20th International Research/Expert Conference
”Trends in the Development of Machinery and Associated Technology”
TMT 2016, Mediterranean Sea Cruising, 24th September - 1st October, 2016

DATA MINING IN CUSTOMER SEGMENTATION

Dražena Gašpar,
Mirela Mabić
Faculty of Economics – University of Mostar
Matice Hrvatske bb, Mostar
Bosnia and Herzegovina

Ivica Ćorić
“Hera” software company
Kralja Petra Krešimira – SPC Rondo, Mostar
Bosnia and Herzegovina

ABSTRACT

The paper analyzes the possibilities of applying data mining techniques to improve customer segmentation. The aim of customer segmentation is to identify and profile lead, average and other company customers and to optimize and tailor future marketing actions so the right message can reach the right customer. Different techniques and models are applied in customer segmentation, like factor and cluster analysis in analyzing company customer data. In order to get required benefits from large data volumes stored in databases or data warehouses and to find hidden relationships between data, authors used cluster analysis, one of the data mining techniques. Data mining is a process of extracting previously unknown and potentially useful and hidden patterns from large databases. The main objective of the paper is to identify the high/medium/low-profit, high/medium/low-value and low/medium/high-risk customers by one of the data mining technique - customer clustering. It presents results of empirical research related to data mining in customer segmentation made in a production company which produces and distributes products like dry fruits, nuts, seeds and cereals for the market of South-East Europe.

Keywords: customer segmentation, data mining, cluster analysis

1. INTRODUCTION

The global economic, social and technological changes forced companies to evolve from product/service-centred strategies to customer-centred strategies. This shift of paradigm put customers at the first place and makes them the most important property of an organisation. In that situation organisations become aware that there cannot be any business prospects without satisfied and loyal customers. The organisations invest a lot of efforts and resources in order to understand each customer individually and use that to make it easier for the customer to do business with them rather than with competitors.

But, the main question is, are they all customers worth such effort? Researchers have tried different methods to calculate the value of individual customers in the hopes of ranking individual clients or segments or even to predict future values, such as can be found in the works of Verhoef and Donkers [13], Hwang et al. [6], Venkatesan and Kumar [11], Gupta and Lehmann [4], Kim et al. [9], Khajvand et al. [8] and Han et al. [5].

All that researches show that a key for a successful business is proper identification of high-profit, low-risk customers, retaining those customers and bring the customers from the lower to higher level group or segment.

Customer segmentation and clustering are two of the most valuable data mining techniques that should be used in organisations and their marketing departments. They use customer data stored in databases or data warehouses to divide customers into segments based on variables as current customer profitability, some measure of risk, a measure of the lifetime value of a customer, and retention probability.

Data mining is considered the most important step in the knowledge discovery process. Data mining is the process of extracting interesting patterns from large amounts of data [14].

This paper focuses on the topic of customer segmentation using data mining techniques. It presents results of empirical research related to data mining in customer segmentation made in a production company which produces and distributes products like dry fruits, nuts, seeds and cereals for the market of South-East Europe. The main task of the research was to identify the high/medium/low-profit, high/medium/low-value and low/medium/high-risk customers by using the data mining technique - customer clustering.

2. CUSTOMER SEGMENTATION AND DATA MINING

Last decades show that data mining is becoming very powerful marketing tool. Data mining, as the term is used in this paper, is the exploration and analysis of large quantities of data (mainly stored in data warehouses) in order to discover meaningful patterns and rules [1]. The most organizations today gather hundreds of terabytes of data from and about their customers. But, the main question is what those organizations learn about their customers from that data? Unfortunately, they often learn very little. Just gathering data is not enough. Data mining techniques offer to the organizations to add intelligence to their data. These techniques enable organizations to exploit the huge amount of existing data about their customers and to get answers on questions like: who are loyal customers, what products to offer to which customers, who are the high/low profitable customers, where to open new shop, what new product/service to offer to their customers and so on.

Clustering is one of the data mining techniques that perform segmentation of heterogeneous customers into a number of more homogeneous subgroups or clusters. Market segmentation is one of the most fundamental strategic planning and marketing concepts wherein grouping of customers is done under different categories such as the keenness, purchasing capability, profitability and interest to buy [7]. As opposed to classification, clustering does not rely on predefined classes. The records are grouped together on the basis of self-similarity [1].

