Management Science Letters 2 (2012) 1683–1688

Contents lists available at GrowingScience

Management Science Letters

homepage: www.GrowingScience.com/msl

A hybrid model using decision tree and neural network for credit scoring problem

Amir Arzy Soltan and **Mohammad Mehrabioun Mohammadi**

Department of Management, University of Tehran, Tehran, Iran

ARTICLEINFO

Article history:
Received December 25, 2011
Received in Revised form
March, 25, 2012
Accepted 24 April 2012
Available online
April 30 2012

Keywords:
BPNN
Neural network
Data mining
Information Technology

ABSTRACT

Nowadays credit scoring is an important issue for financial and monetary organizations that has substantial impact on reduction of customer attraction risks. Identification of high risk customer can reduce finished cost. An accurate classification of customer and low type 1 and type 2 errors have been investigated in many studies. The primary objective of this paper is to develop a new method, which chooses the best neural network architecture based on one column hidden layer MLP, multiple columns hidden layers MLP, RBFN and decision trees and assembling them with voting methods. The proposed method of this paper is run on an Australian credit data and a private bank in Iran called Export Development Bank of Iran and the results are used for making solution in low customer attraction risks.

1. Introduction

One of the primary concerns in financial management is to reduce credit risk as much as possible and credit risk analysis plays an important role on the success of this industry. There are different methods for assessing risk and credit scoring is one of the most popular techniques, which has been widely used among practitioners. Credit scoring has been used for making appropriate decisions on giving loans or credits to customers. There are other techniques used for pattern classification using historical data from real-world case studies called data mining. The results of all these techniques are summarized as a set of rules and regulations, which are normally used for making appropriate decision on loans. Some of these rules and regulations are adjusted by a committee called Basel and all banks and financial institutions are asked to use these regulations.

Many people believe the financial crises happened in 2008 in United States was a result of violation in such regulations. The rules and regulations are also changed to meet market changes and Basel was also modified based on these changes into Basel II. During the past few years, there has been an increase awareness and concerns on how to assign loans to customers more appropriately (Moretto &

* Corresponding author.
E-mail addresses: mehrabioun@ut.ac.ir (M.M. Mohammadi)

doi: 10.5267/j.msl.2012.04.021

Tamborini, 2007; Goodhart, 2011). There are two types of credit scoring: In the first one is called application scoring and according to this method, credit applicants are divided into good and bad risk groups and to different types of credit seekers' financial and personal information are gathered. The second method, on the contrary, uses only current customers' credit information to find customers' payment pattern on loan called behavioral scoring and the focus of this paper is on this type of method (Siddiqi, 2005; Lee et al., 2002; Chuang & Lin, 2009).

Credit scoring technique uses a lists of a number of questions called characteristics for assigment of loans to customers who look for receiving credits. The replies for these questions are processed using different techniques and one of the most popular one if neural network (NN). NN is a flexible method since it allows the characteristics to be interacted in different forms and it consists of one or more groups of connected characteristics.
A single characteristic is normally connected to different characteristics, which make up the whole sophisticated network structure, which outweigh decision trees and scorecards since they do not assume uncorrelated relationships among characteristics. They also do not have any problem from structural instability in the same way as decision trees since they normally cannot depend on a single first question for constructing the whole network. However, the development of the network depends on the qualitative information, which are solicited to detremine the interactions among all different characteristics.

2. Literature review

We will review the literature of credit scoring and the commonly used techniques in feature selection problem. The model can function as a classification model or a correlation model, based on the problem taken into account. For either the classification model or the regression model, it generally consists of two modules. Module 2 utilizes credit scoring related data to train decision tree or neural network model and apply the trained models to new data.

In this section, we explains some of the necessary issues involved with the proposed model of this paper. The first issue is to underestand the idea of using correlation. The primary purpose of linear correlation analysis is to measure the strength of a linear relationship between two variables and they help because they can demonstrate a predictive relationship, which could be exploited in practice (Robert & Johnson, 2011). The correlation can either be positive or negative depending on weathere y increases or decreases as x increases, and there are literally various correlation coefficients, often showed as ρ or r, measuring the degree of correlation. The most commonly used type of correlation is the Pearson correlation coefficient, which is sensitive only to a linear relationship between two variables and can defined as follows,

$$\rho_{X,Y} = \mathrm{corr}(X, Y) = \frac{\mathrm{COV}(X,Y)}{\sigma X \, \sigma Y}$$

