# Data mining methods for Market Segmentation

Ildus Rizaev, Elza Takhavova and Zemfira Zakharova

Kazan National Research Technical University named after A.N. Tupolev-KAI, 10, K. Marx str., Kazan, 420111, Russian Federation
`isr4110@mail.ru`

**Abstract.** Data mining methods give opportunity to solve problems of current interest, market segmentation problem belongs to which. There are different approaches to solve market segmentation problem that differ by used methods. Data mining methods are used to solve classification and clustering problems. There are k-means method, EM algorithm and neural networks which are considered and compared. Deductor platform is used to analyse implementation of clustering algorithms.

**Keywords:** market segmentation, customer preferences, clustering, similarity measure, neural networks.

## 1 Introduction

Development of methods for recording and storing data has currently led to a rapid growth of the amount of collected information. The volumes of data are so impressive that it is simply not realistic for a person to cope with data analysis on their own and the problem to analyze collected information more efficiently is continuously increasing. All modern companies today have the ability to store data in a single database, called the customer base [1]. However, in addition to this, organizations pursue the goal of increasing profitability and reducing costs. The analysis of the customer base remains incomplete if customers are considered as similar people. This can be facilitated by highlighting certain preferences among clients [2-3]. To increase efficiency, it is needed to identify which customer groups exist and then figure out what actions will help attract more customers. To solve this problem, cluster analysis is used. In Data Mining, a common measure for assessing the proximity between objects is a metric or a way of specifying the distance. When using clustering algorithms, problems arise, since the same set of objects can be grouped into clusters in different ways. This led to choose among a large number of clustering algorithms [4-6].

With a large number of customers, it is difficult to build an individual approach, so it is convenient to group them into groups with homogeneous characteristics, which are

---

segments. Clustering can be used to segment and build customer profiles. The efficiency of working with clients increases by taking into account their personal preferences. Clustering can be used in a wide variety of areas; they are retail, banking, telecommunications, insurance, government services and others.

## 2 Materials and methods

Clustering differs from classification in that an output variable is not required and the number of clusters into which a set of data must be grouped may not be known. The output of clustering is not a ready-made answer. A cluster is a group of similar objects. Clustering indicates the similarity of objects which it includes. The resulting clusters require additional interpretation. To determine the similarity of objects, it is needed to set a measure of proximity. The most popular measure of proximity in two dimensions is the Cartesian distance or Minkowski metric [6-7].There are a lot of approaches to solve clustering problem.

The k-means algorithm is one of the simplest, but at the same time, not entirely accurate clustering method [7-8]. The goal of the method is to divide $m$ observations into $k$ clusters, with each object belonging to the cluster with the center (centroid) of which it is closest. In this method, the number of clusters is predefined. This method tends to minimize the total square deviation of cluster points from the centers of these clusters (1):

$$V = \sum_{i=1}^{k} \sum_{x \in S_i} (x - \mu_i)^2 \tag{1}$$

In (1) $k$ is the number of clusters, $Si$ are the resulting clusters, $i$ varies from 1 to $k$ and $\mu_i$ are centers of elements $x$ from cluster $S_i$. The algorithm is performed iteratively until the boundaries of the clusters and the location of the centroids change. The algorithm may take a dozen iterations to execute. The advantage of the algorithm is simplicity of implementation and speed of execution. The disadvantage is the need to initially set the number of clusters and select the initial mass centers.

Kohonen neural networks form a class of neural networks, the main element of which is the Kohonen layer, they are used for data analysis and solving clustering problems [9-10]. The self-organizing Kohonen map is a kind of neural network algorithms, characterized by the fact that it is taught without a teacher, the result depends only on the data structure. Either small random values can be used to initialize the weights, or based on an example of a training sample. The advantage of this method is that the network is trained without a teacher, the implementation is simple, and the corresponding answer is received after passing the data through the layers. The disadvantage of this method is working only with numerical data and the need to preliminary determine the number of clusters.

