ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

An Ameliorated Partitioning Clustering Algorithm for Large Data Sets

Raghavi Chouhan[1], Abhishek Chauhan[2]

MTech Scholar, CSE department, NRI Institute of Information Science and Technology, Bhopal, India [1]
Assistant Professor, CSE department, NRI Institute of Information Science and Technology, Bhopal, India [2]

Abstract: K-medoid clustering algorithm is broadly used for various practical applications. Original K-medoid algorithm use to take initial centroids and medoids arbitrarily that bears on the resulting clusters and it leads to unstable and empty clusters which are meaningless and also amount of iterations can be rather high so K-Medoid is not a substitute for big databases because of its computational complexity. Also the original k-means algorithm is computationally costlier and involves time relative to the product of the number of data items, number of clusters and the number of iterations, the time complexity of K-means is O (tkn) where t is the amount of iterations. Though K-means algorithm usually leads to better outcome, it does not scale well and is not time efficient. Ameliorated k-Medoid clustering algorithm will have the accuracy more than the original one. The new idea for K-medoid algorithm overcomes the deficiency of existing medoid. It initially computes the initial centroids k as per the necessity of user and then provides improved and efficient cluster with no sacrifice on accuracy. It generates steady clusters in order to get better accuracy. It also minimizes the mean square error and improves the quality of clustering, reduces the number of iterations and works on reducing time complexity. The improved k-Medoid clustering algorithm will have accuracy greater than the original one.

Keywords: clustering, K-medoid, data mining, Large data sets

I. INTRODUCTION

With the help of automated statistical analysis or "data mining" techniques, businesses use to discover new trends and patterns of activities that were previously undetected. Data mining is a method of discovering meaningful, new correlation patterns and trends by shifting through vast amounts of data stored in repositories, by means of pattern recognition, statistical and mathematical techniques. Data Mining is mining of information from enormous amount of data. By means of Data mining we can forecast the nature and behavior of every type of data. Clustering is a preprocessing step in every data mining algorithms. A cluster is a group of data instances which are “analogous” to each other and are “different” to data instances in other clusters.

K-Medoids: K-medoid is a conventional partitioning technique of clustering that form k number of clusters for data set of n objects. This k: which is number of clusters required, will be given by user. This algorithm functions on the principle of reducing the sum of dissimilarities among each object and its corresponding reference point. The algorithm arbitrarily chooses the k objects in dataset D as initial representative objects called medoids. We can define medoid as a cluster’s object, whose average difference to every object in the cluster is minimum i.e. it is a most centrally located point in the known data set. After that for all objects in the dataset, it assigns each object to the nearest cluster depending upon the object’s distance to the cluster medoid. After all assignment of a data object to particular cluster the new medoid is determined. The difficulty is that K-Medoids does not produce the same outcome with every run which is overcome in this new improved algorithm.

There has been a spectacular increase in the quantity of information being stored in the electronic format. This gathering of data has taken place at an explosive rate. It was predictable that information is at the heart of the business operations and that decision makers could make the use of data stored to gain the valuable insight into the business. We need clustering at various places as an example like in a market-basket database generally the number of items and therefore the number of attributes in such a database is extremely large so we require a clustering technique that will be able to handle these vast data and can efficiently be able to give effective clusters. A variety of factors such as lengthy iterations, complexity in handling large datasets, formation of unsteady clusters which gives different results on every run causes inaccuracy in outcome, and sometimes ends up with some useless data.

The new method for K-medoid algorithm removes the deficiency of existing K-medoid. It initially calculates the initial centroids k as per requirements of users and then gives better, effective and good cluster without scarifying accuracy and leading to stable cluster formation every time will run the algorithm. Whenever we run K-medoid algorithm in every result for same set of inputs it produces different clusters in output in each run so it lead to formation of unstable clusters. The new improved algorithm generates

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

stable clusters to improve accuracy. It also reduces the mean square error which is defined as the expected value of the squared difference between the estimate and the actual value and thereby improves the quality of clustering. The improved k-Medoid clustering algorithm has the accuracy higher than the original one.

