A Prediction Model for Recognition of Bad Credit Customers in Saman Bank Using Neural Networks

M. Yaghini 1, T. Zhiyan 2, and M. Fallahi 3

1 School of Railway Engineering, Iran University of Science & Technology, Tehran, Iran
2 School of Industrial Engineering, Iran University of Science & Technology, Tehran, Iran
3 School of Railway Engineering, Iran University of Science & Technology, Tehran, Iran

Abstract - The aim of this paper is to present a model based on feed forward neural networks to recognize bad credit customers in Saman Bank. To find an appropriate structure for the proposed neural network model, three different strategies called quick, dynamic and multiple strategies are investigated. The registered data of credit customer in Saman Bank from 2000 to 2008 year is used. To prevent models from over fitting with training data specifications, according to cross validation, we divide existing data set into three subsets called training, testing, and validation set, respectively. To evaluate the proposed model, we compare the result of three different strategies in neural networks with each other and with some common prediction methods such as decision tree and logistic regression. The results revealed that the three-layer neural network based on the back propagation learning algorithm with quick strategy has higher accuracy.

Keywords: Banking, Saman Bank, feed forward neural networks, prediction, bad credit customers

1 Introduction

Since banking industry survival necessitates riskiness, it cannot be prevented and it only can be managed. Risk management is a professional process which its main goal is improving decision quality in all levels of economic institutions including banks, in order to increase wealth of stakeholders. Risk in a financial institution means uncertainty about expected return for assets. One of the main functions of banks is credits granting to real and legal customers. Thus, banks have to minimize probability of any inappropriate decision before granting credit to decrease risk and attract low risk customers.

Credit risk is a result of default probability or probability of loan non-repayment by borrower; this risk is the same as expected loss. Credit risk evaluation is an important topic in financial risk management and has been the major focus of financial and banking industry. Data-mining methods, especially pattern classification, using real-world historical data, are of paramount importance in building such predictive models [1], [2].

Prediction models are classified into two groups. The first group includes models for classifying new credit customers based on their credit risk. The data used for modeling generally consist of financial information and demographic information about the loan applicant. In contrast, the second type of models deals with existing customers and along with other information, payment history information is also used here [3], [4], [5].

In this study, prediction model for new credit customers and prediction of their repayment situation based on neural networks is used. The different learning strategies are used in order to gain high accuracy in neural networks.

The paper is organized as follows. In section II, related literature is reviewed. The collected data from Saman bank customers’ are described in section III. In section IV, an overview of the used neural network and its weighting updating way is provided and then different strategies for finding appropriate structure for the proposed model are discussed. Section V modeling results are described, and conclusions as well as recommendations for future works are presented in section VI.

2 Literature Review

In search for credit risk prediction model with minimum limiting hypothesis, authors have suggested conditional probability models such as linear probability distribution, Logit model and Probit model [3]. Logit and Probit models are more difficult than discriminant analysis models in terms of calculation. The main problem when using these models is using a long and logical time of time series. They are under influence of econometric limitations such as shorter access period to time series of dishonored data. Then expert systems and artificial intelligence were introduced in this field. Neural networks, genetic algorithm and decision trees are among currently available methods in the field [3], [6].

The use of neural networks in business application has been increased recently. Studies indicate that neural networks are an accurate tool for credit risk assessment among others [7], [8]. Lim and Sohn [9] proposed a neural network-based behavioral scoring model, which dynamically accommodates the changes of borrowers’ characteristics after the loans are made. This work suggested that the proposed model could
replace the currently used static model to minimize the loss due to bad creditors. In 2007, an overview of rule extraction techniques for support vector machine was performed to credit risk assessment [10]. This work proposed also two rule extraction techniques taken from the artificial neural networks domain.

In [11] an application of neural networks to credit risk assessment related to Italian small businesses was described. This work presented two neural network systems, one with a standard feed-forward network, while the other with special purpose architecture; and suggested that both neural networks can be very successful in learning and estimating the default tendency of a borrower.

In [12] a hybrid SVM-based credit scoring models were proposed to evaluate an applicant’s credit score from the applicant’s input features. This work used the Australian and German data sets in its implementation. The work in Abdou et al. investigated the ability of neural networks, such as probabilistic neural nets and multi-layer feed-forward nets, and conventional techniques such as, discriminant analysis, probit analysis and logistic regression, in evaluating credit risk in Egyptian banks applying credit scoring models [13]. This work concluded that neural network models gave better average correct classification rates than the other techniques.

