Credit Risk Assessment Based on Flexible Neural Tree Model

Yishen Zhang[1,2], Dong Wang[1,2(✉)], Yuehui Chen[1,2(✉)],
Yaou Zhao[1,2], Peng Shao[3], and Qingfang Meng[1,2]

[1] School of Information Science and Engineering,
University of Jinan, Jinan, People's Republic of China
{ise_wangd, yhchen}@ujn.edu.cn
[2] Shandong Provincial Key Laboratory of Network Based Intelligent
Computing, Jinan 250022, People's Republic of China
[3] School of Mathematics, Dalian University of Technology,
Dalian, People's Republic of China

Abstract. In recent years, as China's credit market continues to expand, a large number of P2P (person-to-person borrow or lend money in Internet Finance) platforms were born and developed. Most of the P2P platforms in China use data mining methods to evaluate the credit risk of loan applicants. Artificial neural network (ANN) is an emerging data mining tool and has good classification ability in many application fields. This paper presents a model of credit risk assessment based on flexible neural tree (FNT), which can reduce the overdue rate and save the analysis time. Overdue and non-overdue sample data are provided by the Jinan Hengxin Micro-Investment Advisory Co., Ltd., and used to build the model. Experiments show that the proposed model is more accurate and has less time cost for the overdue classification of credit risk assessment.

Keywords: Artificial neural network · Credit risk assessment · Flexible neural tree

1 Introduction

Credit loan is an unsecured loan model. In recent years, the credit market has been expanding rapidly in China. On one hand, the rapid development of China's economy has shortened the cycle of capital turnover. On the other hand, because of the improvement of Chinese national consumption capacity, businesses increasingly need high demand for funds, so a large number of P2P Internet inclusive financial platforms came into being. As no complete credit evaluation system like banks in China, P2P platform has small contain ability to non-collateral customers, it obtains better risk prediction results only through the establishment of the corresponding credit risk assessment model. So a large number of platforms are exploring their own methods of credit risk assessment, most of which use data mining approach to try to collect and understand the customer information to better grasp the authenticity and validity of customer information; to evaluation financial situation of customers more reasonable; to predict the business conditions, repayment intention and ability of borrows more accurately.

F. Cong et al. (Eds.): ISNN 2017, Part I, LNCS 10261, pp. 215–222, 2017.
DOI: 10.1007/978-3-319-59072-1_26

The establishment of a good credit risk evaluation model is the biggest challenge to the development of P2P platform and credit market. If the model control is too strict to the customer, the platform will lose some high-quality customers and make it passive in the industry competition. On the contrary, the overdue rate of the platform will continued to rise, which makes financial managers difficult to be responsible and lose credibility. Therefore, it is important to establish the credit risk evaluation model to prevent bad debts happening, to promote the speed of capital flow and to maintain the security and stability of capital. In the field of credit risk assessment, artificial neural networks, genetic programming, genetic algorithms, support vector machines, logistic regression and some hybrid models have achieved gratifying results in terms of performance and precision.

In the past few years, many excellent algorithms and research methods have been tested on the basis of customer information data in the field of credit risk assessment. Khashman used artificial neural network algorithm in Germany customer dataset and achieved the accuracy rate of 83.6% [1]. Bekhet and Eletter applied RBF network algorithm to the Jordanian commercial bank data set, and the test sets had accuracy rate of 86.5% [2]. Wang et al. uses the improved BP neural network algorithm and the accuracy rate is 86% [3]. The traditional Artificial Neural Network has the stationary structure, but Flexible neural tree (FNT) has the special structures which called flexible tree structures, with this characteristic, FNT model can get better property from the learning.

In this paper, a new method based on FNT model was proposed for classification of customer information, and the results in 10-fold cross validation shows our method achieved better performance than the other state-of-arts.

2 Data Collection and Variable Definition

Customer information data can be described from many dimensions. In this paper, we randomly took 300 samples of overdue customers and 300 Negative samples of non-overdue customers all of which were from 2,000 customers of Jinan Hengxin Micro-Investment Advisory Co., Ltd. between 2014 and 2016. In this study, the author chooses 13 dimensions to describe and consider the customer information. The standard of selected dimensions are: (1) do not contain the customer's identity information; (2) exclude the subjective information from the point of view of the actual human audit, such as the use of loans, business models, profits and other objective information which can only be verified by a third party as difficulties to verify and census them.

According to these principles, the selected dimensions can maximize the provided data by customer which objectively and difficulty to forge. The accurate classification based on actual data which can verify and excluding the subjective description. Table 1 shows the variable, values, and definitions of 13 selected dimensions of the study, and the Table 2 shows the examples of datasets.

