ANALYSIS AND IMPLEMENTATION OF MODIFIED K-MEDOIDS ALGORITHM TO INCREASE SCALABILITY AND EFFICIENCY FOR LARGE DATASET

Gopi Gandhi[1], Rohit Srivastava[2]

[1]Student of ME, Computer Science and Engineering, PIET, Gujarat, India
[2]Assistant Professor, Computer Science and Engineering, PIET, Gujarat, India

Abstract

Clustering plays a vital role in research area in the field of data mining. Clustering is a process of partitioning a set of data in a meaningful sub classes called clusters. It helps users to understand the natural grouping of cluster from the data set. It is unsupervised classification that means it has no predefined classes. Applications of cluster analysis are Economic Science, Document classification, Pattern Recognition, Image Processing, text mining. Hence, in this study some algorithms are presented which can be used according to one's requirement. K-means is the most popular algorithm used for the purpose of data segmentation. K-means is not very effective in many cases. Also it is not even applicable for data segmentation in some specific kinds of matrices like Absolute Pearson. Whereas K-Medoids is considered flexible than k-means and also carry compatibility to work with almost every type of data matrix. The medoid computed using k-Medoids algorithm is roughly comparable to the median. After checking the literature on median, we have found a number of advantages of median over arithmetic mean. In this paper, we have used a modified version of k-medoids algorithm for the large data sets. Proposed k-medoids algorithm has been modified to perform faster than k-means because speed is the major cause behind the k-medoids unpopularity as compared to k-means. Our experimental results have shown that improved k-medoid performed better than k-means and k-medoid in terms of cluster quality and elapsed time

Keywords: *Clustering, k-means, k-medoids, Clarans*

--***--

1. INTRODUCTION

Data Mining is a process of identifying valid, useful, novel, understandable pattern in the data. Data Mining is concern with solving problem by analyzing existing data. Clustering is a method of data explorations, a technique of finding patterns in the data that of our interest. Clustering is a form of unsupervised learning that means we don't know in advance how data should be group together [1].

Various Techniques for clustering are as follows [2]

1. Partitioning Method
2. Hierarchical Method
3. Grid- based Method
4. Density-based Method
5. Model-based Method

Among all these methods, this paper is aimed to explore partitioning based clustering methods which are k-means and k-medoids. These methods are discussed along with their algorithms, strength and limitations.

2. PARTITIONING TECHNIQUES

Partitioning techniques divides the object in multiple partitions where single partition describes cluster. The objects with in single clusters are of similar characteristics where the objects of different cluster have dissimilar characteristics in terms of dataset attributes. K-mean and K-medoids are partitioning algorithm [3].

2.1 K-Mean

K-mean algorithm is one of the centroid based technique. It takes input parameter k and partition a set of n object from k clusters. The similarity between clusters is measured in regards to the mean value of the object. The random selection of k object is first step of algorithm which represents cluster mean or center. By comparing most similarity other objects are assigning to the cluster.

Algorithm [4]: The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- K:the number of clusters
- D:a data set containing n object

Output:

- A set of k clusters

Method:

(a) Arbitrarily choose k objects from D as the initial cluster centers.

```
(b) Repeat
    i. Reassign each object to the cluster to which the object is the most similar, Based on the mean value of the objects in the cluster.
    ii. update the cluster means ,i.e., calculate the mean value of the objects for each cluster.
(c) Until no change.
```

2.2 K-Medoid

The k-means method is based on the centroid techniques to represent the cluster and it is sensitive to outliers. This means, a data object with an extremely large value may disrupt the distribution of data[6].

To overcome the problem we used K-medoids method which is based on representative object techniques. Medoid is replaced with centroid to represent the cluster. Medoid is the most centrally located data object in a cluster.

Here, k data objects are selected randomly as medoids to represent k cluster and remaining all data objects are placed in a cluster having medoid nearest (or most similar) to that data object. After processing all data objects, new medoid is determined which can represent cluster in a better way and the entire process is repeated. Again all data objects are bound to the clusters based on the new medoids. In each iteration, medoids change their location step by step.

This process is continued until no any medoid move. As a result, k clusters are found representing a set of
n data objects [3]. An algorithm for this method is given below.

Algorithm [3]: PAM, a k-medoids algorithm for partitioning based on medoid or central objects.

Input:

- K: the number of clusters,
- D: a data set containing n objects.

Output:

- A set of k clusters.

