Binary Journal of Data Mining & Networking 1 (2010) 10-11
http://www.arjournals.org/index.php/bjdmn/index

Research Article

ISSN: 2229 -7170

Efficient k-mean algorithm for large dataset

Ramesh Prasad Aharwal[1*], Manmohan Singh[2]

*Corresponding author:

Ramesh Prasad Aharwal

[1]Department of Mathematics and Computer Science, Govt. P.G. College Bareli (M.P.), India
[2]Department of Computer Science and Engg. BIST Bhopal, (M.P.), India

Abstract

The term data mining is used to discover knowledge from large amount of data. For knowledge discovery many software haven developed, that is known as data mining tools these are statistical, machine learning, And neural networks. K-means and K-medoids are widely used simplest partition based unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters; technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Stored data is used to locate data in predetermined groups called class. Data items are grouped according to logical relationships or consumer preferences called cluster. Data can be mined to identify association. Data is mined to anticipate behavior patterns and trends called sequential patterns.

Keywords:k-mean; algorithm; k-medoids; data mining; Partition method

Introduction

A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering.[1]

Clustering is a dynamic field of research in data mining. Many clustering algorithms have been developed. These can be categorized into partition method, hierarchical method, density based method, grid based method, and model based methods[2].

A partitioning method creates an initial set of k partitions, where parameter k is the number of partitions to construct; then it uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. Typical partitioning methods include k-means, k-medoids, CLARANS, and their improvements. [3]

A hierarchical method creates a hierarchical decomposition of the given set of data objects. The method can be classified as being either agglomerative (bottom up) or divisive (top down), based on how the hierarchical decomposition formed. To compensate for the rigidity of merge or split, the quality of hierarchical agglomeration can be improved by analyzing object linkages at each hierarchical partitioning (such as in CURE and Chameleon) or interesting other clustering techniques, such as iterative relocation [4]

A density based method clusters objects based on the notation of density. It either grows cluster according to the density of neighborhood objects (such as DBSCAN) or according to some density function (such as in DENCLUE). OPTICS is a density based methods that generates an augmented ordering of the clustering structure of the data [5].

A grid based method first quantizes the object space into a finite number of cells that form a grid structure, and then perform clustering on the grid structure. STING is a typical example of a grid-based method based on statistical information stored in grid cells. CLIQUE and Wave Cluster are two clustering algorithms that are both grid-based and density based.[6]

A model-based method hypothesized a model for each of the clusters and finds the best fit of the data to that model. Typical model –based method involve statistical approaches. [7]

Partition Method

Partition clustering algorithm attempts to determine k partitions from a collection of n d-dimensional objects. The aim of partitioning methods is to reduce the variance within each cluster as possible and have large variance between the clusters. For this it uses the iterative relocation technique that attempt to improve the partitioning by moving objects from one group to another with the goal of minimizing square error criterion, which is defined as follow [8].

Let X={x1,x2,.....xn} be the set of n input objects, where xi={xi1,xi2,....xid}. here Xij represent jth feature of object xi. Assume this data is partitioned into groups C= {C1,C2,......Ck} where k is the number of clusters and k<=n.Then the square error criterion is defined as k

$$E = \Sigma \quad \Sigma\, |xi - mi|2$$

$$i = 1 \quad xi \in Ci$$

Where E is the sum of square error for all objects in the database, Xi is the point in space representing a given object, and mi is the mean of cluster Ci

$$nj \quad mi = \Sigma\, Xij \quad l = \text{-------------}\; nj$$

Where Xij is the ith data point in the jth group for i=1,2,…ni and j= 1,2,…k

This criteria tries to make resulting k clusters as compact and as separate as possible.

Problems with Partition Method

User has to specify the number of the clusters and initial cluster centers. Therefore it is very sensitive to initial selection of cluster centers. Accuracy varies according to initial cluster centers.