The EM (Expectation Maximization) algorithm is based, as it were, on maximizing expectations, in which it is assumed that the observation results are distributed according to the normal law in accordance with the Gaussian function [11]. In the EM algorithm, auxiliary hidden variables are introduced, on the basis of which the coefficients are recalculated in order to approximate the parameter vector to maximize the likelihood. Optimal parameters are found by a sequential iterative EM algorithm. This model consists of two steps. The first of them is E-step (Expectation), the values of the likelihood function are found on which. At the second step which is M-step (Maximization), the maximum likelihood estimate is found. The order of execution of the model can be represented as follows.

E-step. Basing on the current values of parameters (2), the vector of hidden variables is calculated ɣ (3).

$$\theta = (\pi_1, \ldots ; \pi_{\text{к}}, \mu_1, \ldots, \mu_{\text{к}} ; \Sigma_1, \ldots, \Sigma_{\text{к}}), \tag{2}$$

$$\gamma_{nk} = \frac{\pi_k N(x_n | \mu_k, \Sigma_k \quad )}{\sum_{j=1}^{K} \pi_j N(x_n | \mu_j, \Sigma_j \quad )} \tag{3}$$

M-step. Based on the current values of the hidden variables, the parameter vector is reevaluated according to (4).

$$\begin{gathered} \mu_k^{new} {=} \Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (x_n - \mu_k^{new})(x_n - \mu_k^{new})^T \\ \pi_k^{new} = \frac{N_k}{N} \\ N_k = \sum_{n=1}^{N} \gamma_{nk}, \delta_{max} > \Delta. \end{gathered} \tag{4}$$

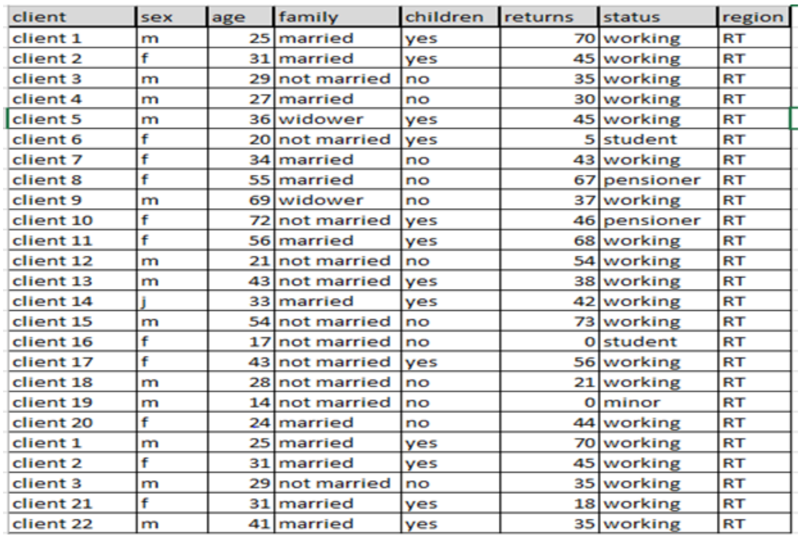
The procedure is stopped if the difference of the hidden variables does not exceed the specified constant (5). This model is based on the methods of mathematical statistics.

$$\delta_{max} = \max\{\delta_{max} | \gamma_{nk} - \gamma_{nk}^0\}. \tag{5}$$

Data mining software tools enable using and compare different approaches to implement and choose the most appropriate method.

## 3 Results

To test the clustering methods, information was prepared with the assignments of the initial data: input, descriptive and output data. The following features were selected as input data: gender, age, marital status, income, store category. For informative purposes full name and purchase amount were used. The rest of the data is highlighted as not being used. For the analysis, the Deductor Studio platform [12] was used, which allows for a comprehensive analysis of enterprise data, predict the indicators of its development, conduct segmentation and search for patterns [6-7]. Initial data is represented in Deductor Studio in the form of a table (Figure 1).



