ADVANCEMENT IN K-MEAN CLUSTERING ALGORITHM FOR DISTRIBUTED DATA

Garima Goel[1]
[1]NCCE/CSE Dept., Israna-Panipat, India
Email: garimagoel06@gmail.com

Abstract*— For business and real time applications large set of data sets are used to extract the unknown patterns which is termed as data mining approach. Clustering and classification algorithm are used to classify the unlabeled data from the large data set in a supervised and unsupervised manner. The assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in any way is termed as cluster analysis or clustering. Through these algorithms the inferences of the clustering process and its domain application competence is determined.the research work deals with K-means and MCL algorithms which are most delegated clustering algorithms.

***Keywords*— Data Mining, Cluster, K-Mean, LIC.**

1.INTRODUCTION

We are living in an era often referred to as the information age. As we believe that information leads to success and power, we have been collecting remarkable amounts of information and thanks to refined technologies such as computers, satellites, etc. for all this. . Initially, with the introduction of computers and resources for mass digital storage, we started collecting and storing all sorts of data in huge amount which contains supervised aswell as unsupervised data. Unfortunately, these enormous collections of data stored on unrelated structures rapidly became irresistible. This disorder has led to the creation of structured databases and database management systems (DBMS). For management of a huge corpus of data and especially for effectual and proficient retrieval of particular information from a large collection whenever needed becomes an important asset and an efficient database management system is required or becomes necessity.
Today from business transactions and scientific data, to satellite pictures, text reports and military intelligence, we have far more information than we can handle. For decision-making information retrieval is simply not enough anymore. Confronted with vast collections of data, now we have created new needs to help us make improved managerial choices. These requirements are automatic summarization of data, pulling out of the "essence" of information stored, and the detection of patterns in raw data.

2. PROBLEM STATEMENT

The k-means method has been shown to be effective in producing good clustering results for many practical applications. However, a direct algorithm of k-means method requires time proportional to the product of number of patterns and number of clusters per iteration. This is computationally very expensive especially for large datasets. The main disadvantage of the k-means algorithm is that a data point can exist in one cluster. Any two clusters can't share one common data point. In this dissertation we present a modifying algorithm that is combination of two algorithms that will show the membership of a data point within the clusters. The basic procedure involves producing all the segmented dataset for 2 clusters up to Kmax clusters, where Kmax represents an upper limit on the number of clusters. Then our measurement of membership function is calculated to determine which data point has how much membership that's value exists between 0 & 1. At o there is no membership but at value 1 there is membership with all clusters. We propose a modifying algorithm for implementing the k means method. Our algorithm produces the same or com.

3.PROPOSED METHODOLOGY

1. Firstly load the data in Matlab code. We will use some parameters to load the data in matlab.
2. After adding up of data we apply the division rule for dividing the data into small packets of the datasets.
3. Then we apply the k-mean clustering algorithm to the resulted packets of datasets formed. The k-mean clustering algorithm is the type of the algorithm of the data mining technique.
4. In data mining, *k*-means clustering is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean.
5. After Appling k-mean clustering algo. We will apply MCL algorithm too, through both clustering algorithms we can find the membership function.
6. After applying both algorithms, we get a membership function plot.

Through resulted membership function we get some graphs as results

4. CLUSTERING IN DATA MINING

We describe about the techniques used by data mining to get KDD. There are basically two methods of data mining Classification and Clustering but there we discuss only Clustering in data mining. We also describe the existing algorithm by combine any two of them we can generate the third algorithm that is called Modifying Algorithm.

Figure 1. Data Mining Cycle

Business Transactions: Every transaction in the business industry is (often) "memorized" for perpetuity. Such transactions are usually time related and can be inter-business deals such as purchases, exchanges, banking, stock, etc., or intra-business operations such as management of in-house wares and assets.

Scientific Data: Whether in a Swiss nuclear accelerator laboratory counting particles, in the Canadian forest studying readings from a grizzly bear radio collar, on a South Pole iceberg gathering data about oceanic activity, or in an American university investigating human psychology, our society is amassing colossal amounts of scientific data that need to be analyzed.

Medical and Personal Data: From government census to personnel and customer files, very large collections of information are continuously gathered about individuals and groups. Governments, companies and organizations such as hospitals, are stockpiling very important quantities of personal data to help them manage human resources, better understand a market, or simply assist clientele. Regardless of the privacy issues this type of data often reveals, this information is collected, used and even shared.

Surveillance Video and Pictures: With the amazing collapse of video camera prices, video cameras are becoming ubiquitous. Video tapes from surveillance cameras are usually recycled and thus the content is lost. However, there is a tendency today to store the tapes and even digitize them for future use and analysis

Satellite Sensing: There are a countless number of satellites around the globe: some are geo-stationary above a region, and some are orbiting around the Earth, but all are sending a non-stop stream of data to the surface. NASA, which controls a large number of satellites, receives more data every second than what all NASA researchers and engineers can cope with. Many satellite pictures and data are made public as soon as they are received in the hopes that other researchers can analyze them.

