International Journal of Mechatronics, Electrical and Computer Technology

Dec. 2013, Special Number, pp 493-503, **ISSN**: 2305-0543

Available online at: http://www.aeuso.org

© Austrian E-Journals of Universal Scientific Organization

Identification of Mazandaran Telecommunication Company Fixed phone subscribers using H-Means and W-K-Means Algorithm

Yaser Babagoli Ahangar[1*], Homayon Motameni[2] and Ramzanali Abasnejad Varzi[3]

[1]Islamic azad university sari branch, Iran

[2]University of Tehran, Science and Research branch, Iran

[3]Science and Technology University, Tehran, Iran

*Corresponding Author's E-mail: ybabagoli@yahoo.com

Abstract

In recent years, companies' interaction with customers has changed significantly. Organizations require a proper understanding of customers and their needs in order to succeed in business. Customer Relationship Management (CRM) using data mining techniques, to discover hidden and valuable knowledge in data and information of organizations, provide the conditions for optimal management of customer relationships, as organizations never lose opportunity for more selling and better customer satisfaction as earning customer loyalty, as well as improve their profitability. One method of customer identification is customer clustering approach. This operation is used when we want to find groups of similar data without having a prediction on pre-existing similarities. This paper proposes a model based on H-Means and W-K-Means algorithm that customers are compared with the optimal number of clusters and clustering results are evaluated in terms of quality.

Keywords: Data mining, clustering, K-Means algorithms

1. Introduction

Customer relationship management allows organizations to better understand their customers and to better understand the differences between them. Companies generally suffer from some of their customers, thus for the same purpose, they have always tried to adopt ways to prevent these problems. However, in many cases it is too late to act prudently and causes the company to be faced with a lot of bills which are not paid because some customers avoid paying the bills and thereby cause a loss of a significant amount of income and capital of the company. One of the methods to identify the customer is customer segmentation approach. Segmentation divides customers into homogeneous clusters. The purpose of clustering is to predict the well-paid and bad-paid customers of the company. So, this problem should be changed to a classification issue in order to classify each customer in a two mentioned groups. One of these segmentation methods is done using different statistical techniques, and data mining. Data mining techniques is one of the fundamental approaches in detecting of customers demography that can be used to obtain a wide range of purposes in different industries.

2. Data mining

There are different definitions of data mining, but the definition below is the most common references to "data mining and knowledge discovery from hidden patterns in a large and complex databases. "Data mining methodology is very strong and with high potential to help organizations so that the most important information in their data warehouses must focus. Data mining helps organizations to explore data on their own systems, patterns, future trends and behaviors discovered and predicted to make a better decision. Over the past decade, data mining algorithms are fast growing exponentially, but the evolution of business data, information and knowledge are the following steps in the table below.

Table1: Steps in the evolution of data mining

Evolution Steps	Business inquiries	Employed technologies	Features
Data collection	How many our total revenue was in the past five years?	Computer – Disks-Types	Analysis of past Data
Data access	How many units were sold in New Gland in march?	Relevant databases	Dynamic analysis of past data at a level
Data Warehousing	How many units have been sold in new Gland on march (in comparison with Boston)?	Multidimensional databases data storage	Dynamic analysis of past data in several Level

The main components of data mining such as statistics, artificial intelligence and machine learning are developing for many years. Today, the development of these techniques, along with high-performance database engines, technology has made this a very practical and effective for data storage environments. If pyramid data is to consider in the following way:

