Implementation of Extended K-Medoids Algorithm to Increase Efficiency and Scalability using Large Datasets

Swarndeep Saket J.
Student (ME CSE IV)
Parul Institute of Engineering and Technology, Vadodara

Sharnil Pandya, PhD
Associate Professor
Parul Institute of Engineering and Technology, Vadodara

ABSTRACT
Clustering techniques are application tools to analyze stored data in various fields. Clustering is a process to partition meaningful data into useful clusters which can be understood easily and has analytical value. The K-Means and K-Medoid Algorithms in their existing structure carry certain weaknesses. For example in case of K-Means algorithm 'deformation' and 'deviations' may arise due to the misbehavior and disruption in the computing process. Similarly in case of K-Medoid Algorithm a lot of iteration is required which consumes huge amount of time and their by reduces the efficiency of clustering. In the present paper, we have proposed a new Modified K-Medoid Algorithm for improving efficiency and scalability for the study of large datasets. The extended K-Medoids Algorithm stand better in terms of execution time, quality of clusters, number of clusters and number of records than the comparative results of K-Means and K-Medoid Algorithm. Extended K-Medoid Algorithm is evaluated using sample real employee datasets and results are compared with K-Means and K-Medoids.

General Terms
Data Mining, Clustering

Keywords
Clustering, k-means, k-Medoids

1. INTRODUCTION
A cluster may be treated as a subset of objects which are similar in nature. It is a 'unsupervised learning process' to group together similar data samples, although, the criteria of classification might differ from each other [1]. To be more precise, a cluster might be defined as collection of data objects with numerous possibilities of classification. In simple words cluster analysis divides data into meaningful groups .Therefore clustering is also called data segmentation. Application of clustering have become a very successful tool for classification of documents, Clustering of web log data, recognition of meaningful patterns of data, undertaking spatial data analysis and even creating thematic maps in GIS and image processing. Various issues of clustering techniques are identification of distance measure, performance and scalability, number of clusters, lack of class labels, structure of database and choosing the initial cluster centers. The main objective of our present research work is to improve the efficiency of extended k-Medoids algorithm by achieving shorter execution time along with varying number of records and clusters. Technically, this may result into reduction of mean square error of creating clusters and an improvement in scalability in the analysis.

The clustering techniques differ in various ways and can be categorized as 'Partitioning techniques', 'Hierarchical techniques', 'Density-based techniques', 'Grid based methods' and 'model-based methods'. But the main focus of the research is on partitioning based Methods.

1.1 Procedure of Cluster Analysis
It is important to understand the basic process of clustering techniques. This has been simplified in the following flowchart. The chart shows how the process starts the given data samples and finally results into formation of clusters, their validation and finally interpretation of results.

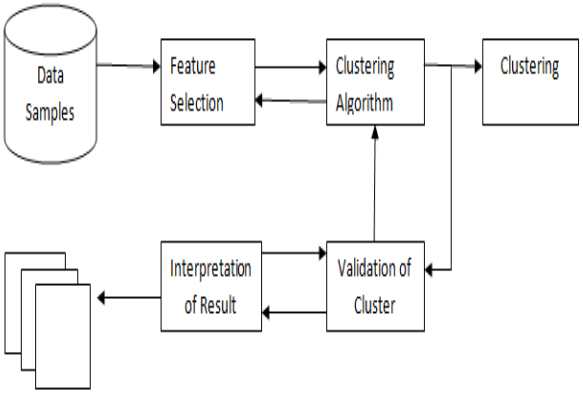

Figure 1: Procedure of Cluster Analysis

1) Feature Selection: In order to reduce the work load and simplify the design process feature selection is important. In this step we have to select the most relevant attributes. It utilizes some transformations to generate useful and novel features from the original ones.

2) Design of Clustering Algorithm: This second step generally starts with the appropriate selection of a 'corresponding proximity measure.' Patterns are grouped according to their resemblance with one another. All clustering algorithm are implicitly connected to define the proximity measure. Clustering algorithms have been developed solve different problems in specific fields. Therefore, it is important to design an appropriate clustering algorithm.