II. LITERATURE SURVEY

A. Clustering

Clustering is an unsupervised learning process where grouping of physical or abstract objects into classes of similar object takes place. A cluster is a group of data objects which are similar to each other within the same cluster and are different to the objects in other clusters. Clustering is also known as data segmentation in certain applications as clustering partitions data sets into groups based on their similarity. Dissimilar to classification, clustering does not depend on predefined classes and class-labeled training examples.

There are lots of clustering methods available, and every one of them may give a different grouping of a dataset. The selection of a particular method will depend on the kind of output required. Some clustering analysis methods are grid based clustering, model based clustering, partition based clustering, density based clustering, and hierarchical clustering.

Partitioning method creates k partitions (clusters) of the known dataset, where all partitions represent a cluster. And each cluster can be represented by a centroid or a cluster representative which is some kind of summary explanation of all the objects present in a cluster.

K-medoids as well as K-means are the best examples of partitioning methods. Both the k-medoids and k- means algorithms are partitional (breaking the dataset up into groups) in addition both attempts to reduce squared error. In the k-means clustering problem, the centroid is not present in the original points in the majority of cases. In comparison to the k-means algorithm k-medoids selects data points as centers which make k-medoids more robust in the presence of noise and outliers than k-means, since a medoid is not as much influenced by outliers or any other extreme values than a mean.

Hierarchical clustering proceeds one after another by either splitting larger clusters, or by merging smaller clusters into larger ones. We can classify hierarchical method as being either divisive or agglomerative, based on how the decomposition is taking place. This agglomerative approach begins with formation of a separate group by each object. It consecutively merges the groups or objects that are near to one another, until a preferred number of clusters are obtained.

The divisive approach, which is also called as top-down approach, begins with all objects in the same cluster. In successive iteration, a cluster is split up into smaller clusters, until a desired number of clusters are obtained. Though in hierarchical methods if a step (merge or split) is complete, it cannot be undone sometimes this may lead to erroneous decisions. DIANA and AGNES are some examples of hierarchical clustering.

Density based clustering methods are based on a local cluster criterion. Clusters are considered as regions in the data space where objects are dense, and are separated by regions where objects density (noise) is less. The general idea is to grow the cluster as far as the density which is the number of objects or data points in the neighborhood exceeds given threshold; that is, for every data point inside a given cluster, the neighborhood of a certain radius need to contain at least a minimum number of points. As a result these regions formed may consist of an arbitrary shape. DBSCAN is a density based method which grows clusters on the basis of a density-based connectivity analysis. OPTICS is one more density-based method which generates an augmented order of the clustering structure of data.

Grid-based clustering algorithm use to divide multidimensional data space into a specified number of cells, and after that clustering operation is applied on it. The major advantage of this approach is its fast processing time, which is normally independent of the amount of data objects along with dependent simply on the number of cells in every dimension of the quantized space. STING (STatistical INformation Grid) is an example of a grid-based method based on statistical information stored in grid cells. It use to divide data space into rectangular cells, and then these cells form a hierarchical structure: and can divide high-level cells into a number of low-level cells. The data statistical information (for example, mean, maximum, minimum, count and data distribution, etc.) of each cell is pre-calculated for the subsequent query processing.

Wave Cluster and CLIQUE (CLustering In QUEst) are clustering algorithms which are both density-based as well as grid-based. CLIQUE is an incorporated algorithm based on grid and density. It divides M-dimensional data space into rectangular cells. If the quantity of data points in a cell is more than a threshold (user input), it will be called as a dense cell. A cluster is the largest collection of dense cells. CLIQUE algorithm will automatically recognize high dimensional space along with high dense data points, and is independent of data input order and data distribution.

B. K-Medoids

K-medoid is a traditional partitioning technique of clustering which clusters *n* objects data sets into *k* clusters. This k: which is number of clusters required is given by user. This algorithm works on the principle of minimize the sum of

dissimilarities among each object and its equivalent reference point. This algorithm arbitrarily selects k objects from dataset D as initial representative objects known as medoids. We can define medoid as an object of a cluster, which is having minimal average dissimilarity to all objects in the cluster i.e. it is the most centrally located point in a known data set. After that for all objects in the given dataset, it assigns every object to its nearest cluster depending upon its distance to the medoid. After each assignment of data object to a particular cluster is done new medoid will be decided.
The difficulty is K-Medoids with every run does not generate the same result, since the resultant clusters depends on initial arbitrary assignments. It is further more robust as compare to k-medoids in the presence of outliers and noise; however it's processing is costlier than k-medoid method. And also, the optimal number of clusters k is difficult to predict, and so it becomes hard for a user with no prior knowledge to state k's value.

