Comparison of Decision Trees and Deep Learning in Personal Credit Classification

Yunteng Li
22120634@bjtu.edu.cn

Economics and Management School, Beijing Jiaotong University, Beijing China

Abstract—The rapid development of the financial industry makes the financial institutions facing severe credit risk problems. The accurate classification of users' credit rating can help banks avoid potential risks and reduce losses. Based on this, this paper makes a comparative study on the application performance of two typical algorithms in the field of machine learning - decision tree and deep learning in bank customer classification. It is found that the decision tree algorithm is suitable for data sets with small amount of data and discrete characteristics while deep learning algorithm is more suitable for the situation of huge amount of data, and has better performance in dealing with the continuity problem.

Keywords-Personal credit; Categorical prediction; Decision tree; Deep learning

1. Introduction

Traditionally, credit refers to the provision of credit by financial institutions to support consumers to buy consumer goods or make consumption activities, so as to solve the contradiction between the existing capacity and consumption demand [1]. The scale of consumer credit in China has been on the rise since 2012, and the service scope of credit consumption has been continuously extended, gradually expanding to education, tourism and other fields. As far as consumer credit is concerned, the data released by the People's Bank of China's RMB Credit Balance Sheet of Financial Institutions shows that the balance in 2017 was about RMB 31.52 trillion, an increase of 25.83% compared with 2016. It is worth noting that, with the rapid development and popularization of credit consumption in China, there are still many problems such as big data security risks [2] and data abuse [3]. At present, banks and other platforms have a long audit process on credit consumption behaviors such as user loans, which makes it difficult to control the overall user fraud risk. Therefore, the research on users' credit risk can not only help banks to quickly improve the efficiency of credit audit, but also reduce the default risk level of users in bank lending and other businesses, which has positive practical significance for the stable development of banks.

The essence of bank customer credit risk assessment is a classification problem, that is, the process of dividing existing bank users into creditworthy users and non-creditworthy users. From the perspective of its development process, it can be roughly divided into three stages. The simple analysis stage, the probability analysis stage and the artificial intelligence stage are the third stage, namely the artificial intelligence stage, which mainly refers to the application of artificial intelligence in credit evaluation. From the beginning of this century until now [4], for the modern trading environment, the credit system is a kind of behavior based on customers' stable repayment ability

BDEIM 2022, December 02-03, Zhengzhou, People's Republic of China

DOI 10.4108/eai.2-12-2022.2328732

and can enjoy corresponding services without immediate payment. Therefore, risk prediction is one of the important references for banks to issue loans.

2. Literature Review

Consumer credit scoring model was first proposed by Durand.D, who took the lead in proposing that the data collected by banks should be quantified by feature vectors, including the basic information of users, credit records and other common information attributes [5]. Alis O.F introduced machine learning algorithm into the field of bank user classification. He used decision tree, deep learning and other methods to evaluate bank user classification [6]. Since the birth of machine learning, the field of financial institutions has been the main battlefield of its application and research. Wang Hongge [7] used logistic regression for credit card fraud detection, with an accuracy of more than 95%. Florentin Butaru et al. [8] concluded by comparing three methods: logistic regression, decision trees, and random forests, that different banks are suitable for different models. Compared with the traditional machine learning algorithm, which needs preprocessing and feature extraction before training, deep neural network can use the cascade of multi-layer nonlinear processing units to automatically extract and transform features. These layers can make neural network discover the nonlinear relationship between data and improve its ability of learning from raw data [9].

3. Methods and Materials

3.1 Introduction of CART decision tree

Decision tree algorithm is a typical machine learning algorithm. It continuously divides nodes to approximate the discrete function values of demand, uses induction algorithm to generate interpretable cotyledon generation method and tree data structure, and uses the decision tree model generated by training to supervise and analyze data sets with similar structures.

CART tree uses Gini value is used as the basis of node splitting, and the category with the highest probability of leaf node is used as the prediction category of this node. When the dependent variable of the data set is continuous and used as regression tree, the minimum variance of the sample is used as the basis of node classification, and the mean value of leaf nodes is used as the predicted value of the node. The calculation method of Gini value is as:

$$\mathrm{Gini}(D)=\sum_{k=1}^{|Y|}\sum_{k!\neq k}p_k p_k^{'}=1-\sum_{k=1}^{|Y|}p_k^2 \tag{1}$$

As can be seen from the above formula, Gini coefficient of data set reflects that two samples are randomly selected from data set D, and their analogy is not the same probability. Therefore, the smaller the Gini coefficient, the smaller the entropy of the data and the higher the purity of the data set.

3.2 Introduction of deep neural networks

We also need to use deep neural network methods in this article. Assuming that there are m samples with n features, the input layer$X \in R^{m\times n}$, this DNN structure contains only one hidden

layer, and the hidden layer contains h neurons, then the weight and deviation of the hidden layer can be expressed as$W_h \in R^{n\times h}$and$b_h \in R^{1\times h}$. If there are q output label values, the weight and deviation parameters of the output layer are $W_0 \in R^{h\times q}$and$b_h \in R^{1\times q}$respectively. Then the output of the hidden layer and the output of the output layer can be calculated by the following formulas (2) and (3):

$$H = XW_{\mathrm{h}} + b_h \tag{2}$$

$$\mathrm{O} = \mathrm{H}W_O + b_o \tag{3}$$

In order to achieve the classification effect, the activation function is introduced to transform each layer (except the input layer) nonlinearly. The activation functions often used in deep learning are ReLU, sigmoid, tanh, softplus, etc. The hidden layer in this paper is softplus function. See below for common activation functions, their formulas and characteristics [10]:

Table 1. Introduction of Common Activation Functions

Function name	***Mathematical formula***
ReLU	$f(x) = \max(0, x)$
Sigmoid	$f(x) = \frac{1}{1+e^{-x}}$
Tanh	$f(x) = \frac{1-e^{-2x}}{1+e^{-2x}}$
Softplus	$f(x) = \log(1+e^{x})$

3.3 Data

The data set used in this paper is obtained from the data set website Kaggle, which mainly describes the customer credit records of a bank in Taiwan Province in 2005.