Tsai and Wu [14] compared performance of single classifier as the baseline classifier with multiple classifiers by using neural networks based on three data sets. This work presented the ensemble classifier outperforms single classifier in a set of data. Setiono et al. used recursive algorithm for extracting classification rules from feed-forward neural networks that have been trained on credit scoring data sets [15].

In [16], a multistage neural network ensemble learning model was proposed to evaluate credit risk at the measurement level. The proposed model consisted of six stages: (1) generating different training data subsets especially for data shortage, (2) creating different neural network models with different training subsets obtained from the previous stage, (3) training the generated neural network models with different training data sets and obtaining the classification score, (4) selecting the appropriate ensemble members, (5) selecting the reliability values of the selected neural network models, and (6) fusing the selected neural network ensemble members to obtain final classification result by means of reliability measurement.

Lin [17] used a three two-stage hybrid models of logistic regression-artificial neural network to construct a financial distress warning system suitable for Taiwan’s banking industry, and to provide an optimal model of credit risk for supervising authorities, analysts and practitioners in conducting risk assessment and decision making. Wang and Huang [18] used back propagation algorithm based on support vector machines for decision making about credit granting to applicants. In [19], a reassigning credit scoring model (RCSM) involving two-stages was proposed. The classification stage is constructing an ANN-based credit-scoring model, which classifies applicants with accepted (good) or rejected (bad) credits. The reassign stage is trying to reduce the Type I error by reassigning the rejected good credit applicants to the conditional accepted class by using the CBR-based classification technique.

Crook et al. compared different methods of credit customers’ classification. They found that neural networks had higher accuracy in prediction [20].

In this study, neural networks of two-, three- and four-layer perceptron under error back propagation learning algorithm are used in order to predict customers’ status. Three strategies, quick, dynamic and multiple, are used for finding appropriate structure for networks. Then accuracy of this network is assessed by decision tree and logistic regression methods. The best model as the proposed model is selected following comparison between results from different methods.

3 Data Understanding

Saman Eqtesad Credit Corporation was established on September 23, 1999 with a share capital of US$ 1.4mln. It opened its first branch on November 22, 1999 and managed to achieve already in its first year of activity a 5% return on equity. In 2002 Saman was the third private financial institution in post-revolutionary Iran to receive a banking license. In this context, the share capital increased to US$ 26mln.

Data set used in this study is related to real customers of Saman bank who have received loan during 2000-2008. This data set includes 20 features including Customer No., Birth date, Entrance date, Gender, Occupation, Education, Marital status, Loan No., Loan status, Repayment way, Beginning date, Final date, Interest, Wage, ISICCODE, ISICEDSC, Type of assurance, Time of contract D., Time of contract M., and Amount of loan.

It should be noted that assurance types received by bank for granting credit to customers includes common documents, declaration, movable properties in bank mortgage, participation bonds, property insurance in bank mortgage, disclaimer, check for premium, certified check, time deposit investment, certified promissory note, leasing bonds, mortgage property documents, property document of civil partnership, shares, bank guarantee, legal agreement, mandate, and credit insurance.

This data set includes 82,093 loan received by real customers. Since all features cannot be used in modeling and due to need for some change in some features, first in this step data are cleaned and new features are constructed so that data are prepared before being used in modeling.

In preliminary examination of features, it was found that since interest and wage features have similar nature and each has many missing values, these two features are integrated. Customer No. feature is omitted because of uniqueness. Results of change and integration of features can be seen in table 1.
### TABLE 1

<table>
<thead>
<tr>
<th>Row</th>
<th>Old Feature</th>
<th>New Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Birth date</td>
<td>Age</td>
</tr>
<tr>
<td>2</td>
<td>Entrance date</td>
<td>Background</td>
</tr>
<tr>
<td>3</td>
<td>Entrance date</td>
<td>Entrance year</td>
</tr>
<tr>
<td>4</td>
<td>Loan No.</td>
<td>City</td>
</tr>
<tr>
<td>5</td>
<td>Loan No.</td>
<td>Type of loan</td>
</tr>
<tr>
<td>6</td>
<td>Beginning date</td>
<td>Beginning year</td>
</tr>
<tr>
<td>7</td>
<td>Final date</td>
<td>Final year</td>
</tr>
<tr>
<td>8</td>
<td>Beginning &amp; Final date</td>
<td>Time if contract</td>
</tr>
</tbody>
</table>

Generally following changes were performed:
- Customer Age feature was extracted from customer Birth date.
- Background and Entrance year features were extracted from customer bank Entrance date feature.
- Regarding explanation mentioned for Loan No., City and Type of loan features were extracted from Loan No. feature.
- Loan contract Beginning year feature was extracted from Beginning date feature and Loan contract Final year feature was extracted from Final date feature.