The 600 samples are based on the statistics in Table 1, and then all the data will processed as "Max_Min standardization" for the next step, and get ready to input to the FNT model, the normalized samples are shown in Table 3. The normalization rule is shown in Eq. (1).

Table 1. Proposed variables for building dataset

Variable	Value	Variable definition
Gender, G	0, 1	0: female 1: male
Degree of education, D	1 to 4	1: graduated from junior high school 2: graduated from high school 3: graduated from junior college 4: get bachelor degree or above
Age, A	Actual value	The number of years the customer has experienced since birth
Marital status, M	0 to 2	0: unmarried 1: married 2: divorce
Account properties, AP	0, 1	0: local registered permanent residence 1: foreign registered permanent residence
The number of years experienced by the company, YC	Actual value	The number of years in which the customer's work unit has been established
Industry categories, IC	0 to 5	Industry category of customer
Job level within the company, JC	0 to 4	0: general staff 1: junior management staff 2: middle management staff 3: senior management staff 4: the founder
Total income, TI	Actual value	Customer's itemized of the savings card which printed by bank, and the difference between total amount and total expenditure in recently six months
Total debt, TD	Actual value	The sum of outstanding loan balance and average usage limit in recently six months shown in the summary of liability information within customer credit report
Housing ownership situation, HS	0 to 2	0: no real estate 1: full purchase 2: mortgage
Vehicle ownership situation, VS	0 to 2	0: no car production 1: full purchase 2: mortgage
Overdue numbers shown in credit reporting, OR	Actual value	The sum of the number of overdue times in the credit transaction details within customer credit report

$$P_{ij}^{'} = \frac{P_{ij} - m_i}{M_i - m_i} \tag{1}$$

where, $P_{ij}^{'}$ is the normalized customer data. P_{ij} is the original customer data. M_i is the maximum value of the dimension i. m_i is the minimum value of the dimension i.

Table 2. Examples

No.	G	D	A	M	AP	YC	IC	JC	TI	HS	VS	OR	TD
1	1	3	48	1	2	4	0	1	150	2	1	3	137.8
2	1	2	36	1	2	10	1	1	30	0	1	0	41.2
3	2	3	49	1	1	2	1	1	48	1	1	2	91.5
4	1	3	54	1	2	11	3	1	102	2	0	4	16.5
5	1	3	36	1	3	7	1	1	20	0	0	3	69.8

Table 3. Normalized samples

No.	G	D	A	M	AP	YC	IC	JC	TI	HS	VS	OR	TD
0	0.5	0.44	0.33	0.5	0.23	0.17	0	0.05	1	1	0.1	0.13	0
0	0.5	0.35	0.33	0.5	0.14	0.17	0	0.09	1	0	0.1	0.09	0
0	0.5	0.23	1	0	0.05	0	0	0.04	1	1	0.4	0.27	1
0	0.5	0.64	0.33	0	0.08	0	0	0.05	0	1	0.1	0.04	1
0	0.5	0.76	0.33	0	0.20	0.17	0	0.15	0.5	0	0	0.19	1

3 Classification Method

3.1 Flexible Neural Tree

Flexible neural tree (FNT) is a special artificial neural network with flexible tree structures. It is proposed by Chen et al. [4, 5] and relatively easy for this model to reach near-optimal structure by using optimization algorithms. The FNT model consists of tree-structural encoding method and specific instruction set, it is also generated by using function set F and terminal instruction set T, described as follows.

$$\mathrm{S} = \mathrm{F} \ \cup \mathrm{T} = \{+_2, +_3 \cdots +_N\} \ \cup \{x_1 \cdots x_n\} \tag{2}$$

where $+_i (i = 1, 2 \cdots N)$ denotes non-leaf nodes with i arguments, the $x_1, x_2 \cdots x_n$ are leaf nodes with none arguments.

Figure 1 shows the output of a non-leaf node which calculated by FNT model. Instruction $+_i$ is also called a flexible neuron operator with i inputs. The output of a flexible neuron +n is calculated as follows and the total excitation of $+_n$ is given by

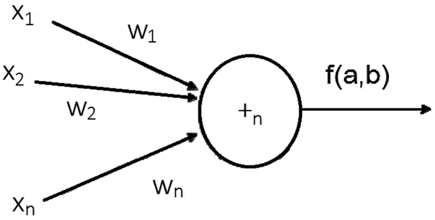

Fig. 1. Non-leaf node of flexible neural tree with a terminal instruction set $\mathrm{T} = \{x_1, x_2, \cdots, x_n\}$

$$net_n = \sum_{j=1}^{n} w_j x_j \tag{3}$$

In Eq. (3), $x_j (j = 1, 2, \cdots, n)$ are the input elements to node $+_n$. The output of the node $+_n$ is then calculated by

$$out_n = f(a_n, b_n, net_n) = e^{-\left(\frac{net_n - a_n}{b_n}\right)^2} \tag{4}$$

A typical FNT model is illustrated in Fig. 2. Its overall output can be computed from left to right by a depth-first method recursively.