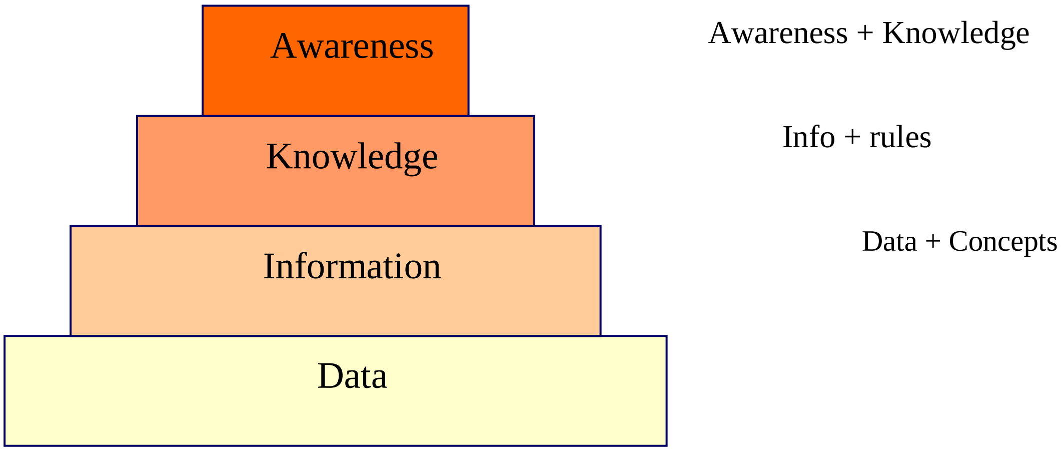

Figure1: Data pyramid

Knowing the definition of data mining and looking up the mentioned pyramid, the role of data mining in companies can be better derived. Data mining, which is caused to reach higher levels of knowledge and the unknown patterns in the organization?

3. What is data clustering?

Clustering is considered often the first and most important steps in the data analysis. Clustering is one branch of unsupervised learning and an automatic process during which the samples are divided into categories whose members are similar to the categories that are called clusters. Therefore, cluster is a collection of objects which are identical with each other and dissimilar with members of objects in other clusters (collections). For similarity, different criteria can be considered for clustering such as distance and Objects that are closer together as a cluster, consider that this type of clustering, called clustering-based distance. For example, the clustering of the input samples shown on the left are divided into four clusters, a similar figure to the right. In this example, each input sample belongs to one of the clusters and there is no sample belonged to more than one cluster.

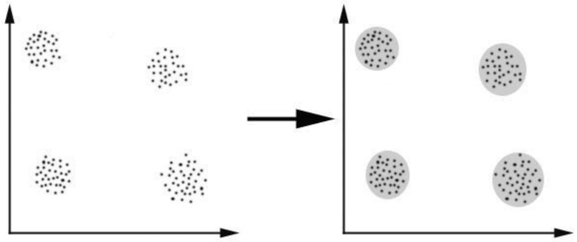

Figure2: samples clustering of Input 1

As another example, consider the following figure. Each of the small circles in the shape of a vehicle (object) that the weight and speed characteristics are specified. Put a circle around each cluster and the cluster label of each ellipse represents. The coordinates of the samples are shown in the feature called feature space. As you can see in Figure, vehicles are divided into three clusters. Each of these clusters can be regarded as a representative. For example, we calculate the average freight vehicles and cargo vehicles can be introduced as representative of the cluster. In fact, clustering algorithms are often so basic that a representative sample is considered as the input samples representative of the degree of similarity, is determined by the clustering of which belongs and After this step, the new representatives of the samples belonging to a cluster is calculated again and representative samples are compared to determine the cluster to which they belong and is repeated until the clusters do not change agents. The purpose of clustering is to find clusters of similar objects in the input samples.

Dec. 2013, Special Number, pp 493-503, **ISSN**: 2305-0543
Available online at: http://www.aeuso.org
© Austrian E-Journals of Universal Scientific Organization

