European Journal of Accounting Auditing and Finance Research
Vol.5, No.7, Pp.30-64, July 2017
Published by European Centre for Research Training and Development UK (www.eajournals.org)

DATA MINING TECHNOLOGY AND ITS ROLE IN DISCOVERING FINANCIAL FRAUD

Yaser Saad Zenad[1], Jawad kadhim Shlaka[1] and Yaser Issam Hamodi[2]

[1]Department of Control and Internal Audit
[3]Department of Computer Engineering, Ministry of Higher Education & Scientific Research
Baghdad, Iraq

ABSTRACT: *The basis of any business - the customer database, which provides information about the client relationship with the company. The increasing complexity of organizational processes and rapidly changing business environment led to strong growth in domestic corporate data companies. In this regard, the increasing interest from the point of view of fraud risk assessments are beginning to provide tools such as data mining (Forensic Data Analytics - FDA), which allows you to narrow sample of suspicious transactions while minimizing the volume of checks. For example, in the field of communication in the database stores information about the conclusion of agreements for the use of services, the time of termination of the contract, a region rate, etc. The analysis revealed 7 out of 31 dentists who deliberately overstate the value of work performed by the insurance.*

K-means algorithm using the algorithm of k-means as 4 clusters formed:

- *Cluster 1: specialized work using expensive additional procedures, the average age of the client - 25, the average cost of services - $ 715;*
- *Cluster 2: minor works without the use of additional procedures, the average age of the client - 21, the average cost of services - $ 286;*
- *Cluster 3: Significant work using expensive additional procedures, the average age of the client - 38, the average cost of services - $ 819;*
- *Cluster 4: Significant work with cheap additional procedures, the average age of the client - 27, the average cost of services - $ 551.*

KEYWORDS: Data Mining, Financial Fraud, Cluster Algorithms, K-Means Algorithm, Data Miner.

INTRODUCTION

The core of any business is customer databases that represent information about relations of clients with the company. Constant complication of organizational processes and rapid changes in business environment result in active growth of volumes of internal corporate data of companies. Therefore, the greater interest from the point of view of fraud risk assessment start to cause such tools as forensic data analytics (FDA), which allows narrowing suspicious transactions sampling at maximum while minimizing the volume of check. For example, in the sphere of communications, the database stores information on the time of service contract

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

execution, time of contract termination, region, tariff, etc. In sales of books – gender, age, purchased books, etc. In internet trading – purchased goods, their quantity, time of purchase, etc. The objective of the analysis is to study the forensic data analytics techniques and its role in detection of fraud, to detect – by the example of dentistry – the dentists purposely overrating their services – potential fraud detection. Tasks of the work.

1) To study methods of forensic data analytics.
2) To characterize Data Mining.
3) To detect potential fraudsters.

Methods of forensic data analytics

There are numerous different methods of forensic data analytics, request modeling, and information collection and processing. Which of them should be used for analysis of your own data and which could be applied in combination with already existing software and infrastructure? Know more about different methods and solutions and learn how to create such solutions with the help of existing software and systems. Study existing tools of forensic data analytics and learn how to determine whether the size and complicacy of your information cause problems in its processing and storage, and how to solve them.

Forensic Data Analytics as Process

In fact, forensic data analytics means processing of information and detection of models and trends therein that help decision making. Principles of forensic data analytics have been known for many years, but with the generation of *big data*, they became even more popular. Big data resulted in explosive growth of popularity of broader methods of forensic data analytics, partly since there appeared much more information, and it becomes more various and vast by its nature and content. When working with big data sets, it is not enough now to use simple and straightforward statistics. Having 30 or 40 million detailed records on purchases, it is not enough to know that two million of them have been made in the same place. In order better to meet the demands of customers, it is necessary to understand if these two million belong to a certain age group, and to know their average income. These business requirements led from simple search and statistical analysis of data to a more complicated forensic data analytics. For solution of business tasks, such a data analytics is required, which allows building a model for description of information and ultimately results in creation of an outcome report. This process is shown on figure (1).

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

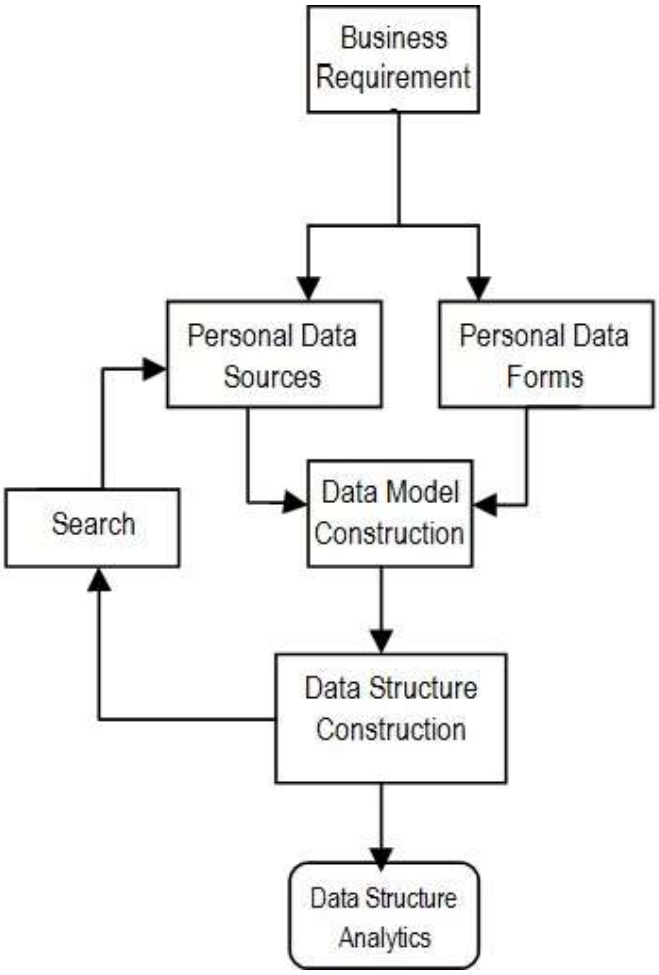


Figure 1. Process chart

Process of data analytics, search and building of model is often an interactive one, since it is necessary to find out and detect different information, which can be derived. It is also necessary to understand how to link, transform and unite them with other data for obtaining a result. After detection of new elements and aspects of data, the approach to detection of sources and forms of data with subsequent correlation of this information with the set result may change.

Tools of forensic data analytics

Forensic data analytics does not merely mean applied database tools or software. Forensic data analytics may be carried out with relatively modest database systems and simple tools, including creation of your own ones, or with the use of prefabricated software packages. Sophisticated forensic data analytics is based on previous experience and algorithms determined with the help of existing software and packages, while different specialized tools are associated with different methods. For example, IBM SPSS, which harks back to statistic analytics and surveys, allows building efficient predictive models upon previous trends and providing exact forecasts. IBM Info Sphere Warehouse ensures data source search, preliminary processing and forensic analytics in one package allowing to derive information from the source base right to the final report. Recently, it became possible to work with very big sets of data and cluster/wide-scaled data processing, which enables even more complicated generalizations of forensic data analytics results by groups and data correlations. Today, an

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

absolutely new range of tools and systems is available, including combined data storage and processing systems. It is possible to analyze very different sets of data, including traditional SQL databases, unprocessed text data, "key/meaning" sets and documental bases. Cluster databases, such as Hadoop, Cassandra, Couch DB and Couch base Server, store and provide access to data by such methods that do not correspond to traditional table structure. In particular, a more flexible form of document base storage gives the data processing a new direction and complicates it. SQL databases regulate structure and strictly adhere to the scheme, which facilitates requests to them and analysis of data with the known form and structure. Documental databases that correspond to standard structure like JSON, or files with a certain machine-readable structure can also be easily processed although this may be complicated due to various and changing structure. For example, Hadoop, which processes absolutely "raw" data, may be difficult to detect and derive information before the start of its processing and correlation.

Basic methods

Some basic methods used for forensic data analytics describe the type of analytics and data restoration operation. Unfortunately, different companies and solutions not always use the same terms, which may cause confusion and apparent complexity. Let's study some key methods and examples of how to use certain tools for forensic data analytics.

Association

Association (or relation) is probably the most well-known, familiar and simple method of forensic data analytics. Simple comparison of two or more elements, often of the same type, is performed for detection of models. For example, when monitoring purchasing habits, it can be noted that cream is usually bought together with strawberry. It is not difficult to create tools for forensic data analytics on the basis of associations or relations. For example, in Info Sphere Warehouse, there is a master that gives configurations of information flows for creation of associations by studying the source of input information, basis for decision making and output information. Figure (2) are provides a respective example for database sample.

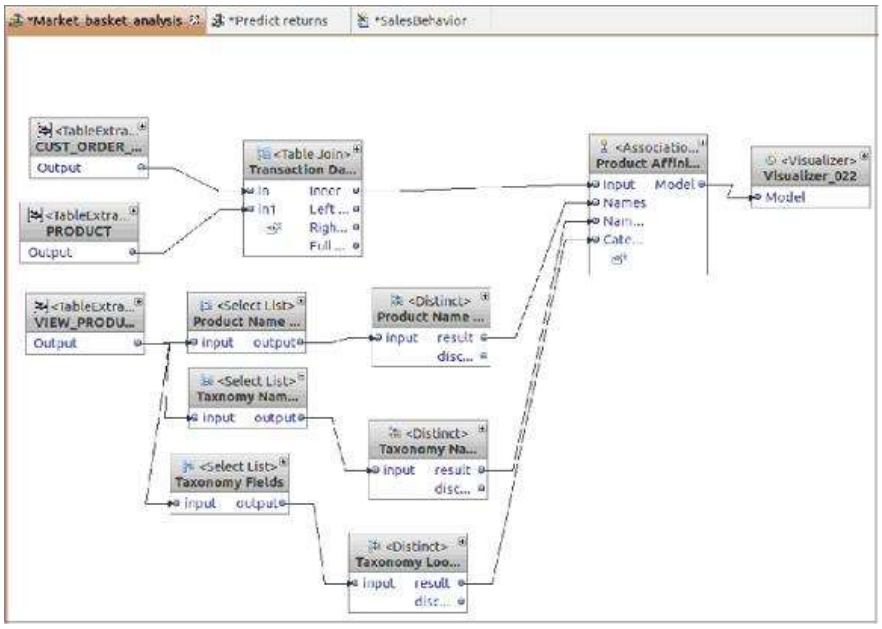


Figure 2. Information flow used for association approach

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

Classification

Classification may be used for obtaining understanding of the type of customers, goods or objects by describing several attributes for identification of a certain class. For example, cars may be easily classified by type (sedan, SUV, cabriolet) by determining different attributes (number of seats, body style, driving wheels). Studying a new car, it may be classified by comparing attributes with the known definition.

The same principles may be applied to customers, for example explain by classifying them by age and social group. Furthermore, classification may be used as input data for other methods.

