Decision Making to Predict Customer Preferences in Life Insurance

Brunela Karamani
Department of Computer Engineering, Polytechnic University of Tirana, Albania

Abstract

Large amounts of data are generated everyday by different system of the modern world. For this competitive world it is very important to extract useful hidden knowledge in these data and more important is to act on the extracted knowledge. The large amount of data requires clustering for collecting only useful dataset. In this paper we have used k-means clustering algorithm on life insurance data set for cluster information and subsequent analysis to predict customer preferences towards life insurance product. Firstly we have presented an overview of k-mean algorithm and then we have analyzed our case of study. WEKA workbench framework is used to display the result of the processed data. We have concluded this paper by outlining the results and conclusion of our work.

Keywords: data mining, clustering, k-means algorithm, data set, life insurance, framework.

1. Introduction

The business world of nowadays is fast and dynamic in nature. It involves a lot of data gathered from different sources and that often are stored in huge Data Warehouses. Transforming these data into useful information called knowledge is the most challenging task for business people. Data mining techniques are used to give a good solution to this situation, which aim to make business able to achieve the above task. According to Kantardzic 2011, the main tasks in the data mining are:

- Classification-discovery of a predictive learning function that will serve later to classify data item into one of several predefined classes
- Regression-discovery of a predictive learning function, which is necessary to map a data item to a real-value prediction variable.
- Clustering is a descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data.
- Summarization is another descriptive task that includes methods for finding compact description for a set (or subset) of data.
- Dependency modelling – means finding a local model that describes significant dependencies between variables.
- Change and deviation detection-discovering the most significant changes in the data.

As data are growing from day to day, the common problem of all is to utilize and understand the large, complex information of huge data sets. Prediction and classification are the two main objectives of data mining. During the prediction process it is possible to predict unknown or future values by using some variables in the data set; in the end of the procedures the model of prediction is returned. During the classification process it is possible to find pattern by describing the data than can be interpreted by the human being. Clustering and outlier processes are required in data mining in order to execute the above processes. In this paper we have used the K-means clustering technique of data mining to describe the data of an insurance life company in Albania. Initially we have presented an overview of the K-means algorithm logic, and then we have explained our case of study in which it is implemented the k-means clustering algorithm. WEKA workbench is used to display the result of the experiments. In the end we have analyzed the conclusions of our work.

2. K-mean algorithm

Clustering is a well-known data mining techniques to find useful pattern from a data in a large database. These patterns are very important for the knowledge workers such as financial analyst, managers, or other similar workers to take right managerial decisions. A simple clustering technique is K-means clustering which is one of the most famous clustering algorithms applied in different types of fields such as Medicine, Psychiatry, Sociology, Criminology, Geology, Market research, etc. This clustering techniques use the K-

means algorithm basis. Initially K initial centroids are chosen, where K is a parameter that represents the number of clusters desired. Then each point is assigned to the closest centroid, and the cluster is composed by each collection of points assigned to the centroid. Based on the points assigned to the cluster each cluster updates its centroid. The process of assignment and update are repeated until no point changes clusters, or until the centroids remain the same. K-means is described by the algorithm below:

K-Means Algorithm
1: Select K points as initial centroids
2: repeat
3: Form K clusters by assigning each point to its closest centroid
4: Recompute the centroid of each cluster
5: until Centroids do not change

The initial centroids for the K-means may be chosen according to a rule or in a random way. When random initialization of centroids is used, different runs of K-means produce different SSEs. The key step of the K-means procedure is to choose the proper initial. The common approach is to choose the initial centroids randomly, but the resulting clusters are often poor. In order to assign a point to the closest centroid, it is needed a measure that quantifies the notion of "closest" for the specific data under consideration. The K-means algorithm calculates the similarity of each point to each centroid; as a result the similarity measures used for K-means are relatively simple.
"Recompute the centroid of each cluster" is the other step of the algorithm. Since the centroid can vary depending on the proximity measure for the data and the goal of the clustering, it is needed this step in order to have a good clustering classification of the data. Once we have the proximity measure and an objective function, the centroid we should choose can often be determined mathematically.
Time and space complexity is another element that should be taken in consideration when working with K-means. Actually the time requirements of K-means are pretty modest, because only the data and the centroids are saved during the process.
The storage required is $O((m + K)n)$ where:

m- the number of points
n- the number of attributes

The time required is also modest, and is often calculated as $O(I*K*m*n)$ where:
I-the number of iteration required for the convergence that are usually a small number
Usually the most typical changes occur during the first little iteration; therefore K-means is linear in m and is efficient as well as simple provided that K, is significantly less than *m*.
Besides the goods sides, K-means has a problem when handling empty clusters. Empty clusters can be obtained if no points are allocated to a cluster during the assignment step. To solve this situation a strategy is needed to choose a replacement centroid. One solution may be to choose the point that is farthest away from any current centroid, or another solution may be to choose the replacement centroid from the cluster that has the highest SSE.