Problems with k-medoids clustering algorithm
The algorithm is simple and has nice convergence but there are number of problems with this. Some of the weaknesses of k-medoids are

- When the amount of data is not so large, initial grouping is determining the cluster significantly.

Based on distance we get circular shaped cluster.

- The number of cluster, K, must be determined beforehand which sometimes get hard to predict beforehand.
- By using the same data, which is entered in a different order we may get different cluster if the amount of data is few.
- Experiments show that outliers can result into a problem and may force algorithm to recognize false clusters.
- As we assume that each attribute has the same weight so it gets difficult to know which attribute contributes more to the grouping process

An Improved K-medoids Clustering Method for Near-duplicated Records Detection in reference [1], how to resolve the problem of detecting near-duplicated records in K-medoid clustering method is proposed in this paper. It consider each record in database as a separate data object, it uses weights of attributes and edit-distance method to get similarity value between records, and then it detects duplicated records by forming clusters of these similarity values. This algorithm can adjust the number of clusters automatically by comparing similarity value with preset similarity threshold, and it also avoids a large numbers of I/O related operations which is used by conventional "sort/merge" algorithm for sequencing. Through experiment, it is proved that this algorithm use to have high availability and good detection accuracy

With reference to improved K-Medoids Clustering Based on Cluster Validity Index and Object Density [2], Clustering classifies different groups of objects through formation of subsets called as clusters by partitioning of data sets. Algorithms like k-medoids and k-means are acting as roots for clustering. However traditional k-medoids clustering algorithm suffered from many limitations. First limitation is that it should have previous knowledge about cluster number which is parameter k. Next, it initially needs to do the arbitrary k selection of representative objects and if these k-medoids are not chosen properly then it will be difficult to obtain natural cluster. Third limitation is that it also depends on the input dataset's order. With the help of cluster validity index first limitation was removed. Than for the other two limitations related to conventional k-medoids, an enhanced k-medoid algorithm is proposed. They had proposed a novel method for the initial representative object selection rather than arbitrary selection of medoids from initial k objects. This technique depends on objects density. They found objects set which were heavily populated and then they chose medoids from these obtained set**.** Initial medoids taken from these k data objects are then utilized in clustering process. Validity of the proposed algorithm is proved using diet structure and iris dataset in order to find natural clusters in this datasets.

In the Analysis of Initial Centers for k-Means Clustering Algorithm [11] Data Analysis uses to play a significant role in understanding different events. Cluster Analysis is broadly used data mining technique for discovery of knowledge. Clustering is having wide applications in the field of Pattern Matching, Artificial Intelligence, Compression, Image Segmentation etc. Clustering is the method of finding group of objects so that objects belonging to one group will be similar to each other and those objects will be different from the objects in another group. K-Means clustering algorithm is one among the popular algorithms which gained lot of attraction because of its ease of implementation and simplicity. Because of arbitrary selection of k-initial centers K-Mean algorithm's efficiency is limited. That is why; we surveyed different approaches for initial selection of centers for k-Means algorithm. There is also a comparative analysis of Data Clustering with Modified k-Means Algorithm using MATLAB R2009b and Original K-Means algorithm. As the similarity measure Euclidean distance is chosen for implementation and results are evaluated.

In [10] Omar Kettani, Benaissa Tadili, Faycal Ramdani, suggested that in data mining, the k-means algorithm is among the most commonly and widely used technique to solve clustering problems because of its good performance and simplicity. However, one of the major drawbacks of k-

means algorithm is that its performance and accuracy are dependent to the initial choice of clustering centers, which use to get generated arbitrarily. In order to overcome the drawback of this algorithm, they proposed a simple deterministic method which is based on nearest neighbor search and k-means procedure so in order to improve results of clustering. Experimental results made on a variety of data sets reveal that the proposed method is more accurate than the standard K-means algorithm.