For example, for determination of classification, it is possible to apply decision trees. Cauterization allows using common attributes of different classifications for the purpose of defining clusters.

Cauterization

When studying one or several attributes or classes, it is possible to group certain data elements together and obtain a structured conclusion.

At simple level, cauterization uses one or several attributes as a basis for determination of a cluster of similar results. Cauterization is useful for determination of different information since it correlates with other examples, thereby it is possible to see where similarities and ranges correlate with each other.

Cauterization methods work in both directions. It may be assumed that there is a cluster at certain point, and then to use your own identification criteria in order to check this.

The graph provided on Figure 3 shows an example. The age of a customer here is compared to the price of the purchase. It is reasonable to expect that people between twenty and thirty years old (before marriage and children), as well as between 50 and 60 years old (when children have already left the family) have a higher available income.

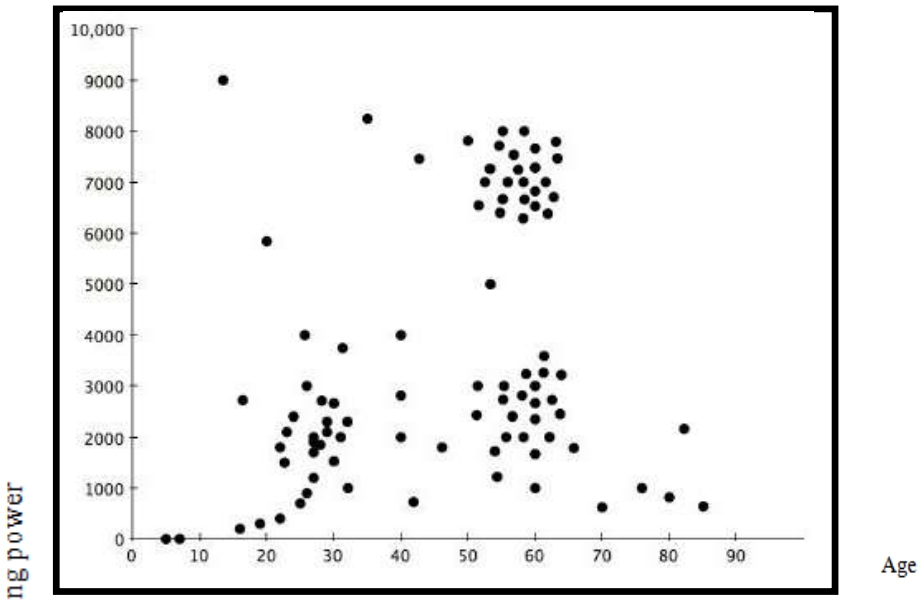


Figure 3. Cauterization

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

This example shows two clusters: one within the range $2,000/20-30 years old, and another – within $7,000-8,000/50-65 years old. In this case, we made a hypothesis and checked it in a simple graph, which may be drawn up with the help of any applicable software for construction of graphs. More complicated combinations require a complete analytical package, particularly, if it is necessary automatically to make decision on the basis of information about the nearest neighbor. Such building of clusters constitutes a simplified example of so-called image of the nearest neighbor. Individual purchasers may be differentiated by their literal proximity to each other on the graph. It is very probable that purchasers from the same cluster also share other common attributes, and this assumption may be used for search, classification and other types of analytics of data set members.

Cauterization method may be applied in reverse direction: to detect different artifacts by considering certain input attributes. For example, recent study of four-digit PIN codes determined clusters of numbers within the range1-12 and 1-31 for the first and the second pairs. If showing these pairs on the graph, it is possible to see clusters related to the dates (birthdays, anniversaries).

Forecasting

Forecasting is a broad topic, which covers the area from equipment component failure prediction to fraud detection and even company's profit forecasting. In combination with other methods of forensic data analytics, forecasting includes analysis of trends, classification, correlation with model and relations. By analyzing past events or examples, it is possible to forecast the future. For example, by using data on credit card authorization, it is possible to unite decision tree analytics of previous transactions of a person with classification and correlation with historic models for detecting fraudulent transactions. If air ticket purchase in the US coincides with transitions in the US, it is quite probable that such transactions are actual.

Sequential models

Sequential models, which are often used for long-term data analytics, are a useful method for detection of trends or regular repetition of similar events. For example, purchasers' data allows determining that in different seasons they buy different sets of products. Upon this information, the basis purchase forecasting application, basing on the frequency and story of purchases, can automatically predict that specific products will be added to this purchase.

Decision trees

Decision tree connected with most of other methods (first and foremost with classification and forecasting) may be used either as a selection criterion, or for support of selection of specific data within the general structure. The decision tree starts with a simple question, which has two answers (sometimes, more). Each answer leads to the next question helping to classify and identify data or make forecasts. Figure (4) shows an example of failure state classification.

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

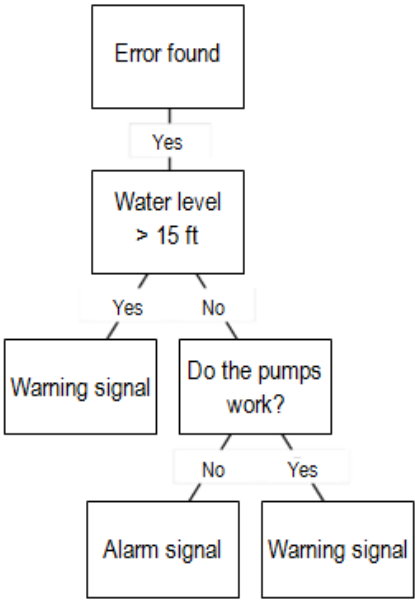


Figure 4. Decision tree

Decision trees are often used with properties information classification system and with forecasting systems, where different forecasts may be based on the past historic experience that enables building a decision tree structure and obtain the result.

Combinations

It is very rare that only one of these methods is used in practice. Classification and cauterization are similar methods. By applying cauterization for determination of the nearest neighbors, it is possible to clarify classification in addition. Decision trees are often applied for building and detecting classifications that may be tracked in historic periods for determination of sequences and models.

Processing with memorization

With all basic methods, it is often essential to record and then analyze the obtained information. For some methods, it is absolutely obvious. For example, when building sequential models and learning for forecasting, historical data from different sources and information examples are analyses. In other cases, this process may be more explicit. Decision trees are rarely built just once and are never forgotten. When new information, events and data points are detected, it may be necessary to build additional branches or even absolutely new trees. Some of these processes can be automated. For instance, building of a forecasting model for credit card fraud detection means determination of probabilities, which may be applied for the current transaction, with subsequent updating of this model with addition of new (confirmed) transactions. Then, this information is registered, and thereby, the next time, the decision can be made faster.

Characteristics of data mining

Data Mining means analysis and detection by a “machine” (algorithms, artificial intelligence tools) of latent knowledge in raw data that has not been known before, nontrivial, practically

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

useful, and available for interpretation by people. (Founder of Data Mining, Pyatetsky-Shapiro.)Data mining methods allow solving many tasks faced by analysts. The basic of them are classification, regression, search for association rules and cauterization, detection of untypical observations.

- **Detection of untypical observations**: Detection of untypical observation in data, which are of "special" interest, or detection of errors, which should be eliminated for further analysis.
- **Classification**: The task is to determine the affiliation of an object with a class by its properties. It should be noted that multiple classes of this task, with which the object may be affiliated, is not known in advance.
- **Regression**: Search for a function that describes dependence by object properties with the least error.
- **Association rules**: Using association rules, a store manager can define goods, which are more often bought together (i.e. if someone buys good 1, he will buy also good 2), and use this information for marketing campaigns.
- **Cauterization**: Detection of latent structure in data or observations that are similar one way or another.

Verbal assignment of task

An insurance company has contracts with 10 clinics, where 100 dentists work. The database consists of 20,000 patients. With the use of SQL, we upload those dentists from the database, who rendered the insured services more than 25 times. In total, there were 30 such dentists found. The sale length was 1,000 – the total number of works fulfilled by them under insurance. Total amount of insurance was 500,000$.

The sample includes the information on services rendered by dentist under insurance:

- Age of the client.
- Type of works.
- Additional services.
- Cost of rendered services in $.
- Dentist's personal number.

It is necessary to find out dentist, who purposely overrate their services – to detect potential fraudsters (fraud detection).

Data structure

Data were uploaded from the database of the company working in medical insurance. The data are reports on services provided to clients by different dentists within one session. Only those dentists are considered, the number of works by which performed under insurance exceeds 25.Sample size – 1,000 observations.

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

A fragment of the table is provided below:

Fraud detection in the work if dentists

	1 Age of the client	2 Type of works	3 Additional services	4 Cost of rendered services in $	5 Dentist's personal number
1	30	3	3	1003	16
2	27	1	1	157	5
3	30	3	3	1016	17
4	40	3	2	735	20
5	16	2	1	320	21
6	20	3	3	134	28
7	16	2	1	450	24
8	17	2	1	331	20
9	38	3	2	705	16
10	13	1	2	275	8
11	30	1	1	588	4
12	21	3	3	139	25
13	25	2	2	844	1
14	27	2	2	914	6
15	13	1	2	279	9
16	29	3	3	984	13
17	24	1	1	148	16
18	13	1	2	286	11
19	24	1	1	147	17
20	25	1	1	150	14
21	17	2	1	336	25
22	37	3	2	694	15
23	33	2	3	629	7
24	26	2	2	890	4
25	31	3	3	1033	0
26	32	3	3	1082	10
27	21	2	1	385	24
28	46	3	2	828	28
29	34	2	3	650	9
30	29	1	1	165	0

This table shows the information about 1,000 of works performed by different dentists.

Rows show Works performed, and columns show the following parameters:

- Age of the client input in table.
- Type of works (1 – insignificant, 2 – significant, 3 – special).
- Additional services (1 – none, 2 – inexpensive additional services, 3 – expensive) for all services.
- Cost of rendered services in $.
- Dentist's personal number (used for exact identification of the dentist).

For instance, the first row in the table show the information on the work performed – Age of the client 35, significant work has been performed with application of expensive additional procedures; the work was performed by Dentist 10.

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

Main approaches to fraud detection

Most of methods applied for fraud detection solve the task of classification. They require availability of objects, for which it is known in advance to which of two classes they belong – Fraud or Non-fraud(and quite a big number for building a qualitative model).Such methods belong to supervised learning class. In our task, it is necessary to detect potential fraudsters when having no information, to which of the classes specific observations belong. Cauterization unlike classification requires no information of affiliation with the class, and, consequently, belongs to unsupervised learning. The task of cauterization is solved at the initial stages of study. Its solution helps better to understand data and their nature. Big advantage of cluster analysis is that it enables dividing objects not by one but by a whole set of attributes.

Advantages of Data Mining technology application

Data mining technology unlike individual cauterization and classification methods enables:

- Automatically to determine optimal number of clusters.
- To work with data bulk.
- Requires no objects, for which it is known in advance, to which of classes they belong.
- In-Place Database Processing.