3) Validation of Clusters: Different approaches lead to different clusters and are used for same algorithm. The correctness of clustering algorithm results is verified using predefined criteria and techniques these assessments should be objective and have no preferences to any algorithm. It is used for finding pattern in noise.

Cluster validation is based on three categories: external indices, internal indices and relative criteria. These are defined on the basis of three types of clustering structures: partition clustering, hierarchical clustering and individual clusters. Tests for a situation where no clustering structure

exits in the data are also considered [2].External indices are based on some pre specified structure. It is used as a standard to validate the clustering solutions. Internal indices are no dependent on prior knowledge. Relative indices place the emphasis on the comparison of different clustering structures to provide a reference [3].

4) Result Interpretation: The main goal of clustering is to provide the users with meaningful insights into the original data. They can effectively solve the problems encountered.

2. LITERATURE SURVEY

2.1 K-Means

K-Mean is first developed by James Macqueen in 1967. A cluster is represented by its centroid, which is usually the mean of points within a cluster. "The objective function used for k-means is the sum of discrepancies between a point and its centroid expressed through appropriate distance "[4]. The k-means algorithm has the following important properties:

1. 1 .It is efficient in processing large datasets.
2. It often terminates at local optimum.
3. It works only on numerical values.
4. The clusters have convex shapes.

Algorithm steps [5]:

a) The technique requires arbitrary selection of choose k objects from D as the initial centers, where k is the number of clusters and D is the data set containing n objects.

b) Repeat the first step.

c) Reassign each object to the cluster to which object is most similar in nature.

d) Calculate the mean value of the objects for each cluster.

e) Until no change

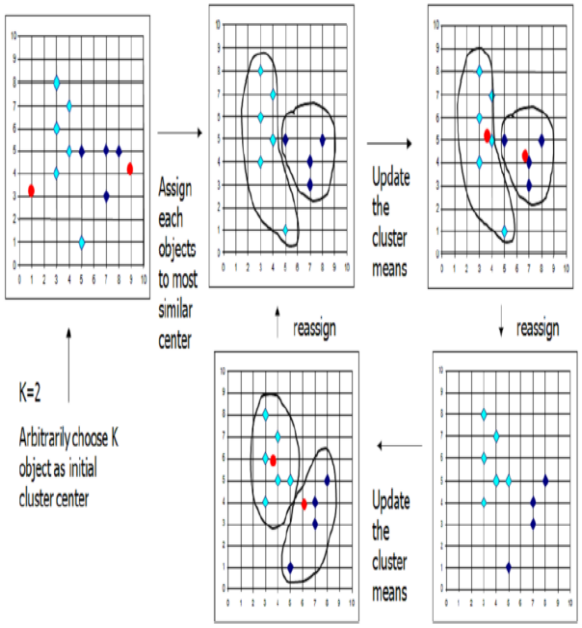

Figure 2: Working of K-Means Algorithm [7]

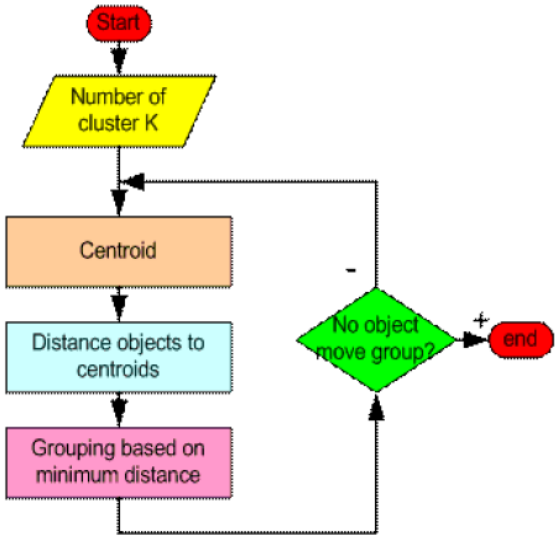

Figure 3: Flow chart of K-Means Algorithm [8]

2.2 K-Medoids Algorithm

K-Medoids Algorithm which is also known as Partition Around Medoids (PAM) is developed by Kaufman and Rousseuw in 1987. It is based on classical partitioning process of clustering. The algorithm selects k-medoid initially and then swaps the medoid object with non medoid thereby improving the quality of cluster. K-Medoids Algorithm is comparatively robust than K-Mean particularly in the context of outliers. It can be defined as that object of a cluster, instead of taking the mean value of the object in a cluster according to reference point.