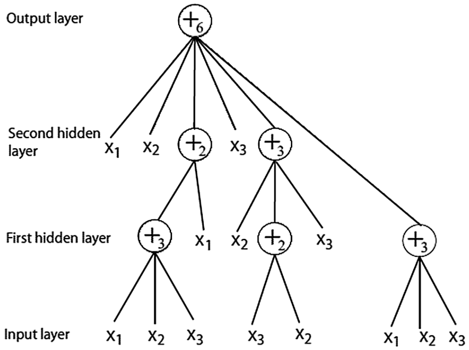

Fig. 2. Typical representation of FNT with function instruction set $\{+_2, +_3, +_4, +_5, +_6\}$ and terminal set $\{x_1, x_2, x_3\}$, which has four layers.

General learning algorithm of FNT

- Step 1. Initialize the values of parameters used in the particle swarm optimization (PSO) algorithms. Set the elitist program as NULL and set the fitness value as the biggest positive real number. Create the initial population.
- Step 2. Construct optimization using PSO algorithm, in which the fitness function is calculated by root mean square error (RMSE).
- Step 3. If the better structure has found, then go to step 4, otherwise go to step 2.
- Step 4. Optimize parameters using PSO algorithm.
- Step 5. If the maximum number of local search is reached, or no better parameter vector is found for a significantly long time (100 steps), then go to step 6; otherwise go to step 4.
- Step 6. If the satisfied solution is found, then stop; otherwise go to step 2.

3.2 Prediction Assessment

In statistical analysis, two methods can be used to check the effectiveness of the classifier in applications, namely, independent dataset tests and 10-fold cross validation tests. For 10-fold cross validation, the full training set will be separated equally into 10 subset. Each subset will regarded as test data set to compute the overall accuracy (OA) of the model trained by the rest of full training data set. In addition, Sensitivity (Sens) and Specificity (Spec) are also used to evaluate the performance of classifier.

4 Discussion and Results

In this study, the FNT model was used to perform a 10-fold cross validation of a data set containing 600 sample data, i.e. 540 training samples and 60 testing samples were used for each experiment and were performed on each data set. The results show that the average accuracy of the test set is 88.32% (Table 4). In the Table 4, "T" is abbreviation of "trail", "D" is abbreviation of "data", "OA" is abbreviation of "Overall", "A-acc" is abbreviation of "Average accuracy rate" and "acc" is abbreviation of "accuracy rate", the values of "A-acc" and "acc" are percentages.

Table 4. The part of results of FNT model in 10-fold cross validation

		T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	A-acc
D0	miss	11	7	5	9	3	10	10	7	4	5	88.17
	acc	83	88	92	85	95	83	83	88	93	91	
D1	miss	9	10	8	8	7	5	5	6	9	8	87.50
	acc	85	83	87	87	88	91	91	90	85	86	
D2	miss	8	7	10	6	9	4	9	8	5	4	88.33
	acc	87	88	83	90	85	93	85	86	91	93	
D3	miss	7	9	6	9	7	9	7	8	6	4	88.00
	acc	88	85	90	85	88	85	88	86	90	93	
D4	miss	9	8	3	3	8	6	7	4	9	5	89.67
	acc	85	87	95	95	87	90	88	93	85	91	
D5	miss	10	10	8	5	6	6	5	7	9	5	88.17
	acc	83	83	87	92	90	90	91	88	85	91	
D6	miss	9	5	9	7	10	8	9	10	5	7	86.83
	acc	85	92	85	88	83	86	85	83	91	88	
D7	miss	8	9	11	5	3	7	7	4	9	6	88.50
	acc	87	85	82	92	95	88	88	93	85	90	
D8	miss	7	10	9	8	6	9	7	5	5	5	88.17
	acc	88	83	85	87	90	85	88	91	91	91	
D9	miss	9	4	7	6	4	3	8	6	5	9	89.83
	acc	85	93	88	90	93	95	86	90	91	85	
OA												88.32

We compared the average accuracy, sensitivity and specificity between our model and other methods. The results are shown in Table 5, we can see that our method has higher accuracy compared to other method, and the specificity is slightly better than the others. Another point to make is this: the sensitivity value of Improved BP Neutral

Network method is 91.6%, and this value was calculated by once experiment result form with 14 positive simples and 6 negative samples, totally 20 simples. The proportion of positive samples is much higher, so the sensitivity value also high, besides the sensitivity index is mentioned there only and no mention of any other place, so this value is included in Table 5 for reference.