In-Place Database Processing

In-Place Database Processing (IDP) is a developed technology of database access developed in Stat Soft for achieving high efficiency of the direct interface between external server data and analytical functionality of STATISTICA products.

IDP technology has been developed to help ensuring access to data in big DB using single-step process that does not require creation of local copies of data. IDP significantly increases efficiency of STATISTICA; in particular, it is well adapted to tasks of data mining and research Data analytics.

Reason for high speed of IDP

High speed of IDP technology against traditional methods is stipulated not only by the fact that IDP allows STATISTICA to refer directly to DB for data and skip excessive step of data import and creation of a local data file, but also due to its “multi-task” (asynchronous and distributed processing) architecture. In particular, IDP uses resources (several processors) of the DB server for performing operations with requests, derivation of records and their transmission to the computer, while STATISTICA immediately processes these records as soon as they arrive.

Cauterization algorithms

K-Means algorithm

It divides numerous elements of vector space into k-numbers of clusters known beforehand. The action of algorithm is such that it tries to minimize the mean square deviation at points of every cluster. The main idea is that the mass centre is recalculated at every iteration for every cluster, which was received at the previous step, and then the vectors are divided into clusters

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

again in accordance with that, which of new centers appears to be closer by the chosen metrics. The algorithm is completed, when cluster centers change at specific iteration.

EM algorithm

The core of the idea of EM algorithm is the assumption that distribution of source set distribution is a linear combination of subsets having normal distribution. The aim of the algorithm is to decompose (divide) the set into subsets, as well as to assess distribution parameters for every subset, which maximize log-likelihood function used as a model quality measure. Normal distribution parameters – expectancy and dispersion.

Descriptive analysis

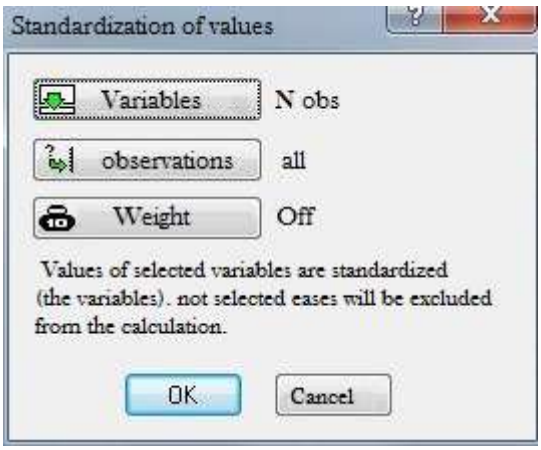
Varaiable	**Observ. No.**	Mean	Minimum	Maximum	St. dev.
Age of the client	1000	26,5040	7,0000	56,000	9,3611
Cost of provided services in $	1000	537,5040	123,0000	1270,000	342,5043

The table shows that continuous variables Age of the client and Cost of services have different variability (dispersion) – Standard deviation (last column).

In cauterization, it is very important that variables have the same variability (dispersion). For this purpose, we use Standardizations procedure.

- Select the tab Data.
- Select Standard is Dialogue window opens.
- Select variables – Age of the client, Cost of services.

Click OK.

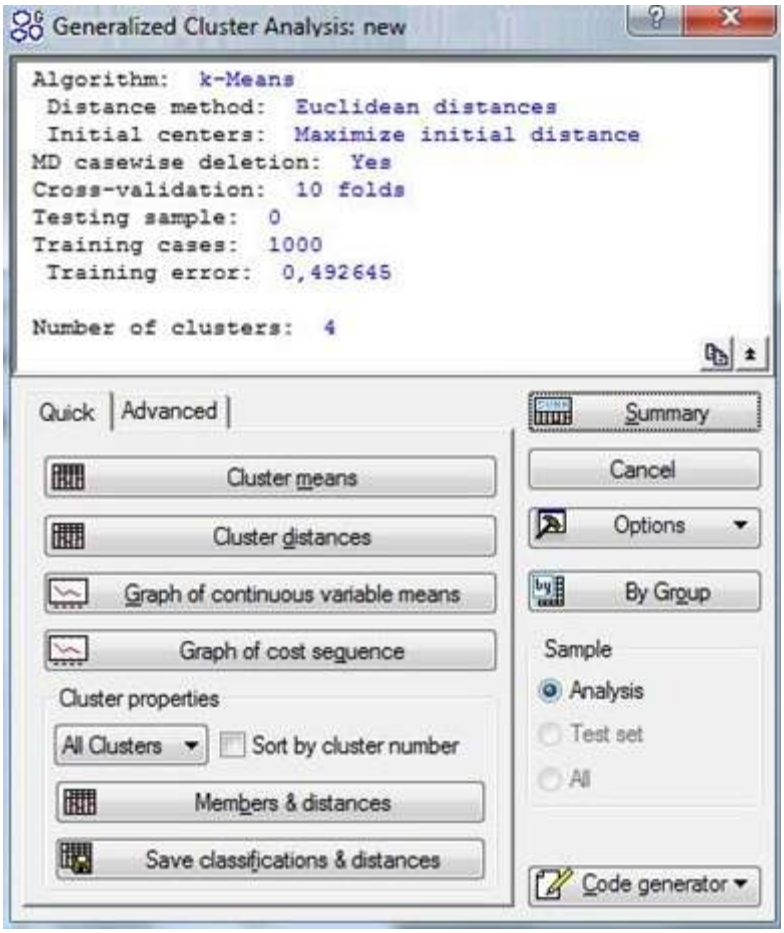


ISSN 2053-4086(Print), ISSN 2053-4094(Online)

To the beginning:

K-Means cauterization.

Result analysis–tab Quick.



In the upper part of the dialogue window, a working area is located, where main cauterization characteristics are shown:

- Algorithm – K-Means.
- Distance method –Euclidian.
- Initial centers – Maximum distance between clusters.
- MD case wise deletion –Yes.
- Cross-validation – 10 times.
- Testing sample – 0.

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

- Training cases – 1,000.
- Training error – 0.492645.
- Number of cluster – 4.

In the tab Quick (Fast), it is possible to view the following results:

Analysis results: Cluster description

Cluster Means are:

Cluster	Age of the client Mean	Type of works Mean	Additional procedures Mean	Cost of provided services in $ Mean	Number of observations	Observation percentage
1	25,00800	3	3	714,8520	250	25,00000
2	21,15750	1	1	285,7675	400	40,00000
3	38,40500	3	2	809,1650	200	20,00000
4	27,38667	2	2	551,0067	150	15,00000

Rows show cluster numbers. Columns show variables, selected at the beginning of analysis. The last column shows the share of observations in every cluster. The following clusters were obtained:

- Cluster 1: Special work with the use of expensive additional procedures, average age of clients – 25, average cost of services – 715$;
- Cluster 2: Insignificant work without use of additional procedures, average age of clients – 21, average cost of services – 286$;
- Cluster 3: Significant work with the use of expensive additional procedures, average age of clients – 38, average cost of services – 819$;
- Cluster 4: Significant work with the use of cheap additional procedures, average age of clients– 27, average cost of services – 551$.

Cluster distance

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 1	0,000000	1,464964	1,039958	1,422238
Cluster 2	1,464964	0,000000	1,527129	1,438623
Cluster 3	1,039958	1,527129	0,000000	1,049391
Cluster 4	1,422238	1,438623	1,049391	0,000000

For example, the distance between Cluster 1 and Cluster 2 – 1.465 (upon Euclidian metrics).

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

Cluster objects and distances:

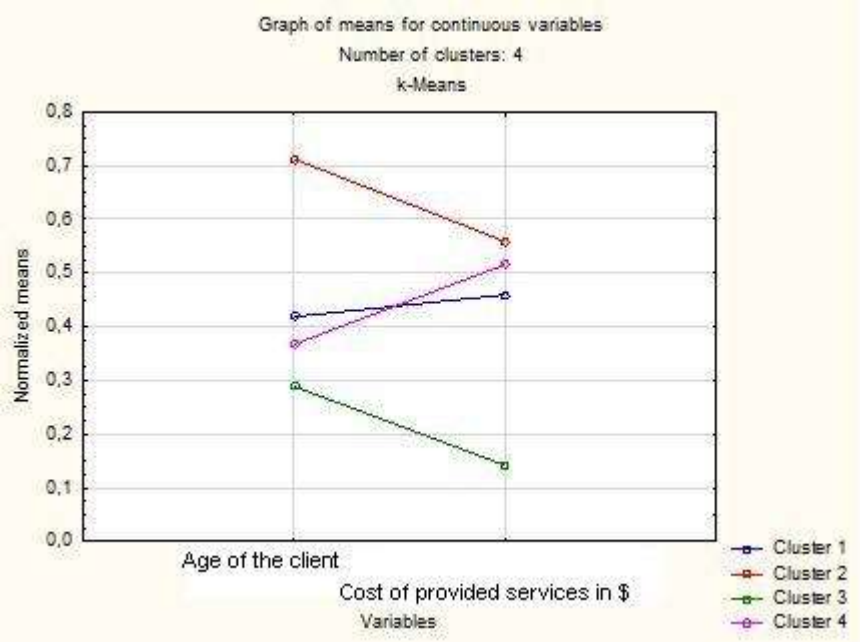
Case No.	Final classification	Additional procedures	Type of works	Age of the client	Cost of provided services in $	Distance to centered
1	4	3	2	0,90759	0,35765	1,016450
2	2	1	1	0,48029	-1,06715	0,224022
3	3	2	1	0,05299	1,12844	1,031600
4	4	2	2	-0,16067	0,87443	0,254053
5	1	3	3	0,48029	1,45252	0,304730
6	2	1	2	-1,22892	-0,67008	1,008051
7	2	1	2	-0,90844	-0,55913	1,003449
8	2	1	1	-0,05384	-1,12263	0,152201
9	1	3	3	0,26664	1,27734	0,240995
10	2	1	2	-1,33574	-0,71679	1,010627
11	3	2	1	0,15981	1,18100	1,028836
12	3	2	3	1,12124	0,44524	0,107777
13	2	1	1	0,15981	0,09196	0,283683
14	2	2	1	-1,44257	-0,74307	1,013766
15	2	1	1	-0,26749	-1,14014	0,134172
16	2	1	1	0,48029	0,19999	0,343941
17	3	2	3	1,44172	0,57954	0,071614
18	2	1	2	-0,80162	-0,52701	1,002894
19	2	1	1	-0,16067	-1,12846	0,141261
20	3	2	1	0,15981	1,24231	1,031057
21	1	3	3	-0,58797	-1,16642	0,509531
22	2	1	1	0,05299	-1,11095	0,163769
23	1	3	3	0,37346	1,39121	0,280001
24	3	2	3	1,22806	0,48319	0,092927
25	3	2	1	0,15981	1,15764	1,028074
26	1	3	3	-0,58797	-1,16642	0,509531
27	4	2	2	-0,05384	1,03793	0,299503

Basing hereon, it is possible to understand, which observations belong to which cluster.

