ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology
Volume 1, Issue 3, May2012

EFFICIENT K-MEANS CLUSTERING ALGORITHM USING RANKING METHOD IN DATA MINING

Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur

Lovely Professional University

Phagwara- Punjab

***Abstract*— Clustering is an essential task in Data Mining process which is used for the purpose to make groups or clusters of the given data set based on the similarity between them. K-Means clustering is a clustering method in which the given data set is divided into K number of clusters. This paper is intended to give the introduction about K-means clustering and its algorithm. The experimental results of K-means clustering and its performance in case of execution time is discussed here. But there are certain limitations in K-means clustering algorithm such as it takes more time for execution. So in order to reduce the execution, time we are using the Ranking Method. And also shown that how clustering is performed in less execution time as compared to the traditional method. This work makes an attempt at studying the feasibility of K-means clustering algorithm in data mining using the Ranking Method.**

***Index Terms*—Clustering, K-means Clustering, Ranking method.**

I. INTRODUCTION

In today's highly competitive business environment Clustering play an important role. As K- means Clustering is a method for making groups of the data set or the objects that are having similar properties. In this paper the II section includes the introduction part of Clustering and section III contain the related study or the literature survey about K-Means clustering algorithm. IV Section contains introduction about K-means Clustering algorithm and also examples of this algorithm. This Section also includes how in K-means algorithm the distance between the objects and mean is calculated and the methods of selecting initial points in K-means Clustering algorithm. Section V contains main steps in K-means clustering algorithm, then Section VI includes introduction about our proposed method Ranking Method and what is the need of Ranking method. Now section VII includes the introduction about the tool used for the implementation. Section VIII includes the results of both K-means clustering and Ranking Method and also shown the execution time taken by both algorithm during the clustering process. Section IX includes the research diagram which shows how in this research, the step by step implementation is done. Then the last Sections X includes the conclusion.

II. CLUSTERING

Mainly Clustering is the method which includes the grouping of similar type objects into one cluster and a cluster which includes the objects of data set is chosen in order to minimize some measure of dissimilarity. Clustering is a type of unsupervised learning not supervised learning like Classification. In clustering method, objects of the dataset are grouped into clusters, in such a way that groups are very different from each other and the objects in the same group or cluster are very similar to each other. Unlike Classification, in which predefined set of classes are presented, but in Clustering there are no predefined set of classes which means that resulting clusters are not known before the execution of clustering algorithm. In this these clusters are extracted from the dataset by grouping the objects in it.

Types of Clustering Algorithms

- Hierarchical Clustering Algorithm
- K-means Clustering Algorithm
- Density Based Clustering Algorithm
- Self-organization maps (SOM)
- EM clustering Algorithm

III. RELATED WORK

K. A. Abdul Nazeer, M. P. Sebastian (2009) "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm" In this paper an improvement in K-means clustering is shown. In this paper in the first phase of K-means clustering algorithm, the initial centroids are determined systematically so as to produce clusters with better accuracy [1]. The second phase makes use of an efficient way for assigning data points to clusters.

D. Napoleon, P. Ganga lakshmi (2010) "An Efficient K-Means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points" In this paper the uniform distribution of the data points is discussed that how this approach reduce the time complexity of the K-means clustering algorithm [2]. By using this approach the elapsed time is reduced and the cluster is of better quality. In this a very good method is used for finding the initial centroid. In this initially, the distance between each data points is computed.

Madhuri A. Dalal1 Nareshkumar D. Harale 2 Umesh L.Kulkarni (2011) "An Iterative Improved k-means Clustering" discuss an iterative approach which is beneficial in reducing the number of iterations from k-mean algorithm , so as to improve the execution time or by reducing the total number of distance calculations [3]. So Iterative improved K-means clustering produces good starting point for the K-means algorithm instead of selecting them randomly. And it will lead to a better cluster at the last result.

IV. K-MEANS CLUSTERING ALGORITHM

K-means clustering is a well known partitioning method. In this objects are classified as belonging to one of K-groups. The results of Partitioning method is a set of K clusters,
each object of data set belonging to one cluster. In each cluster there may be a centroid or a cluster representative.
In case where we consider real-valued data, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases.

Example: A cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset.

K-means is a data mining algorithm which performs clustering of the data samples. As mentioned previously, clustering means the division of a dataset into a number of groups such that similar items falls or belong to same groups. In order to cluster the database, K-means algorithm uses an iterative approach.

The input in this case is the number of desired clusters and the initial means and also produces final means as output. These mentioned initial and final means are the means of clusters. If in the algorithm requirement is to produce K clusters then there will be K initial means and final means also.