Clustering techniques can be classified, based on the logic used for deriving clusters, into following groups: partition techniques, hierarchical techniques, density based techniques and grid techniques.

Partition techniques develop a subdivision of the given dataset into a predetermined number K of non-empty subsets. K-means and K-medoids are two of the best-known partition algorithms [12].

Hierarchical techniques carry out multiple subdivisions into subset, based on a tree structure and characterized by different homogeneity thresholds within each cluster. Hierarchical algorithms do not require the number of clusters to be predetermined. They can be subdivided into two main groups: agglomerative and divisive techniques [12].

Density-based techniques derive clusters from the number of observations locally falling in a neighbourhood of each observation [12]. Density based method are of two types: Density based Connectivity and Density based Functions. Density based Connectivity is related to training data point and DBSCAN and DBCLASD comes under this while Density Functions is related to data points to computing density functions defined over the underlying attribute space and DENCLUE comes under this [10].

Grid techniques first derive a discretization of the space of the observations, obtaining a grid structure consisting of cells. Subsequent clustering operations are developed with respect to grid structure [12]. The most popular grid algorithms are: STING, OptiGrid, GRIDCLUS, GDILC and WaveCluster.

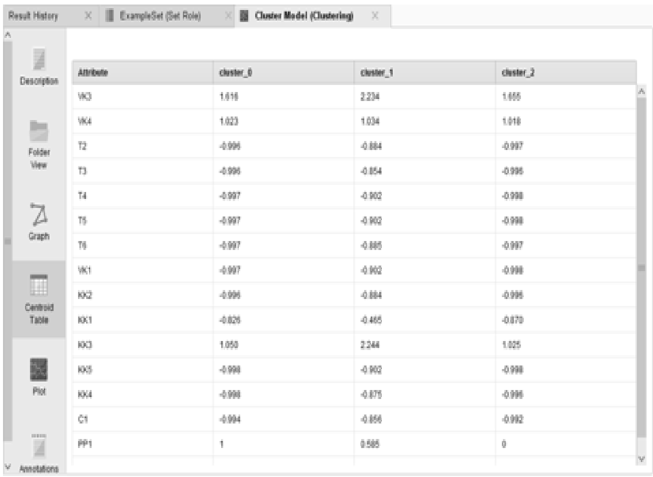
In this paper is used partition technique and K-means algorithm, one of the most used clustering algorithms. The “K” in its name refers to the fact that algorithm looks for a fixed number of clusters which are defined in terms of proximity of data points to each other [1].

3. CLUSTERING – SEGMENTATION OF CUSTOMERS

The K-means clustering based on customer profitability was tested using empirical data from a company which produces and distributes products such as dry fruits, nuts, seeds and cereals for the

South-East European market. The data source is a data warehouse generated by the company for the period of 2012 to 2015. The following variables (attributes) were used in K-means clustering: year to which data refers, customers to which data refers, customer size (1-small, 2-medium, 3-big), customer origin (1-domestic, 2-EU, 3-non-EU), total costs of customer's order processing and fulfilment costs in observed year, total cost of shipping orders costs in observed year, total costs of purchase and warehousing costs for delivered products to customer in observed year, total costs of raising purchase orders to suppliers for delivered products to customer in observed year, total costs of receiving shipments from suppliers for delivered products to customer in observed year, total revenue realized by the customer in observed year, total number of deliveries to customer in observed year, total number of different places where products were delivered to customers in observed year, net margin realized in customer trade in observed year, total number of returns of good in customer trade in observed year, total value of returns of good in customer trade in observed year, total value of the discount given to customers in customer trade in observed year and the average time of delay in customer payments in observed year (1- +less than average, 2-more than average) [3].

As software tool was used Rapid Miner Studio version 7.1., an open source predictive analytics platform. The number of cluster was set at 3. The goal was to see what kind of 3 groups existed in data set related to customer profitability. The results of clustering process are shown at Figure 1, Figure 2, Figure 3 and Figure 4.