Determination of essential factors

- Let's consider the continuous variables first.
- Let's draw up a graph of average age of clients and cost of services in each cluster.

ISSN 2053-4086(Print), ISSN 2053-4094(Online)



The graph shows that average age of clients and cost of services in Cluster 2 are maximum in comparison with other clusters.

- Let's carry out dispersion analysis for determination of factors influencing affiliation of an object with the cluster.

	Between SS	df	Within SS	df	F	p value
Age of the client	461,4554	3	537,5446	996	285,0055	0,00
Cost of provided services in $	409,1656	3	589,8344	996	230,3069	0,00

The dispersion analysis shows that variables Age of the client and Cost of provided services influence affiliation of an object with the cluster, since p value is less than 0.05. In other words, both factors – Age of the client and Cost of provided services – are essential.

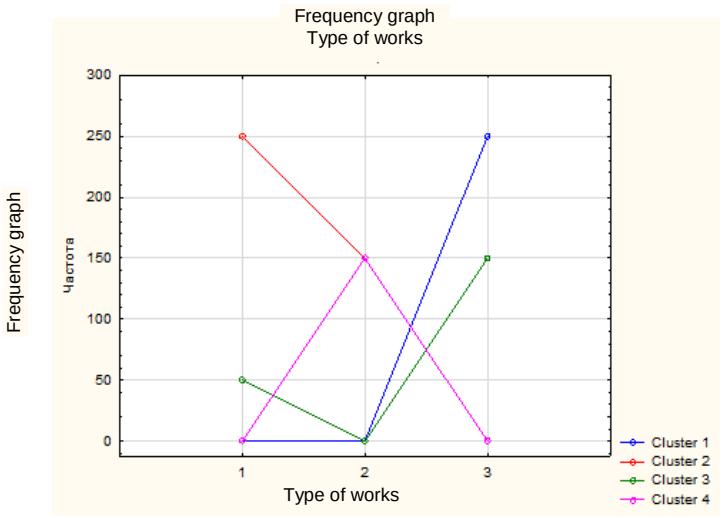
- Let's consider categorical variables. We will draw up Frequency tables and Frequency graphs for categorical variables (type of works, additional procedures) for each cluster
- Type of works.

Frequency table

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
1	0	250	50	0	300
2	0	150	0	150	300
3	250	0	150	0	400

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

Frequency graph:



The graph shows that:

A. Cluster 2 has the biggest number of insignificant works.

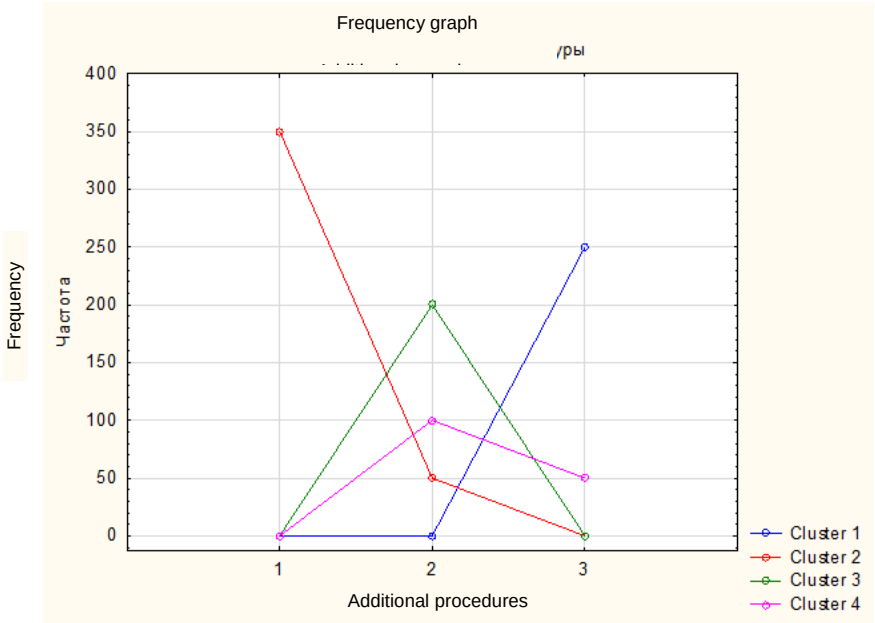
B. Cluster 1 has the biggest number of special works.

Additional procedures.

Frequency table:

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
1	0	350	0	0	350
2	0	50	200	100	350
3	250	0	0	50	300

Frequency graph:



ISSN 2053-4086(Print), ISSN 2053-4094(Online)

The graph shows that:

i. Cluster 2 has the biggest number of works performed without use of additional procedures.

ii. Cluster 3 has the biggest number of works performed with the use of cheap additional procedures.

iii. Cluster 1 has the biggest number of works performed with the use of expensive additional procedures.

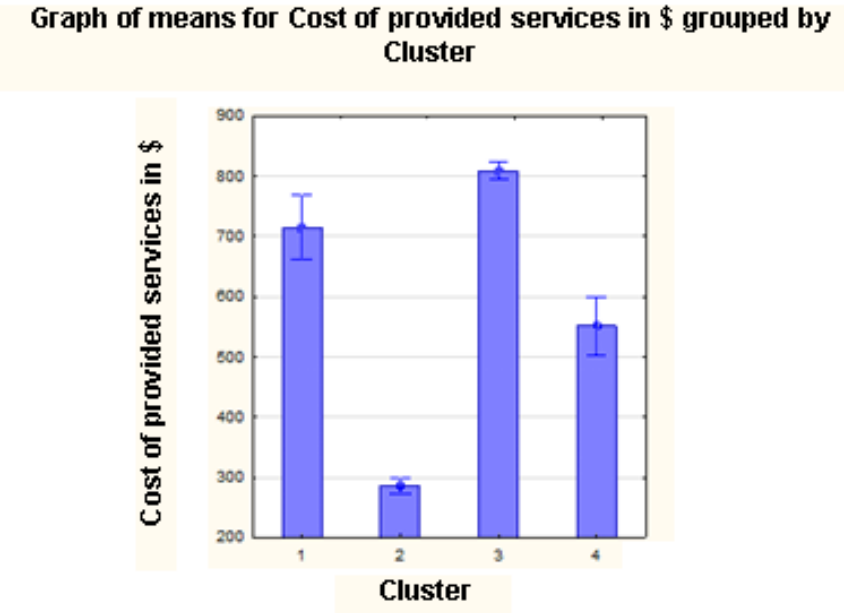
Let's determine what variables provide essential influence on affiliation with the cluster. Let's use Chi-square criteria for categorical variables:

Additional procedures	df	Chi-square	p value	G-square	p value
	6	1543,651	0,00	1699,764	0,00
Type of works	6	1156,250	0,00	1423,615	0,00

The table shows that type of works and additional procedures influence the affiliation with the cluster, since p value is less than 0.05. In other words, factors type of works and additional procedures are essential. K-Means algorithm launch is described in Appendix 1.

Potential fraud detection

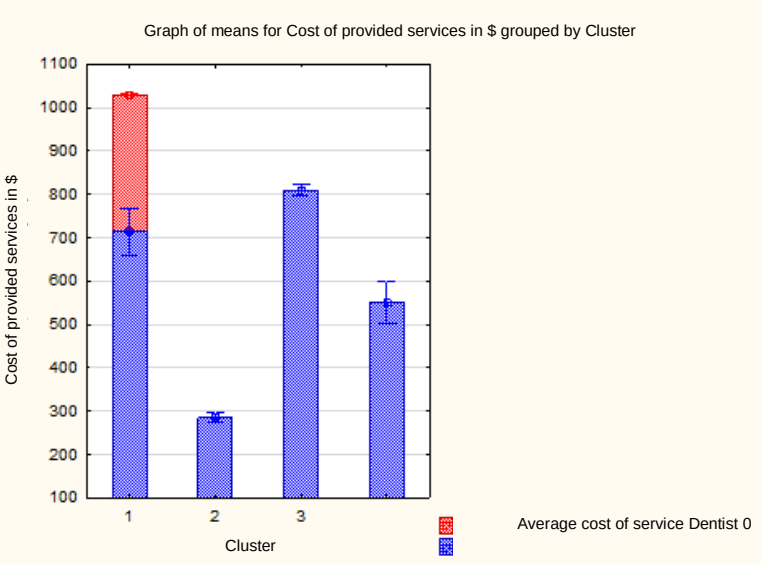
We are interested in those dentists who overrate their services. In order to detect such dentists, it is necessary to compare average total cost of provided services for each cluster and average cost of provided services of a dentist in each cluster. Graph of total average cost of provided services in each cluster:



ISSN 2053-4086(Print), ISSN 2053-4094(Online)

Dentist 0

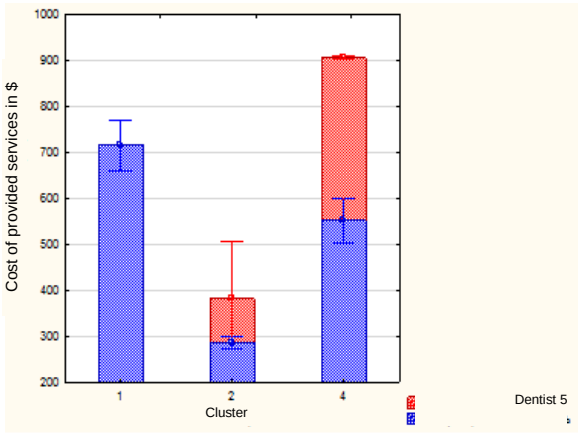
Graph of total average cost of provided services of Dentist 0 and Graph of total average cost of provided services in each cluster:



The graph shows that in Cluster 1, cost of services of Dentist 0 is significantly higher than total average cost of services. Dentist 0 is a potential fraudster.

Dentist 5

Graph of total average cost of provided services of Dentist 5 and Graph of total average cost of provided services in each cluster:

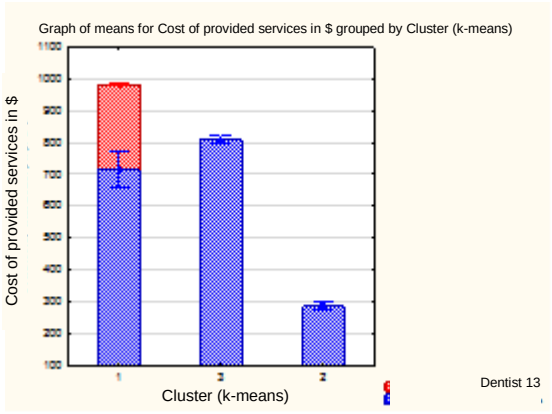


The graph shows that average cost of services of Dentist 5 in Cluster 4 and Cluster 2 is significantly higher than total average cost of services in this cluster. Dentist 5 is a potential fraudster.