In the next step, data were examined in terms of quality and validity, and various methods were used for cleaning dirty data. Manual value replacement methods were used for Age, Background and Time of contract features because of missing and noisy values. Occupation feature was omitted from data set due to high amount of missing values and lack of access to these values. Missing values of other features were replaced using C&RT decision tree algorithm.

Since the goal for this study is predicting customers’ loan status, Fig 1 shows way of value distribution for this variable. As it can be seen from diagram, most loans are in settled and active status, and limited numbers of loans are in risky status of doubtful debts, outstanding, and past due.

![Fig. 1. Histogram diagram of loans status](image)

Fig. 2 indicates relationship between loans status feature with repayment way. The figure imply that most settled loans are related to loans with all at once and installment repayment ways. Fig 3 indicates that most settled, active, outstanding, past due and doubtful debts are related to customers entering to bank in 2004.

![Fig. 2. Relationship between loan status feature and repayment way](image)

![Fig. 3. Histogram diagram of entrance year variable](image)

### 4 The neural network model

#### 4.1 Neural Network Topology

Artificial neural network is regarded as a main technique in supervised learning and unsupervised learning in data mining. Neural network used in this work is feedforward neural network, which is also called multilayer perceptron. For network training, error back propagation algorithm and momentum were used. Activation function used for this network, was Sigmoid function, which is as follows:

\[
\sigma(x) = \frac{1}{1 + e^{-x}}.
\]

All weights are randomly initialized in [-0.5, 0.5]. At the all weights were selected randomly in ,beginning of training
\[-0.5 \leq w_{ij} \leq 0.5\] distance in network. The weights are updated as follows:

\[
\Delta w_{ij}(t+1) = \eta \delta_{pj}o_{pi} + \alpha \Delta w_{ij}(t),
\]

(2)

where, parameter \(\eta\) is learning rate, \(\delta_{pj}\) is propagated error, \(o_{pi}\) is \(i\)th output neuron for sample \(p\), parameter \(\alpha\) is momentum, and \(\Delta w_{ij}(t)\) is change extent in previous iteration. \(\infty\) is assumed as constant in learning process, but \(\eta\) value changes during iterations of learning process. First, it is given as an initial value, and then it reduces by \(\eta_{low}\) logarithmically. When \(\eta\) is less than \(\eta_{low}\), then its value is set as \(\eta_{high}\). That is \(\eta(t-1) < \eta_{low}\), then \(\eta(t) = \eta_{low}\). Logarithmic deduction function for learning rate parameter is as follows:

\[
\eta(t) = \eta(t-1). \exp(\log(\eta_{low}/\eta_{high})/d),
\]

(3)

where, \(d\) is the value set by user and is called \(\eta\) decline. This process continues until the end of learning period. \(\delta_{pj}\) error value for outer layers is calculated as follows:

\[
\delta_{pj} = (t_{pj} - o_{pj})o_{pj}(1-o_{pj}),
\]

(4)

for other layers, it is as follows:

\[
\delta_{pj} = o_{pj}(1-o_{pj}) \sum_{k} o_{pj}w_{kj},
\]

(5)

where, \(t_{pj}\) is expected output for prediction. Network weights are updated after each output prediction [21].

### 4.2 Neural Network Structure

One challenging problem in neural networks is finding optimal structure for network based on available data. To overcome this problem, various strategies were used in this work, and finally the most accurate method was selected.

1) **Quick Strategy**

In fast strategy, only one neural network of three-layer perceptron type is trained separately. This network has a hidden layer with \(\max ((\Theta_{j} + \Theta_{o})/20, 3)\) neurons by default, where \(\Theta\) is neuron numbers in input layer and \(\Theta_{o}\) is the neuron numbers in output layer [21].

