Processing of large amounts of data on a credit scoring example using neural network technology

K. K. Nurlybayeva & G. T. Balakayeva
Al-Farabi Kazakh National University, Kazakhstan

Abstract

Nowadays there is the growing problem of mining large amounts of data. This article is dedicated to the issue of development of a credit scoring model as an example of processing large volumes of data. Some data mining algorithms are described in the paper. Three methods have been used for the experiment; namely, logistic regression, decision trees and neural networks. All of them have been applied for the modeling of credit scoring. According to the results of a comparative analysis, neural networks have been selected as a technology for the credit scoring model design. The main aim is to choose the best method of data mining and construction of predictive credit scoring without using expensive software, together with the ability for self-learning and updating. To implement and achieve the goal, the following tasks have been undertaken: collecting and preparing the initial data, analysis and selection of available technologies and methods of solution, to determine the most suitable method of data mining to build a credit scoring system, and now the project is on the way to creating an expert system.

Keywords: data mining, logistic regression, decision trees, neural networks, scoring model, credit scoring system.

1 Introduction

Nowadays, the problem of mining large amounts of data is growing every day. This article is dedicated to the issue of the development of credit scoring as an example of processing large volumes of data. Some data mining algorithms are used in this paper. According to the results of a comparative analysis, neural network technology is more preferable for the construction of credit scoring.

WIT Transactions on The Built Environment, Vol 134, © 2013 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/SAFE130161

A key indicator of the effectiveness of the bank is the ability to manage credit risk, which is the main banking risk. A significant part of the income generated by banks takes the credit activity. The assessment of the potential benefits in relation to the default probability has a special importance and relevance at the present time

2 Credit scoring

At the moment there are a lot of technologies in the banking industry which are applied when analysing the creditworthiness of borrowers. One of them and the most basic is credit scoring. Scoring is an assessment of the solvency of bank customers. In addition, scoring can be identified as a mathematical or statistical model, based on the credit history of previous clients who are already consumers of the credit institution services, where the bank can determine the probability that the loan will be returned on time by the customer; that is, the diagnosis of the probability of bankruptcy of the potential borrower when considering its lending. Thus, the client is evaluated by scoring and assigned a rating. The main tool is a scorecard – a mathematical model that allows for the comparing of the characteristics of the borrower with numerical values and ending up with a scoring ranking result.

3 Credit scoring types

Several types of scoring are distinguished: credit (or application) scoring, which is the obtaining of the credit rate of the potential borrower on the basis of some of its characteristics, which are primarily contained in the questionnaire of the borrower; behavioral scoring – a dynamic estimate of the expected behavior of the client to repay the loan, based on data of the transactions history on its accounts is used, in particular, for the prevention of debt. In addition, another type of scoring is produced which is a collector scoring (scoring penalties, collection scoring), designed for the selection of the priority of “bad” borrowers in arrears, and the areas of work in the recovery of their debt, as well as the scoring of persons who are applying for a loan. The latter type of scoring in domestic practice is often referred to as the screening of potential borrowers.

4 Application scoring

Credit scoring (application scoring) is a way to assess the client’s creditworthiness, which is based on numerical and statistical methods and concluded by assigning points; a rating, according to the filling of a questionnaire developed by credit risk assessors; risk managers. The calculated score system makes the decision to approve or deny a loan [1].

The main purpose is to choose the best method of data mining and the construction of predictive credit scoring without using expensive software, together with the ability for self-learning and updating. To implement and achieve the goal, the following tasks have been undertaken: collecting and preparing the initial data, analysis and selection of available technologies and methods of solution, to determine the most suitable method of data mining to build a credit scoring system, and now the project is on the way to creating an expert system.

Market analysis to date indicates that there are ready-made products in credit scoring: the software SAS Enterprise Miner, PolyAnalyst, STATISTICA Data Miner, Oracle Data Mining, analytical platform Deductor [3]. For many credit institutions, ready scoring systems are too expensive. A quality product in the banking industry is used mainly in the major lending institutions and banks. In small commercial banks or units, credit scoring may not be used because of the resistance of field staff and a commitment to the traditional methods of borrowers' creditworthiness assessment. Therefore, for such commercial organizations there is a need to create a system of credit scoring which is not overloaded with many ways of data analysis and pattern identification, but with one of the best methods and the most user-friendly interface.