Linear correlation analysis is different from association rule mining or market basket analysis (Chiang, 2007; Zhou & Yau, 2007), since it identifies relationships among attribute groups and is suited to making general inferences about a domain. Association rule analysis, in contrast, detects the relationships among the attributes' instances and is useful for deriving local properties of the instances (Chang, 2007). Linear correlation analysis has been widely implemented in data mining and feature selection issues, Taniguchi & Haraguchi (2006) developed an algorithm to find high correlated instances in large databases. The results of paired item sets with high correlation in one database is already known as discovery of correlation, which results huge dimension reduction and minimizing time costs (Chang, 2007). Chiang et al. (2005) described an automatic linear correlation discovery methodology, which adopts statistical measurement functions to discover correlations from databases' attributes. Their methodology automatically pairs attribute groups having potential linear correlations, measures the linear correlation of each pair of attribute groups, and confirms the

discovered correlation so proposed methodology facilitates linear correlation discovery for databases with a large amount of data. Several attempts have been made to correlation discovery (Chua et al., 2002; Taniguchi & Haraguchi, 2006; Pan et al., 2004), these approaches developed new methods for finding paired item sets with high correlation in one database, which yields better data management and data analysis. These studies used these three targets: 1- identifying appropriate linar correlation among pairs 2- seeking potentially significant itemset pairs 3- derive applicable knowledge from data analysis.

An Artificial Neural Network (ANN) is an information processing paradigm and the idea comes from the way biological nervous systems such as the brain, process information operates. The key factor of this paradigm is the structure of the information processing system composed of a large number of highly interconnected processing elements called neurons working in unison to solve particular problems (Zhou & Yau, 2007; Tsai & Wu,2008). A neural network (NN) features different interconnected nodes serving as signal receivers and senders, the network architecture designed to describe connections among the nodes, and the training algorithm associated with finding values of network parameters (weights) for a particular network. ANN has been implemented in different applications such as engineering, science, education, social research, medical research, business, finance, forecasting and related fields (Zhang et al., 1998; West, 2000; Lee & Chen, 2002).

A neural network is a system, which consists of highly inter-connected, interacting processing components based on neuro-biological techniques. NNs process data through the interactions of a relatively large number of processing items and their connections to some external inputs. Backpropagation neural networks (BPN) is a gradient steepest descent training technique utilized paradigm to date in business applications and it is a network, which includes a number of neurons connected by links. The nodes in the network can be classified as three various layers: the input layer, the output layer, and one or more hidden layers. The nodes in the input layer receive input information from external sources and the nodes in the output layer give the target output signals (Wang & Xu, 2010). For the gradient descent algorithm, the step size, called the learning rate, is crucial since smaller learning rates normally slow down the learning process before convergence (Talukder & Casasent, 2001).

A decision tree is a graphical scheme of a procedure for classifying or assessing an alternative of interest. By graphical representation, they clearly explain how to reach a decision, and they are capable of constructing from labeled instances. There are two well-known techniques for constructing decisiontrees, which are C4.5 and CART. A decision tree is normally built recursively in a top-down manner. If a set of labeled instances is relatively pure, then the tree is a leaf, with the assigned label being that of the most frequently occurring class in that set. Otherwise, a test is built and placed into an internal node, which constitutes the tree so far. The test classifies a partition of the instances according to the outcome of the test as applied to each instance. A branch is built for each block of the partition, and for each block, a decision tree is built, recursively.

NNs provide a new approach for feature extraction using hidden layers and classification, Besides, existing feature extraction and classification algorithms are able to be mapped into NN architectures for effective hardware implementation. For classification or prediction, BPNN is the most widely used neuralnetwork method and it is adopted for the propsoed study of this paper.

3. Proposed model

In order to verify the accuracy and effectiveness of the proposed two-stage credit scoring model using linear correlation and artificial NN, one private bank in Tehran, Iran is used in this study, There are totally 223 housing loan customers in the dataset with 45 good credit customers while the remaining 178 are bad credit customers. The 10% relative ratio of bad credit customers to total customers is

very close to the national standard in Iran and hence should be a representative dataset in verifying the feasibility of the proposed scheme. Each bank customer in the dataset contains 46 independent variables which can be summarized in Table 1 and the dependent variable is the credit status of the customer-good or bad credit. The correlation models will be implemented using the popular SPSS software (SPSS, 2010) . All the modeling tasks are implemented on an IBM PC with Intel Pentium four 800 MHz CPU processor. So 13 variables has been ommiteed and after splitting 4 variables there has beeen finally 43 variable that can predict credit scoring.