2) **Dynamic Strategy**

In dynamic strategy, network structure changes during learning process and neurons are added to network so that network efficiency increases and reaches to optimal accuracy. This stage includes two sub stages for finding appropriate structure, and then final network training. In order to find appropriate structure, first a network is constructed with two hidden layers each having two neurons. Initial learning rate is set as \(\alpha=0.05\) and \(\alpha=0.9\), and network is trained for one iteration. A copy of network is developed, one is called right side network and the other one is called left side network. Then a neuron is added to the second hidden layer of right network, and again both networks are trained for one iteration and total error is measured on both networks. If left side network has less error, it is maintained and one neuron is added to first hidden layer of right side network. If right side network has less error, left side network is replaced by a version of right side network, and again one neuron is added to second hidden layer of right side network. Both networks are trained for one iteration, and this process is repeated so that a termination condition is reached. For learning rate setting in each iteration, two vectors are calculated; first is motion vector \(M(t)\), based on weights change in one iteration, and the second one is change vector \(C(t)\), based on current iteration momentum. Vectors \(C(t)\) and \(M(t)\) are calculated as follows:

\[
M(t) = 2[W(t)-W(t-1)].
\]

(6)

\[
C(t) = 0.8C(t-1)-M(t).
\]

(7)

Where, \(W(t)\) is current iteration weight vector and \(W(t-1)\) is previous iteration weight vector. magnitude ratio vector for these two vectors is defined as follows:

\[
m(t) = \|M(t)\|/\|C(t)\|.
\]

(8)

this is learning acceleration indicator. If it is less than, learning is slowing down and learning rate is multiplied by 1.2, and if it is higher than 5, learning is accelerating and learning rate is multiplied by \(4/m(t)\). Once appropriate structure is found for neural network by dynamic strategy, final network should be trained based on error back propagation method. The network is trained with an initial learning rate of \(\alpha=0.02\) and \(\alpha=0.09\) [21].

3) **Multiple Strategy**

In this strategy, multiple networks are trained simultaneously in parallel until it is reached to termination constraint, and the network with highest accuracy is selected. It is accomplished in this way: first, multiple networks are constructed with hidden layer, with neurons in hidden layer varies between 3 to maximum number of neurons in input layer. Then multiple networks with two hidden layers are constructed per every network with one hidden layer, where the number of neurons in first hidden layer is exactly equal to number of single-layer network’s. but the number of neurons in second hidden layer varies and can be 2, 5, 10, 17 and so on, we can have maximum number of neurons in first hidden layer in the second hidden layer, and it is possible to train the networks by constructing them [21].
5 Predicting Customer Repayment Status

One of main problems in training prediction models is model overfitting with current training data characteristics. This causes wrong and less accurate prediction for new values. To prevent models from overfitting with training data specifications, according to cross validation, we divide existing data set into three subsets called training, testing and validation set, respectively, regarding high amount of data. Then training was stopped periodically and network was evaluated using validating data following every training period. Thus, we are able to determine beginning of overfitting.

In present database, customer loans status include 5 statuses: settled, active, past due, outstanding, and doubtful debts. The output layer of neural network consists of 5 neurons, and each of these statuses is assigned as a neuron in output layer. The proposed model can predict these statuses.

Input layer variables in the proposed neural network include: Age, Entrance year, Gender, Educational level, Marital status, Loan type, Repayment way, Interest, ISICCODE, Time of contract, Amount, Assurance type 1-19, and Assurances sum. Concerning the number of statuses for ISICCODE, Time of contract, Amount, Assurance type 1-19, Marital status, Loan type, Repayment way, Interest, and Assurance sum. Containing the number of statuses for each of these variables, input layer of neural network includes 61 neurons.

Results for comparing predictions by two-, three- and four-layer neural networks with three strategies using input and output neurons can be seen in tables 2 to 4. All model results were obtained by SPSS Clementine 12.0 software.

Decision tree and multinomial logistic regression techniques were used for evaluating obtained results. Chi-squared Automatic Interaction Detector (CHAID) algorithm is used for constructing decision tree. CHAID algorithm is an effective classification technique, which utilizes capability of statistic test as a measure for evaluating a possible predicting characteristic value. It determines similar values regarding target variable, and integrates them and keeps non-identical values. Then it selects the best predictor so that the first branch of tree is formed; each node is composed of a group of similar values for selected characteristic. This process continues until tree matures. Used statistic test depends on measuring level of key characteristic. If target characteristic is continuous, F test is used, if it is discrete, Chi-Square test is used [25]. Table 5 shows results for prediction by decision tree and logistic regression.

Comparison of above tables suggests that the highest prediction accuracy is for three-layer neural network with quick strategy. Quick and multiple strategies require more memory as well as longer time for algorithm execution. The least prediction accuracy is for four-layer neural network with quick strategy. As it can be seen from results, one cannot select a specific method definitely as the best method; rather there is need for analysis on accuracy, time and required memory for each method.

