An Enhanced K-Medoid Clustering Algorithm

Archna Kumari
Department of Computer Science &Engineering
RGPV, Indore, M.P.
kumara.archana14@gmail.com

Pramod S. Nair
Department of Computer Science &Engineering,
RGPV, Indore, M.P.
pramodsnair@yahoo.com

Sheetal Kumrawat
Department of Computer Science &Engineering,
RGPV, Indore, M.P.
sheetal2692@gmail.com

Abstract – Data mining is a technique of mining information from the raw data. It is a non trivial process of identifying valid and useful patterns in data. Some of the major Data Mining techniques used for analysis are Association, Classification and Clustering etc. Clustering is used to group homogenous kind of data, but it is different approach from classification process. In the classification process data is grouped on the predefined domains or subjects. A basic clustering technique represents a list of topics for each data and calculates the distance for how accurately a data fit into a group. The Cluster is helpful to get fascinating patterns and structures from an outsized set of knowledge. There are a lots of clustering algorithms that have been proposed and they can be divided as: partitional, grid, density, model and hierarchical based. This paper propose the new enhanced algorithm for k-medoid clustering algorithm which eliminates the deficiency of existing k-medoid algorithm. It first calculates the initial medoids 'k' as per needs of users and then gives relatively better cluster. It follows an organized way to generate initial medoid and applies an effective approach for allocation of data points into the clusters. It reduces the mean square error without sacrificing the execution time and memory use as compared to the existing k-medoid algorithm.

Keywords- *Data Mining, Clustering, Partitional Clustering, K-Medoid, Enhanced K-Medoid Algorithm.*

*****

I. Introduction

Data Mining is the extraction of information from large amounts of data to view the hidden knowledge and to facilitate its use in the real time applications. It is a significant process in which useful and valid patterns are identified in a data. It tends to work on the data and best techniques are developed to arrive at reliable conclusion and decisions massive amounts of data. There are many techniques used in Data mining process for data analysis such as clustering, association, classification etc [3, 11]. Clustering is among one of the most effective techniques for the analysis of data. Most of the application is utilizing the cluster analysis methods for categorizing data. Clustering is used to group same kind of data, but it is a different approach from classification process. In the classification process data is grouped on the predefined domains or subjects. Clustering plays an important role in the analysis space within the field of knowledge mining. The Cluster could be a method of partition a collection of knowledge in an exceedingly significant sub category known as clusters. It helps users to grasp the natural grouping of cluster from the info set. Its unattended classification which means it's no predefined categories. Information is sorted into clusters in such the simplest way that information of an equivalent cluster are similar and people in alternative teams are dissimilar. It aims to reduce intra-class similarity whereas to maximize interclass difference. The Cluster is helpful to get fascinating patterns and structures from an outsized set of knowledge. There are a lots of clustering algorithms that have been proposed and they can be divided as: partitional, grid, density, model and hierarchical based [9]. The process of Partition based clustering algorithm firstly generates the value of k. Here k is denoting the number of partitions that are wanted to be created. After that it applies an iterative replacement procedure that tries to get better results [7]. Some the important partitional clustering algorithms are k-Means [5,6,], and k-medoids [8]. In the K-mean algorithm, the centroid is defined as the mean of the cluster points. But in the K-medoid clustering algorithm, uses the object points as the representative point to make a cluster center. The disadvantage of K-Medoid does not generate the same result with each run, because the resulting clusters depend on the initial random assignments. As an attempt to solve the problem, the K-Medoid algorithm is used which is an unsupervised learning algorithm. Thus the propose work is dedicated to enhanced algorithm for k-medoid clustering algorithm which eliminates the deficiency of existing k-medoid algorithm. It first calculates the initial medoids 'k' as per needs of users and then gives relatively better cluster.