| client | sex | age | family | children | returns | status | region |
|---|---|---|---|---|---|---|---|
| client 1 | m | 25 | married | yes | 70 | working | RT |
| client 2 | f | 31 | married | yes | 45 | working | RT |
| client 3 | m | 29 | not married | no | 35 | working | RT |
| client 4 | m | 27 | married | no | 30 | working | RT |
| client 5 | m | 36 | widower | yes | 45 | working | RT |
| client 6 | f | 20 | not married | yes | 5 | student | RT |
| client 7 | f | 34 | married | no | 43 | working | RT |
| client 8 | f | 55 | married | no | 67 | pensioner | RT |
| client 9 | m | 69 | widower | no | 37 | working | RT |
| client 10 | f | 72 | not married | yes | 46 | pensioner | RT |
| client 11 | f | 56 | married | yes | 68 | working | RT |
| client 12 | m | 21 | not married | no | 54 | working | RT |
| client 13 | m | 43 | not married | yes | 38 | working | RT |
| client 14 | j | 33 | married | yes | 42 | working | RT |
| client 15 | m | 54 | not married | no | 73 | working | RT |
| client 16 | f | 17 | not married | no | 0 | student | RT |
| client 17 | f | 43 | not married | yes | 56 | working | RT |
| client 18 | m | 28 | not married | no | 21 | working | RT |
| client 19 | m | 14 | not married | no | 0 | minor | RT |
| client 20 | f | 24 | married | no | 44 | working | RT |
| client 1 | m | 25 | married | yes | 70 | working | RT |
| client 2 | f | 31 | married | yes | 45 | working | RT |
| client 3 | m | 29 | not married | no | 35 | working | RT |
| client 21 | f | 31 | married | yes | 18 | working | RT |
| client 22 | m | 41 | married | yes | 35 | working | RT |

**Fig. 1.** Source data table.

Segmentation analysis was performed using three k-means methods, Kohonen maps and EM clustering. When using the k-means method, the initial data was assigned as above was said: input, informative and output data. The choice of the number *k* can be based on theoretical justification or intuition. The number of clusters was set equal to 5. Result cluster profiles are shown in Figure 2. They display statistical information on clusters as a percentage. The data “category of stores” (this column in the table in the Figure 1 is not shown) includes such stores as grocery, construction, furniture, household, pharmacies, etc., a total of 51 records. Figure 2 shows this category in the highlighted part of the table on the right. In total, five clusters were identified by categories: shops, marital status, gender, income and age.

The main difference between Kohonen's self-organizing maps and the k-means method is that in it all neurons (class center and node) are ordered into some structure. Let's continue the example with clients for certain preferences. The same document for processing was selected. The same input and information data were leaved. The results of the execution of the algorithm are displayed on maps and a separate map is built for each input parameter. The convenience of the map (Figure 3) is that a user can click on the cluster number and see the location of the cluster on other maps that were built using the input values. For example, cluster 0 includes men, with an average age of 36 years, with marital status “married”, with an income of 50 thousand, and chose the following categories of stores: construction, food.

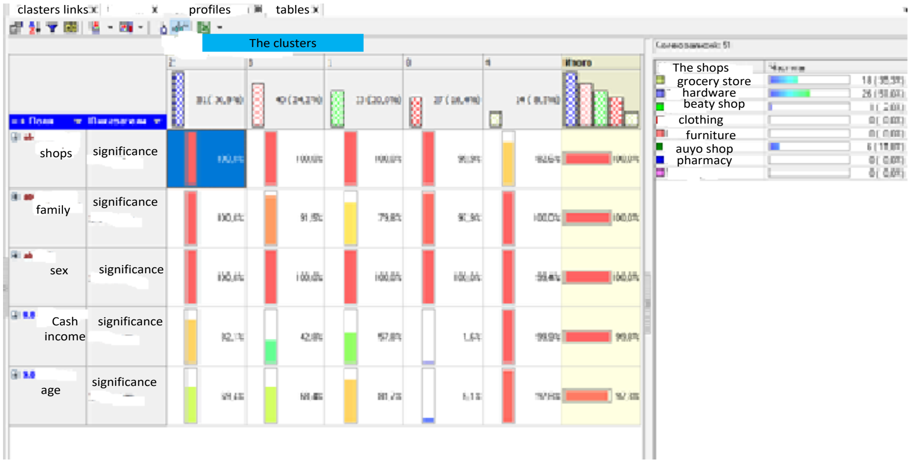


**Fig. 2.** Cluster profiles.

For EM clustering, the same sample, leaving the same input data and settings were chosen too. Clustering with this method allows user to choose automatic or fixed definition of clusters at the step of setting parameters. According to the connection of the clusters, it can be said, that the system has allocated the maximum number of clusters that were set.

