Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Clustering Method for categorical and Numeric Data sets

Simmi Bagga[1]

[1] Sant Hira Dass Kanya Maha Vidyalaya, Kala Sanghian, Distt Kapurthala,Punjab,India

Received: 21 August 2011 Accepted: 19 September 2011 Published: 29 September 2011

Abstract

Many issues concerned with clustering process are due to large datasets involves. In clustering computation become expensive when there are large data sets involved and work efficiently when there is limited number of cluster with relatively small data set. This paper will present a new technique for clustering for large datasets. That will work efficiently equally with large data set as well as with small data sets. The main idea behind this method is to divide the whole process in two steps. The first step uses a cheap approximate distance measure that divide the data into overlapped subsets we call it stubs. Then in second step clustering is performed for measuring exact distances only between points that occur in common stubs. The stub based clustering approach reduces computation time over a traditional clustering and also increases its efficiency.

Index terms— Clustering, stubs, cetroid, K-Means Clustering.

1 I. INTRODUCTION

rom the last few years, the development of Information technology become very fast. So there has been widespread change in the adoption and utilization of new technologies in every field. That generated huge amount of data in the various fields. So there has been widespread change in the adoption and utilization of new technologies in every field that generated huge amount of data in the various fields.

To handle such large amount of data having complex structure effectively and efficiently for decision making, we use the concept of Data Mining. Data Mining is the process of identifying valid, useful and understandable patterns in data. There are various techniques have been developed and used in Data Mining including association, classification, clustering, prediction and sequential patterns etc. Clustering techniques are used for combining group that are similar to each other. Each cluster should be different from other clusters.

Clustering is one of the oldest and effective techniques of Data Mining.

Basically clustering algorithms are divided in to two types these are Partitional and Hierarchical. (a) Agglomerative algorithm starts with object as a separate cluster and then merges group according to a distance measure. These algorithm stops when all object converts in to single object or at the point where user wants to stop. These algorithms follow bottom-up merging. (b) Divisive algorithms works opposite to agglomerative strategy. They start with one group of all objects and then split groups into smaller ones. This process repeats until each object falls in one cluster, or according to the user desire. These types of algorithm follows top down merging. Further there are various categories of clustering algorithm. These categories are mainly focused on specific kind of data set or with some specific problems.

i. Density-Based Clustering: Density Based clustering algorithms group objects according to the functions that deal with density objectives. Density is defined as number of objects resides near the data objects. In this approach cluster grows longer as the density increases i.e. number of object in neighborhood increases. These are mainly hierarchical in nature.

ii. Grid-Based Clustering: Grid based Clustering algorithms deals with the spatial data i.e. about the objects related to space. Spatial data includes structure of objects in space, its relationships and properties. In these types of algorithm, we quantize data in to cells. Then we with only with those objects that's belongs to cell.

iii. Model-Based Clustering:

These algorithm deals with the approximations of model i.e. deals with the various parameters of the model that best fit the data. These types of methods can either be partitional or hierarchical depends on the iv. Categorical Data Clustering:

These algorithms are mainly developed for data where numerical-oriented, distance measures cannot be applied. These are very close to both partitional and hierarchical methods.

Clustering algorithms become computationally expensive when the data set is to be large. There are mainly three reasons in which the data set can be large:

? When large number of elements in the data set involved ? When data element can have many features.

? When many clusters to be discover from the data set.

Recent cluster algorithms have focused on the efficiency issues. K-means clustering algorithm is very efficient when we start by finding good initial points, but is not efficient when the size of cluster become large. There is no such algorithm works efficiently when any of above three kinds of data is there in data set because in that case we have millions of data element with many features and cluster.

