An Effective Procedure for Mining Frequent Item sets From a Massive Data Set Using Clustering Technique

Devilal Birla
Research scholar
Mewar University Chittorgarh (Raj)
devbirla@gmail.com

Abstract: Recurrent item set mining has been a heart preferred theme for data mining researchers for over a span. A large amount of nonfiction has been dedicated to this research and marvelous progress has been made, ranging from efficient and ascendable algorithms for frequent item set mining in transaction databases to numerous research frontiers, such as sequential pattern mining, structured pattern mining, correlation mining, associative classification, and frequent pattern-based clustering, as well as their broad applications. In this paper, we develop a new technique for more efficient frequent item set mining. Our method scans the database only one time whereas the previous algorithms scan the database more than one time. In this way our proposed algorithm will reduce the complexity of frequent pattern mining. We present efficient techniques to implement the new approach.

1. Introduction

Clustering is a method of alignment similar type of data. Clustering are two types Partitioned clustering and Hierarchical clustering. Further classifications are show. Clustering Algorithms

1. Partitioning Methods
   I. Relocation Algorithms
   II.Probabilistic Clustering
   III.K-medoids Methods IV.K-means Methods V.Density-Based Algorithms a. Density-Based Connectivity Clustering
   b. Density Functions Clustering
2. Hierarchical Methods
   I. Agglomerative Algorithms
   II. Divisive Algorithms
3. Grid-Based Methods
4. Methods Based on Co-Occurrence of Categorical Data
5. Constraint-Based Clustering
6. Clustering Algorithms Used in Machine Learning
   I. Gradient Descent and Artificial Neural N/Networks
   II. Evolutionary Methods
7. Scalable Clustering Algorithms

8. Algorithms For High Dimensional Data

I. Subspace Clustering

II. Projection Techniques

III. Co-Clustering Techniques

The data clustering is an unsupervised clustering for finding classes, concepts, or group of patterns automatically. This is very useful method applied in various applications. Various researchers have proposed different methods to achieve clustering. K-means is one of them; it is a part of partitions clustering method. Kmeans is a very common clustering algorithm used in a variety of grouping applications.

Few different datasets have been used to prove the concept of time reduction in novel approach.

2. Literature survey:

2.1 Introduction to the Background and Related Works

The chapter deals with relevant literatures review of various topics that fall under data or object clustering. The first part discusses traditional clustering notations and formulations, different similarity criteria, quality measures to assess the clustering solutions, and lastly some of the well-known clustering algorithms along with their computational difficulty. The unsupervised clustering is an unlabeled collection of documents. The plan is to cluster the documents without additional knowledge or involvement so that documents within a cluster are more similar than documents between clusters. Traditional clustering techniques can be categorized into two major groups as partitioned and hierarchical.

2. Recomputed centroids of newly assembled groups [8] and Berkhin [12] stated that online k-means performs better than the batch edition in the field of content document collections. Primarily k documents from the quantity are selected randomly as the initial centroids. Then, iteratively documents are assigned to their nearest centroids and centroids are updated incrementally, i.e., after the assignment of a document to its nearest centroids. When no reassignments of documents occur the calculation stops.

2.2 Data Clustering Concept

The clustering task is to divide a dataset into meaningful collections or clusters so that points within a cluster are similar to one another (high intra cluster similarity) but differ from points in other clusters (low inter cluster similarity). The theme has been surveyed extensively under various disciplines in the past three decades.

1. The data clustering problem can be formulated dataset of n objects, each having dimension d,

2. The dataset is partitioned into divisions or clusters.

3. The quality of the produced clusters is evaluated using different outer and inner feature events.

2.3 K-means Clustering Algorithm

The classical k-means algorithm is measured as an effective clustering algorithm in producing good clustering results for many practical applications [1], [4], [7]. The algorithm is an iterative procedure and requires the number of clusters k to be given a priori. The initial partitioning is randomly generated, that is, the centroids are randomly initialized to some points in the area of the space. K-means divides the dataset into k non overlapping areas recognized by their centroids based on objective function condition where objects are assigned to the closest centroids. The most widely used objective function condition is the distance condition. This section describes the original k-means clustering algorithm. Euclidean distance is generally considered to determine the distance between data points and the centroids. When all the points are included in some clusters, the first step is completed and an early grouping is done.

Step 1: Select objects randomly. These objects represent initial group centroids. K

Step 2: Assign each object to the group that has the closest centroid.

