"Dimitrie Cantemir" Christian University
Knowledge Horizons - Economics
Volume 9 , No. 1, pp. 43–47
P-ISSN: 2069-0932, E-ISSN: 2066-1061
© 2017 Pro Universitaria
www.orizonturi.ucdc.ro

METHODS USED IN DATA MINING

Ana-Maria Ramona STANCU[1] , Mihaela MOCANU[2]
Academy of Economic Studies, Bucharest, Romania, E-mail ana_maria_ramona@yahoo.com
"Dimitrie Cantemir" Christian University, Bucharest, Romania, E-mail rmocanu99@yahoo.fr

Abstract *Data mining is the process of discovering information in data warehouses and can be described as a unifier between Statistics, Artificial Intelligence and databases. Data mining techniques allow information extraction and making forecasts starting from historical data. In the first part of the paper we describe the methods used in Data Mining, then implement data in ODM and study the regression model.*

Key words:
classification, clustering, methods, model, regression
JEL Codes:
C8, C82

1. INTRODUCTION

The knowledge-based discovery methods are based on two ways in order to meet the objectives, namely: the first way is based on forecasts and the other one is based on data description. The method based on description studies the relationships within them and interprets the data, and the methods based on predictions put emphasis on data behavior. The techniques used are usually based on induction, the model learning rules from a set of data which is then tested on new data model until the acceptable model in terms of results accuracy is achieved.

The researched methods are:

a) Classification. Classification is one of the most popular operations in Data Mining and it is used for business issues such as: customer renunciation rate, risk management and ads related to the content of a site. The classification consists in grouping the cases based on a predictable attribute. Each case contains a set of attributes, out of which one is the classification attribute (predictable attribute). The operation consists in finding a model that describes the predictable attribute as a function of other attributes taken as input values. Data mining algorithms that require a data set on which to perform a training-instruction operation are called *supervised algorithms*.

b) Clustering. Clustering is a statistical method used for grouping multi-dimensional data.Namely, it is useful to perform summarization of large amounts of information, each group contains several items with similar characteristics

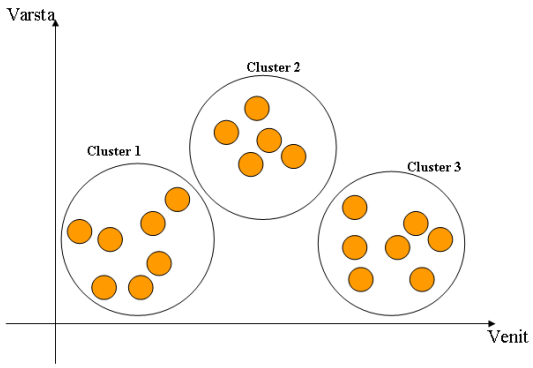

Figure 1 - People clusters according to the age and income

For example, if we have the attributes: age and income, then the segmentation algorithm gathers in data sets, as follows:

- Cluster 1: includes young population with a low income;
- Cluster 2: includes average age population with income;
- Cluster 3: includes older population with a low income;

Segmentation is an uncontrolled data mining operation, there is no attribute that may lead to the instruction process, all input parameters are treated equally. Most clustering algorithms build their model by means of iterations which stop when the model is fully covered, that is when the boundaries of these segments are stabilized.

c) Association. The association also called "shopping basket analysis" is another operation used in Data Mining. The best example of business problem analysis that uses association is a table of sales transactions and the identification of those elements that are most often found in the same "shopping basket". The basic use of the association is to identify the common sets of products and rules for cross-selling. In terms of association, each product or each pair attribute-value is considered an *item*. The association has two goals: to find the most frequent sets of items and rules of association. Most algorithms meet these targets by scanning the original data set several times. The frequency threshold is defined by the user before the model processing.