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

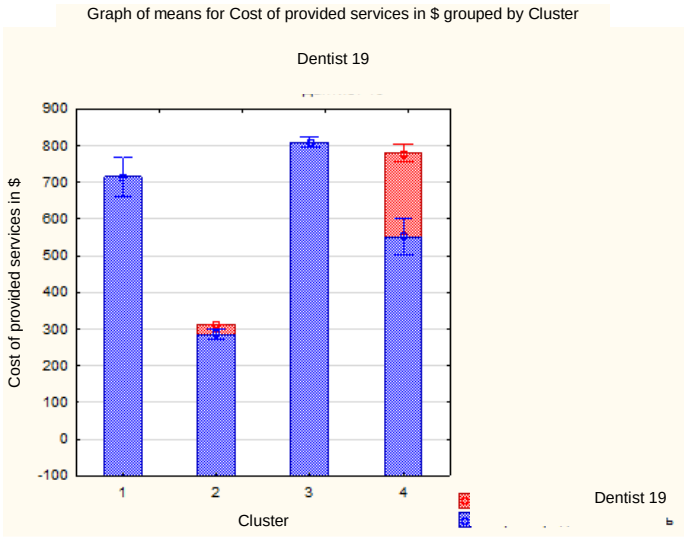
Dentist 13

Graph of total average cost of provided services of Dentist 13 and Graph of total average cost of provided services in each cluster:



The graph shows that average cost of services of Dentist 13 is significantly higher than total average cost in this Cluster 1. Dentist 13 is a potential fraudster.

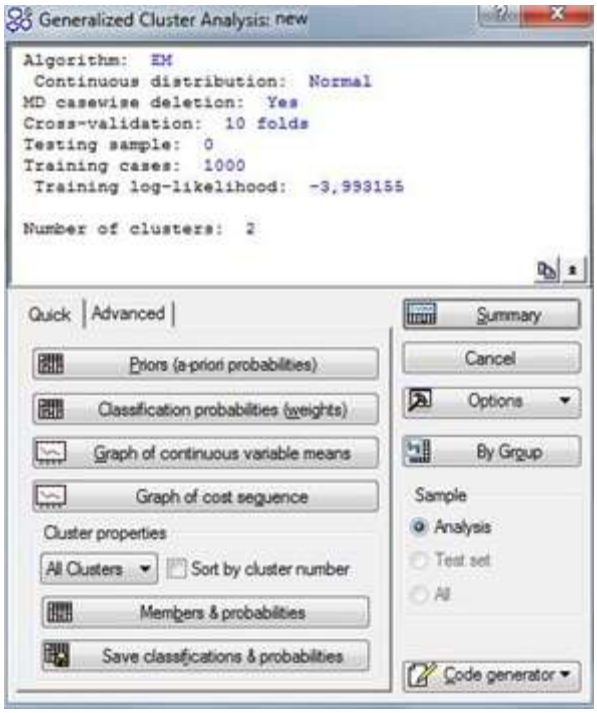
Dentist 19



The graph shows that average cost of services of Dentist 19 is significantly higher than total average cost in Cluster 4. Dentist 19 is a potential fraudster.

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

Em Algorithm (Result analysis – tab Quick)



EM cauterization gave only 2 clusters

Result analysis: Cluster description

Cluster	Age of the client Mean	Cost of provided services in $, Mean	Type of work, Mean	Additional procedures, Mean
1	34,12000	882,1756	3	2
2	20,27273	255,5000	2	2
Total	26,50400	537,5040	2	2

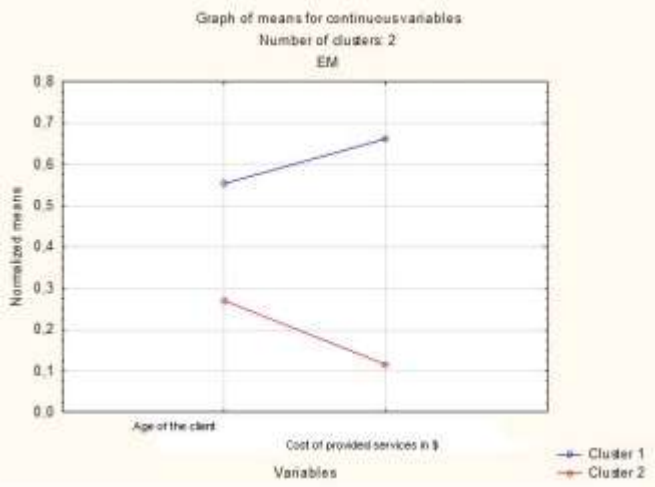
The following clusters were obtained:

Cluster 1: Special works with the use of cheap additional procedures, average age of clients – 34, average cost of services – 882$;

Cluster 2: Significant and insignificant works with the use of cheap additional procedures, average age of clients – 26.5, average cost of services – 256$.

- Determination of essential factors
- Let's consider the continuous variables first.
- Let's draw up a graph of average age of clients and cost of services in each cluster.

ISSN 2053-4086(Print), ISSN 2053-4094(Online)



- Let's carry out dispersion analysis for determination of factors influencing affiliation of an object with the cluster.

	Between SS	df	Within SS	df	F	p value
Age of the client	541,5678	1	457,4322	998	1181,562	0,00
Cost of provided services in $	828,5691	1	170,4309	998	4851,891	0,00

The dispersion analysis shows that variables Age of the client and Cost of provided services influence affiliation of an object with the cluster, since p value is less than 0.05. In other words, both factors – Age of the client and Cost of provided services – are essential.

- Let's consider categorical variables.

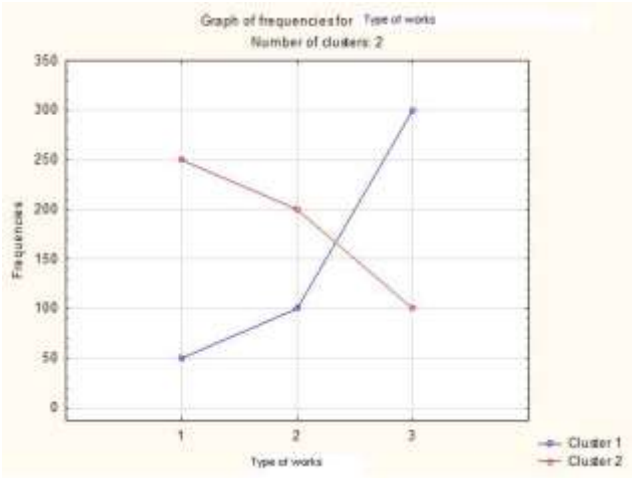
We will draw up Frequency tables and Frequency graphs for categorical variables (type of works, additional procedures) for each cluster.

- Type of works.
    1. Frequency table

	Cluster 1	Cluster 2	Total
1	50	250	300
2	100	200	300
3	300	100	400

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

2. Frequency graph:



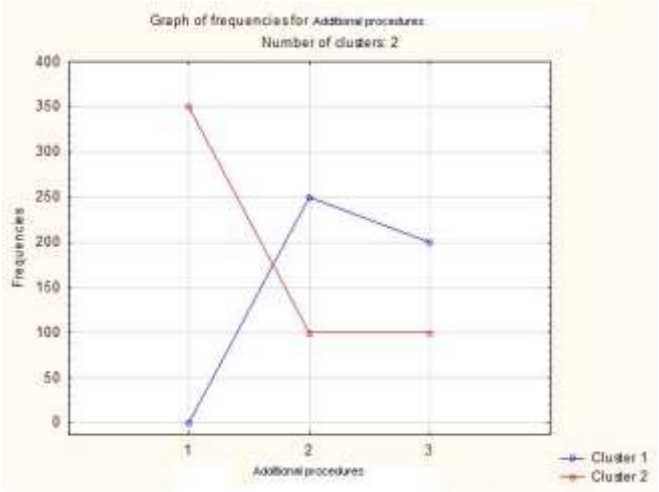
3. The graph shows that:

1. Cluster 2 has the biggest number of insignificant and significant works.
2. Cluster 3 has the biggest number of special works.

- Additional procedures.
  1. Frequency table:

	Cluster 1	Cluster 2	Total
1	0	350	350
2	250	100	350
3	200	100	300

  2. Frequency graph:



ISSN 2053-4086(Print), ISSN 2053-4094(Online)

3. The graph shows that:

    a. Cluster 1 has the biggest number of works performed without use of additional procedures.

    b. Cluster 2 has the biggest number of works performed with the use of expensive and inexpensive additional procedures.

- Let's determine what variables provide essential influence on affiliation with the cluster. Let's use Chi-square criteria for categorical variables:

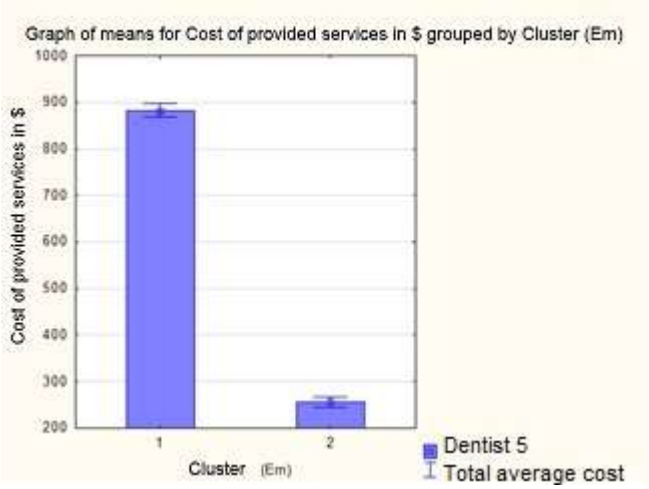
	df	Chi-square	p value	G-square	p value
Type of works	2	259,2593	0,00	274,1643	0,00
Additional procedures	2	442,0394	0,00	575,5804	0,00

The table shows that type of works and additional procedures influence the affiliation with the cluster, since p value is less than 0.05. In other words, factors type of works and additional procedures are essential.

Module launch algorithm is described in appendix 2

Potential fraud detection

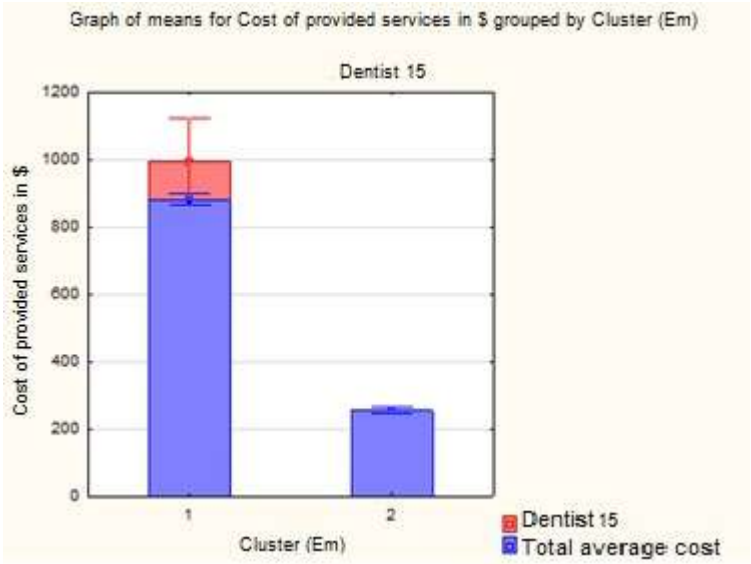
Graph of total average cost of provided services in each cluster:



Dentist 15

Graph of total average cost of provided services of Dentist 15 and Graph of total average cost of provided services in each cluster.