Method:

```
Step 1: Initialize k by random selection from n data points of data matrix X.
Step 2: Calculate distance between data points n and medoid k.
    a. Element by Element binary operations to compute X-mean(X,2) ,Where X denotes data matrix and return A.
    b. Compute A^2 and return S and then convert sparse matrix (S) to full matrix (S).
    c. Compute -2*A(T)*A and return D and then convert sparse matrix (D) to full matrix (D) .Here T means transpose matrix of A
    d. Add D with S and return D.
    e. Add D with S(T) and return D. Here T means transpose matrix of S.
    f. Select minimum from D.
Step 3: For each data point o, swap it with medoid k, and compute the total cost C.
Step 4: Compute minimum cost c from total cost C.
Step 5: Add c to Final Matrix.
Step 6: Repeat step 2 to 5 for all data points.
```

2.3 Modified K-Medoids

Proposed k-medoids algorithm has been modified to perform faster than k-means because speed is the major cause behind the k-medoids unpopularity as compared to k-means. Our experimental results have shown that improved k-medoid performed better than k-means and k-medoid in terms of cluster quality and elapsed time.

Algorithm: The Modified k-medoid algorithm for partitioning,

Input:

- K number of cluster
- D data set containing n object.

Output:

- Set of K clusters

Method:

```
Step 1: Initialize k by random selection from n data points of data matrix X
Step 2: Calculate distance between data points n and medoid k
    a. Element by element binary operation:
        I. v+v(T) Centered dot (also called dot product) and return Y (T stands for transpose matrix of v)
        II. Compute 2*(X(T)*X) where X is data matrix, X(T) is transpose of data matrix X and return Z
        III. Compute Y-Z and return D
    b. Select minimum from D by obtaining a random sample from D(n,k) where n is number of columns and k is medoid.
Step 3: For each data point o, swap it with medoid k, and compute the total cost C
Step 4: Compute minimum cost c from total cost C
Step 5: Add c to Final Matrix
```

3. COMPARISON BETWEEN K-MEAN, K-MEDOIDS AND MODIFIED K-MEDOIDS ALGORITHMS WITH DIFFERENT PARAMETER

3.1 Time Line Chart

Table 1: Time line chart of K-Mean, K-Medoids and Modified K-Medoids Algorithms

Number Of Cluster	Time Execution (in Seconds)		
	K-mean	K-medoid	Modified K-medoid
2	0.2016	0.1454	0.0324

3	0.2228	0.2204	0.0420
4	0.3029	0.2946	0.0660
5	0.4282	0.3628	0.0819

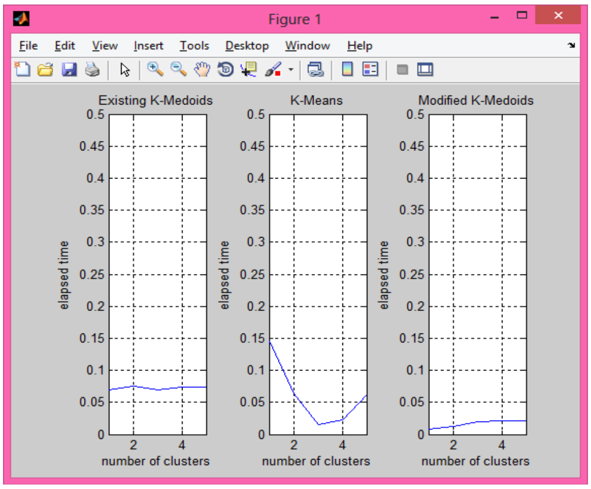


Fig 1: Time line Chart

Comments: Above Figure shows comparison of execution time between K-mean, K-medoid and Modified K-medoid algorithms. As graph shows that when number of cluster is less, Modified k-medoid takes less time to execute than k-mean and K-medoid algorithm. At the most number of cluster is increased; Execution time taken by Modified K-medoid algorithm is less than K-mean and K-medoid algorithm.

3.2 Cluster Quality, Crierian Error and Dunns index of K-mean, K-medoid and Modified K-medoid:

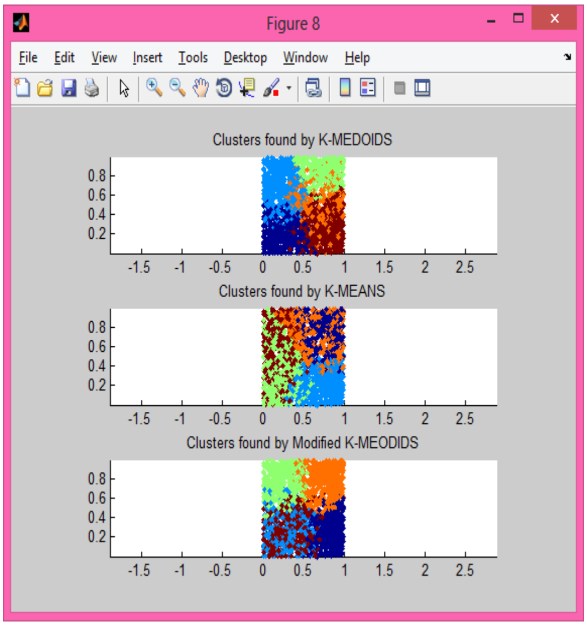
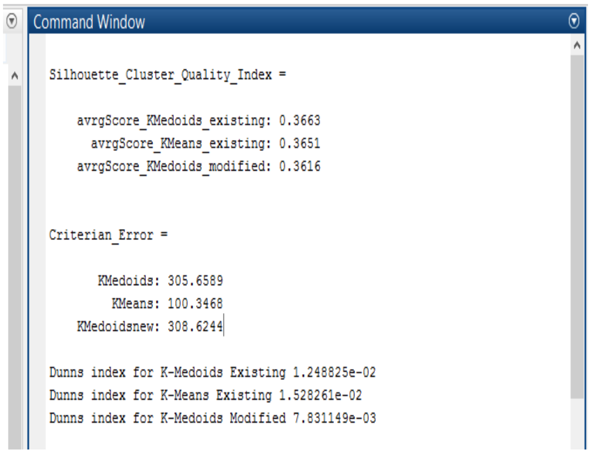