II. Literature Survey

The detailed study has been done to identify the drawbacks and possible solutions to resolve the limitations of existing system. Performance of repetitive cluster algorithms depends very much on the choice of cluster centre which is set at each step. In this section, the brief outlook of various algorithms is given. The issues related to these algorithms and also the many kinds of approaches used by authors to resolve those problems are discussed. [13] In this paper the projected algorithm calculated the space matrix one time and used it for locating novel medoids at every unvarying step. The experimental results are compared with the similar kind of existing algorithms. The output illustrates that the new algorithm takes an appreciably less time in calculation with equivalent performance compared to the PAM clustering algorithm. The advantage is in the calculation of the mean to finalize the centre points in each iteration of the cluster point. The proposed algorithm finds the centre points in a less time as compared with the k-mean algorithm. [12] The paper describes that the k- medoid algorithm is chosen as the object points as the initial medoids that ensures the clustering process better than the K-mean.The k-means and k-mediods algorithms are computationally expensive as consider the time parameter**.** The new approach in the proposed algorithm reduces the shortcomings of exiting k mean algorithm. First of all, it determines the first centroids for k as per needs of users and so offers higher, effective and stable cluster. It additionally takes a minimum time in execution as a result of it already eliminates the super numeracy distance computation by the victimization of the previous iteration. [11] The Paper explains clustering algorithm is that the methodology of organizing items the process of clustering involves creating groups of the data that is clustered or classes so that data within a cluster may have more resemblance when compared to each other, but are very different to the values in the different clusters. It has taken two clustering algorithms like k-means clustering algorithm and k-medoid clustering algorithm. But, the standard k-medoid algorithm rules do suffer from several shortcomings. The number of clusters must be defined before starting the process of K-medoid. The selection of k representative objects are not chosen properly in the initial stage the whole clustering process will move towards wrong direction and finally leads to clustering accuracy degradation. Third one is also responsive to the arrange of the input data points. The main drawback was eliminated from the exploitation of cluster validity index. [6] The Paper explains that the mean sq. error of clusters may be reduced by creating changes in creation of initial centroids. If the initial centroid is chosen systematically and properly than far better clusters are created. During this algorithm, first, the space between each and every data point is determined. Currently using that calculation, initial centroids are created by taking the points in the same set that has minimum distance to one another.

III. PROPOSED ALGORITHM

Distance Calculation:

Calculation of the path length between two clusters involves some or all elements of the clusters. To get the similarity or the relativity between the components of a population a common metrics of distance between two points is created. Euclidean metric is the most common distance measure which explains the space or distance between two points p = $(p_1,p_2,\ldots\ldots)$ and q = $(q_1,q_2, \ldots.)$

$$d = [\sum(p_i - q_i)^2]^{1/2} \text{.........Eq (1)}$$

K-Medoids algorithm:

The basic strategy of K-Medoids clustering algorithm is to find k clusters in n objects by first arbitrarily finding a representative object (the Medoids) which is the most centrally located object in a cluster, for each cluster. Each remaining object is clustered with the Medoid to which it is the most similar. K-Medoids method uses representative objects as reference points instead of taking the mean value of the objects in each cluster. The algorithm takes the input parameter k, the number of clusters to be partitioned among a set of n objects. A typical k-medoids algorithm for partitioning based on Medoid or central objects is as follows [12]:

Input:
K: The number of clusters
D: A data set containing n objects

Output: A set of k clusters that minimizes the sum of the dissimilarities of all the objects to their nearest medoid.
Method: Arbitrarily choose k objects in D as the initial representative objects;

```
Repeat:
 1. Assign each remaining object to the cluster with the nearest medoid;
 2. Randomly select a non medoid object Orandom;
 3. Compute the total points S of swap point Oj with Oramdom
 4. If S < 0 then swap Oj with Orandom to form the new set of k medoid
 5. Select the configuration with the lowest cost
 6. Repeat steps 2 to 5 until there is no change in the medoid
```

Proposed algorithm:

Proposed Approach of classical partition is primarily based on the partitional clustering rule. The process of declaring the k for number of clusters will remain intact in the proposed method also. The main difference in the proposed approach is in the way it chooses the starting k object points.

The enhanced K-Medoids algorithm, initializes the cluster medoids by selecting the initial medoid points. The implementation result shows it is better performing in perspective of the time taken for the clustering process of the entire objects. The memory usage in this approach is also comparatively less as per the implementation result.

Initially the distance between the origin and the object points are calculated. Then the object points are sorted in the ascending order. The ordered objects now divided into k clusters. The medoids of k clusters will be calculated as using the distance measures.

Algorithm:

Input:
A = $\{a_1, a_2,\ldots\ldots,a_n\}$ // set of n items
Let k be the number of desired clusters
Output:
Let k be set of clusters

```
Steps:
        Let O be the origin point with attribute values 0.
        Calculate distance between each data point and origin.
        Sort the data points in ascending order of the value obtained in step 2.
        Partition the sorted data points into k equal sets.
        Assign first elements of those partitions to be initial medoids. Assign them as medoids.
```

```
6. Compute the distance of each data-point $a_i$ ($1<=i<=n$) to all the medoids $m_j$ ($1<=j<=k$) as $d(a_i, m_j)$
7. Repeat
8. Find the closest medoid $m_j$ for data points in ai and assign $a_i$ to cluster j.
9.  Set ClusterId[i]=j. // j:Id of the closest cluster.
10. Set NearestDist[i] = $d(a_i, a_j)$.
11. For each cluster j (1 <= j <= k), again calculate medoids.
12. For each $a_i$,
    12.1 calculate the distance with current nearest medoid.
    12.2 The data point stays in the same cluster if the current nearest distance is less or equal.
  Else
    12.2.1 For every medoid mj (1<=j<=k) compute the distance d ($a_i$, $m_j$).
  End for;
    Until the convergence criteria is met.
```

IV. RESULTS ANALYSIS

The given section includes the performance analysis of the implemented algorithms for the k-medoid. The performance of algorithms are evaluated and compared in this chapter.