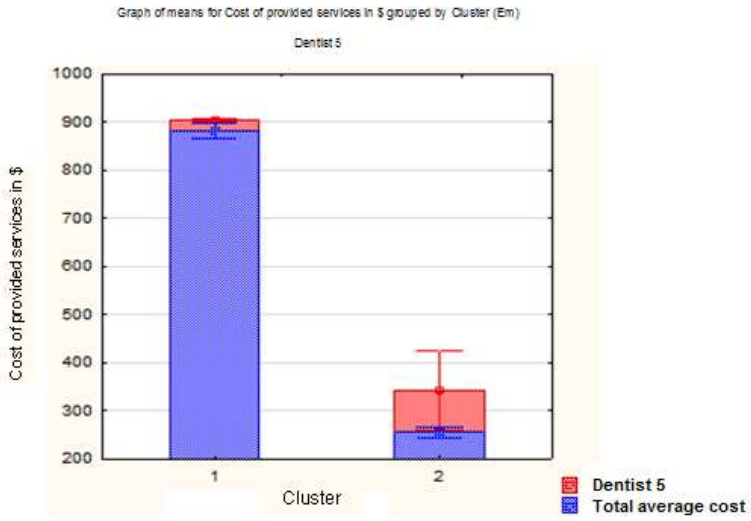
ISSN 2053-4086(Print), ISSN 2053-4094(Online)



The graph shows that in Cluster 1, cost of services of Dentist 15 is significantly higher than total average cost of services. Dentist 15 is a potential fraudster.

Dentist 5

Graph of total average cost of provided services of Dentist 5 and Graph of total average cost of provided services in each cluster.

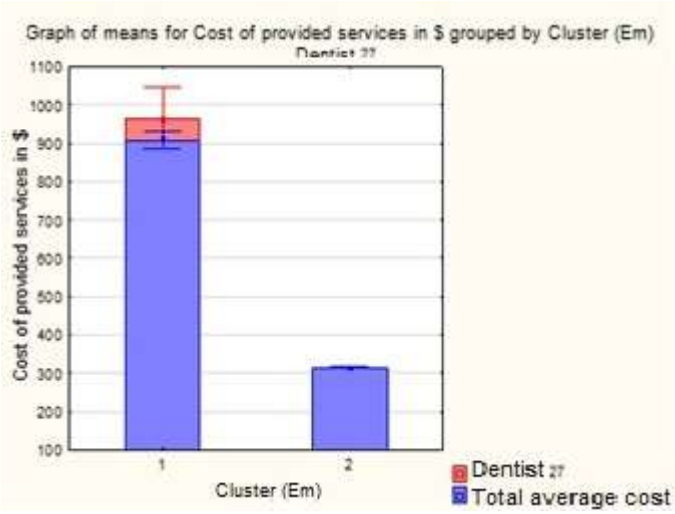


The graph shows that average cost of services of Dentist 5 in Cluster 2 is significantly higher than total average cost of services in this cluster. Dentist 5 is a potential fraudster.

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

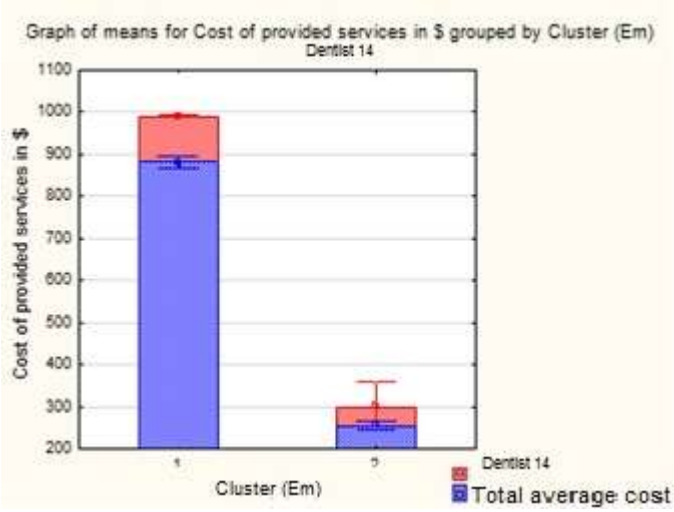
Dentist 27

Graph of total average cost of provided services of Dentist 5 and Graph of total average cost of provided services in each cluster:



The graph shows that average cost of services of Dentist 27 is significantly higher than total average cost in this Cluster 1. Dentist 13 is a potential fraudster.

Dentist 14

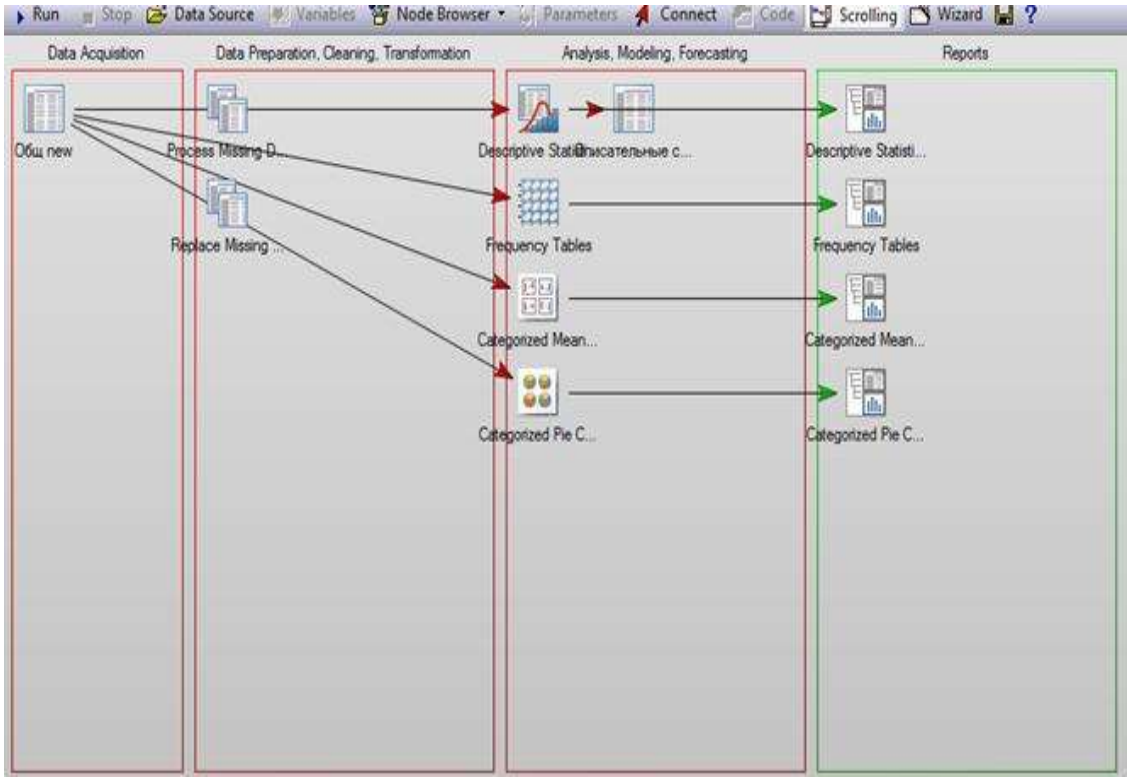


The graph shows that average cost of services of Dentist 14 is significantly higher than total average cost in Cluster 1 and Cluster 2. Dentist 14 is a potential fraudster.

Automation

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

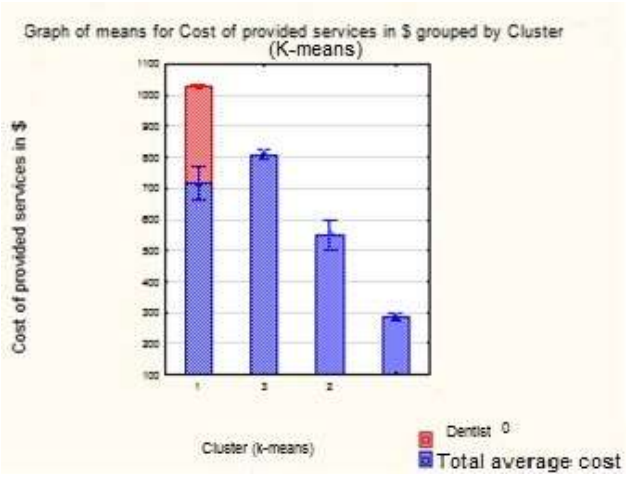
For automatic computation of descriptive statistics and graph construction in STATISTICA Data Miner, there is a project construction module. A fragment of interactive construction is shown below.



At the first step we input variables that will be subject to analysis (in the first red rectangle). At the second step, date cleaning and filtering takes place – the process of analysis of missed data and replacement of missed data by the mean (in the second red rectangle). At the third step, descriptive statistics, frequency tables, mean graph in every cluster, diagrams are computed (in the third red rectangle).

The last rectangle (green) – there are results of analysis. You can view the obtained results in it.

Dentist 0



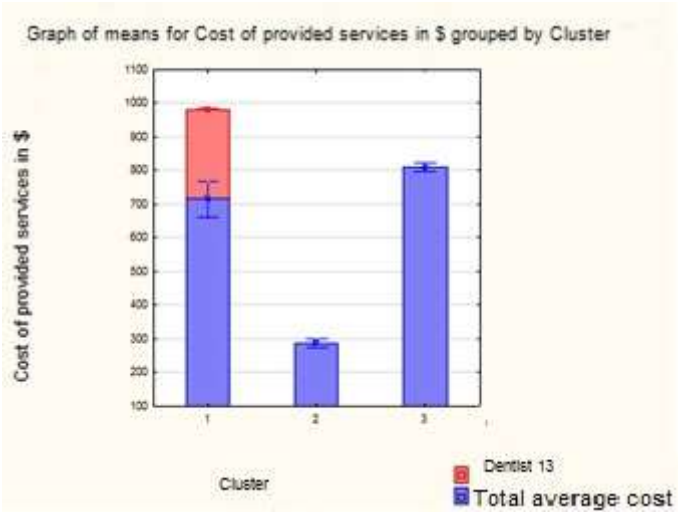
ISSN 2053-4086(Print), ISSN 2053-4094(Online)

Significantly overrates his services for provision of special services with the use of expensive procedures.