Attribute	cluster_0	cluster_1	cluster_2
VK3	1.616	2.234	1.655
VK4	1.023	1.034	1.018
T2	-0.996	-0.884	-0.997
T3	-0.996	-0.854	-0.996
T4	-0.997	-0.902	-0.998
T5	-0.997	-0.902	-0.998
T6	-0.997	-0.885	-0.997
VK1	-0.997	-0.902	-0.998
KK2	-0.996	-0.884	-0.996
KK1	-0.826	-0.465	-0.870
KK3	1.050	2.244	1.025
KK5	-0.998	-0.902	-0.998
KK4	-0.998	-0.875	-0.996
C1	-0.994	-0.856	-0.992
PP1	1	0.585	0

Figure 1.The centroid table

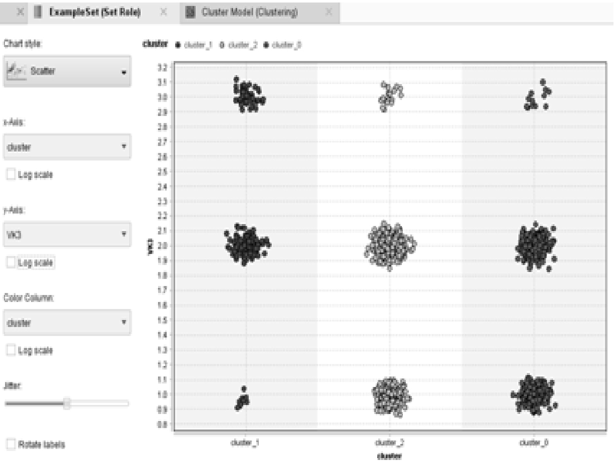
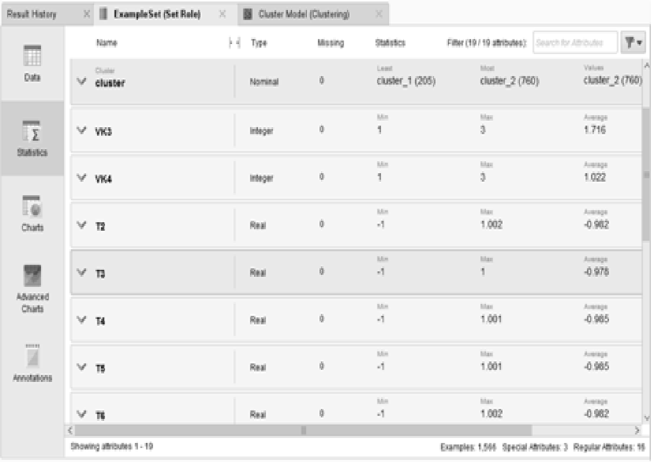


Figure 2. Graphical presentation of clusters



Name	Type	Missing	Statistics		
Cluster cluster	Nominal	0	Least cluster_1 (205)	Most cluster_2 (760)	Values cluster_2 (760)
VK3	Integer	0	Min 1	Max 3	Average 1.716
VK4	Integer	0	Min 1	Max 3	Average 1.022
T2	Real	0	Min -1	Max 1.002	Average -0.982
T3	Real	0	Min -1	Max 1	Average -0.978
T4	Real	0	Min -1	Max 1.001	Average -0.985
T5	Real	0	Min -1	Max 1.001	Average -0.985
T6	Real	0	Min -1	Max 1.002	Average -0.982