The most powerful tools to assess the risk are self-learning algorithms, with the ability to adapt. Four algorithms are analyzed for the solution of the data mining problems of credit scoring: logistic regression, association rules, decision trees, neural networks.

The scoring system must be "trained" on several samples, and if the samples are small, we can use the models which will "propagate" or simulate this sample, based on available data. It is clear that training the system is also an effective method but it is impossible to have the same results as with real data analysis but this approach will certainly be much more effective than a "ready" system [2].

Introduction of scoring in the banking system practice is needed both by the banks in terms of confidence in the return of the loan by the borrower, and for borrowers, where the scoring system significantly reduces the time for making a decision on the bank loan.

The analysis of credit scoring is based on questionnaires. Analyzed data in this case are presented in the form of a regular table, which contains precedents; in our case, 1000 are used. Due to the secrecy of bank information and the inability to find the necessary data from credit bureaus because of the cost and the absence of the access to it to build and test, systems were used to test data that are used for similar applications. Additional facts used by the analyst, such as the date, the ratio of liabilities/income, age, length of stay in the region, marital status, education, work experience in the last place, the level positions, credit history, delinquencies on loans over the previous 60 days, the table also contains a column that indicates the answer: to give a loan or not.

On the basis of experiments, the method is selected for implementation.

4.1 Data mining methods

Methods which are used in the experiment:

$$S = p_1\overline{X}_1 + p_2\overline{X}_2 + \cdots + p_n\overline{X}_n$$

where S – the value of a generalized estimate of the object; $\overline{X}_1$, $\overline{X}_2$, …, $\overline{X}_n$ – normalized values of the factors that affect the analyzed characteristics of the evaluated object; $p_1, p_2, \ldots, p_n$ – weights that characterize the significance of the relevant factors for the experts. This model is used in the paper.

- Logistic regression

$$\log\,(p/(1-p)) = w_o + w_1x_1 + w_2x_2 + \ldots + w_nx_n$$

- Decision trees
- Neural networks

$$S = \sum_{i=1}^{n} w_i\,x_i$$

where n – number of neuron's inputs
x_i – value of the i-th neuron's inputs
w_i – weight of the i-th synapse

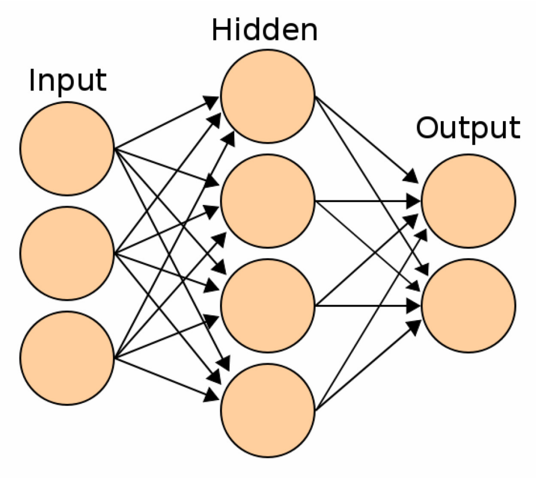

Figure 1: An example of neural networks.

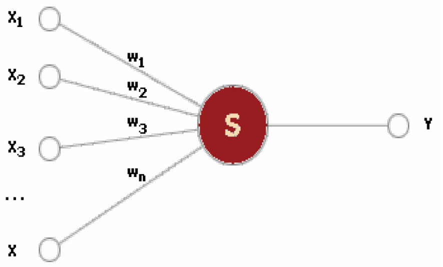

Figure 2: Presentation of the model of neural networks.

where n – number of neuron's inputs
x_i – value of the i-th input
w_i – weight of the i-th synapse
S – neuron
Y – output

The value of axon of a neuron is determined by the formula

$$Y = f(S) \tag{2}$$

where f – an activation function

Usually

$$f(x) = \frac{1}{1+e^{-\alpha x}} \tag{3}$$

The main advantage of this function is the fact that it is differentiable on the whole x-axis and has a very simple derivative:

$$f'(x) = \alpha f(x)(1 - f(x)) \tag{4}$$

4.1.1 Back-propagation neural networks

A back-propagation neural network is a powerful tool for searching patterns, forecasting, qualitative analysis.