3. Case Study

A portfolio database in a life insurance company contains a set of life insurance product purchased by customers. A life insurance company's funds are collected by way of premiums, where every premium represents a risk that is covered by that premium. The source dataset was extracted from an insurance company database in Albania.
The database was 14'180 records covering a time period of one year (1st January to 31st December 2011), where each record contains 34 fields of data (for example fields about personal data customer id, name, surname, gender, address, city, state, marital status). The first task was to remove from the database those variables which were irrelevant to the task at hand (for example office at life insurance policy was taken, name and surname of insurer). The second task is the process of performing various transformations on data (for example the birth date to age). Simple k-means algorithm is applied on life insurance data set for cluster information and subsequent analysis to predict customer preferences towards life insurance product. We considered 10 relevant attributes for analysis. The 10 attributes of the dataset for the experiment after preprocessing taken for analysis are shows below:

	Name	Value type	Full Name	Descriptions
Attribute 1	age	numerical	age of person	values from 19 to 65
Attribute 2	gender	qualitative	gender of person	1-male and 2-female
Attribute 3	marital	qualitative	marital status	1-single; 2-married; 3-other
Attribute 4	loads	qualitative	insurance history	1-yes 2-no
Attribute 5	no.persons	qualitative	No of persons in charge	{1, 2, 3, 4, 5, 6}
Attribute 6	volume	numerical	amount of loan to be insured	
Attribute 7	maturity	qualitative	Duration in month	2-up to 60 months; 3-from 60-120 months; 4-over 120 months
Attribute 8	occupation	qualitative	occupation of insurer	1-skilled employee; 2-official 3-Self employment 4-unemployment
Attribute 9	risk	qualitative	Risk profession of insurer	1-low 2 medium 3-high
Attribute 10	product	qualitative	Product Type	1-First Category; 2 Second Category; 3-Third category; 4 -Fourth Category

Figure 1 The 10 attributes of the dataset

The WEKA -"Waikato Environment for Knowledge Analysis" tool has been considered for the purpose of

analysis and test results and it is used for data mining. Data mining finds valuable information hidden in large volumes of data. WEKA is a collection of machine learning algorithms for data mining tasks, written in Java and it contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

The key features of WEKA are it is open source and platform independent. After the pre-process tools we choose and modify the data being acted on. In the figure 2 below we have show the correlation between the class "product" and other attributes {age, gender, marital, loads, number of persons, maturity, occupation and risk}.

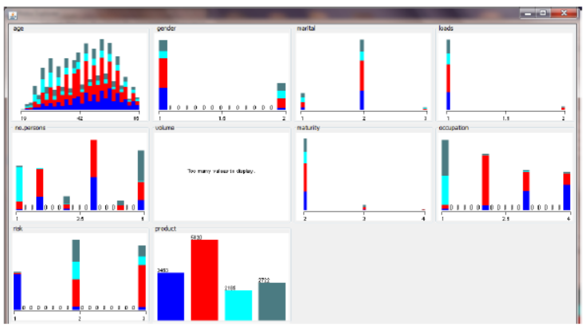

Figure 2 The correlation between the class "product" and other attributes

Results of simple – k means clustering on dataset is given in figure 3.

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Relation: lifeinsurance2

Instances: 14180

Attributes: 10

{age; gender; marital; loads; no.persons; volume; maturity; occupation; risk; product}

Test mode: evaluate on training data

=== Model and evaluation on training set ===

kMeans

======

Number of iterations: 18

Within cluster sum of squared errors: 30392.038654916567

Missing values globally replaced with mean/mode

Cluster centroids: Cluster#

Attribute	Full Data (14180)	0 (0.12)	1 (0.23)	2 (0.32)	3 (0.12)	4 (0.21)
age	43	39	47	45	40	45
gender	1.278	1.3965	1.0011	1	1.1443	2
marital	1.8496	1.6835	1.9047	1.8462	1.7106	1.9634
loads	1.0401	1.0247	1.0007	1.0792	1	1.0545
no.persons	3.5781	1.0008	3.79	3.64	5.40	3.63
volume	25,000	30,000	50,000	40,000	30,000	35,000
maturity	2.0813	2.0006	2	2.14	2.14	2.0901
occupation	2.2187	1.0106	3.3969	2.4996	1.01	1.8622
risk	2.1554	2.2659	1.2107	2.5836	2.30	2.3877
product	II category	II Category	I category	III Category	II Category	II Category

Time taken to build model (full training data): 406.24 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	2036 (14%)
1	3923 (28%)
2	3965 (28%)
3	2481 (17%)
4	1775 (13%)

Figure 3 Results of simple – k means clustering

The first column gives you the overall population centroid. The other five columns give the centroids for cluster 0 to 4 respectively. Each row gives the centroid coordinate for the specific dimension.