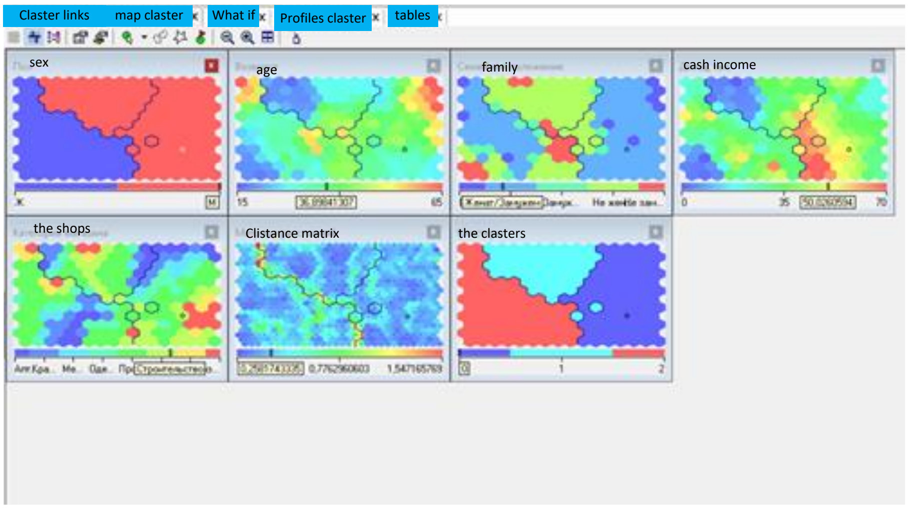
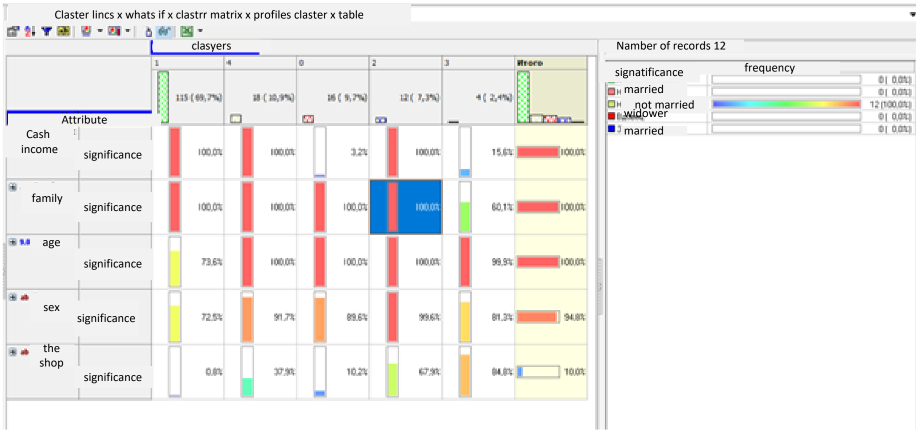


**Fig. 3.** Visual representation of the Kohonen map.

The choice of the method of initial initialization was made. If it is chosen at random, we get 5 clusters. We can see that cluster 1 includes both men and women with the status of married / married or widower aged 23 to 65, including all categories of stores (Figure 4). For example, cluster # 2 included 70% of women and 30% of men aged 16 to 21, single and unmarried, interested in such categories of stores as food, beauty, pharmacy, clothing, household.

Let's configure for a fixed number of clusters. The choice of the method of initial initialization of clusters is given. Let's choose it at random, we get 5 clusters. We can see that cluster 1 includes both men and women with the status of married / married or widower aged 23 to 65, including all categories of stores (Figure 4).

For example, cluster 2 included 70% of women and 30% of men aged 16 to 21, single and unmarried, interested in such categories of stores as food, beauty, pharmacy, clothing, household. When the method of initial initialization of clusters is chosen "from the training set", we get 5 clusters (Figure 4). In this case, cluster 2 includes unmarried women from 16 to 20 years old who are interested in the categories of the store: food, clothing, household and pharmacy.