K-Medoids Algorithm [6]:

Input: The number of clusters K and a database containing n objects**.**

Output: A set of k clusters

Method: The following steps are recommended by Tagaram Soni Madhulatha [9]

1. The algorithm begins with arbitrary selection of the K objects as mediod points out of n data points (n>K).
2. After selection of the K mediod points, associate each data object in the given data set to most similar mediod.
3. After selection, we randomly select non-mediod object O.
4. Calculate the total cost of swapping for non mediod object O
5. If S>0, then swap initial mediod with the new one.
6. Repeat steps until there is no change in the mediod.

3. PROPOSED ALGORITHM

3.1 Modified K-Medoids Algorithm:

The algorithm is being proposed by keeping in mind that K-medoid Algorithm is not efficient for large datasets. In extended K-Medoids algorithm, they are using 'Manthaan distance' to find out grouping of clusters and randomly select the medoid points. The goal of the proposed algorithm is to make efficient and scalable for large datasets using manthaan distance instead of Euclidean distance. It will improve the accuracy, efficiency and scalability for K-Medoid Algorithm.

Algorithm: The Extended K-Medoid algorithm for partitioning.

Method:

Step 1: First randomly select k of the n data points as the Medoids. Then calculate initial cluster centers.

Step 2: It will perform the grouping of clusters c1,c2…ck in a single group

Step 3: Assigning each remaining objects to nearest clusters using manthaan distance.

Step 4: Each non-medoid data point o, they will swap m and o and also compute the total cost of the configuration.

Step 5: After swapping process, select the configuration with the lowest cost.

Step 6: Repeat step 3 to 5 until there is no change in the medoid.

4. EXPERIMENTAL SETUP AND RESULTS

4.1 Experimental Setup

The implementation of algorithm was carried out in front end ASP.NET framework using C# language and back end is SQL server 2008. ASP.Net C# is a simple, modern and object-oriented programming language developed by Microsoft. This language is based on C and C++ programming language. This C# language was developed by Anders Hejlsberg and his team during the development of .NET Framework. It is a structural language and also produces efficient programs. This C# programming language is compiled on a various types of computer platforms. It is a component oriented language and also easy to understand and learn. C# language is a part of .NET framework.

4.2 Experimental Results

In the present Paper, the most representative algorithm K Medoids, K-Means and proposed algorithm were analyzed based on their basic approach for large data set.

Table 1: Number of clusters and execution time (in milliseconds)

Number of Clusters	Execution Time K-Mean Algorithm	Execution Time K-Mediod Algorithm
2	21348	20325
3	42170	40380
4	68510	67210
5	74360	73340

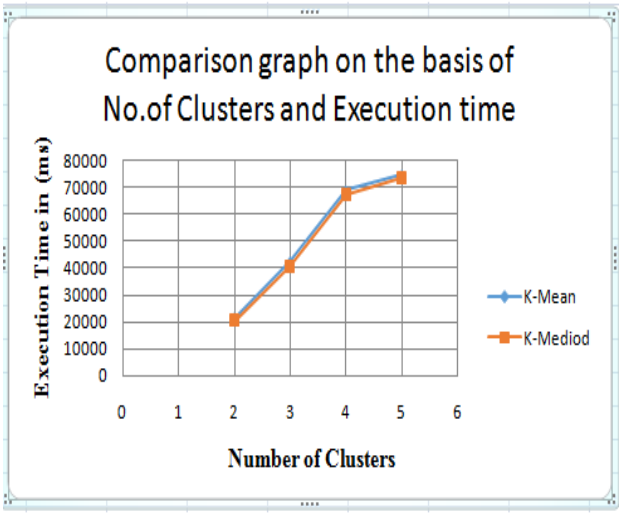

Figure 4: Comparison Graph

It is cleared from the table 1 and relevant graph figure 4 that irrespective of number of clusters the execution time taken by K-Mediod algorithm is generally less than that of K-Mean algorithm.