A) Mean Square Error

Figure 1 and table 1 show Mean square error comparison results, which has been carried out on the same size of Iris datasets. The mean square error is calculated the difference between the instances of each cluster and their cluster center. Smaller values indicate a cluster of higher quality. The values of the graph are represented using table 1 where the amount of Mean Square Error of the proposed algorithm is given in the last column and the first column contain existing k-medoid algorithm. In the similar ways the given graph as given in figure 1 contains the comparative mean square error of all the algorithms.

In this figure red color shows the proposed algorithms performance and the green color shows the result of existing k-medoid algorithm. For demonstrating the performance of the system Y axis contains the Mean Square Error and X axis contains the number of cluster.

TABLE 1: MEAN SQUARE ERROR PERFORMANCE

Number of Clusters	K-Medoid (MSE)	Proposed K-Medoid Algorithm(MSE)
5	77.8	54.98
10	36.62	34.52
15	27.75	25.14
20	23.99	20.96
25	23.96	18.15

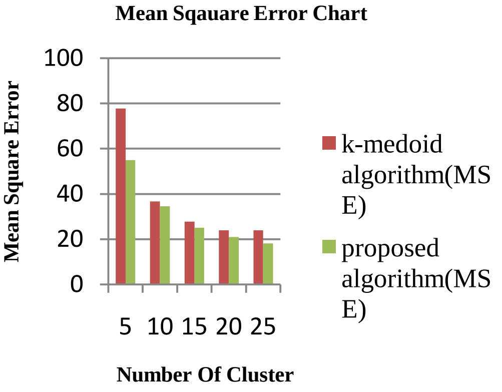

Figure 1: Mean Square Error Comparison Chart

B) Execution Time

The comparative time utilization of the proposed and existing k-medoid algorithms is given using figure 2 and table 2. In this graph the horizontal axis contains the number of clusters and the vertical axis contains execution time in terms of milliseconds. In this diagram the red colour demonstrates the output of the existing k-medoid clustering algorithm and the green colour demonstrates the output of proposed k-medoid algorithm. According to the comparative results analysis the performance of the proposed technique shows the less time consuming as compared to the original k-medoid algorithm. The new algorithm gives better performance without much increment in execution time. The comparison is done on iris dataset.

TABLE 2: COMPARISONS BETWEEN ALGORITHM WITH NUMBER OF CLUSTER AND EXECUTION TIME OF IRIS DATASET

Number of Cluster	K-Medoid (Time in ms)	Proposed K- Medoid Algorithm(Time in ms)
5	33.9879	16.8096
10	41.3256	33.2097
15	106.4943	52.2604
20	164.0374	69.0788
25	209.2598	72.3591

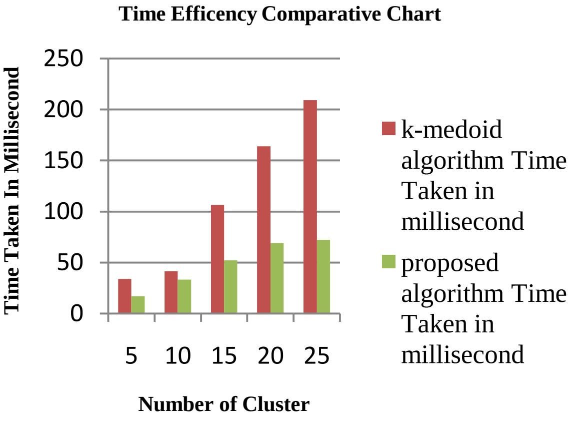

Figure 2: Graphs Represent Number of Clustering and Execution Time Comparison for Iris Dataset

C) Memory Used

Memory use of the system also termed as the space complexity in terms of algorithm performance. That can be calculated using the following formula:

Memory consumption=

Total memory - Free memory

The amount of memory consumption depends on the amount of data residing in the main memory, therefore, that affects the computational cost of an algorithm execution. The comparison between all the algorithms that is existing k-medoid and proposed k-medoid algorithm is done on the basis of memory required for the execution of the algorithm. The experimental results of K-Medoid and proposed clustering algorithm are shown in the figure 3. In the graph the memory use of the proposed k-medoid algorithm and the existing k-medoid algorithm is shown. The experiments suggests For reporting the performance of figure 3 Y axis contains the use of memory consumption during experimentations and the X axis shows the number of clusters. According to the results the proposed algorithm demonstrates similar behaviour even if the size of clusters increases.