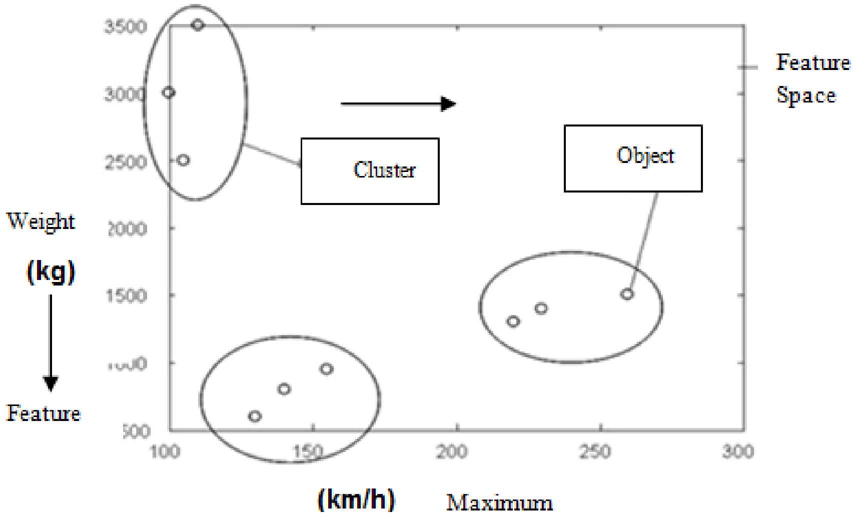

Figure3: Vehicle Clustering

4. Partitioning methods

Suppose a set of n have not been classified. A method of partitioning is making the data into k partitions such that each partition has at least one data and k ≤ n. If a partition is kind of hard, data can only be a member of a cluster, and in a fuzzy partitioning, a data can be used in a cluster (with different degrees of membership). The famous algorithm for hard partitioning method is W-k-Means and H-Means algorithm. In the algorithm, the mean of each cluster is shown by the mean values of the data which is contained in. These algorithms are suitable for few or moderate amounts of data. For clustering high-volume data using partitioning methods, such algorithms must be developed.

5. Clustering algorithm

All of Clustering Algorithm are repetitive routine for a fixed number of clusters, the following estimates are predicted:

- to gain points as cluster centers: in fact, the mean of the points is belonged to each cluster.
- assigning each data to a cluster: it might have been the least distance to the cluster center.

In the simple kind of this method, the required number of clusters is randomly chosen as points. Then, data regarding the proximity (similarity) is attributed to one of these clusters, thus resulting relatively new clusters.

By repeating this procedure, new centers can be calculated in each repetition by data averaging and re-attributing data to new clusters. This process continues until the change in the data is not obtained. The objective function is as follows.

$$J = Min \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} \quad K_j \right\|^2 \qquad (1)$$

Measure the distance between points "‖ ‖" and K_j is the j-th cluster center.

The beneath algorithm is considered for this method:

1) K points are selected as the initial cluster centers.
2) Each data sample is attributed to the cluster that has the least distance to the data.
3) After all data belonging to one of the clusters is calculated for each cluster, a new point is calculated as the center (Average points belonging to each cluster).
4) Steps 2 and 3 are repeated until the cluster centers do not change.

6. W-K-Means clustering algorithm

• In the W-K-Means algorithm, the data are based on the weight. It means that if our data is x1, x2, ..., xn , then they gain the weights of w1, w2, ..., wn. Clustering quality will be greatly affected during this approach with initialized weights. After data preparation, data weight is designed to provide more information for W-K-Means algorithms to improve accuracy. In the algorithm W-K-Means, average is calculated as follows:

$$\bar{\mathbf{x}} = \frac{\sum \boldsymbol{Wi*Xi}}{\boldsymbol{n}} \qquad (2)$$

Where Xi and Wi are input data and weight of each of the input data respectively. Cluster center is determined by the formula and the data are placed in the cluster that is less than the distance to the cluster center. Other steps are repeated.

7. H-Means Clustering Algorithm

In the harmonic mean clustering algorithm, K-Means algorithm is performed by work steps in which they are calculated as the mean of cluster centers as follows:{ We then split the data into two clusters using the following formula to calculate the cluster center.}

$$\bar{\mathbf{x}} = \frac{\boldsymbol{n}}{\sum \frac{\mathbf{1}}{\boldsymbol{Xi}}} \qquad (3)$$

8. Introduction to Data Collection

It is Obvious that in any data mining project, gaining valuable and useful results, depend on a rich data. If the data does not contain interesting features, the results will not be very useful and reliable. However, since the data mining technology is new and the vast majority of industries and company executives are not familiar enough with, as a result, they will refuse to access for data mining, due to security concerns.

These difficulties in obtaining data initially created. The data used in this thesis is related to the bills of fixed telephone subscriber of Mazandaran Telecommunication Company in 1391 (solar). The data records are 6569336 involving 1320113 Customer payment information related to fixed phones in the province of Mazandaran in seven phases. Data was as text and EmEditor software is used for data observation. Due to the large volume of data, Notepad or WordPad software are not practical.

• Each record includes information such as the date fixed telephone billing, bill payment history, amount of issued bill and the bank code which the bill is paid for.

• Initial examination of the data showed that, although having more characteristics in common, such as age, educational level, joint, joint-type (commercial, residential, etc.) could obtain better results, however, present data is suitable for data mining. The history records are also in an acceptable level.