Step 3: When all objects have been assigned, recalculate the positions of the centroids. k

Step 4: Repeat Steps 2 and 3 until the centroids no longer move.

The k-means algorithm is the most extensively studied clustering algorithm and is generally effective in producing good results. The major drawback of this algorithm is that it produces different clusters for different sets of values of the initial centroids. Quality of the final clusters heavily depends on the selection of the initial centroids. The 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.

Algorithm 1: The k-means clustering algorithm

Input:

D = {d1, d2...dn} //set of n data items.

k // Number of desired clusters

Output:

A set of k clusters.

Steps:

1. Randomly or arbitrarily choose k data items from D as initial centroids;
2. Repeat Assign each item di to the cluster which has the closest centroid;
3. Calculate new mean for each cluster;
4. Until convergence criteria is met.

K-means algorithms have clustering datasets properly and fine the centroids according to cluster As shown in Algorithm 1, the original kmeans algorithm consists of two phases: One

for determining the initial centroids and the other for assigning data points to the nearest clusters and then recalculating the cluster means. The second phase is accepted out repetitively until the clusters get stabilized, i.e., data points stop crossing over cluster boundaries or no move any object.

2.3.1 Given Approach by Mr. Fang Yuan

Mr. Fang Yuan have [11] proposed a systematic method for finding the initial centroids. The centroids obtained by this method are consistent with the allocation of data. Hence it produced clusters with better accuracy, compared to the original k-means algorithm. However, Yuan's method does not suggest any improvement to the time complication of the k-means algorithm.

2.3.2 Given Approach by Mr. A M Fahim

Fahim[14] proposed an efficient method for assigning data points to clusters. The original k-means algorithm is computationally very expensive because each iteration computes the distances between data points and all the centroids. Fahim's approach makes use of two distance functions for this purpose- one similar to the k-means algorithm and another one based on a heuristics to reduce the number of distance calculations. But this method presumes that the initial centroids are determined randomly, as in the case of the original kmeans algorithm. Hence there is no guarantee for the accuracy of the final clusters.

2.3.3 Modified Approach by Mr. K. A. Abdul Nazeer and M. P. Sebastian

In the improved clustering method discussed in this paper, both the phases of the original k-means algorithm are modified to improve the accuracy and efficiency [9]. The improved method is outlined as Algorithm 2. Algorithm 2: The improved method Input:

D = {*d1, d2,.....,dn*} // set of n data items

k // Number of desired clusters

Output:

A set of k clusters.

Steps:

1. Set m = 1;
2. Compute the distance between each data point and all other data points in the set D;
3. Find the closest pair of data points from the set D and make a separate set for centroid. Another set of a data point set Am (1<= m <= k) which contains these two data points, Delete these two data points from the set D, because they separate for initialize centroid;
4. Find the data point in D that is closest to the data point set Am, Add it to Am and delete it from D;
5. Repeat step 4 until the number of data points in Am reaches 0.75*(n/k);
6. If m<k, then m = m+1, find another pair of data points from D between which the distance is the shortest, form another data-

point set Am and delete them from D, Go to step 4;

7. For each data-point set Am (1<=m<=k) find the arithmetic mean of the vectors of data points in Am, these means will be the initial centroids.

In the first phase, the initial centroids are determined systematically so as to produce clusters with better accuracy [11]. The second phase makes use of a variant of the clustering method discussed by Fahim[14]. 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. Algorithm 2 describes the method for finding initial centroids of the clusters [11].

2.3.4 Example

For better understanding K-means algorithm tray to give a example. In this example consider the five objects or data points see in following Table 2.2.

Table 2.2 Objects and their Coordinates

Objects	X_i	X_j
A	2	4
B	8	2
C	9	3
D	1	5
E	8.5	1

In Table 2.2 five data points A, B, C, D and E, X_i and X_j are coordinates of the data points. Coordinates are showing the location of the data points. Calculate the distance between each data points. See in next table 2.3. Table: 2.3 Distance between Data Points

1. Suppose objects are A, B, C, D, E as a data points.

2. D = {6.325, 7.071, 1.414, 7.159, 1.414, 7.616, 1.118, 8.246, 2.062, 8.500}

3. Set Am = {1.118}, Am(1<=m<=k)

4. Set A1 = {1.118}

5. Closest to Am from D than add in Am and delete from D ,

6. Am = {1.118, 2.062}

7. D = {6.325, 7.071, 1.414, 7.159, 1.414, 7.616, 8.246, 8.500}

8. If m<k then m = m+1 find another pair of data points from D which distance is the shortest.

9. Each data points are set Am, according to k.

10. All data points of Am, these means will be the initial Centroids. The first step is determining the distance between each data-point and the initial centroids of all the clusters. The data-points

are then assigned to the clusters having the closest centroids. This results in an initial grouping of the data-points. For each data point, the cluster to which it is assigned and its distance from the centroids of the nearest cluster (nearest distance) is noted. Inclusion of data-points in various clusters may lead to a change in the values of the cluster centroids.