Problem Analysis

The basic problem in k-means algorithm is that, it not produces analytic solution. The K-means algorithm gave better results only when the initial partition was close to the final solution. The main problem in clustering algorithm is the object can have hundreds of attributes that have to be taken into consideration for clustering. One of the key issues is how to reduce this number to achieve an efficient algorithm. The main feature of clustering in a data mining application is the high number of objects that have to be clustered. Thus the processing time or the memory requirement of the algorithm can be huge, that has to be reduced using some heuristics. Validating the resulting clusters is also a hard task. In case of low dimensionality, when the clusters can be represented visually, the validation can be made by a human, but in case having large number of objects with high dimensionality statistical methods have to be used and indices have to be defined which can be computationally expensive.

Proposed Work

Industry have a great demand for scalable algorithms which not only works for large dataset but also works for complex data types such as mixed data objects. Proposed K-mean algorithm is efficient and scalable algorithm for large dataset with various attributes. We design and implemented a modified version of K-mean algorithm and evaluated its performance under two different implementation strategies for initial selection of mean. The proposed work Instead of initial centroids are selected randomly, for the stable cluster the initial centroids are determined systematically. It calculates the Euclidean distance between each data point and selects two data-points between which the distance is the shortest and form a data-point set which contains these two data-points, then we delete them from the population. Now find out nearest data point of this set and put it into new set. The numbers of elements in the set are decided by initial population and number of clusters systematically. These ways find the different sets of data points. Numbers of sets are depending on the value of k. After finding the initial centroids, it starts by forming the initial clusters based on the relative distance of each data-point from the initial centroids. These clusters are subsequently fine-tuned by using a heuristic approach, thereby improving the efficiency.

Conclusion

This procedure is based on the optimization formulation and a novel iterative method. According to the above numerical experiment results, the proposed method is an effective Clustering based method. It can be applied to many different kinds of clustering problems or combined with some other data mining techniques for getting more promising results for applications. . It guarantees that our proposed algorithm never generate empty cluster. From experiment we observe that proposed algorithm give more accuracy for dense dataset rather than sparse dataset
Our clustering algorithm serves as a good benchmark to monitor the progression of student's performance in institute. If So, even the method of selecting initial cluster described in the proposed method is good enough to use when considering both the performance and the execution time.

References

[1]. Ray S, Pakhira MK. Clustering of Scale Free Networks Using a K-medoid Framework. International Conference on Computer and Communication Technology (ICCCT)-2009.

[2]. Fahim AM, Salem AM. "Efficient enhanced k-means clustering algorithm", Journal of Zhejiang University Science, 2006;1626 – 1633,.

[3]. Friedrich Leisch, Bettina Gr un. "Extending Standard Cluster Algorithms to Allow for Group Constraints", Compstat 2006, Proceeding in Computational Statistics, Physicaverlag, Heidelberg, Germany,2006

[4]. Velmurugan T. Santhanam T. "Computational Complexity between K-Mean and K-Medoids Clustering Algorithms for Normal and Uniform Distribution of Data Points", Journal of Computer Science. 2008;6(3):363-368,.

[5]. Raut Shital A, Sathe SR. A Modified Fastmap K-Means Clustering Algorithm for Large Scale Gene Expression datasets. International Journal of Bioscience, Biochemistry and Bioinformatics, 2009;1(4).

[6]. Kumar P, SiriKrishanwasan. Comparative Analysis of K-mean Based Alogorithms", International Journal of Computer Science and Network Security. 2007;10 (4).

[7]. Fahim AM, Saake G, Salem AM, Torkeyand FA, Ramadan MA. K-Means for Spherical Clusters with Large Variance in Sizes", world Academy of Science Engineering and Technology 2008;45.

[8]. Dechang Pi, Xiaolin Qin and Qiang Wang, "Fuzzy Clustering Algorithm Based on Tree for Association Rules", International Journal of Information Technology. 2006;12(3).