9. Implementation of W-K-Means algorithm

- First, the data are divided into four clusters, and then the formula will calculate the mean of each cluster as the cluster center:

$$\bar{\mathbf{x}} = \frac{\sum \boldsymbol{Wi*Xi}}{\boldsymbol{n}} \tag{4}$$

C # source code

```
foreach (DataRow row in resultp2)
            {
               sum_price2 += Convert.ToDouble(row[2]) * Convert.ToDouble(row[1]);
               sum_date2 += Convert.ToDouble(row[1]);
               month_No++;
            }
            sum_price2 = sum_price2/sum_date2;
```

program 1

After calculating cluster centers, the distance of each data to clusters centers is calculated and then the data will be placed in a cluster that have minimal distance to its center. This process will continue until the data does not move between clusters.

10. Results of K-Means implementation

Table2: The first phase of the K-Means algorithm based on amount of bill

Input Data	The area of Clustering payment	Number of data
P=1320112	35000<p1<250000	103860
	250000<p2<1550000	1082
	1550000<p3<8000000	41
	1000<p4<35000	1215129

Data is placed in 4 clusters that are shown in the second column. In the third column, the number of data is shown in each cluster. WK-Means algorithm given in payment for the data values held to achieve better results and because a late paying amount equal to bill payment value was provided therefore only one level of payment will be clustered into four groups.

Note that the second column shows the range of paying in four clusters and the third column represents the number of data in each cluster, we can conclude that more than 90 % of payments is between 1000 and 35000 Tomans.

11. Implementation of H-Means algorithm

• First, the data are divided into four clusters, and then the formula will calculate the mean of each cluster as the cluster center:

$$\bar{\mathbf{x}} = \frac{n}{\sum \frac{1}{Xi}} \qquad (5)$$

C # source code

```
foreach (DataRow row in resultp1)
            {
               sum_price1 += 1 / (Convert.ToDouble(row[2]) + 1);
               sum_date1 += Convert.ToDouble(row[1]);
               month_No++;
            }
            sum_price1 = month_No / sum_price1;
```

program 2

After calculating cluster centers, the distance of each data to clusters centers is calculated and then the data will be placed in a cluster that have minimal distance to its center. This process will continue until the data does not move between clusters.

12. Results of H-Means implementation

Table3: The first phase of the H-Means and H-Means algorithm based on amount of bill

Input Data	The area of Clustering payment	Number of data
P=1320112	373870	50000<p1<8000000
	946242	1000<p2<50000

We then split the data into two clusters using the formula to calculate the cluster center.

Both P1 and P2 cluster achieved using H-Means algorithm. P1 data is the amount of paid bills between 50000 and 8000000 Tm as well as the number of data in this cluster is equal to 373870 customer. Cluster P2 is related to the amount of bill payment data between Tm 1,000 to 50,000 Tm as well as the number of subscribers (customers) is 946,242. In the next step, we repeat the process again except this time, clustering is based on the time difference of bill payment and bill issuing.

Table4: The second stage of the H-Means algorithm based on the payment term

P1=373870	0<p11<20	201290
	20<p12<346	172580
P2=946242	0<p21<20	517760
	20<p22<346	428482

first column of each payment based on paid money, has two categories that clustered on late payments, displayed in the second column and the number of customers in each cluster is shown in the third column.

Results and discussion

Considering the clustered data, the following conclusions can be reached regarding clusters of fixed telephone subscribers of Telecommunication Company of Mazandaran province.

1- 92/042 % of subscriber's bill is less than Tm 35,000 in each period.

2- 65/086 % of subscribers that amount is less than Tm 28,000 and paid their bills on time and considered as well-paid subscriber.

3- Only 0/003 % of subscribers are considered as Bad-payer.

4- 64/705% of subscribers, who their bills are over 1750000 Tm, pay their bills during 50 days.

5- 0/26 % of consumer's phone bill is more than 140,000 Tm.

ACKNOWLEDGEMENTS

The professors of Islamic Azad University, Sari, Mazandaran University of Science and Technology and the University of Noshirvani Babylon Telecom province appreciation and thank

Reference

[1] M. Berry, G. Linoff,, John Wiley and Sons, "Data Mining Techniques: For Marketing, Sales, and Customer Support", pp. 123-154, 1997

[2] A. Berson, S. Smith, K. Thearling, McGraw-Hill, "Building Data Mining Applications for CRM", pp. 86-94, 1999

[3] S. B. Kotsiantis, P. E. Pintelas، "Recent Advanced in Clustering: A Brief Survey", WSEAS Transactions on Information Science and Applications 1, pp. 73-81، 2004.

[4] H. Ian, F. Eibe, Morgan Kaufmann: "Data mining practical machine learning tools and techniques", pp. 213-225, 2005