Table-2: Number of records and execution time (In milliseconds)

Number of Records	Execution Time K-Mean Algorithm	Execution Time K-Mediod Algorithm
300	94242	82232
400	112371	110301
500	138523	129202
600	164362	139561

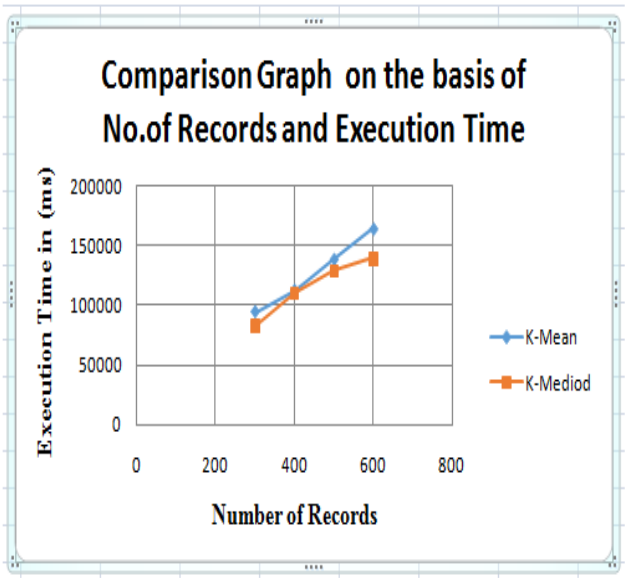

Figure 5: Comparison Graph on the basis of Number of records and execution time

Above table and figure shows the comparison between K-mean and K-Medoid Algorithms. As the graph shows that irrespective of number of records and the execution time

taken by K-Mediod algorithm is generally less than that of K-Mean algorithm. At the most number of records is increased than the execution time taken by K-Medoids is less than the K-Means Algorithm.

Table 3: Number of clusters and execution time (in milliseconds)

Number of Clusters	Execution Time K-Mediod Algorithm	Execution Time Extended K-Mediod Algorithm
2	20325	19311
3	40380	38595
4	67210	65916
5	73340	72321

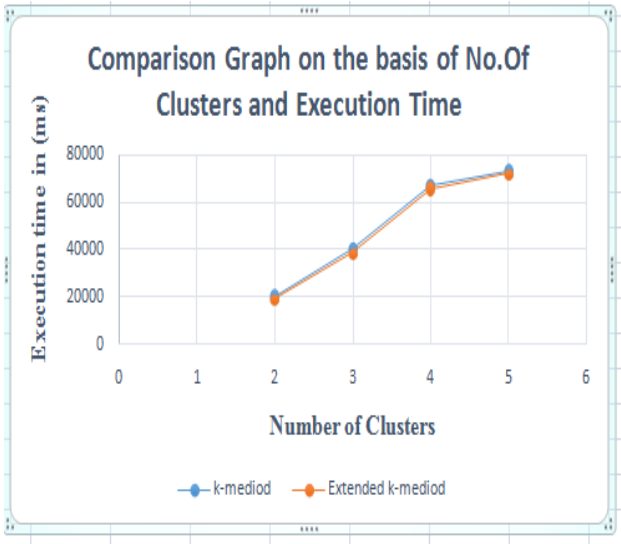

Figure 6: Comparison graph on the basis of number of clusters and execution time

It is cleared from the table 3 and relevant figure 6 that irrespective of number of clusters the execution time taken by Extended K-Mediod algorithm is generally less than that of K-Mediod algorithm.