After termination of this clustering algorithm, each object of dataset becomes a member of one cluster. The cluster is determined by searching throughout the means for the purpose to find the cluster having nearest mean to the object. Cluster with shortest distanced mean is cluster to which examined object belongs. In case of K-means algorithm, it tries to group the data items in dataset into desired number of clusters. To perform this task well it makes some iteration until some converges criteria meets. After each iteration, recently calculated means are updated such that they become closer to the final means.

And at final, the algorithm converges and then stops performing iterations. Expected convergence of K-means algorithm is illustrated in the figure 2. In this example the algorithm converges in three iterations. The initial means which may be gathered randomly are represented by Blue points. Purple points are for the intermediate means. Finally, red points describe the final means which are also the results of K-means clustering.

A. Measurement of Distance Between Objects And Means

In order to measure the distance between objects and means different K-means clustering techniques can be used. Most popular distant metric that used is Euclidean Distance. Euclidean distance is represented as the square root of addition of squared differences between corresponding dimensions of object and the mean or cluster centroid.
Euclidean distance is the most common distance metric which is most commonly used when dealing with multi-dimensional data.

B. Selection of Initial Means

Basically the selection of initial means is up to the developer of clustering system what he/she wants. But this selection of initial means is independent of K-means clustering, because these initial means are inputs of K-means algorithm. In some cases, it is preferred to select initial means randomly from the given dataset while some others prefer to produce initial points randomly. As known that selection of initial means affects both the execution time of the algorithm and also the success of K-means algorithm.
Certain strategies are introduced to gather better results that are considering the initial means.

a) The simplest form of these strategies is that, in order to execute K-means algorithm with different sets of initial means considered and then select the best results. But this strategy is hardly feasible when dataset is large and especially for serial K-means.
b) Another strategy that is used to gather better clustering results is to use refine initial points method. If in case, it is possible to begin K-means with initial means which are closer to final means, then it is strongly possible case that the number of iterations that the clustering algorithm needs to converge will decrease. It also lessens the time required for conversion and also increases the accuracy of final means.

So there are different ways that are used for evaluating clustering results. This is upto the developers of clustering systems who need to decide on which criteria to use, in order to select the best results for clustering.

V. STEPS OF K-MEANS CLUSTERING ALGORITHM

K-Means Clustering algorithm is an idea, in which there is need to classify the given data set into K clusters, the value of K (Number of clusters) is defined by the user which is fixed. In this first the centroid of each cluster is selected for clustering and then according to the chosen centriod, the data points having minimum distance from the given cluster, is assigned to that particular cluster. Euclidean Distance is used for calculating the distance of data point from the particular centroid.

This algorithm consists of four steps:

1. Initialization
In this first step data set, number of clusters and the centroid that we defined for each cluster.

2. Classification
The distance is calculated for each data point from the centroid and the data point having minimum distance from the centriod of a cluster is assigned to that particular cluster.

3. Centroid Recalculation
Clusters generated previously, the centriod is again repeatly calculated means recalculation of the centriod.

4. Convergence Condition
Some convergence conditions are given as below:
4.1 Stopping when reaching a given or defined number of iterations.
4.2 Stopping when there is no exchange of data points between the clusters.
4.3 Stopping when a threshold value is achieved.

5. If all of the above conditions are not satisfied, then go to step 2 and the whole process repeat again, until the given conditions are not satisfied.

Main advantages:

1. K-means clustering is very Fast, robust and easily understandable. If the data set is well separated from each other data set, then it gives best results.
2. The clusters do not having overlapping character and are also non-hierarchical in nature.

Main disadvantages:

1. In this algorithm, complexity is more as compared to others .
2. Need of predefined cluster centres.
3. Handling any of empty Clusters: One more problem with K-means clustering is that empty clusters are generated during execution, if in case no data points are allocated to a cluster under consideration during the assignment phase.

VI. RANKING METHOD

With regards to Clustering, ranking operations are a natural way to estimate the likelihood of the occurance of data items or the objects. So we propose evaluating ranking overall design of database for student data in order to form the clusters. So Ranking function introduce new opportunities to optimize the results of K-means clustering algorithm.

A. Need of Ranking Method

Search of relevant records or similar data search is a most popular function of database to obtain knowledge. There are certain similar records that we want to fall in one category or form one cluster. That`s why, we need to rank the more relevance student marks by a ranking method and to improve search effectiveness.
In last, related answers will be returned for a given keyword query by the created index and better ranking strategy. So I have applied this Ranking method with K-means clustering method because this method is also having the property to find relevant records. So it is also helpful in creating clusters that are having similar properties between all data points within that cluster.