Fig 2: Cluster Generation



```
Silhouette_Cluster_Quality_Index =

    avrgScore_KMedoids_existing: 0.3663
      avrgScore_KMeans_existing: 0.3651
    avrgScore_KMedoids_modified: 0.3616


Criterian_Error =

       KMedoids: 305.6589
         KMeans: 100.3468
    KMedoidsnew: 308.6244

Dunns index for K-Medoids Existing 1.248825e-02
Dunns index for K-Means Existing 1.528261e-02
Dunns index for K-Medoids Modified 7.831149e-03
```

Fig 3: Result Analysis

Comments: Above Figure shows result analysis of k-mean, k-medoid and Modified k-medoid. When number of cluster increase, Silhouette cluster quality index of modified k-medoid improves better than k-mean and k-medoid. Criterian error of modified k-medoid is decrease than k-mean and k-medoid. Dunns index also increase. So modified k-medoid is better than k-mean and k-medoid.

3.3 Other Cluster Evaluation Matrices Index

Table 2: Cluster Evaluation Matrices Index for Modified K-Medoids Algorithms

Number Of Cluster	**Davies-Bouldin Index (DB)**	**Calinski-Harabasz Index (CH)**	**Krzanowski - Lai Index (KL)**
2	1.6482	665.77	640.67
3	1.3310	675.80	630.35
4	1.1674	689.38	616.53
5	1.0699	741.89	588.00

Comments: Davies-Bouldin Index (DB) is a matrix for evaluating clustering algorithms. Lower value is considered as good value. Result shows that when number of cluster increase, value of DB decrease. Calinski-Harabasz Index (CH) use to compare clustering solutions obtained on the same data. It is a distance calculation based cluster quality index. The higher the value, the "better" is the solution. Table shows that number of cluster increase, value of CH also increase. Krzanowski and Lai Index (KL) computed on feature dimensionally of the input matrix. The index is calculated on the basis of within-group dispersion matrix generated by clustering algorithm. Result shows that when number of cluster increase, value of KL decrease.

4. IMPLEMENTATION

The implementation of algorithm was carried out in MATLAB programming Language. MATLAB (matrix laboratory) is a multi-paradigm numerical computing environment. MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces.

5. CONCLUSIONS AND FUTURE WORK

Data mining is one of the largest and challenging areas of research with the major topic "Clustering". In this research work, we have implemented existing k-means, existing k-medoids and proposed modified k-medoids algorithms. All of the mentioned algorithms has been implemented using MATLAB environment. Modified k-medoids have been performed better than k-means and existing k-medoids on the larger data sets in the terms of elapsed time and clustering quality in our experimental results. A number of performance parameters has been computed for the purpose of effective and valuable comparison between three. Total time, Dunn's index, Davies-Bouldin index, Calinski-Harabasz index, Krzanowski and Lai, silhouette Indices has been computed to verify the performance of the modified k-medoids over exising k-medoids and k-means. The experimental results have proved the modified k-medoids better in all aspects of the performance analysis. In the future, this work should be enhanced to perform much better in terms of cluster quality or elapsed time. Additionally, there is also a space of improvement in the selection and plotting of the indices selection in the final result analysis.

REFERENCES

[1] Saurabh Shah & Manmohan Singh "Comparison of A Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid algorithm", International Conference on Communication Systems and Network Technologies, 2012.

[2] T. Velmurugan,and T. Santhanam, "A Survey of Partition based Clustering Algorithms in Data Mining: An Experimental Approach" An experimental approach Information. Technology. Journal, Vol, 10,No .3 , pp478-484,2011.

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

[4] "Data Mining Concept and Techniques" ,2nd Edition, Jiawei Han, By Han Kamber.

[5] Jiawei Han and Micheline Kamber, "Data MiningTechniques", Morgan Kaufmann Publishers, 2000.

[6] Abhishek Patel,"New Approach for K-mean and K-medoids algorithm", International Journal of Computer Applications Technology and Research,2013.

[7] http://www.cs.put.poznan.pl/jstefanowski/sed/DM-7clusteringnew.pdf

[8] http://www.ccs.neu.edu/home/mirek/classes/2012-S-CS6220/Slides/Lecture4-Clustering.pdf