In addition to identifying the common sets of items based on a threshold frequency, most association algorithms also find the rules of association. An association rule has the form: (A, B) => C with a probability *p*, where A, B, C are common sets of items. In the specialized literature, Data Mining, this probability is called *confidence*. The probability is a value that the user must specify before the training of an association model.

d) Regression. Regression is similar to classification, and the main difference between the two models is that, in case of regression, the predictable attribute is a continuous number. The regression techniques have been studied for hundreds of years in the field of statistics. Linear regression and logistic regression are the most widely used regression methods. Other regression techniques are regression trees and neural networks.

e) Forecasting. Forecasting is another important method in Data Mining, and the input values are time series containing ordered auxiliary observations, and forecasting techniques work with general trends and periodicities, and the most commonly used technique is ARIMA (the Auto Regressive Integrated Moving Average model).

The following figure contains two curves: the thickened curve shows the real evolution of Microsoft shares for a period, and the thin curve is a time series model which was obtained by the forecasting technique.

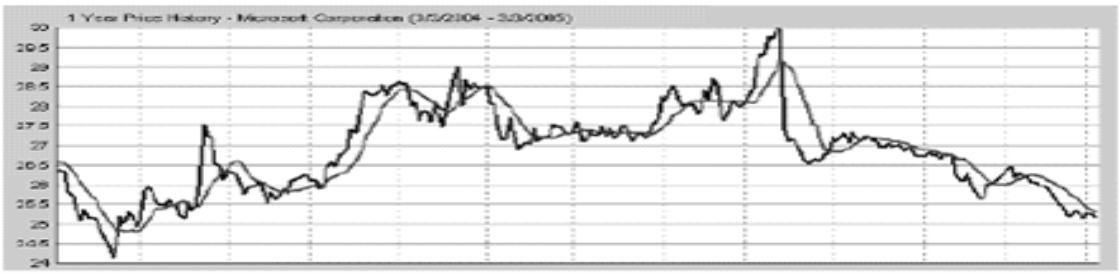

Figure 2 - The actual and forecasted evolution of Microsoft shares

f) The sequential analysis. *The sequential analysis* is used for finding patterns in a discrete series. A sequence consists of a number of discrete values. The sequence analysis is a relatively new Data Mining method and it becomes important for two reasons: the analysis of the DNA and of the Websites' files. Currently, there are several sequences analysis techniques such as the Markov chains.

g) The deviation Analysis. The deviation analysis aims at finding those rare cases that behave differently from the majority. It is also called fraud detecting as it relates to the detection of those behaviors that differ from the commonly observed

behaviors, repeatedly. This method is used to detect credit card fraud. The identification of abnormal cases out of the millions of transactions is a real challenge. Other applications are: computer networks breakage detection, error analysis in production, etc. This operation is still in the research stage because there is deviation in standard techniques for deviation analysis. Usually, for this model the analysts develop modified variants of decision trees or neural network algorithms. In order to generate significant rules, it is necessary to determine sets of abnormal cases within the involved sets.

2. REGRESSION ANALYSIS WITH ODM

The application we developed on a regression in Oracle Data Mining is called **w_regresion_10**, application performed using the *table* R_Operations from which we selected the field Id_Subsidiary from the Data Analysis which we grouped according to Data_Operation, and from class regression we chose Amount_Credit according to Code_Client.

In the first stage of the application we visualized the algorithms REGR_GLM and REGR_SVM after which we found out the company's impact on the people surveyed according to the *loan amount* granted.

After applying the algorithm GLM on the loan granted by a bank, the customer takes into account the *debit amount* granted (the coefficient is 95,98) and the bank employee (the coefficient is 107.37). (Figure 3)