VII. TOOLS USED FOR K-MEANS CLUSTERING ALGORITHM IMPLEMENTATION

The tools that are used for the implementation of this improved k-means clustering algorithm incorporated with threshold value and also for Ranking Method is the Visual Studio 2008 using C#.

VIII. RESULTS

A. K-Means Clustering Results

In this case, clusters are created in K-means clustering algorithm, using the concept of threshold value. Graph that is given below shows the number of clusters that are made on the basis of the threshold value. On the basis of the centroid the clusters are formed.
This graph is made on the basis of the values x and y, which values are taken on the both axis of the graph. The Euclidean distance is calculated between both the centroid and the data points. Each cluster is shown with different color in order to distinguish between them.

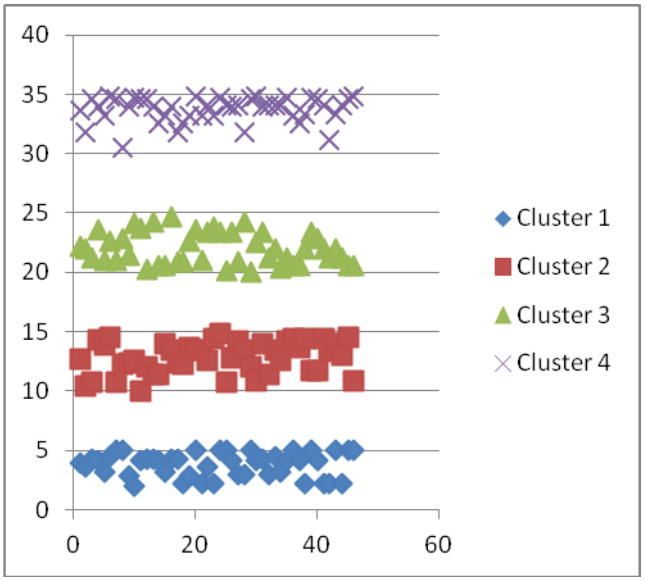

Fig 1 Graph Shows Clusters in K-means Clustering Algorithm

B. K-Means Clustering Results using Ranking Method

Graph below shows the accuracy and performance of ranking method. In this case, clusters are formed on the basis of rank that is calculated by applying ranking method. The execution time also reduces as compared to K-means clustering algorithm and it is used over large data set. As shown in graph, the clusters are created with accuracy and well differentiated from each- other.

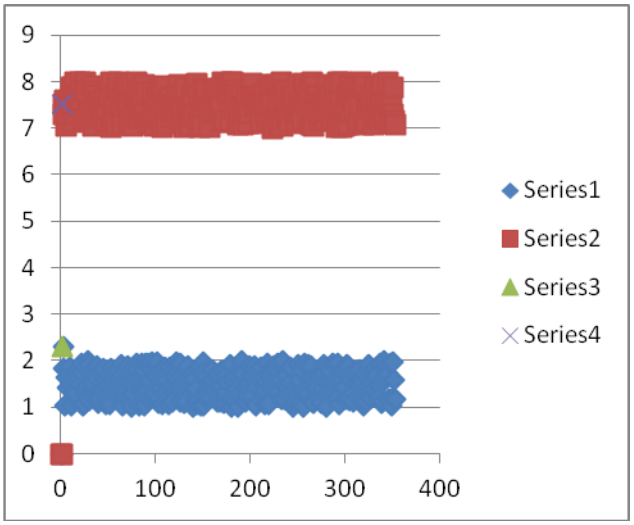

Fig 2 Graph of K-means clustering algorithm by applying Ranking Method

C. Execution Time Analysis For K- Means Clustering Algorithm

Execution time analysis for K-means clustering algorithm is done on the basis of the number of records that are considered for clustering and how much time is taken by this whole process.

As if the number of records are100 that are considered for clustering, then it takes execution time 132ms and so on for all records. So in this way using different number of records, the execution time differentiation is shown.

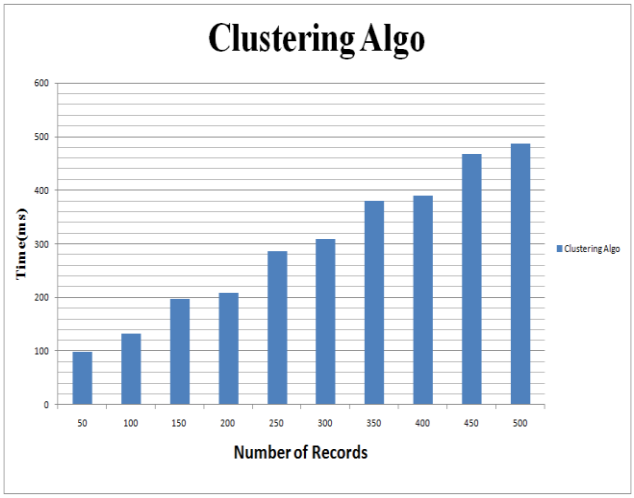

Fig 3 Execution time of K-means clustering algorithm.