In this paper we introduce a new clustering technique that work well even in the case of large data set and large clusters. The main idea of this new technique is to perform clustering in two steps. In first step, divide the data into overlapping subsets called stubs, then in final step in which expensive distance measured only among points that occur in a common stub. First step can be performed very quickly and roughly where as second step is little rigorous. During the first step we mainly built stub by using approximate distance measure, the second step can be performed by any standard algorithm of clustering. The computation time and complexity is saved by the approximate distance measure used to create stubs. This computation is saved by eliminating all of the distance comparisons among the data set. We have found small accuracy increases due to the usage of two different distance measures. Clustering based on stubs applied to many different standard clustering algorithms, like K-means, Greedy Agglomerative Clustering and Expectation-Maximization.

2 II STUB BASED CLUSTERING

The basic idea behind the stub based algorithm is that one can reduce the number of distance computations for clustering. The first step cheaply partitioning the data set into overlapping subsets called stubs and then in second step the distance is measuring only among the data points that belong to a common subset. The stub based technique uses two different sources to cluster items: a cheap and approximate similarity measure. For example, in first step we just measure approximate distances among the data set like proportional comparison between data set and calculate the similarities between the data set that will reduce the size of the data set up to some extent. This step can be performed cheaply and in little time. In the final step, the more accurate similarity measures are performed that is more expensive in nature. In this step detailed distance measurement is performed.

We divide this clustering process into two steps. In the first step we use the short distance measure to create some number of overlapping subsets, called stubs. These are calculated by the proportionate measure. Stub is just a subset of the data elements find by approximate measure of its similarities. These distances are the distance performed from a central point. A data element can appear in more than one subset or stub but every data element must belong to at least one stub. Stubs are created with the intention that data element appearing in common stub may be far different that they could not possibly be in the same cluster. The method used for calculating the distance to create stubs is approximate. There exist many overlapping stubs in the data set, because we choose a large enough distance to ensure each and every data element should belong to any stub. These stubs are just made by measuring approximate distance and it's a very cheap method of calculating and it reduces the size of data set.

In the second step, we perform traditional clustering algorithm on that filtered stubs, such as Kmeans, Greedy Agglomerative Clustering or by using any accurate distance measure algorithm. The main restriction impose in this method is that we do not calculate the distance between two points that do not belong to same stub. This restriction is imposed because we assume the distance between the two different stubs to be infinite. The expensive distance measurements will only be made between the same stubs. This is will overall reduce the number of calculation.

If the first step is not properly performed that is if stubs are not properly made then it degrade the performance of the second step also. So stubs should be created carefully. If stubs are not too large and do not overlap much, then we cannot avoid expensive computation for clustering. The constraints imposed on the clustering imposed by the stubs may not lose accuracy but will increase computation efficiently. If distance to a cluster is measured to the centroid of the cluster, then clustering accuracy will be preserved exactly.

For every cluster, there exists a set of stubs. Expensive distance measurements will only made between pairs of data points in the same stubs.

3 a) Creating Stubs

In the case of stub based clustering, user will be able to focus domain-specific features in order to design a short distance measures. User efficiently creates stubs using these measures. For example, if we have large data

of patients of number of hospitals that contain information of diagnosis, treatments and payment histories. A cheap measure calculates the similarity between diagnoses of the patients. Result might be 1, if they have the similarities and 0 if they do not have any similarity. In this case stub creation is small and the common diagnosis results fall in the same stub. If the same patient falls in multiple diagnoses then he will fall into multiple stubs and also some stubs will overlap. The small number features are sufficient to built stub, even if the data item may have thousands of features.

4 b) Cheap Distance Measurement

There are various methods to calculate the cheap distance measure. One of the methods for distance measure for text is based on the inverted index. An inverted index is in the form of sparse matrix in which, each word can directly access the list of documents containing that word. When we want to find all documents according to some, we need not to measure the distance to all documents, but only have to examine the list of documents associated with each word in the query. The documents which have no words in common with the query will never be considered. The inverted index can efficiently calculate a distance metric.

Using the above distance measure stub can creates as follows. Start with a list of the data items, and with two distance thresholds, lets say T1 and T2, where T1 > T2. Pick a data point from the list and approximately measure distance to all points. Put all similar data points to the distance threshold T1 into a stub. Remove from the list all points that are within distance threshold T2. Repeat until the list is empty. The inverted index can be applied to real-valued data item.