Table 4.Number of records and execution time (In milliseconds)

Number of Records	Execution Time K-Mediod Algorithm	Execution Time Extended K-Mediod Algorithm
300	82232	70222
400	110301	108231
500	129202	119881
600	139561	114760

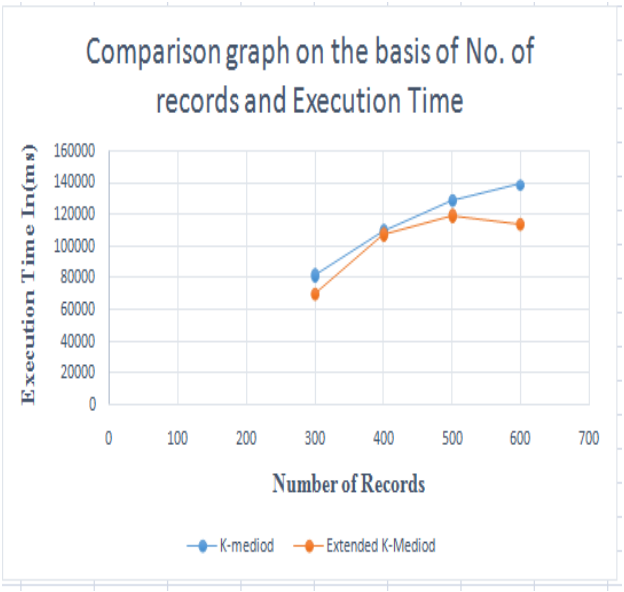

Figure 7: Comparison Graph

Above table and figure shows the comparison between K-Medoid Algorithm and Modified K-Medoids Algorithms. As the graph shows that irrespective of number of records the execution time taken by Extended K-Mediod algorithm is generally less than that of K-Mediod algorithm. The extended K-Mediod performs better than K-Mediod algorithm in most of the cases.

5. CONCLUSION AND FUTURE SCOPE

In the present paper, we have proposed a modified K-Mediod algorithm for improving efficiency and scalability for the study of large datasets. The result from number of clusters and records shows that the extended K-Mediod Algorithm has better performance in terms of execution time, quality of clusters .number of clusters and number of records than K-Means and K-Mediod Algorithms. Extended K-Mediod Algorithm is evaluated using sample real employee datasets and results are compared with K-Means and K-Medoids Algorithms.

In the future work, the comparison is made of the extended K-Medoids Algorithm with other algorithms in order to substantiate and bring improvement in the study. They can use the different techniques to further enhance the efficiency and scalability by reducing the execution time.

6. ACKNOWLEDGMENTS

The author would like to thank the entire college of Parul Institute of Engineering and Technology, in particular My Co-author Dr. Sharnil Pandya for the guidance during this research.

7. REFERENCES

[1] J. Kleinberg, (2002) “An impossibility theorem for clustering,” in Proc. Conf. Advances in Neural Information Processing Systems, 2002,vol. 15, pp. 463–470.

[2] A. Gordon, C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. Bock, and Y. Bada, Eds (1998) “Cluster validation,” in Data Science, Classification, and Related Methods. New York: Springer-Verlag, 1998, pp. 22–39

[3] P. Indira Priya and Dr. D.K. Ghosh (2013),” A Survey on Different Clustering Algorithms in Data Mining

Technique", (IJMER) International Journal of Modern Engineering Research, Jan-Feb 2013, Vol. No-3, Issue-1,pp. 267-274.

[4] Pradeep Rai and Shubha Singh (2010), "A Survey of Clustering Techniques", International Journal of Computer Applications (0975-8887) ,October 2010, Vol. 7-No. 12, pp. 1-5,

[5] Jiawei Han and Micheline Kamber,(2000) "Data Mining Techniques", Morgan Kaufmann Publishers, 2000.

[6] Shalini S Singh & N C Chauhan,(2011) ,"K-means v/s K-medoids: A Comparative Study", National Conference on Recent Trends in Engineering & Technology, 2011.

[7] .J. Han, M. Kamber, and M. Kauffman (2006), *Data Mining: Concepts and Techniques*, 2nd ed., 2006.

[8] Dr. Aishwarya Batra, "Analysis and Approach :K-Means and K-Medoids Data Mining Algorithms", 5th IEEE International Conference on Advanced Computing & Communications Technologies (ICACCT-2011), ISBN 81-87885-03-3.

[9] Tagaram Soni Madhulatha (2011),"Comparison between K-Means and K-Medoids Clustering Algorithms", Communications in Computer and Information Science, 2011.