Text Reports and Memos (e-mail Messages): Most of the communications within and between companies or research organizations or even private people, are based on reports and memos in textual forms often exchanged by e-mail. These messages are regularly stored in digital form for future use and reference creating formidable digital libraries.

The World Wide Web Repositories: Since the inception of the World Wide Web in 1993, documents of all sorts of formats, content and description have been collected and inter-connected with hyperlinks making it the largest repository of data ever built. Despite its dynamic and unstructured nature, its heterogeneous characteristic, and its very often redundancy and inconsistency, the World Wide Web is the most important data collection regularly used for reference because of the broad variety of topics covered and the infinite contributions of resources and publishers.

5.ALGORITHM

There are several major *data mining procedures* have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns. We will briefly examine those data mining techniques with example to have a good overview of them. Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in *market basket analysis* to identify what products that customers frequently purchase together. Based on this data businesses can have corresponding marketing campaign to sell more products to make more profit. The main techniques that we will discuss here are the ones that are used 99.9% of the time on existing business problems. There are certainly many other ones as well as proprietary techniques from particular vendors - but in general the industry is converging to those techniques that work consistently and are understandable and explainable. One of the most popular heuristics for solving the k-means problem is based on a simple iterative scheme for finding a locally minimal solution. This algorithm is often called the k-means algorithm. There are a number of variants to this algorithm, so, to clarify which version we are using, we will refer to it as Lloyd's algorithm. (More accurately, it should be called the generalized Lloyd's algorithm since Lloyd's original result was for scalar data.).

Genetic algorithms (GA's) work on a coding of the parameter set over which the search has to be performed, rather than the parameters themselves. These encoded parameters are called *solutions* or *chromosomes* and the objective function value at a solution is the objective function value at the corresponding parameters. GA's solve optimization problems using a population of a fixed number, called the *population size*, of solutions.

The main disadvantage of the k-means algorithm is that the number of clusters, *K*, must be supplied as a parameter. In this dissertation we present a simple validity measure based on the intra-cluster and inter-cluster distance measures which allows the number of clusters to be

determined automatically. The basic procedure involves producing all the segmented dataset for 2 clusters up to *Kmax* clusters, where *Kmax* represents an upper limit on the number of clusters. Then our validity measure is calculated to determine which is the best clustering by finding the minimum value for our measure. The validity measure is tested for LIC dataset for which the number of clusters in known.

6.THE MODIFYING ALGORITHM

In this section, we describe the modifying algorithm. As mentioned earlier, the algorithm is based on storing the multidimensional data points in a kd-tree. For completeness, we summarize the basic elements of this data structure. Define a box to be an axis-aligned hyper-rectangle. The bounding box of a point set is the smallest box containing all the points. A kd-tree is a binary tree, which represents a hierarchical subdivision of the point set's bounding box using axis aligned splitting hyper planes. Each node of the kd-tree is associated with a closed box, called a cell. The root's cell is the bounding box of the point set. If the cell contains at most one point (or, more generally, fewer than some small constant), then it is declared to be a leaf. Otherwise, it is split into two hyper rectangles by an axis-orthogonal hyper plane. The points of the cell are then partitioned to one side or the other of this hyper plane. (Points lying on the hyper plane can be placed on either side.) The resulting sub cells are the children of the original cell, thus leading to a binary tree structure. There are a number of ways to select the splitting hyper plane. One simple way is to split orthogonally to the longest side of the cell through the median coordinate of the associated points. Given n points, this produces a tree with on nodes and O.log n. depth.

7. RESULTS AND ANALYSIS

7.1. RESULT

STEP 1 Run MATLAB Platform and browse the folder where we have put the code file.

STEP 2 Select the code file which wants to run and press F5 button to run the program. We can also run by using Debug tab in menu bar and then click on Run.

STEP 3 After run the program we have to select the dataset on which we want to implement algorithm. The dataset is classified by colouring the different colour to different columns as described below:

- This colour shows the Policy Number in Dataset.
- This colour shows the Initial Name in Dataset.
- This colour shows the Premium Amount in Dataset.
- This colour shows the Age in Dataset.
- This colour shows the Agent Code in Dataset.

STEP 4 MF Plot for dataset 1 before implement algorithm.

STEP 5 Click on Start button and we get the 5 clusters which represent the similar data in different -2 groups as shown in figure 2.