For each cluster, the centroids are recalculated by taking the mean of the values of its data-points. Up to this step, the procedure is almost similar to the original kmeans algorithm except that the initial centroids are computed systematically. On the other hand, if the new centroids of the present nearest cluster are more distant from the data-points than its previous centroids, there is a chance for the data-point getting included in another nearer cluster. In that case, it is required to determine the distance of the data-point from all the cluster centroids. The new nearest cluster needs to be identified and the new value of the nearest distance is recorded. The loop is repeated until no more data-points cross cluster boundaries, which indicates the convergence criterion. The heuristic method described above results in significant reduction in the number of computations and thus improves the efficiency.

Objects	A	B	C	D	E
A	0	6.325	7.071	1.414	7.159
B		0	1.414	7.616	1.118
C			0	8.246	2.062
D				0	8.500
E					0

6. Conclusion

The paper, discuss the K-means clustering algorithm using centroids method with the Random, Uniform and Minimum Distance Data Points of initial centroids. All the methods of initial centroids are based on the method criteria data points. It is tested on three different datasets with Euclidean distance. The K-means clustering is a common algorithm to find clusters in given dataset. Euclidean distance is a default calculating distance between two data points and objects for two or multidimensional objects. First the distance between each data points is calculated. After calculating distance, the Minimum distance of the data points is calculated and the initial centroids are selected according to clusters. The Minimum Distance Data Point method will more accurate and efficient than Uniform method or Random method. The execution time of all three methods is calculated for different datasets and variation in number of clusters, number of data points and number of dimension is noted.

References:

1. A. Jain, M. Murty, and P. Flynn, “Data Clustering: A Review,” ACM Computing Surveys, Vol. 31, pp 264- 323, 1999.

2. R. Xu, “Survey of Clustering Algorithms,” IEEE Transactions on Neural Networks, Vol. 16, Issue 3, pp 645-678, 2005.

3. R. Duda, D. Stork, and P. Hart, *Pattern Classification*, John Wiley, 2001.

4. J. Hartigan, *Clustering Algorithms*, John Wiley & Sons, New York, NY, 1975.

5. Z. Huang, "Extensions to the K-means Algorithm for Clustering Large Datasets with Categorical Values," Data Mining and Knowledge Discovery, Vol. 2, pp 283-304, 1998.

6. A. Pujari, *Data Mining Techniques*, Universities Press, 2001.

7. J. Hartigan, and M. Wong, "A K-means Clustering Algorithm," Applied Statistics, Vol. 28, pp 100-108, 1979.

8. M. Steinbach, G. Karypis, and V. Kumar, "A Comparison of Document Clustering Techniques," KDD Workshop on Text Mining, 1999.

9. K. Nazeer, and M. Sebastian, "Improving the Accuracy and Efficiency of the K-means Clustering Algorithm," Proceedings of the World Congress on Engineering, Vol. 1, 2009

10. J. Chaturvedi, and P. Green, "K-modes Clustering," Journal of Classification, Vol. 18, pp 35-55, 2001.

11. F. Yuan, Z. Meng, H. Zhang, and C. Dong, "A New Algorithm to Get the Initial Centroids," Proceedings of the 3rd International Conference on Machine Learning and Cybernetics, pp 26-29, 2004.

12. P. Berkhin, "Survey of Clustering DataMining Techniques," Research paper, Accrue Software, http://www.accrue.com/products/researchpapers.html, 2002.

13. E. Forgy, "Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classification," Biometrics, Vol. 21, pp 768-780, 1965.

14. A. Fahim, A. Salem, A. Torkey, and M. Ramadan, "An Efficient Enhanced Kmeans Clustering Algorithm," Journal of Zhejiang University, Vol. 10, pp 1626- 1633, 2006.

15. M. Dunham, *Data Mining Introductory and Advanced Topics*, Low Press, 2004.