From Aloysius George in Efficient High Dimension Data Clustering using Constraint-Partitioning K-Means Algorithm [8], with the growth in data size, clustering of high dimensional databases is having a challenging task that must satisfy the necessities of both result quality and computation efficiency. In order to attain both requirements, feature space has received much significance amongst data mining researchers over original data space clustering. Accordingly, they performed data clustering of large dimension dataset by using Constraint-Partitioning K-Means clustering algorithm which was not able to properly fit so as to cluster large dimensional data sets in terms of efficiency and effectiveness, because of the inherent sparse of large dimensional data and thus resulted in producing inaccurate and indefinite clusters. Thereby, they carry out two steps for clustering large dimension dataset. Firstly, they perform dimensionality reduction on the large dimension dataset by using Principal Component Analysis for data clustering as a preprocessing step. Then, they combined the dimension reduced dataset to the Constraint-Partitioning K-Means clustering algorithm in order to produce accurate and good clusters. And then performance of this approach was evaluated by using high dimensional datasets such as Ionosphere dataset and Parkinson's dataset. The experimental results proved that the proposed approach is very efficient in producing precise and accurate clusters.

According to Experimental study of Data clustering using k-Means and modified algorithms [7], the k- Means clustering algorithm which is an old algorithm and has been hugely researched due to its simplicity of implementation and ease. Clustering algorithm is very useful and has a broad attraction in investigative data analysis. This paper presents experimental study results of various approaches to k-Means clustering, thus comparing results on different datasets by using some modified algorithms which are implemented using MATLAB R2009b and original k-Means. And then results are calculated on the basis of some performance measures such as no. of iterations, no. of points misclassified, accuracy, Silhouette validity index and execution time.

Harmony *K*-means algorithm for document clustering [5], to have a fast and high quality document clustering is an important task in enhancing web crawling, search engine results, organizing information, and filtering or information retrieval. Recent studies reveal that the most generally used partition-based Clustering algorithm, which is the *K*-mean algorithm, is more appropriate for larger datasets. However, a local optimal solution can be generated by *K*-means algorithm. In this paper they proposed a New Harmony *K*-means Algorithm (HKA) which is dealing with document clustering that is based upon Harmony Search (HS) optimization process. It is also proved with the help of finite Markov chain theory that the Harmony *K*-means Algorithm gets converged to the global optimum. To exhibit the speed and effectiveness of Harmony *K*-means Algorithm, they have applied Harmony *K*-means Algorithm on various standard datasets. They are also comparing the Harmony *K*-means Algorithm with other model-based document and meta-heuristic clustering methods. Experimental results reveal that the Harmony *K*-means algorithm converges to the finest known optimal faster than any other method and also the quality of clusters are comparable.

Extensions to the *k*-Means Algorithm for Clustering Large Data Sets with Categorical Values [4], the *k*-mean algorithm is well recognized for its effectiveness in clustering big data sets. However, running only on numerical values prohibits k-means algorithm to be used in order to cluster real world data which including categorical values. In this paper they are presenting two algorithms which will be extending the *k*-mean algorithm to definite domains and to the domains with mixed categorical and numeric values. The *k*-mode algorithm is using an easy dissimilarity matching measure in order to deal with categorical objects, in which it replace the means of clusters by modes, along with uses a frequency-based method to revise modes in the clustering procedure so as to reduce the cost function for clustering. By means of these extensions the *k*-mode algorithm enables the categorical data to be clustered in a manner comparable to *k*-mean. From the description of a combined dissimilarity measure the *k*-prototypes algorithm, further integrate the *k*-modes and *k*-means algorithms in order to let the clustering of objects described by mixed categorical and numeric attributes. They used the well recognized credit approval and soybean disease data sets to exhibit the clustering performance for the two algorithms. There experiments on these two real world data sets along with half a million objects each show that the two algorithms are efficient when clustering large data sets, which is quite critical to data mining applications.