During this phase we implement a supervised NN, which is based on the BPNN due to its ease of implementation and the availability of necessary dataset for training and validating this supervised learner. The NN input layer has 24 neurons based on the number of the applicant's attributes numerical input values; each input neuron receives a normalized numerical value. There is one hidden layer containing h neurons; the number of hidden neuron depends on the NN framework. We apply our investigation using three neural models, including ANN-1 MLP one hidden layer, ANN-2 MLP multiple hidden layer and ANN-3 RBFN . The optimum number of hidden neurons h in all three framework, which assures meaningful training while keeping the time expenditure to a minimum, was obtained after several experiments involving the adjustment of the number of hidden neurons from one to 50 neurons. The output layer has one single neuron, which applies binary output data representation; '0' for accepting or '1' for rejecting a credit application. A threshold value of 0.5 is implemented to distinguish people with good and bad credit. If the output result of the NN is greater than or equal to 0.5, the presented case is assigned to one class of good, accept; otherwise it is assigned to the other class of bad, reject. Hence,

Applicant i has a good credit if : NN out>=0.5

Applicant i has a bad credit if : NN out < 0.5

Table 1 shows details of our variables used for the proposed study of this paper.

Table 1
Research variables

	Definition		Definition		Definition		Definition		Definition
X_1	Mineral and industrial products	X_{11}	Short term finance facility	X_{21}	Target market risks	X_{31}	Former period sales	X_{41}	Last three years average exports
X_2	Agriculture products	X_{12}	Long-term financial facilities	X_{22}	Company history	X_{32}	Two pre-sale	X_{42}	Last three years average imports
X_3	Chemical and oil product	X_{13}	Total debt	X_{23}	Executives history	X_{33}	Current term assets	X_{43}	Property type
X_4	Infrastructure services	X_{14}	Capital	X_{24}	Partnership and cooperative	X_{34}	Assets prior	X_{44}	Property age
X_5	Tax return	X_{15}	Stakeholders wage	X_{25}	Exchange LLP	X_{35}	assets Two- term ago	X_{45}	Loan type
X_6	Audit organization	X_{16}	Gross profit	X_{26}	PJS	X_{36}	Stakeholders current wage	X_{46}	Credit worthiness
X_7	Valid audit	X_{17}	Finance costs	X_{27}	Limited liability	X_{37}	Stakeholders prior period wage		
X_8	Current assets	X_{18}	Net profit	X_{28}	LLP	X_{38}	Stake holder's two ago wage period		
X_9	Non-current assets	X_{19}	Domestic market	X_{29}	Experience with banks	X_{39}	Current circulation of creditor accounts		
X_{10}	Total assets	X_{20}	Outside market	X_{30}	Current period sales	X_{40}	Current Accounts weighted Average		

In our model, NNout(i) is the output of the NN model obtained when the attributes of the i^{th} case (applicant) are presented to the network. This is basically the output credit decision associated with applicant i. So after running Clementine for each of architectures with different hidden layers, different nodes and different learning schemes we find 3 best NN architecture listed in Table 2 .

Table 2
Neural network architectures

NN model	NN type	Input layer nodes	First hidden layer nodes	Secound hidden layer nodes	Output layer nodes	Training- testing- validation ratio	Accuracy
1	MLP	42	10		1	(60-20-20)	81.7219
2	MLP	42	2	2	1	(50-30-20)	82.9707
3	RBFN	42	10		1	(60-20-20)	81.9991

After analysisng different decision tree architectures we find the best decesion tree architeture illustated in table and after that we ensembled NN architectures with the best decisin tree architecture with voting, weighted voting and maximum weighted voting strategy that has illustrated in table, the best accuracy for proposed model is found as 91.44 .

