ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2 Issue 8, August 2013

Compression Record Based Efficient k-Medoid Algorithm to Increase Scalability and Efficiency

Archana Kumari[1], Hritu Bhagat [2],

[1] *Department of Computer Engineering,Medicaps institute of technology and management ,Indore ,*

[2] *Department of Computer Engineering,Medicaps institute of technology and management ,Indore ,*

Abstract: *- Clustering analysis is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes. K-medoid clustering algorithms are widely used for many practical applications. Original K-medoid algorithm select initial centroids and medoids randomly that affect the quality of the resulting clusters and sometimes it generates unstable and empty clusters which are meaningless. The original k-means algorithm is computationally expensive and requires time proportional to the product of the number of data items, number of clusters and the number of iterations. Improved k-Medoid clustering algorithm has the accuracy higher than the original.*

Keywords: *Data Mining, Clustering, k-mediod.AI,*

I. Introduction

The huge amount of data collected and stored in databases increases the need for effective analysis methods to use the information contained implicitly there. Clustering is important data mining technique to extract useful information from various high dimensional datasets. A wide range of clustering algorithms is available in literature and still an open area for researcher. One of the primary data analysis tasks is cluster analysis, which helps the user to understand the natural grouping or structure in a dataset. Therefore, the development of improved clustering algorithms has been received much attention. The goal of a clustering algorithm is to group the objects of a database into a set of meaningful subclasses [3].

Clustering has been a widely studied problem in a variety of application domains including data mining and knowledge discovery [10], data compression and vector quantization [11], Clustering is the process of partitioning or grouping a given set of patterns into disjoint clusters. pattern recognition and pattern classification [7], neural networks, artificial intelligence, and statistics. Existing clustering algorithms can be broadly classified into hierarchical and partitioning clustering algorithms [17].

Hierarchical algorithms decompose a database *D* of *n* objects into several levels of nested partitioning (clustering), represented by a dendrogram, i.e., a tree that iteratively splits *D* into smaller subsets until each subset consists of only one object. There are two types of hierarchical algorithms; an agglomerative that builds the tree from the leaf nodes up, whereas a divisive builds the tree from the top down. Partitioning algorithms construct a single partition of a database *D* of *n* objects into a set of *k* clusters. Optimization based partitioning algorithms typically represent clusters by a prototype. Objects are assigned to the cluster represented by the most similar prototype.

This is done such that patterns in the same cluster are alike, and patterns belonging to two different clusters are different. An iterative control strategy is used to optimize the whole clustering such that, the average squared distances of objects to its prototypes are minimized. These clustering algorithms are effective in determining a good clustering, if the clusters are of convex shape, similar size and density, and if their number *k* can be reasonably estimated. Depending on the kind of prototypes, one can distinguish k-means, k-modes and k-medoids algorithms. In k-means algorithm [8], the prototype, called the center; is the mean value of all objects belonging to a cluster. The k-modes algorithm [16] extends the k-means paradigm to categorical domains. For k-medoids algorithms [7], the prototype, called the "medoid"; is the most centrally located object in the cluster. The algorithm CLARANS, introduced in [20], is an improved k-medoids type algorithm restricting the huge search space by using two additional user-supplied parameters.

It is significantly more efficient than the well known k -medoids algorithms PAM and CLARA, presented in [7].Among clustering formulations that are based on minimizing a formal objective function, perhaps the most widely used and studied is k-means clustering. Given a set of *n* data points in real d-dimensional space, *Rd*, and an integer *k*, the problem is to determine a set of *k* points in *Rd*, called centers, so as to minimize the mean squared distance from each data point to its nearest center. Although the k-means method has a number of advantages over other data clustering techniques, it also has drawbacks; it converges often at a local optimum [2], the final result depends on the initial starting centers. Many researchers introduce some methods to select good initial starting centers; you can see [5] and [6]. Other researchers try to find the best value for the parameter *k* that determines the number of clusters or the value of *k* must be supplied by the user. You can see [22] and [21]. In recent years, many improvements have been proposed and implemented in the K-means method; you can see [9]. The k means clustering algorithm attempts to determine *k* partitions that optimize a certain criterion function. The average square error criterion, defined in (1), is the most commonly used (*cj* is the mean of cluster *Ci*, *n* is the number of objects in the dataset).

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

The average square-error is a good measure of the within cluster variation across all the partitions. Thus, the average square error clustering tries to make the *k* clusters as compact and separated as possible, and works well when clusters are compact clouds that are rather well separated from one another [12].

II. Related Work

.

