorithm is widely used for clustering data. But this algorithm does not always give good results, because the accuracy and efficiency of the resulting clusters depend on the selection of initial centroids. This paper presents an improved k-means algorithm which uses a systematic method to finding initial centroids and an efficient way for assigning data items to appropriate clusters. This algorithm does not have dead unit problem. This algorithm ensures the clustering of data in less time without sacrificing the accuracy of clusters. The results do not depend on the ordering of data and computational efforts are minimized by using the threshold value. Our experimental results show that the proposed algorithm produces better results than that of k-means algorithm.

VII. ACKNOWLEDGEMENT

In this world every work is done with the help of someone. We are really very grateful to find such personalities. Those personalities directly or indirectly involved and influence our work. We are very grateful to all those persons who have been our support and strength during every phase of dissertation. I am highly indebted to Asst. Prof. Avani Jadeja, for her valuable guidance and supervision regarding my dissertation topic as well as for providing necessary and important information regarding the thesis. We would also like to thank and appreciate our HOD Intrajeet Rajput to help us and provide us every help which is needed during the dissertation phase. I would love to express my gratitude towards my lovely parents for their endless co-operation and encouragement in completion of this thesis. My thanks and appreciations also go to with my dear friends who have helped me out with their abilities.

REFERENCES

[1] M. S. V. K. Pang-NingTan, "Data mining," in *Introduction to data mining*, Pearson International Edition, 2006, pp. 2-7.
[2] R. W. Stanforth, "Extending K-Means Clustering for Analysis of Quantitative Structure Activity Relationships (QSAR)," 2008.
[3] Han, Jiawei, Kamber, Micheline. (2000) Data Mining: Concepts and Techniques. Morgan Kaufmann
[4] M. S. V. K. Pang-NingTan, "Data mining," in *Introduction to data mining*, Pearson International Edition, 2006, pp. 487-496.
[5] M. S. V. K. Pang-NingTan, "Data mining," in *Introduction to data mining*, Pearson International Edition, 2006, pp. 496-508
[6] IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002.
[7] Proceedings of the Eighteenth International Conference on Machine Learning, 2001, p. 577-584.
[8] 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
[9] Journal of Information & Computational Science 10: 1 (2013) 193–199 Available at http://www.joics.com
[10] University of Agder, 2012 Faculty of Engineering and Science Department of ICT
[11] Malay K. Pakhira "Clustering Large Databases in Distributed Environment" IEEE International Advance Computing Conference (IACC 2009) held at Patiala, India from 6-7 March 2009.
[12] Madhuri A. Dalal, Naresh Kumar D. Harale, Umesh L. Kulkarni "An Iterative Improved K-means clustering" ACEEE Proc. Of Int. Conf. on Advances in Computer Engineering 2011, DOI: 02.ACE.2011.02.183, 2011, Pg.25-28.
[13] Jirong Gu, Jieming Zhou, Xianwei Chen "An Enhancement of K-means Clustering Algorithm" IEEE International conference on Business Intelligence and Financial Engineering DOI:10.1109/BIF.2009.204 Pg.237-240.
[14] Flame data set http://cs.joensuu.fi/sipu/datasets/flame.txt
[15] http://en.wikipedia.org/wiki/Cluster_analysis
[16] ISSN: 2319-7080 International Journal of computer science and communication engineering volume 2 issue 1.