TABLE 3: COMPARISON BETWEEN PROPOSED AND EXISTING SYSTEM BASED ON NUMBER OF CLUSTERS AND MEMORY REQUIREMENT

Number of Cluster	**K-Medoid (Memory in bytes)**	**Proposed K-Medoid (Memory in bytes)**
5	204596	204652
10	204544	201960
15	204576	212792
20	212768	210132
25	212788	220984

Figure 3: Graphs Represent Number of Clusters and memory required Comparison for Iris Dataset

The comparison between the algorithms that is k-medoid and proposed k- medoid algorithm is done on the Iris data set which contains 150 data points with five attributes.

Table 4 describe the performance summary of both algorithms. According to obtained result the proposed algorithm is able to clustering the data points. Therefore the proposed algorithm based on partitional clustering algorithm is adoptable and efficient

TABLE 4: PERFORMANCE PARAMETERS

S.No.	Parameters	K- Medoid Algorithm	Proposed K-Medoid Algorithm
1.	Mean Square Error	High	Low
2.	Execution Time	Low	High
3.	Memory Use	Comparatively High	Comparatively Low

V. CONCLUSION AND FUTURE WORKS

An enhanced k-medoid algorithm which is new approach of classical partition based clustering algorithm improves the execution time of k-medoid algorithm. The results conclude that, the proposed implementation of the k-medoid algorithm is better performed as compare with the K-medoid. From experiment it is observed that the proposed algorithm gives better performance in all parameters.

Proposed Approach of classical partition is primarily based on the partitional clustering rule. The process of declaring the k for number of clusters will remain intact in the proposed method also. The main difference in the proposed approach is in the way it chooses the starting k object points.

work enhances the system's performance with the use of other types of attributes in data set. Proposed system's performance was evaluated and it was concluded that it was a better option to calculate the distance and arrange these in ascending order rather than calculating the distance form origin.

REFERENCES

[1] L. Kaufman and P. J. Rousseau, “Finding Groups in Data: an Introduction to Cluster Analysis”, John Wiley & Sons, 1990.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn, “ Data Clustering: A review”. ACM Computing Surveys, Vol. 31 No. 3, pp.264– 323, 1999.

[3] J. Han and M. Kamber. “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, August 2000.

[4] Rui Xu and Donlad Wunsch, “Survey of Clustering Algorithm”, IEEE Transactions on Neural Networks,Vol. 16, No. 3, May 2005.

[5] Sanjay Garg, Ramesh and Chandra Jain, “Variation of K-Mean Algorithm: A study for High Dimensional Large Data Sets”, Information Technology Journal,Vol. 5, No. 6, pp.1132 – 1135, 2006.

[6] K. A. Abdul Nazeer and M. P. Sebastian “Improving the Accuracy and Efficiency of the K-Means Clustering Algorithm” Proceedings of the World Congress on Engineering , Vol.1, pp.1-3, July 2009.

[7] T. Velmurugan and T. Santhanam, “A Survey of Partition Based Clustering Algorithms in Data Mining: An Experimental Approach”, Journal, Vol. 10, No. 3, pp. 478- 484, 2011

[8] Shalini S Singh and NC Chauhan, “K- means v/s K-medoids: A Comparative Study”, National Conference on Recent Trends in Engineering & Technology, 2011.

[9] Rui Xu and Donlad Wunsch, “Survey of Clustering Algorithm”, IEEE Transactions on Neural Networks, Vol. 16, No. 3, May 2005.

[10] M. S. Chen, J. Han and P. S. Yu., “Data Mining: An Overview from a Database Perspective”, IEEE Transactions on Knowledge and Data Engineering, Vol. 8, pp. 866-883, 1998

[11] Bharat Pardeshi and Durga Toshniwal, "Improved K-Medoid Clustering Based On Cluster Validity Index and Object Density", IEEE, 2010.

[12] Abhishek Patel and Purnima Singh, "New Approach For K-mean and K-Medoids Algorithm", International Journal of Computer Applications Technology and Research, 2013

[13] H.S. Park, and C.H. Jun, "A Simple and Fast Algorithm for K-Medoids Clustering", Department of Industerial and Management Engineering POSTECH, 2009.

[14] R. Fisher, “UCI Machine Learning Repository”, 1936. https://archive.ics.uci.edu/ml/datasets/Iris