Claster lincs x whats if x clastrr matrix x profiles claster x table

clasyers

| Attribute | | 1 | 4 | 0 | 2 | 3 | Итого |
|---|---|---|---|---|---|---|---|
| | | 115 (69,7%) | 18 (10,9%) | 16 (9,7%) | 12 (7,3%) | 4 (2,4%) | |
| Cash income | significance | 100,0% | 100,0% | 3,2% | 100,0% | 15,6% | 100,0% |
| family | significance | 100,0% | 100,0% | 100,0% | 100,0% | 60,1% | 100,0% |
| age | significance | 73,6% | 100,0% | 100,0% | 100,0% | 99,9% | 100,0% |
| sex | significance | 72,5% | 91,7% | 89,6% | 99,6% | 81,3% | 94,8% |
| the shop | significance | 0,8% | 37,9% | 10,2% | 67,9% | 84,8% | 10,0% |

Namber of records 12

| signatificance | frequency | |
|---|---|---|
| | | 0 ( 0,0%) |
| married | | 0 ( 0,0%) |
| not married | | 12 (100,0%) |
| widower | | 0 ( 0,0%) |
| married | | 0 ( 0,0%) |

**Fig. 4.** EM clustering. From the training set.

## 4 Discussion

Considered algorithms have fast cluster detection speed. However, the simplest for the user the clustering algorithm by Kohonen Maps became. In this algorithm, isn't needed to specify the number of clusters at the input, the number of clusters determining is provided by the method. Also, in addition, this method has its own special display method, in the form of colored maps. On the maps, it is convenient immediately to determine where which cluster is located and what data is included in this cluster. Also, this algorithm has good resistance to noisy data. Thus, with the help of Kohonen maps, the analyst can obtain more accurate segmentation results.

## 5 Conclusion

To choose the method on the context of solving market segmentation problem comparative analysis of three clustering methods was performed. K-means, EM-clustering method, and Kokhonen's neural networks method were considered and analyzed applied for market segmentation. Kokhonen's neural networks method was recommended as preferable because accuracy is the most critical requirement for the subject areas linked with marketing sphere. Databases information about customers gives opportunity to group clients by different preferences and take interests these groups into account when making decisions in marketing to improve business processes.

## References

1. Pavlov, B.P., Garifullin, R.F., Mingaleev, G.F., Babushkin, V.M.: Key technologies of digital economy in the Russian Federation. In: Proceedings of the 33rd International Business Information Management Association Conference, IBIMA 2019: Education Excellence and Innovation Management through Vision, 2020, 3401-3407 (2019).
2. Kotler Ph.: Fundamentals of Marketing. Short course Dialectics, Moscow; St. Petersburg (2019).
3. Kotler Ph., Keller K.L.: Marketing management.12 edn. Piter, Saint Petersburg (2006).
4. Barsegyan, A.A., Kupriyanov, M.S., Stepanenko, V.V., Kholod, I.I.: Data analysis technology: Data Mining, Visual Mining, Text Mining, OLAP. 2 edn. BHV-Petersburg, Saint Petersburg (2007).
5. Mandel, I.D.: Cluster analysis. Finance and statistics, Moscow (1988).
6. Ian, H. Witten, Eibe Frank, Mark, A. Hall.: Data Mining: Practical Machine Learning Tools and Techniques. 3rd edn, Morgan Kaufmann (2011).
7. Dyuk, V., Flegontov, A., Fomina, I.: Application of Data Mining technologies in the scientific, technical and humanitarian areas. Izvestia: Herzen university journal of humanities & sciences, Saint Petersburg: Russian state. A.I. Herzen Pedagogical University, 138, 77-84 (2011).
8. Mirkes E.M.: K-means and K-medoids: applet. University of Leicester (2011).
9. Kohonen, T.: Self-Organizing Maps: third extended edn: Springer-Verlag, New York (2001).
10. Wasserman, F.: Neural Computing. Theory and Practice: Mir, Moscow (1992).
11. Hastie T., Tibshiran, R., Friedman, J.: The The EM algorithm. The Elements of Statistical Learning: Springer, New York (2001).
12. Rizaev, I.S., Takhavova, E.G.: Solution of the Problem of Classification of Vehicles on the Basis of Statistical Estimates of Data. Proceedings 12th International Scientific and Technical Conference "Dynamics of Systems, Mechanisms and Machines", Dynamics 2018, 8601417 (2019).