Dentist 5

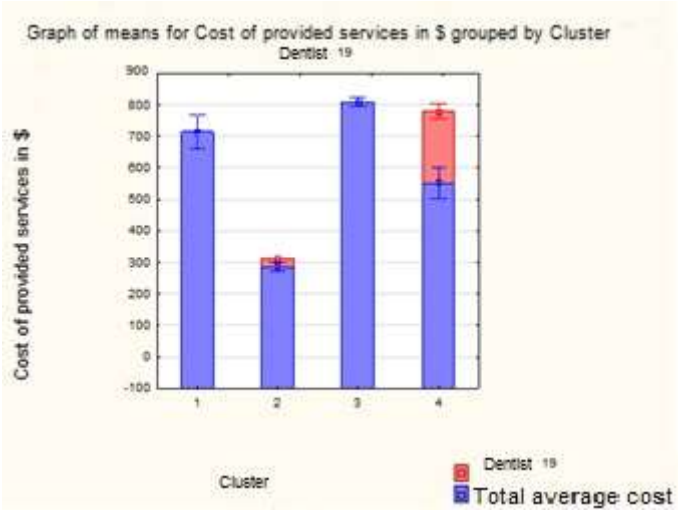
Significantly overrates his services for provision of insignificant services without use of additional procedures and the cost of works for provision of significant services with the use of expensive works.

Dentist 13



Significantly overrates his works for provision of special services with the use of expensive procedures.

Dentist 19



ISSN 2053-4086(Print), ISSN 2053-4094(Online)

Significantly overrates his works for provision of significant services with the use of expensive works.

CONCLUSION

The analytics showed that 7 dentists of 31 purposely overrate the works performed under insurance.

With the help of k-means algorithm, 4 clusters were generated:

Cluster 1: Special work with the use of expensive additional procedures, average age of clients – 25, average cost of services – 715$.

Cluster 2: Insignificant work without use of additional procedures, average age of clients – 21, average cost of services – 286$.

Custer 3: Significant work with the use of expensive additional procedures, average age of clients – 38, average cost of services – 819$.

Cluster 4: Significant work with the use of cheap additional procedures, average age of clients – 27, average cost of services – 551$.

After comparison of average cost of services in each cluster with average cost of services of each dentist, 4 dentists were detected, who are potential fraudsters.

Significantly overrates works for provision of significant and insignificant services with the use of cheap additional works.

As a result, 7 potential fraudsters were detected: Dentist 0, Dentist 5, Dentist 13, Dentist 14, Dentist 15, Dentist 19, and Dentist 27.

Applied techniques for analytics of data of insured events allow computing, how much work of certain dentists differs from the norm. Important issues are solved: How many fraudulent dentist are there? How much money are subject to the risk of activities of the latter?

Cluster algorithms (K-Means algorithm and EM algorithm) are convenient tools for answering the set questions.

REFERENCES

Bushinskaya, M.G. Indices of Thwarting Fraud Investigation / Law and Justice. 2005, No. 12.
Bushinskaya, M.G., Dubrovin, S.V. Concealment of Crime as a Method to Thwart Fraud Investigation / Law and Justice. 2006, No. 1.
Churyak, K.N. Deception as a Method of Committing Crime in History / Herald of the Ufa Law Institute of the Russian Interior Ministry. 2002, No. 3.
Guyva, O.A. Fraud: Use of Special Knowledge in Investigation / “Black Holes” in Russian Law. 2007, No. 5.
Han J., Kamber M., Data mining: Concepts and Techniques. – Morgan Kaufmann Publishers. – 2001.

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

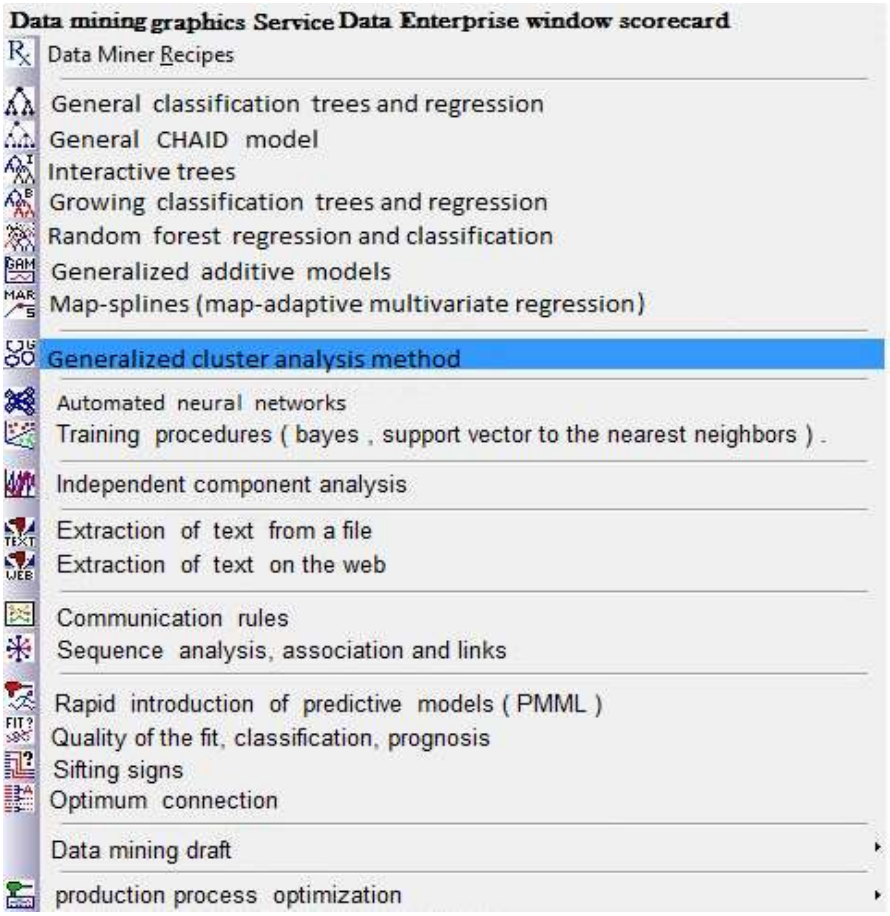
Ilyin, I.V. Historical Development of the Criminal Law Concept of Fraud in Russian Law / History of State and Law. 2007, No. 3.
Kochkina, S.V. Subjective Aspects of Fraud / Herald of the Barnaul Law Institution of the Russian Interior Ministry. 2006, No. 10.
Kolitsin, Ye.V. On Some Issues of Fraud / Herald of the Moscow University. 2006, No. 1.
Konar A., Artificial intelligence and soft computing: behavioral and cognitive modeling of the human brain. – CRC Press LLC. – Boca Raton, Florida/ – 2000.
Kramin, Ye.S. On Methods to Raise Efficiency of Preventing Fraud in the Consumption Market / Investigator. 1999, No. 9.
Kurinov, B.A. On Some Scientific and Legal Aspects of Criminalistic Description of Fraud / State and Law. 2005, No. 12.
Lanovoy, A.F. Method of Fraud: Criminalistic Analysis of the System of Fraudster's Acts and its Terminological Designation / Russian Investigator. 2007, No. 6.
Ledyayev, A.P. Classification of Criminalistically Significant Indices of Organized Fraud / Russian Judge. 2006, No. 9.
Limonov, V.N. Criminal Law Assessment of Fraud / Journal of Russian Law. 2002, No. 12.
Mikautadze, S.R. Structure and Dynamics in the Development of Professional Crime in Modern Russia / Russian Investigator. 2006, No. 8.
Mikhal, O. Complexity of Fraud Qualification / Criminal Law. 2007, No. 6.
Minayeva, A.V. Specific Implementation of Individual Measures of Investigating Fraud Cases / Law and Justice. 2006, No. 1.
Mitra S., Acharya T., Data Mining. Multimedia, Soft Computing, and Bioinformatics. – John Wiley & Sons, Inc. – Hoboken, New Jersey. – 2003.
Navalikhin, A.A. Thwarting of Fraud Investigation and Criminalistic Methods of its Prevention. Tyumen Law Institute of the Russian Interior Ministry, 2008.
Petukhov, B.V. On the Concept of Fraud / Lawyer. 2004, No. 3.
Popov, V.A. Criminalistics. Legal Literature, M., 2002.
Schepalov, S.V. Fraud is an Intentional Infliction of Property Damage / Russian Justice. 2003, No. 1.
Semenov, V.M. On the Concept of Stolen Item // Russian Investigator. M., Lawyer, 2005, No. 9. P. 34-37.
Shagiakhmetov, M.V. Specifics of Investigating Major Frauds / Legality. 1999, No. 12.
Stepanov, R.G. Data Mining Technology: Intellectual Data Analysis. Kazan, 2008. 58 pages.
Taratunin, B.K. Fraud or Robbery? // Criminal Process. 2005, No. 2.
Yeliseyev, V.V. Specific Individuality of the Fraud Victim / Modern Law. 2006, No. 8.
Zavidov, B.D. Criminal Law Analysis of Fraud / Advocate. 2002, No. 6.
Zeynalov, M.M. Controversial Issues in Determination of Fraud Object / Modern Law. 2007, No. 12.

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

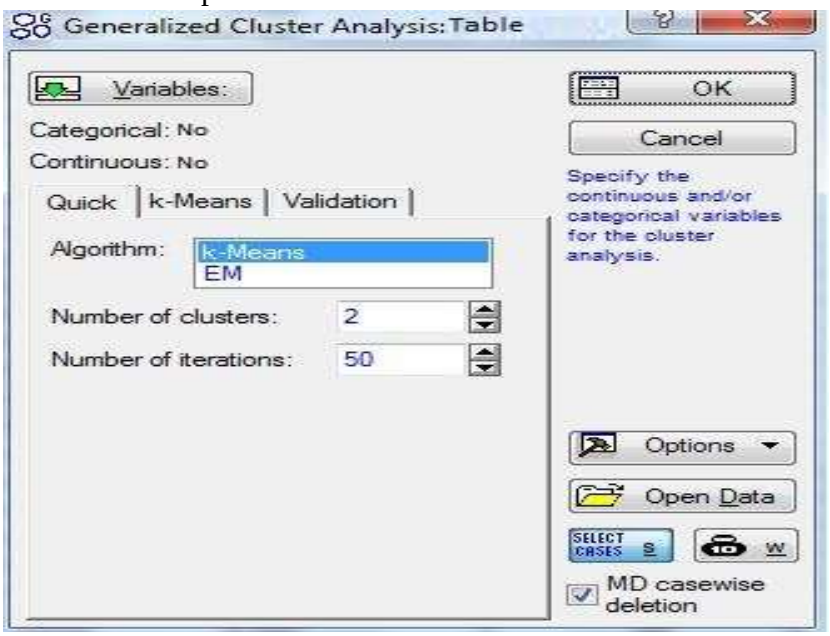
Appendix 1. K-Means algorithm launch algorithm

Step 0 (Module)

Open the tab Data Miner and select module Generalized EM and k-Means Cluster Analysis.