Table 3
Decision tree results

	Testing- training ratios								
	(90-10)	(80-20)	(70-30)	(60-40)	(50-50)	(40-60)	(30-70)	(20-80)	(10-90)
Accuracy	78.38	77.93	81.08	81.018	85.59	83.33	87.82	87.84	**89.19**

Table 4
Ensemble models

Ensemble model	Strategy	Accuracy
1	Voting	75.23
2	Weighted voting	83.71
3	Maximum weighted voting	**91.44**

We have integrated 223 training data sets with 43 possible factors associated with credit scoring and used the decision tree algorithm to calculate the error rate report and obtained its tree-shaped structure, as shown in Table 3. Therefore, there has been substantial difference between NN models, decisin tree models and ensembled model, so in case of validation section we considered output results with logestic regression and support vector machie, and results has been illustrated in table, which can inform us that proposed model has the best accuracy and minimun type I and type II error.

Table 5
Support vector machine results (accuracy= 79.73)

Real category	Classified category with support vector machine model	
	0	1
0	19(8%)	27(12%)
1	8(4%)	168(76%)

Table 6
Logistic regression results (accuracy= 83.78)

Real category	Classified category with logistic regrssion model	
	0	1
0	0(0%)	45(20%)
1	0(0. 9%)	175(79%)

4. Conclusion

This paper presented an empirical study on some neural network models for credit risk assessment under various learning schemes and implemented a simple method to use credit evaluation system. In

our approach we trained three models of a three-layer supervised NN; based on the back propagation learning algorithm, under nine learning schemes. We also described in detail the criteria and considerations to decide upon an optimum learning scheme and neural model. Furthermore, we proposed a simple but sophisticated technique of normalizing the input data to use for the proposed model. The proposed method of this paper was run on an Australian credit data and a private bank in Iran called Export Development Bank of Iran and the results were used for making solution in low customer attraction risks. In conclusion, the credit risk evaluation NN model performs reasonable well when using the maximum weighted voting learning scheme has 91.44% accuracy and 8% type I error and 0.4% type II error that has substantial difference with other classification models.

References

Chuang, C. L., & Lin, R.H. (2009). Constructing a reassigning credit scoring model. *Expert Systems with Applications*, 36, 1685-1694.

Chiang, R., Chua, C.E., Lim, E. (2005). Linear correlation discovery in databases: a data mining approach. *Data & Knowledge Engineering*, 53, 311-337.

Chang, H.J., et al. (2007). An anticipation model of potential customers' purchasing behavior based on clustering analysis and association rules analysis. *Expert Systems with Applications*, 32, 753-764.

Chua, C.E.H., Chiang, R.H.L., & Lim, E.P. (2002). An intelligent middleware for linear correlation discovery. Decision Support Systems, 32, 313-326.

Goodhart, C. (2011). *The Basel Committee on Banking Supervision*. Cambridge University.

Lee, T.S. et al. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, 23, 245-254.

Lee, T.-S., & Chen, N.-J. (2002). Investigating the information content of non-cash-trading index futures using neural networks. *Expert Systems with Applications*, 22, 225-234.

Moretto, M., & Tamborini, R. (2007). Firm value, illiquidity risk and liquidity insurance. *Journal of Banking & Finance*, 31, 103-120.

Pan, J.Y., Yang, H.J., Faloutsos, C., & Duygulu, P. (2004). *Automatic multimedia cross-modal correlation discovery*. In Proceedings of the 10th ACM SIGKDD Conference KDD 2004. Seattle, WA, 653-658.

Robert, P.J.K., & Johnson, R. (2011). Elementary Statistics. 11th Edition ed.: Brooks/Cole, (2011).

Siddiqi, N. (2005). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. Wiley.

Taniguchi, T., & Haraguchi, M. (2006). Discovery of hidden correlations in a local transaction database based on differences of correlations. *Engineering Applications of Artificial Intelligence*, 19, 419-428.

Tsai, C.-F., & Wu, J.-W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, 34, 2639-2649.

Talukder, A., & Casasent, D. (2001). A closed-form neural network for discriminatory feature extraction from high-dimensional data. *Neural Networks*, 14, 1201-1218.

Wang, J., & Xu, Z. (2010). New study on neural networks: the essential order of approximation. *Neural Networks*, 23, 618-624.

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, 27, 1131-1152.

Zhang, G. et al. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14, 35-62.

Zhou, L., & Yau, S. (2007). Efficient association rule mining among both frequent and infrequent items. *Computers & Mathematics with Applications*, 54, 737-749.