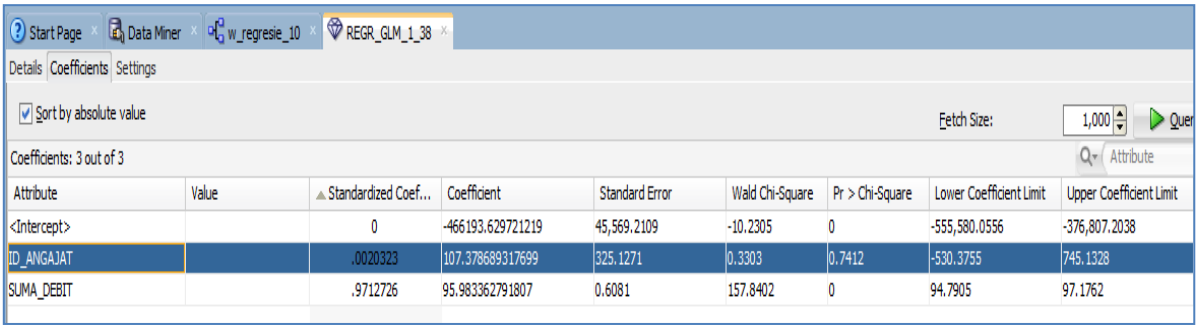

Figure 3 –GLM model visualization

Namely,

Attribute	Standard coefficient	Coefficient	Standard error	Wald Chi square	Pr > Chi square	Lower coefficient limit	Upper coefficient limit
Id_Employee	.00203	107.37	325.12	0.33	0.74	-530.37	745.13
Amount_Debit	.097127	95.98	0.60	157.84	0	94.79	97.17

Table 1 – Re-writing GLM model visualization

In the second stage we compare SVM and GLM algorithms having tested them.

In the first stage we studied the performance of the three algorithms and noticed that SVM and GLM algorithms have Mean Actual Value = 1.124, and GLM algorithm has predictive confidence (is 80.59) higher than the SVM predictive confidence (is 77.97).

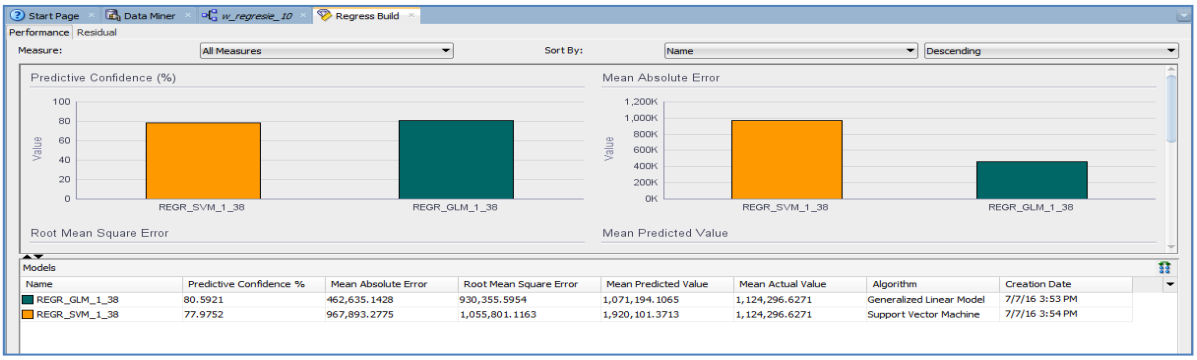

Figure 4 – Comparison of algorithms performances

Namely,

Model	Predictive Confidence %	Mean Absolute error	Root Mean Square Erros	Mean Predicted Value	Mean Actual Value
GLM	80,5921	462635,14	930355,59	1071194,10	1124296,62
SVM	77,9752	967893,27	1055801,11	1920101,37	1124296,62

Table 2 – Re-writing the algorithms performances

After testing the GLM algorithm – We residually notice the fact that staring from values higher than 20,000, there is a different distribution of points; so, a single model can be created by sharing the set of data according to the loan amount exceeding or not this amount.

It can be noticed that the predictive confidence (80,59) of the GLM algorithm is higher than the of the SVM algorithm which is 77,97, and the average absolute error of the GLM algorithm (is 462,635) is lower than that of the SVM algorithm which is 967,893.