Showing attributes 1 - 19

Examples: 1,566 Special Attributes: 3 Regular Attributes: 16

Figure 3.Clustering statistic

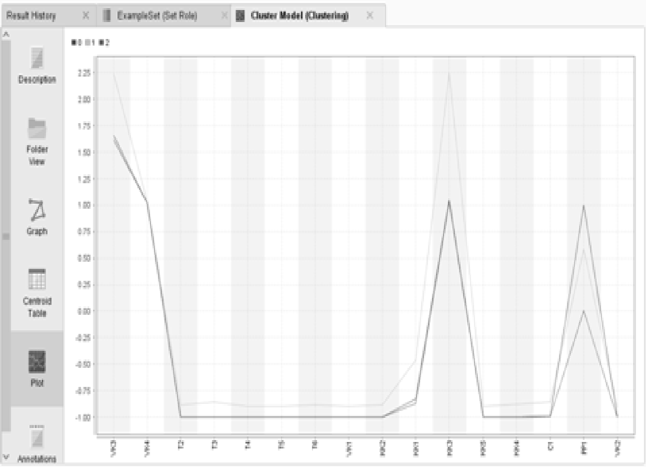


Figure 4.Clusters – plot view

The result of clustering was 601 customers in cluster 0, 205 customers in cluster 1 and 760 customers in cluster 2. The centroid table (Figure 1.) and plot view (Figure 4.) show the differences and similarities between clusters 0, 1 and 2. The cluster 1 consists of mostly medium companies (attribute VK3) which bought higher number of different products (attribute KK2) delivered on higher number of different places (attribute KK3) with lower average payment delay.

The main difference between clusters 0 and 2 is average payment delay (Figure 4). This three clusters corresponding to three segments defined as high level, medium and low-profit customers [2].

4. CONCLUSION

The K-means clustering, based on customer profitability, presented in this paper is a good starting point for future research of using data mining techniques in customer segmentation. The aim of future research should be testing the importance of used variables and finding possibilities for change/extend the set of variables. Namely, in presented clustering example, the differences between clusters were small, especially related to costs. The authors are aware of the necessity to test the model on a greater number of customers' data and on more similar companies, so their further research will go in that direction.

5. REFERENCES

[1] Berry, M.J.A, & Linoff,G.S. (2004). Data mining techniques: for marketing, sales, and customer relationship management, 2nd edition, Wiley Publishing, Inc., Indianapolis, Indiana, USA.

[2] Gašpar, D., Ćorić, I., & Mabić, M. (2015). Data Mining in Customer Profitability Analysis. Advances in Economics and Business 3(12), 552-559, 2015

[3] Gašpar, D., Markić, B. & Ćorić, I. (2012). Machine Learning in Customer Profitability Forecasting. 16 th International Research/Expert Conference "Trends in the Development of Machinery and Associated Technology" TMT 2012, Dubai, UAE, 10-12 September.

[4] Gupta,S., & Lehmann, D. (2006). Managing customers as investments. Upper Saddle River, NJ: Wharton School Publishing.

[5] Han,S.H.,Lu,S.X.,& Leung,S.C.H. (2012). Segmentation of telecom customers based on customer value by decision tree model. Expert Systems with Applications, 39, 3964–3973.

[6] Hwang, H., Jung,T.,& Suh,E.(2004). An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry. Expert Systems with Applications,26, 181–188.

[7] Kashwan,K.R., Velu,C.M. (2013). Customer Segmentation Using Clustering and Data Mining Techniques. International Journal of Computer Theory and Engineering, Vol. 5, No. 6.

[8] Khajvand, M., Zolfaghar,K., Ashoori,S.,& Alizadeh,S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behaviour: Case study. Procedia Computer Science, 3, 57–63.

[9] Kim,S.Y., Jung, T.S.,Suh,E.H.,&Hwang,H.S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. Expert Systems with Applications, 31, 101–107.

[10] Nagpal,P.B.& Mann,P.A. (2011). Comparative Study of Density based Clustering Algorithms. International Journal of Computer Applications (0975 – 8887), Volume 27– No.11.

[11] Venkatesan,R.,& Kumar,V. (2004). A customer lifetime value framework for customer selection and resource allocation strategy. Journal of Marketing, 68, 106–125.

[12] Vercellis, K. (2009). Business Intelligence: Data Mining and Optimization for Decision Making, John Wiley and Sons, USA.

[13] Verhoef,P.C.,& Donkers,B.(2001). Predicting customer potential value an application in the insurance industry. Decision Support Systems, 32, 189–199.

[14] Ziafat,H.& Shakeri,M. (2014). Using Data Mining Techniques in Customer Segmentation. Int. Journal of Engineering Research and Applications ISSN : 2248-9622, Vol. 4, Issue 9 (Version 3), pp.70-79.