In cluster 0 have 2,036 records or 14% of data set and represents customers with male gender, average age 39, marital status married, number of persons in charge 1, with insurance history, prefer the product type second category with amount 25'000 euro, duration up to 10 years and risk classification medium.

In cluster 1 have 3,923 records or 28% of data set and represents customers with male gender, average age 47, marital status married, number of persons in charge 4, with insurance history, prefer the product type first category with amount 50'000 euro and duration up to 10 years and risk classification low.

In cluster 2 have 3,965 records or 28% of data set and represents customers with male gender, average age 45, marital status married, number of persons in charge 4, with insurance history, prefer the product type third category with amount 40'000 euro and duration up to 10 years and risk classification high.

In cluster 3 have 2,481 records or 17% of data set and represents customers with male gender, average age 40, marital status married, number of persons in charge 5, with insurance history, prefer the product type second category with amount 30'000 euro and duration up to 10 years and risk classification medium.

In cluster 4 have 1,775 records or 13% of data set and represents customers with female gender, average age 45, marital status married, number of persons in charge 4, with insurance history, prefer the product type first category with amount 35'000 euro and duration up to 10 years and risk classification medium.

The clustering model shows the centroid of each cluster and statistics on the number and percentage of instances assigned to different clusters.

Cluster centroids are the mean vectors for each cluster and can be used to characterize the clusters. Numbers are the average value of everyone in the cluster.

Each cluster shows us a type of behavior in customers, from which conclusions are drawn.

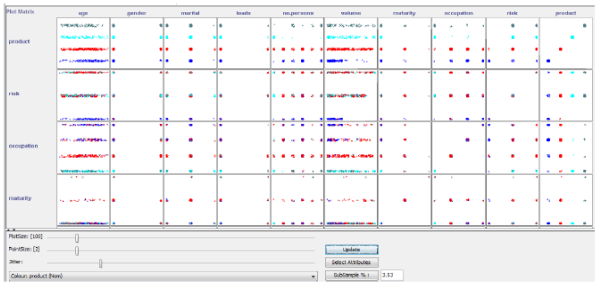

Figure 4 The distribution of five type of product for each cluster

For the chosen data set the cluster number as the x-axis, the instance number as the y-axis and the "product" attribute as the colour dimension. This will result in a visualization of the distribution of five type of product for each cluster. The observed results as clusters 0, 3 and 4 are dominated by second category of product type, while clusters 1 and 2 are dominated by first and third category respectively. In this case, by changing the type of colour dimension to other attributes, their distribution within each of the clusters can be observed.

4. Conclusion

Analysis of insured behaviour enables companies to improve support of their customer oriented processes, in order to improve the overall performance of the life insurance company.

K-means algorithm methodology has enormous contribution for researchers to extract the hidden knowledge and information. In this paper we have used k-means clustering algorithm to identify the type of product based preferences towards life insurance product in addition to other attributes.

We presented the segmentation of customer generated by clustering model, and then we analyzed and explained the results of the processed data, and we have displayed the results using the WEKA framework.

5. References

[1] Koteeswaran, S., P. Visu and J. Janet A Review on Clustering and Outlier Analysis Techniques in Datamining American Journal of Applied Sciences 9 (2): 254-258, 2012 ISSN 1546-9239

[2] Yanchang Zhao R and Data Mining: Examples and Case Studies http://www.RDataMining.com April 2013

[3] Mahendiran, N. Saravanan, N. Venkata Subramanian and N. Sairam Implementation of K-Means Clustering in Cloud Computing Environment Research Journal of Applied Sciences, Engineering and Technology 4(10): 1391-1394, 2012, ISSN: 2040-7467

[4] Swasti Singhal, Monika Jena A Study on WEKA Tool for Data Preprocessing, Classification and Clustering, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, May 2013

[5] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2nd , 2006

[6] Weka: http://www.cs.waikato.ac.nz/~ml/weka/

[7] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Morgan Kaufmann, 2011

[8] M. Panda and M. Patra. A novel classification via clustering method for anomaly based network intrusion detection system.International Journal of Recent Trends in Engineering, 2:1–6, 2009.