A back-propagation neural network is composed of several layers of neurons, each neuron layer i is associated with each neuron layer $i+1$.

In general, the problem of neural networks training is to find some functional relationship $Y = F(X)$, where X – input, and Y – the output vectors. In general, such a task with a limited set of input data, has an infinite number of solutions. There is a task of minimization of the error of the neural networks by means of the least squares method:

$$E(w) = \frac{1}{2}\sum_{j=1}^{p}(y_j - d_j)^2 \tag{5}$$

where y_j – the value of the j-th output neural networks
d_j – a target value of the j-th output
p – the number of neurons in the output layer

Weights are changed at each iteration by formula:

$$\Delta w_{ij} = -\mu \frac{\partial E}{\partial w_{ij}} \tag{6}$$

where μ – parameter, which defines the speed of the training

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_j} * \frac{\partial y_j}{\partial s_j} * \frac{\partial s_j}{\partial w_{ij}} \tag{7}$$

where y_j – the value of the j-th neuron output
S_j – weighted sum of the input signals, defined by formula (1)

There is

$$\frac{\partial S_j}{\partial w_{ij}} \equiv x_i \tag{8}$$

where x_i – value of the i-th neuron

The definition of the first multiplier of formula (7)

$$\frac{\partial E}{\partial y_j} = \sum_k \frac{\partial E}{\partial y_k} * \frac{\partial y_k}{\partial S_k} * \frac{\partial S_k}{\partial y_j} = \sum_k \frac{\partial E}{\partial y_k} * \frac{\partial y_k}{\partial S_k} * w_{jk}^{(n+1)} \tag{9}$$

where k – number of the n+1 layer's neurons

Let us assume that

$$\delta_j^{(n)} = \frac{\partial E}{\partial y_j} * \frac{\partial y_j}{\partial S_j} \tag{10}$$

Then we can define the recursive function for the determination of the n-th layer, if the (n+1) th layer is defined:

$$\delta_j^{(n)} = \left[\sum_k \delta_k^{(n+1)} * w_{jk}^{(n+1)}\right] * \frac{dy_j}{dS_j} \tag{11}$$

The last layer is determined by formula

$$\delta_j^{(N)} = \left(y_i^{(N)} - d_i\right) * \frac{dy_i}{dS_i} \tag{12}$$

Then formula (6) is defined as follows

$$\Delta w_{ij}^{(n)} = -\mu\delta_j^{(N)} * x_i^n \tag{13}$$

Neural network learning algorithm:

1. Determine the inputs and ouputs of the neural network
2. Calculate the output layer by formula (12) and calculate the changes of the weights of the output layer N by formula (13)
3. Calculate for the other layers of NN, *n = N-1..1* by formulae (11) and (13) respectively
4. Adjust the weights of the neural networks:

$$w_{ij}^{(n)}(t) = w_{ij}^{(n)}(t-1) + \Delta w_{ij}^{(n)}(t) \tag{14}$$

5. If the error is essential, then go to step 1.

5 Conclusion

The aim of the paper is to describe the neural network model developed for the credit scoring system for the banking system of Kazakhstan. On the basis of this experiment, we can conclude that the instance of error in the neural network is lower than with other data mining techniques. It is important for the Bank to achieve the maximal accuracy when determining the reliability of the borrower, so the system of the credit scoring using a neural network has the best structure, which is determined through experiments.

Thus, a developed scoring system based on a neural network will allow the credit institution to obtain an effective competitive advantage to maintain and improve their competitive position in the market and survive against competitors for a long time. As an advantage of the neural network system, it may be noted that when the state of the market and the banking system changes it is possible to adapt the model to the new realities. This requires training of the neural network on a new sample of clients. The analysis of the possible automatic updates of the credit scoring system parameters by using the neural networks methods will be done in future work where parallel computing techniques will be implemented also for the comparison of time consumption.

References

[1] Melikova A.V. “Optimization of before credit processing. Effective decision-making”, ed. “Bank lending”, 2007

[2] Pishulin A. “Credit scoring system: necessities and advantages”, ed. “Financial Director”, 2011

[3] http://www.credits.ru