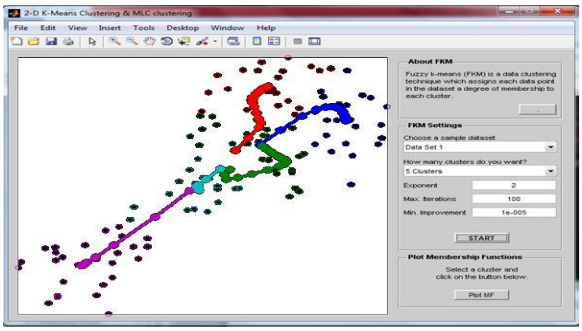

Figure 2. CLUSTER ON THE BASIS OF SIMILAR PROPERTIES

STEP 6 Select any one cluster by click any one and MF Plot generate after implementation of algorithm as shown in figure 3.

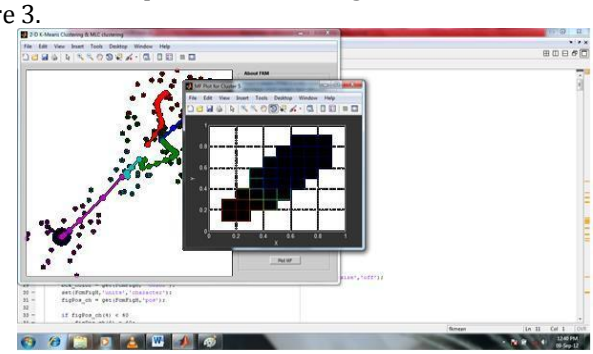

Figure 3. MF PLOT IS GENERATED FOR THE CLUSTER WHICH REPERSENT POLICY NUMBER

7.2. ACCURACY

Accuracy of cluster 5 is described by plot the graph time Deviation and Data Accuracy as shown in figure 4.

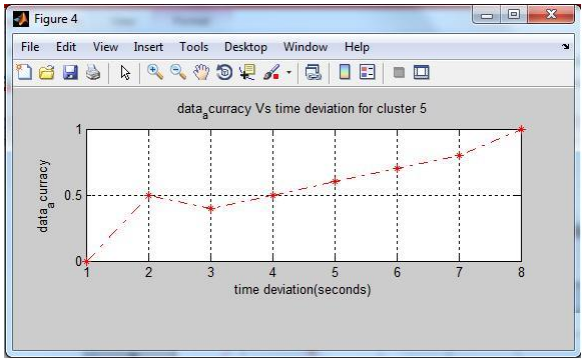

Figure 4. Accuracy V/S Time Deviation for Cluster 5

Data Accuracy	0.0	0.5	0.4	0.5	0.6	0.7	0.8	1
Time Deviation (seconds)	1	2	3	4	5	6	7	8

TABLE 1: Accuracy Vs Time Deviation for cluster 5

8. CONCLUSION

From all the above calculations we come to the conclusion that the K-Mean algorithm is an excellent algorithm when we are dealing with a small or medium sized data. It simply provides good performance vector every time. A direct algorithm of k-means method requires time proportional to the product of number of patterns and number of clusters per iteration. This is computationally very expensive especially for large datasets. The main disadvantage of the k-means algorithm is that the number of clusters, *K*, must be supplied as a parameters.So, to overcome from this we have made a modified algorithm which is combination of K-means and MCL.

9. REFERENCES

[1] *Ben-Hur, D. Horn, H.T. Siegelmann, and V. Vapnik. "A support vector clustering method". In International Conference on Pattern Recognition, 2000.*

[2] *Ahmed S, Coenen F, Leng PH "Tree-based partitioning of date for association rule mining". Knowl Inf Syst 10(3):315–331, (2006).*

[3] *Banerjee A, Merugu S, Dhillon I, Ghosh J "Clustering with Bregman divergences". J Mach Learn Res 6:1705–1749, (2005).*

[4] *Bonchi F, Lucchese C "On condensed representations of constrained frequent patterns". Knowl Inf Syst 9(2):180–201, (2006).*

[5] *Breiman L "Addison-Wesley, Reading. Republished in Classics of mathematics". SIAM, Philadelphia, (1991).*

[6] *Breiman L "Prediction games and arcing classifiers". Neural Comput 11(7):1493–1517, (1999).*

[7] *Breiman L, Friedman JH, Olshen RA, Stone CJ "Classification and regression trees". Wadsworth, Belmont, (1984).*

[8] *Johannes Grabmeier and Andreas Rudolph "Techniques of Cluster Algorithms in Data Mining" Received November 12, 1998; Revised May 23, 2001*

[9] *U. Fayyad, G.Piatetsky-Shapiro and P.Smyth. From data mining to knowledge discovery in databases.Ai Magazine, Volume 17, pages 37-54, 1996.*

[10] *In Year 1997, Pavel BerkhinHo performed a work,"Survey of Clustering Data Mining Techniques"*

[11] *In Year 2001, Glenn Fung performed a work,"A Comprehensive Overview of Basic Clustering Algorithms".*