The number of correct predictions in each status (output neurons) by selected neural network is given in table 3. In addition, table 4 indicates results for status prediction using four-layer neural network with fast strategy and first learning schema having least accuracy among networks.

<table>
<thead>
<tr>
<th>Strategy for network structure</th>
<th>The number of layers</th>
<th>Prediction accuracy in validating data (%)</th>
<th>testing data (%)</th>
<th>training data (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quick</td>
<td>Two</td>
<td>85.77</td>
<td>86.15</td>
<td>85.64</td>
</tr>
<tr>
<td>Quick</td>
<td>Three</td>
<td>86.33</td>
<td>86.71</td>
<td>86.36</td>
</tr>
<tr>
<td>Quick</td>
<td>Four</td>
<td>84.70</td>
<td>85.78</td>
<td>85.55</td>
</tr>
<tr>
<td>Dynamic</td>
<td>Four</td>
<td>85.47</td>
<td>85.98</td>
<td>85.97</td>
</tr>
<tr>
<td>Multiple</td>
<td>Four</td>
<td>86.20</td>
<td>86.46</td>
<td>86.16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Status</th>
<th>Active</th>
<th>Doubtful debts</th>
<th>Outstanding</th>
<th>Past due</th>
<th>settled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>7302</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5291</td>
</tr>
<tr>
<td>Doubtful debts</td>
<td>0</td>
<td>3518</td>
<td>331</td>
<td>513</td>
<td>1</td>
</tr>
<tr>
<td>Outstanding</td>
<td>1</td>
<td>770</td>
<td>732</td>
<td>1227</td>
<td>1</td>
</tr>
<tr>
<td>Past due</td>
<td>1</td>
<td>495</td>
<td>351</td>
<td>1435</td>
<td>0</td>
</tr>
<tr>
<td>settled</td>
<td>2131</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>57975</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Status</th>
<th>Active</th>
<th>Doubtful debts</th>
<th>Outstanding</th>
<th>Past due</th>
<th>settled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>7903</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4690</td>
</tr>
<tr>
<td>Doubtful debts</td>
<td>0</td>
<td>3099</td>
<td>140</td>
<td>1124</td>
<td>1</td>
</tr>
<tr>
<td>Outstanding</td>
<td>0</td>
<td>428</td>
<td>191</td>
<td>2112</td>
<td>1</td>
</tr>
<tr>
<td>Past due</td>
<td>0</td>
<td>30</td>
<td>88</td>
<td>22182</td>
<td>0</td>
</tr>
<tr>
<td>settled</td>
<td>3265</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>56841</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Prediction accuracy in validating data (%)</th>
<th>testing data (%)</th>
<th>training data (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision tree</td>
<td>83.71</td>
<td>83.70</td>
<td>83.23</td>
</tr>
<tr>
<td>Decision tree</td>
<td>83.70</td>
<td>83.72</td>
<td>83.23</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>84.95</td>
<td>85.33</td>
<td>84.67</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>84.90</td>
<td>85.25</td>
<td>84.62</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>85.02</td>
<td>85.42</td>
<td>84.64</td>
</tr>
</tbody>
</table>
Comparison of results for decision tree and logistic regression with neural networks shows that neural networks have higher accuracy in predicting customers’ repayment status. It should be mentioned that there is not much difference in accuracy in methods used in this work, but they differ in processing time and model construction considerably. Finally, three-layer neural network model with quick strategy was selected as the model for this work.

6 Conclusion

In this work, a model based on neural networks with high accuracy is suggested for predicting credit customers’ repayment status in Saman bank. Three networks with different layer and neuron numbers as well as three strategies were used so as to determine model optimal structure. Credit customers’ data during 2000-2008 was used for prediction model construction. For evaluating the proposed model, obtained results were compared to results for decision tree and logistic regression. Results indicate that three-layer neural network with quick strategy has the highest accuracy. Using the proposed model it makes possible to have an accurate prediction of customer repayment status, and acting for appropriate planning for reducing bank credit risk. Thus, bank can minimize probability of any inappropriate decision before granting credit to decrease risk and attract low risk customers.

There are various fields for future study. In order to improve prediction accuracy, creative methods such as genetic algorithm or simulated annealing for finding optimal neural network structure can be used. Creative algorithms can be used for increasing network’s learning speed. In addition, integrated models can be applied for accurate prediction of customer repayment status.

7 References