In the table that also shows the number of records and the clustering execution time taken by K-means clustering algorithm is shown. As if the number of records are 50, the the execution time will be 98ms and so on. With the help of this type of tables we can easily calculate the performance.

Table 1. Execution time for K-means clustering

Records	Execution Time for Clustering Method
50	98
100	132
150	198
200	209
250	287
300	309
350	380
400	390
450	467
500	487

D. Execution Time Analysis for using Ranking Method

In this graph x-axis represents number of records and y-axis represents time that are included during ranking method when applied on clustering approach.

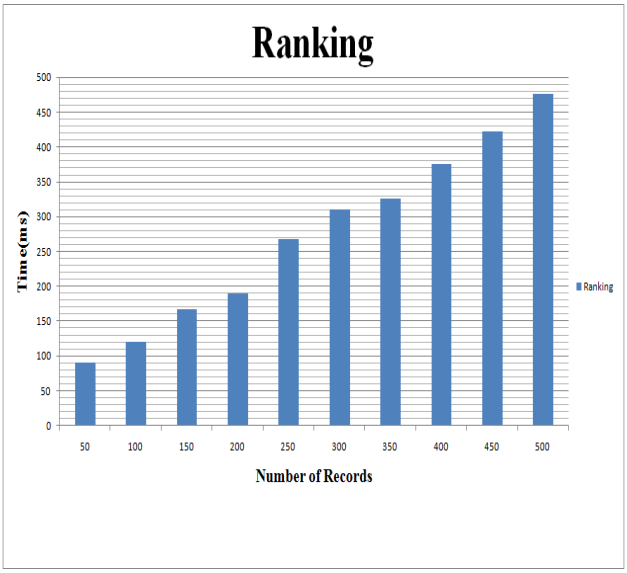

Fig 4. Execution time using ranking method for clustering.

The execution time for ranking method is less. So this is an appropriate approach applied for clustering method.
As in case of only K-means clustering for 50 records take the execution time that is 98ms, but in this case of Ranking method, for the purpose of executing same number of records, it takes 91ms.

And the main table that shows the execution time for the Ranking method for each particular records

Table 2 Execution time table for Ranking method

Records	Execution time for Ranking Method
50	91
100	121
150	167
200	190
250	267
300	310
350	326
400	376
450	422
500	476

IX. RESEARCH DIAGRAM

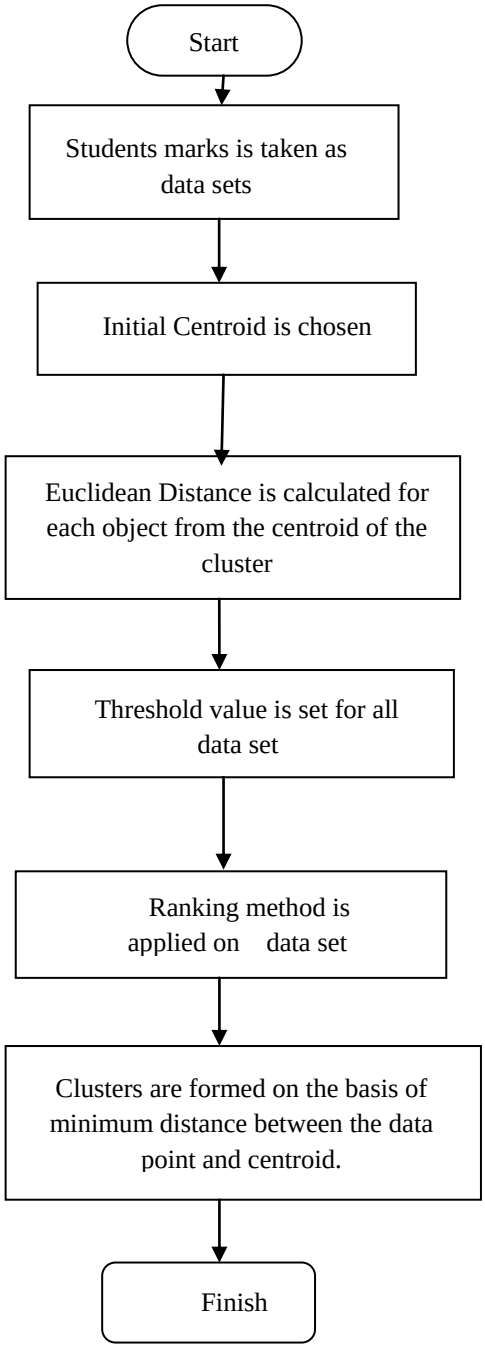

Fig 5 Research Diagram for K-means clustering algorithm.