Feature selection for k-means clustering stability: theoretical analysis and an algorithm [3],because of small input deviations learning algorithm's steadiness is a very important property, as this shows that the resultant models are robust in presence of data sample fluctuations and noisy features. The qualitative character of the steadiness property hardens the growth of stability optimizing, practical, data mining algorithms as a number of issues come up, like: how stability can be efficiently linked by intrinsic data properties,

or how "much" stability is sufficient. They gathered these issues and then explored the outcome of increase in steadiness in continuous k-mean clustering problem. There study is based on both statistical arguments and mathematical optimization that balance each other and let the firm understanding of the algorithm's steadiness properties. Interestingly, they derived that steadiness maximization as expected introduces a transaction between variance and cluster separation, leading to the selection of features which is having a high cluster separation index and is not falsely exaggerated by the features variance. This proposed algorithmic system is based on a Sparse PCA approach to facilitate the selection of features that will increase steadiness in a greedy manner. In their study, they also analyzed various properties of Sparse PCA related to steadiness which promotes Sparse PCA as a feasible feature selection means for clustering. The practical significance of this proposed method is verified in the context of cancer research, where they considered the difficulty to detect latent tumor biomarkers using microarray gene expression data. And the application of their technique on a leukemia dataset reveals that the transaction among variance and cluster separation leads to the feature selection analogous to significant biomarker genes. Few of them are having relative small variance and cannot be detected without the direct optimization of steadiness in Sparse PCA based k-mean. Other than the qualitative assessment, they had also verified there approach like a feature selection technique for *k*-mean clustering which is using research datasets of four cancer. The experimental results demonstrate that the practical usefulness of their framework as a feature selection mechanism for clustering.

Clustering analysis is a descriptive job that attempts to recognize groups of homogeneous objects based on their attributes values. K-medoid clustering algorithms are broadly used for several practical applications. Original K-medoid algorithm use to select initial centroids and medoids arbitrarily which affects the quality of resultant clusters and also at times it produces empty and unstable clusters that contains no meaning. The original k-means algorithm is computationally very expensive and it also need time which is equal to the product of the number of clusters, number of data items and the number of iterations. Enhanced k-Medoid clustering algorithm is having the accuracy greater than the original.

CONCLUSION

This new approach will enhance the efficiency of original k-medoid algorithm in dealing with comparatively larger data which is an important issue nowadays due to regular increase in the size of data in every field and together with this it will also improve its speed in doing so and also the resultant clusters which is very important part of the entire process will also get improved.

REFERENCES

[1] Ying Pei, Jungang Xu, Zhiwang Cen, Jian Su ,"IKMC: An Improved K-medoids Clustering Method for Near-duplicated Records Detection", 2009 IEEE conference.

[2] Bharat Pardeshi1 and Durga Toshniwal2," Improved K-Medoids Clustering Based on Cluster Validity Index and Object Density",2010 IEEE conference.

[3] Dimitrios Mavroeidis · Elena Marchiori," Feature selection for k-means clustering stability: theoretical analysis and an algorithm",29 May 2013 Springer.

[4] ZHEXUE HUANG," Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values", Data Mining and Knowledge Discovery 2, 283–304 (1998)

[5] Mehrdad Mahdavi · Hassan Abolhassani," Harmony K-means algorithm for document clustering", 11 December 2008 Springer Science+Business Media, LLC 2008.

[6] Swagatika Devi,Trloknath Pandey, Alok Kumar Jagdev, "Performance Improvement of disk based k-means over k-means on large datasets", Journal of Theoretical and Applied Information Technology 31st March 2013. Vol. 49 No.3.

[7] Dr. M.P.S Bhatia1 and Deepika Khurana ," Experimental study of Data clustering using k-Means and modified algorithms", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.3, May 2013.

[8] Aloysius George," Efficient High Dimension Data Clustering Using Constraint-.Partitioning K-Means Algorithm"IAJIT July 28, 2011

[9] 'PavellBerkhin' Survey of Clustering Data Mining Techniques",

[10] Omar Kettani, Benaissa Tadili, Faycal Ramdani," A Deterministic K-means Algorithm based on Nearest Neighbor Search",International Journal of Computer Applications (0975 – 8887) Volume 63– No.15, February 2013

[11]M.P.S Bhatia,Ph.D,Deepika Khurana,"Analysis of Initial Centers for k-Means Clustering Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 71– No.5, May 2013 .

[12] Kyriakos Mouratidis , Dimitris Papadias,Spiros Papadimitriou, "Tree-based partition querying: a methodology for computing medoids in large spatial datasets" Springer-Verlag 2007,12 April 2007

[13] Gholamhosein Sheikholeslami, Surojit Chatterjee, Aidong, Zhang," *WaveCluster*: a wavelet-based clustering approach for Spatial data in very large databases" Springer-Verlag 2000