B. 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 [11]. 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 [16]. The algorithm takes the input parameter k, the number of clusters to be partitioned among a set of n objects [6]. A typical KMedoids algorithm for partitioning based on Medoid or central objects is as follows:

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 $O_{random}$;
    3. Compute the total points S of swap point $O_j$ with
       $O_{ramdom}$
    4. If S < 0 then swap $O_j$ with $O_{random}$ to form the
       new set of k medoid
Until no change;
```

III. Modified K-Mediod Algorithm

A. Modified K-mean algorithm:

The K-mediod algorithm is a popular clustering algorithm and has its application in data mining ,image segmentation, bioinformatics and many other fields[14].This algorithm works well with small datasets.In this paper we proposed an algorithm that works well with large datasets. Modified k-mean algorithm avoids getting into locally optimal solution in some degree, and reduces the adoption of cluster -error criterion.

Modified K –Medoid Algorithm

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

```
1 Method:  Arbitrarily choose k objects in D as the initial representative objects;
2 For each data-point $d_i$, find the closest centroid cj and assign $d_i$ to cluster j Select initial
k=1;
3 Do
4 OldMSE=MSE;
5 MSE1=0;
6 For j=1 to k
7 mj=0; nj=0;
8 endfor
9 For i=1 to n
10 For j=1 to k
11 Compute squared Euclidean distance d2(xi, mj);
11.1 Repeat assigns each remaining object to the cluster with the nearest medoid;
11.2 If D < 0 then swap O j with O random to form the new set of k medoid
12. Randomly select a non medoid object O random;
12.1 compute the total points S of swapping object Oj with O ramdom;
12.2 Until no change
 12.3 end for
13 Find the closest centroid mj to xi;
14 mj=mj+xi; nj=nj+1;
15 MSE1=MSE1+d2(xi, mj);
16 endfor
17 For j=1 to k
18 nj=max (nj, 1); mj=mj/nj;
19 endfor
20 MSE=MSE1; while (MSE<OldMSE)
```

Sub Algorithms: calculate Distance