X. CONCLUSION

The proposed work represents ranking based method that improved K-means clustering algorithm performance and accuracy. In this we have also done analysis of K-means clustering algorithm by applying two methods, one is the existing K-means clustering approach which is incorporated with some threshold value and second one is ranking method applied on K-means algorithm and also compared the performance of both the methods by using graphs. The experimental results demonstrated that the

proposed ranking based K-means algorithm produces better results than that of the existing k-means algorithm.

FUTURE WORK

In future, in case of clustering the marks of students from different-2 databases are considered by using the concept of Query redirection. By using the Query redirection approach we can easily cluster the large amount of data from distributed environment as from different databases. So if this approach is considered, then the performance of K-means clustering algorithm is improved for large samples of data set that are also distributed in nature.

ACKNOWLEDGMENT

Foremost, I would like to express my sincere gratitude to Ms. Jaspreet Kaur Sahiwal who gave her heart whelming full support in the completion of this research paper with her stimulating suggestions and encouragement to go ahead in all the time. She has always been a source of inspiration and confidence for me. She has beaconed light to me as a guide at all stages of preparation of my Research work.
I express my thanks from the core of my heart to my parents and friends for encouragement, cooperation and also help in challenging circumstances. At last I am very thankful to my GOD who has given me this golden opportunity to do M.Tech as well as to do research work.

REFERENCES

[1] K. A. Abdul Nazeer & M. P. Sebastian" Improving the Accuracy and Efficiency of the K-Means Clustering Algorithm" .Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, London, U.K, July 1 - 3, 2009.
[2] D. Napoleon & P. Ganga lakshmi, "An Efficient K-Means Clustering Algorithm for Reducing Time Complexity using Uniform Distribution Data Points", IEEE, 2010.
[3] Madhuri A. Dalal & Nareshkumar D. Harale "An Iterative Improved k-means Clustering" Proc. of Int. Conf. on Advances in Computer Engineering, 2011.
[4] Paul S. Bradley & Usama M. Fayyad, "Refining Initial Points for K-Means Clustering", 15th International Conference on Machine Learning, ICML98.
[5] Osama Abu Abbas "Comparison of various clustering algorithms" The International Arab Journal of Information Technology, Vol. 5, No. 3, July 2008.
[6] Jirong Gu & et.al, "An Enhancement of K-means Clustering Algorithm ", IEEE International Conference on Business Intelligence and Financial Engineering, 2009.
[7] Dost Muhammad Khan & Nawaz Mohamudally "A Multiagent System (MAS) for the Generation of Initial Centroids for k-means clustering Data Mining Algorithm Based on Actual Sample datapoints", IEEE, 2009.
[8] Malay K. Pakhira, "Clustering Large Databases in Distributed Environment ", IEEE 2009 WEE International Advance Computing Conference (IACC 2009) Patialae, India, 6-7 March 2009.
[9] Shi Na & et.al,"Research on k-means Clustering Algorithm", IEEE Third International Symposium on Intelligent Information Technology and Security Informatics, 2010.
[10] Jaehui Park, Sang-goo Lee "Probabilistic Ranking for Relational Databases based on Correlations" ACM 2010.

AUTHORS PROFILE

1)Navjot Kaur: Received her Bachelor degree in Information Technology and and currently doing M.Tech from Lovely Professional University, Jalandhar in department of Computer science and Information Technology. She has published her paper in ICTRCTA 2012 titled" Comparison of various Clustering Algorithms". Currently doing research on K-Means Clustering algorithm enhancement using Ranking Method in Data Mining . And research interests are in Database and Data Mining.

2) Jaspreet Kaur Sahiwal: Received her Bachelor Degree from Guru Nanak Dev University, Amritsar. She is an Assistant Professor at Lovely Professional University at department of Computer Science and Technology. She has completed her masters in Database. Her research interests are in Data Warehouse and Data Mining.

3) Navneet Kaur: Received her Bachelor Degree and currently doing her Masters (M.Tech) in Information Technology from Lovely Professional University. Currently doing research on Keyword Search in Distributed Database environment and also published one International publication ICTRCTA 2012 titled" Review study of Keyword Search over Relational Database using the Ranking method". Her research interests are in Database.