Table 5. The comparison of our method and other methods

Algorithm	Accuracy (%)	Sens (%)	Spec (%)
Improved BP neutral network	86	91.6	62.5
Radial basis function scoring model	86.5	84.2	87.9
Artificial neural networks	83.6	Null	Null
This method (average)	88.32	85.67	92.79

5 Conclusion

In this study, we proposed a redesigned and redefined customer information feature dimension and FNT model for the field of credit risk assessment. Compared with other methods, the method proposed in this study has different degrees of improvement in various evaluation indexes, while the validity of the FNT model is proved. In the future, we will continue to improve the algorithm method and search for more effective classifiers in order to obtain better classification accuracy in this field.

Acknowledgements. This research was supported by the National Natural Science Foundation of China (Grant No. 61302128, 61573166, 61572230, 61671220, 61640218), the Youth Science and Technology Star Program of Jinan City (201406003), the Shandong Distinguished Middle-aged and Young Scientist Encourage and Reward Foundation, China (Grant No. ZR2016FB14), the Project of Shandong Province Higher Educational Science and Technology Program, China (Grant No. J16LN07), the Shandong Province Key Research and Development Program, China (Grant No. 2016GGX101022).

References

1. Khashman, A.: Neural network for credit risk evaluation: investigation of different neural models and learning schemes. Exp. Syst. Appl. **37**(9), 6233–6239 (2010)
2. Bekhet, H., Eletter, S.: Credit risk assessment model for Jordanian commercial banks: neural scoring approach. Rev. Dev. Financ. **4**(1), 20–28 (2014)
3. Wang, L., Chen, Y., Zhao, Y., Meng, Q., Zhang, Y.: Credit management based on improved BP neural network. IHMSC **1**, 497–500 (2016)
4. Chen, Y., Yang, B., Dong, J., Abraham, A.: Time-series forecasting using flexible neural tree model. Inf. Sci. **174**, 219–235 (2005)
5. Yang, B., Chen, Y., Jiang, M.: Reverse engineering of gene regulatory networks using flexible neural tree models. Neurocomputing **99**, 458–466 (2013)
6. Abdou, H., Pointon, J., El-Masry, A.: On the applicability of credit scoring models in Egyptian banks. Banks Bank Syst. **2**(1), 4–19 (2007)

7. Bensic, M., Sarlija, N., Zekic-Susac, M.: Modeling small-business credit scoring by using logistic regression, neural networks and decision trees. Intell. Syst. Account. Financ. Manag. **13**(3), 133–150 (2005)
8. Blanco, A., Mejias, R., Lara, J., Rayo, S.: Credit scoring models for the microfinance industry using neural networks: evidence from Peru. Exp. Syst. Appl. **40**(1), 356–364 (2013)
9. Heiat, A.: Comparing performance of data mining models for computer credit scoring. J. Int. Financ. Econ. **12**(1), 78–83 (2012)
10. Koh, H., Tan, W., Goh, C.: A two-step method to construct credit scoring models with data mining techniques. Int. J. Bus. Inf. **1**(1), 96–118 (2006)
11. Jagric, V., Kracun, D., Jagric, T.: Does non-linearity matter in retail credit risk modeling? Financ. uver-Czech J. Econ. Financ. **61**(4), 384–402 (2011)
12. Wu, C., Guo, Y., Zhang, X., Xia, H.: Study of personal credit risk assessment based on support vector machine ensemble. Int. J. Innov. **6**(5), 2353–2360 (2010)
13. Xie, T., Yu, H., Wilamowski, B.: Comparison between traditional neural networks and radial basis function networks. In: Proceedings of 2011 IEEE International Symposium on Industrial Electronics, pp. 1194–1199 (2011)
14. Yap, P., Ong, S., Husain, N.: Using data mining to improve assessment of credit worthiness via credit scoring models. Exp. Syst. Appl. **38**(10), 1374–1383 (2011)
15. Memarian, H., Balasundram, S.: Comparison between multi-layer perceptron and radial basis function networks for sediment load estimation in a tropical watershed. J. Water Resour. Prot. **4**, 870–876 (2012)