```
1 For i=1 to n
Compute squared Euclidean distance
d2(xi, Clusterid[i]);
If (d2(xi, Clusterid[i])<=Pointdis[i])
Point stay in its cluster;
2 Else
3 For j=1 to k
4 Compute squared Euclidean distance
d2(xi, mj);
5 endfor
6 Find the closest centroid mj to xi;
7 mj=mj+xi; nj=nj+1;
8 MSE=MSE+d2 (xi, mj);
9 Clustered[i] =number of the closest centroid;
10 Pointdis[i] =Euclidean distance to the closest
centroid;
11 endfor
12 For j=1 to k
13 mj=mj/nj;
14 endfor
```

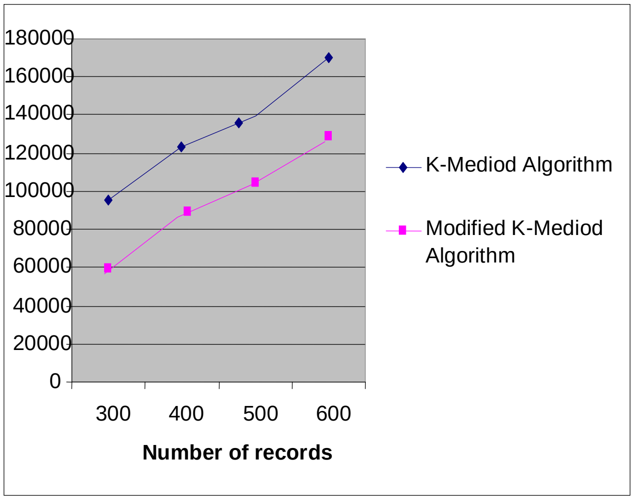
IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this Paper, the most representative algorithms K-Mediod, and proposed algorithm Modified K-Mediod were examined and analyzed based on their basic approach for large data set, using student class dataset. The best algorithm in each category will found out based on their performance. Comparison between K-Mediod and Modified K-Mediod algorithm with Number of Clusters and Execution Time (in milliseconds) is shown in the table.

TABLE II: K-MEDOID AND MODIFIED K-MEDIOD PERFORMANCE

Number of Records	Time taken to execute (In millisecond) K-Mediod Algorithms	Time taken to execute (In millisecond) Modified K-Mediod Algorithm
300	95672	59735
400	123272	89332
500	139826	106243
600	170231	128338

Graph shows the comparison between K-medoid and Modified K-mediod algorithm on the basis of various numbers of clusters and execution time. Modified K-Mediod gives better performance in comparison to K-Medoid algorithm.



V. Conclusion

From the experimental results,it is observed that the comparison between K-mediod and Modified K-mean algorithm shows that when number of clusters is less, Modified K-mediod takes less time to execute than the K-mediod and if the number of clusters is more, then it is again true that modified K-mediod takes less time to execute than the K-mediod.

References

[1] S. A. Raut, S. R. Sathe, and A. Raut, "Bioinformatics: Trends in GeneExpression Analysis," *proceedings of 2010 International ConferenceOn Bioinformatics and Biomedical Technology*, 16-18 April 2010,Chengdu, China.

[2] S. A. Raut, S. R. Sathe, and A. P. Raut, "Gene Expression Analysis-AReview for large datasets," *Journal of Computer Science andEngineering*, vol.4, Issue 1, November 2010.

[3] Xiong, H., J. Wu and J. Chen, 2009. K-Meansclustering versus validation measures: A datadistribution perspective. IEEE Trans. Syst., Man,Cybernet.PartB,39:318-331.http://www.ncbi.nlm.nih.gov/pubmed/19095536.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[4] Berkhin, P., 2002. Survey of clustering data miningtechniques. Technical Report, Accrue Software,Inc.http://www.ee.ucr.edu/~barth/EE242/clustering_survey.pdf

[5] MacQueen, J.B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In Proc. of 5th Berkley Symposium on Mathematical Statistics and Probability, Volume I: Statistics, pp. 281–297.

[6] S. Ray, and R. H. Turi, "Determination of number of clusters in k-meansclustering and application in colour image segmentation," InProceedings of the 4th International Conference on Advances in PatternRecognition and Digital Techniques, 1999, pp.137-143.

[7] G. Sheikholeslami, S. Chatterjee, and A. Zhang, "Wave-Cluster: AMulti-Resolution Clustering Approach for Very Large SpatialDatabases," Proc. 24th Int. Conf. on Very Large Data Bases. New York,1998, pp. 428-439.

[8] R. Sibson, "SLINK: an optimally efficient algorithm for the single-linkcluster method," The Comp. Journal, 16(1), 1973, pp. 30-34.

[9] T. Zhang, R. Ramakrishnan, and M. Linvy, "BIRCH: An Efficient DataClustering Method for Very Large Databases," Proc. ACM SIGMODInt. Conf. on

[10] Zhang Y. , Mao J. and Xiong Z.: An efficient Clustering algorithm, In Proceedings of Second International Conference on Machine Learning and Cyber netics, November 2003.

[11] A. Gersho, a nd R. M. Gray *Vector Quantization and Signal Compression*, Kluwer Academic, Boston, 1992.

[12] S. Guha , R. Rastogi, and K. Shim, "CURE: An Efficient Clustering Algorithms for Large Databases," Proc. ACM SIGMOD Int. Conf. on Management of Data, Seattle, WA, 1998, pp. 73-84.

[13] V. Hautamaeki , S. Cherednichenko , I. Kaerkkaeinen , T. Kinnunen, And P. Fraenti, "Improving K-Means by Outlier Removal," SCIA 2005,LNCS 3540, 2005, pp. 978-987.

[14] V. Hautamaeki , I. Kaerkkaeinen, and P. Fraenti, " Outlier detection using k-nearest neighbourgraph," In: 17th International Conference on Pattern Recognition (ICPR 2004), Cambridge, United Kingdom.

[15] A. Hinneburg, and D. Keim, "An Efficient Approach to Clustering in Large Multimedia Databases with Noise," Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining, New York City, NY. 1998.

[16] Z. Huang, " A fast clustering algorithm to cluster very large Categorical data sets in data mining," Proceedings of the SIGMOD Workshop on Research Issues o n Data Mining and Knowledge Discovery, Dept. of Computer Science, The University of British Columbia, Canada, 1997,pp. 1-8.

[17] A. K. Jain, and R . C. Dubes, *Algorithms for Clustering Data*, PrenticeHall, 1988.

[18] L. Kaufman, a nd P. Rousseeuw, " Finding Groups in Data: An Introduction to Cluster Analysis," Wiley, 1990.

[19] J.B. MacQueen, " Some methods for classification and analysis of multivariate observations. Proc. 5th Symp. Mathematical Statistics and Probability, Berkelely, CA, Vol(1), 1967, pp. 281297.

[20] R.T. Ng, and J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining," Proc. 20th Int. Conf. on Very Large Data Bases,Morgan Kaufmann Publishers, San Francisco, CA, 1994,

[21] D. T. Pham , S. S. Dimov, and C. D. Nguyen, "Selection of *k* in K-Means clustering," Mechanical Engineering Science, vol(219), 2004.

[22] S. Ray, and R. H. Turi, " Determination of number of clusters in k-mean clustering and application in colour image segmentation," In Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques, 1999, pp.137-143.

[23] G. Sheikholeslami, S. Chatterjee, and A. Zhang, " Wave-Cluster: A Multi - Resolution Clustering Approach for Very L arge Spatial Databases," Proc. 24th Int. Conf. on Very Large Data Bases. New York,1998, pp. 428-439

[24] R. Sibson, "SLINK: an optimally e fficient algorithm for the single-linkcluster method," The Comp. Journal, 16(1), 1973, pp. 30-34.