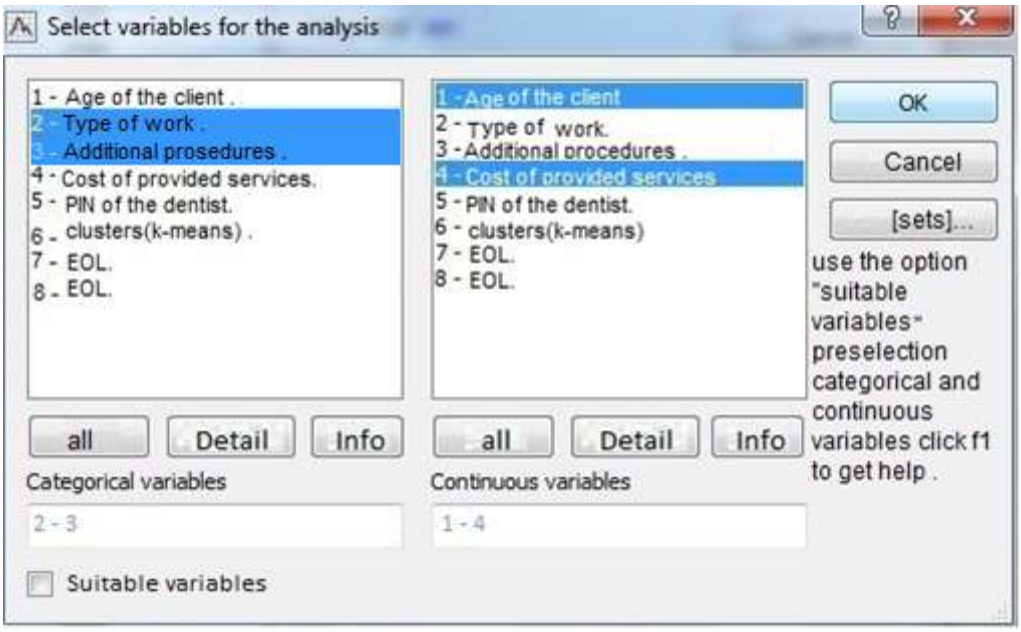
Dialogue window will open:



ISSN 2053-4086(Print), ISSN 2053-4094(Online)

Step 1 (Selection of variables).

Click Variables.



As categorical variables, we will select:

- Type of work.
- Additional procedures.

As continuous variables:

- Age of the client.
- Cost of provided services in $.

Step 2 (Cauterization parameter setting).

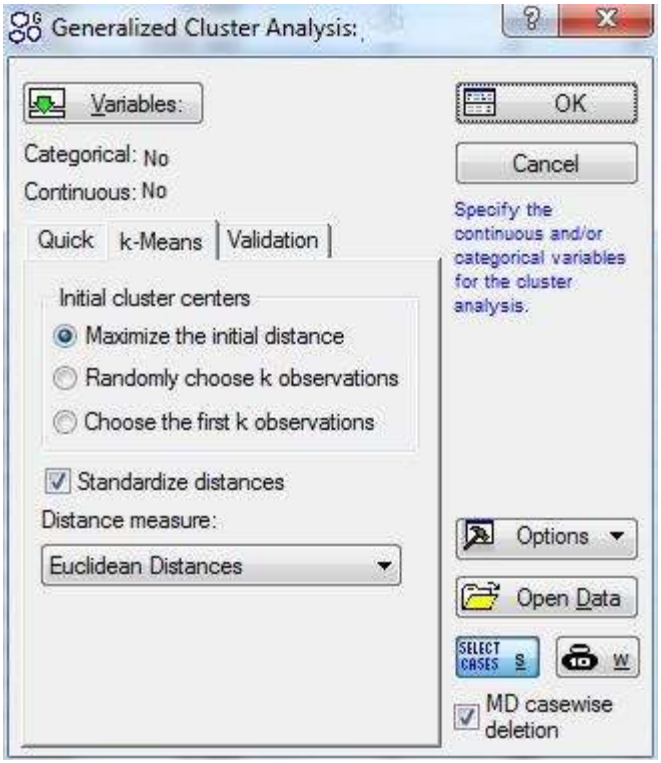
In the tab Quick, we will select:

- K-Means.
- Number of cluster – 2
- Number of iterations – 50.

As shown on the figure above.

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

Let's go to the tab k-means:



The following parameters are set in this tab:

- Initial cluster centers.
- Connection measurement.

Leave on default. Initial cluster centers will be determined so as they have the maximum distance between them. Connection measurement (metrics in multidimensional space) – Euclidian.

Step 3 (Validation).

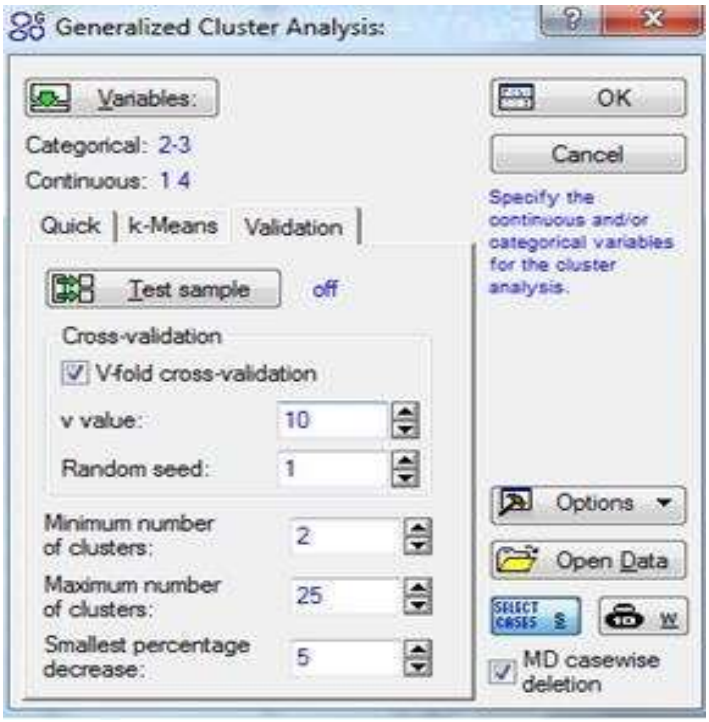
In the tab Validation, let's tip the box "cross validation". The rest parameters shall be left unchanged. Click OK.

ISSN 2053-4086(Print), ISSN 2053-4094(Online)

Appendix 2. EM algorithm launch algorithm

Repeat Step 0, Step 1 (Appendix 1).

Step 2 (Cauterization parameter setting).

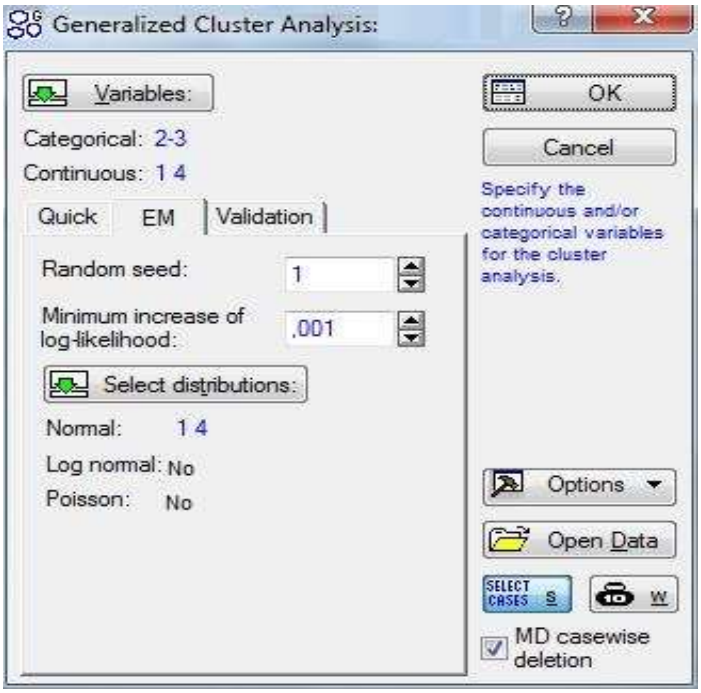


In the tab Quick' we will select:

1. EM algorithm.
2. Number of cluster – 2.
3. Number of iterations – 50.

Let's go to the tab EM:

ISSN 2053-4086(Print), ISSN 2053-4094(Online)



The following parameters are set in this tab:

- Random seed.
- Minimum increase of log-like hood.

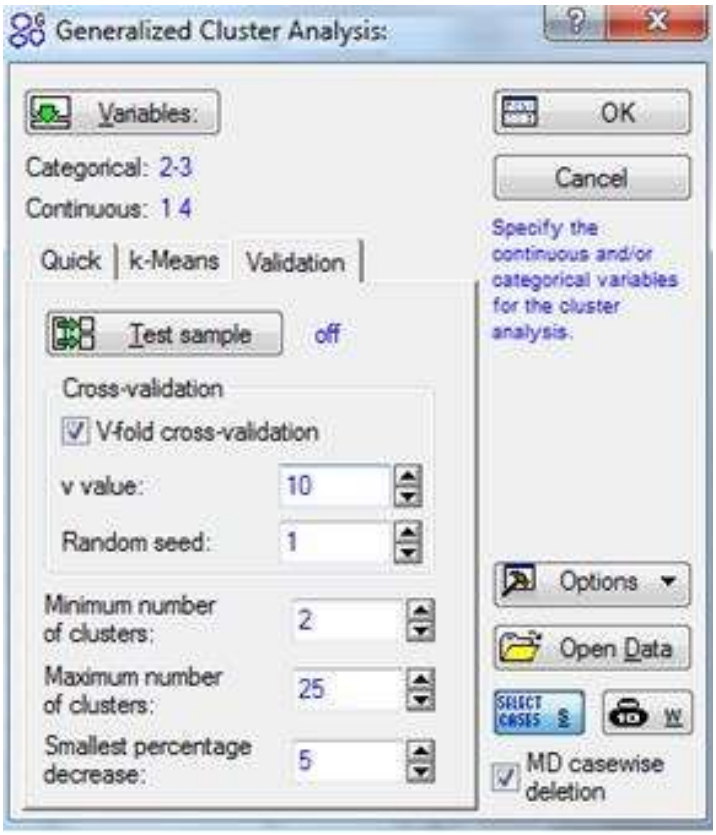
Let's leave them on default.

For continuous variables, we will select normal as distribution.

Step 3 (Validation).

In the tab Validation, let's tip the box cross validation. The rest parameters shall be left unchanged. Click OK.

ISSN 2053-4086(Print), ISSN 2053-4094(Online)



ISSN 2053-4086(Print), ISSN 2053-4094(Online)