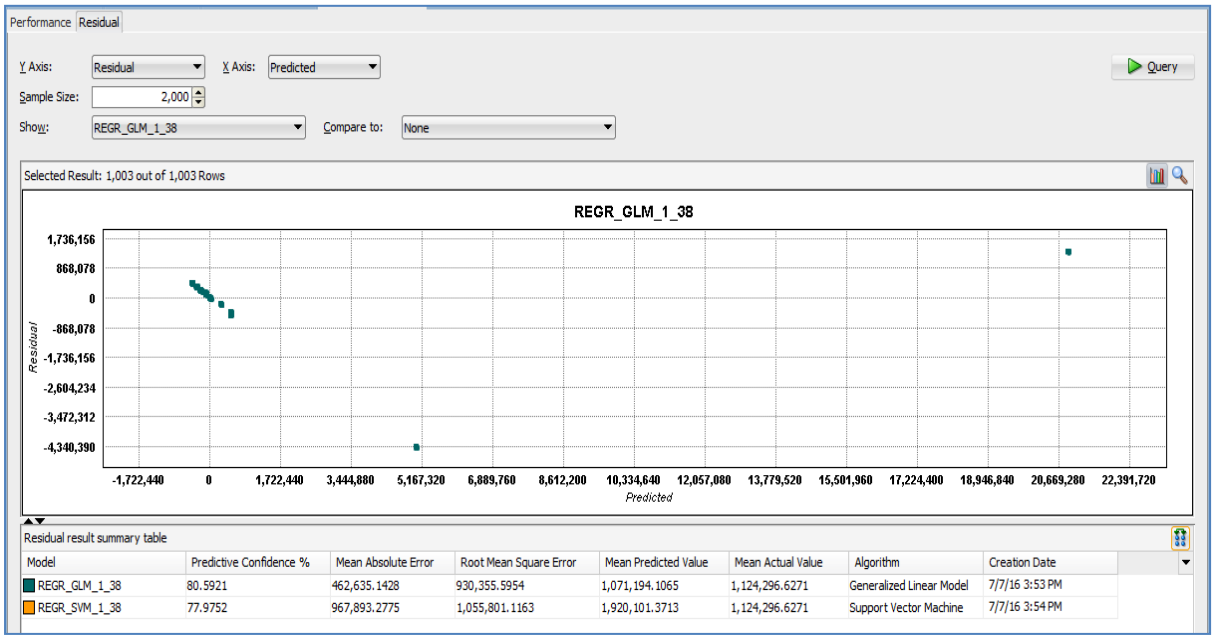

Figure 5 –Algorithms comparison – residual

Namely,

Model	Predictive Confidence %	Mean Absolute error	Root Mean Square Erros	Mean Predicted Value	Mean Actual Value
GLM	80,5921	462635,14	930355,59	1071194,10	1124296,62
SVM	77,9752	967893,27	1055801,11	1920101,37	1124296,62

Table 3 – Re-writing the Algorithms comparison - residual

In this application, after the application of the regression algorithms, we wanted to find out if the Amount _Debit's charged by the bank is important when appealing to its service. After visualizing the GLM model we noticed that both the bank employee (coefficient 107.37%) and the debit amount (coefficient 95.98%) are important factors when appealing to a banking service.

3. CONCLUSIONS

Depending on the specific case, some Data Mining techniques are more effective than others even when for solving a problem there is only a single option. In the implemented model we wanted to find out if the loan amount and the bank branch are important factors in choosing the bank. According to the results, after applying the model we saw that the bank branch and the loan amount are important factors when a person uses the services of a bank.

4. REFERENCES

1] Berry M.J., Linoff G.S., *Data mining Techniques: For marketing, sales, and customer support*, John Wiley & Sons, Inc., (1997)

[2] Quinlan J.R., *Decision trees and decision-making*, IEEE Transactions on Systems, Man and Cybernetics, vol. 20, no. 2, (1990), pp. 339-346

[3] Gh. M. Panaitescu, *Transmiterea si codarea informatiei [Information Transmission and Coding]*, Course notes, Oil and Gas University, Ploiesti, Department of Automation, Computers and Electronics, 2015

[4] Gh. M. Panaitescu, *Transmiterea si codarea informatiei [Information Transmission and Coding]*, Course notes, Oil and Gas University, Ploiesti, Department of Automation, Computers and Electronics, 2015