5 c) K-Means Stubs

One can also use the stubs idea to speed up prototype based clustering methods like K-means. K-Mean is well known partitioning based method of clustering. It is simple and iterative method works around one artificial point which represent the average location of the cluster is called Centroid.

This algorithm takes an input a number of clusters that is the k from. Means is an average location of all the members of a cluster. In this algorithm we have to partition n object set to k clusters. Cluster is measured on the basis of its mean or average location. The basic idea behind this algorithm is to first randomly select k of the object that is centroid of cluster. Using this k of clusters, we optimize intra cluster similarity and inter-cluster dissimilarity. Each remaining object, the most similar object is assigned to cluster based the distance between object and centroid or cluster mean.

Then we compute new mean. This process repeats until the whole function overages. This is an iterative method in which we always redefine the center point or centroid until cluster detection is finished.

This approach is basically based on prototypes that are associated with the stubs that contain them. The prototypes are only influenced by data inside the associated stubs. After creating the stubs, we decide how many prototypes will be created for each stub. Then we place prototypes into each stub. For each prototype, we find the stubs that contain in it (computed by using cheap distance measures) and then calculate distances from that prototype to points within those stubs.

K-means algorithm not gives just similar results for stub. In K-means each data point is assigned to a single prototype. As long as the cheap and expensive distance measures are sufficiently similar that the nearest prototype is within the boundaries of the stubs that contain that data point, then the same prototype will win.

6 III. COMPUTATIONAL COMPLEXITY

We can simply say that computational time is saving using stub based method. This technique has two step and mainly done two types of comparisons. One is relative fast step where stubs are created. The other one is a slow clustering process in which we apply K-Mean clustering algorithm. If we create stubs by using inverted index method then there is no need to perform pair wise distance comparison.

In the case of K-means or Expectation-Maximization, clustering without stubs requires O (nk) distance comparisons per iteration of clustering. Consider the K-mean method with stubs where each cluster belongs to one or more stubs. Assume that clusters have the same overlap factor f as data points do. Then, each cluster needs to compare itself to the fn/c points in f different stubs.

7 IV. CONCLUSION

Clustering large data sets having large cluster is a very tedious task. It is very expensive and inefficient to deal with large data set having large number of clusters using traditional clustering methods. The goal of this paper is to describe a new stub based method for clustering that takes relatively less computation time and performs result more effectively in the case of large data set. In this paper we described how can we create stubs and how can we apply traditional cluster methods like Kmean clustering method to perform effective clustering. [1] [2]

[1] © 2011 Global Journals Inc. (US) Global Journal of Computer Science and Technology Volume XI Issue XVIII Version I 50

[2] © 2011 Global Journals Inc. (US)


Global Journals Inc.
Global Journal of Computer Science & Technology

Figure 1:

[He et al. ()] 'An Efficient Algorithm for Clustering Categorical Data'. Z He , S Xu , Deng . *Journal of Computer Science and Technology* 2002. 17.

[Tajunisha ()] 'An efficient method to improve the clustering performance for high dimensional data by Principal Component Analysis and modified K-means'. Saravanan Tajunisha . *International Journal of Database Management Systems (IJDMS)* 2011. 3.

[Anderberg ()] M R Anderberg . *Cluster Analysis for Application*, 1973. Academic Press.

[Hartigan ()] *Clustering Algorithms*, John A Hartigan . 1975. New York: John Wiley.

[Grossman et al. ()] 'Data Mining Standards Initiatives'. Robert L Grossman , Hornick , . F Mark , Gregor Meyer . *Communications of the ACM* 2002. 45 (8) p. .

[Jiawei and Kamber ()] *Data Mining: Concepts and Techniques*, Han Jiawei , Micheline Kamber . 2001. Sanfransico, CA: Morgon Kaufmann.

[Singh and Simmi ()] 'Three Phase Iterative Model of KDD'. G N Singh , Bagga Simmi . *International Journal of Information Technology and Knowledge Management* 2011. 4 (2) p. .