The data set used in this paper has already replaced the classification attributes with numerical values, so there is no need to convert the classification information into digital information that can be recognized by the machine. In this paper, OneHotEncoder, StandScaler, fit_transform and other methods are mainly used to process the feature matrix, so that it can avoid data analysis problems such as virtual variable traps. Finally, the data is standardized,so that it can better meet the requirements of machine learning algorithms.

4. Results and Discussion

4.1 CART decision tree analysis results

The effectiveness of decision tree is evaluated mainly through the calculation of micro-average, macro-average, weighted average and sample average by precision, recall, f1_score and other methods of prediction results. The experimental results under the f1-score evaluation standard of accuracy and regression rate and the resulting decision tree are as follows:

Table 2. Experimental Results of Decision Tree when Training Set Accounts for 80%

	precision	***recall***	***f1-score***	***support***
microavg	0.82	0.81	0.81	24000
macroavg	0.74	0.63	0.66	24000
weightedavg	0.80	0.81	0.79	24000
samplesavg	0.81	0.81	0.81	24000

Table 3. Experimental Results of Decision Tree when Training Set Accounts for 50%

	precision	***recall***	***f1-score***	***support***
microavg	0.82	0.81	0.81	24000
macroavg	0.74	0.63	0.66	24000
weightedavg	0.80	0.81	0.79	24000
samplesavg	0.81	0.81	0.81	24000

Table 4. Experimental Results of Decision Tree when Training Set Accounts for 20%

	precision	***recall***	***f1-score***	***support***
microavg	0.82	0.81	0.81	24000
macroavg	0.74	0.63	0.66	24000
weightedavg	0.80	0.81	0.79	24000
samplesavg	0.81	0.81	0.81	24000

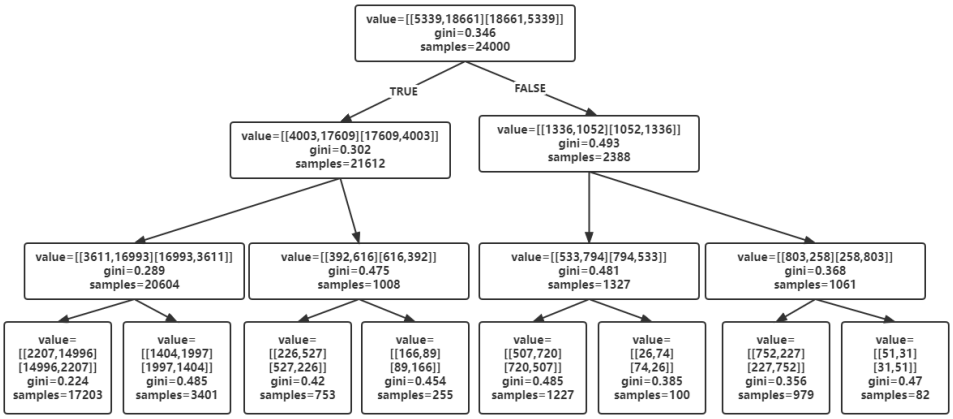

Figure 1 Resulting decision tree.

when the proportion of training sets in the decision tree CART algorithm is gradually reduced, precision, recall and f1-score in the evaluation methods of data prediction are slightly increased, while the macro average is slightly decreased, but the results are basically the same, indicating that the decision tree Cart algorithm is not suitable for a large range of data sets, and can play a better role in smaller data sets with obvious feature classification. It can be seen that the CART algorithm is more suitable for projects with small amount of data in bank data and higher requirements for forecasting process. If it can be matched with certain manpower assistance, it can not only reduce workload but also achieve higher work efficiency.

4.2 Deep neural network analysis results

The effectiveness of the deep learning multilayer vector machine model is mainly evaluated by ACC function, Loss function and other methods, in which the Loss function is the loss function, which is used to estimate the gap between the output and the real value of the model and guide the optimization of the model. According to the data mentioned above, the deep neural network analysis was carried out, and the Loss value and accuracy of the analysis results were evaluated

$$\hat{\theta} = arg\min_{\theta} \frac{1}{N} \sum_{1}^{N} L[y_i, f(x_i; \theta) + \lambda \phi(\theta)] \tag{4}$$

Table 5. DNN Experimental Results when Training Set Accounts for 80%

Output neuron	***epoch***	***time***	***Total loss***	***accuracy***
neuron1	100	1.330s	0.44527	0.8102
neuron2	100	1.332s	0.44253	0.8175
neuron3	100	1.328s	0.44064	0.8266

Table 6. DNN Experimental Results when Training Set Accounts for 50%

Output neuron	***epoch***	***time***	***Total loss***	***accuracy***
neuron1	100	0.677s	0.42714	0.8348
neuron2	100	0.680s	0.44000	0.8263
neuron3	100	0.679s	0.41259	0.8154

Table 7. DNN Experimental Results when Training Set Accounts for 20%

Output neuron	***epoch***	***time***	***Total loss***	***accuracy***
neuron1	100	0.241s	0.47726	0.8035
neuron2	100	0.244s	0.48089	0.7982
neuron3	100	0.245s	0.47251	0.8010

In addition, through SCALARS tool in tensorboard, the changes of accuracy of neural network in different training processes are shown and drawn into line graphs.

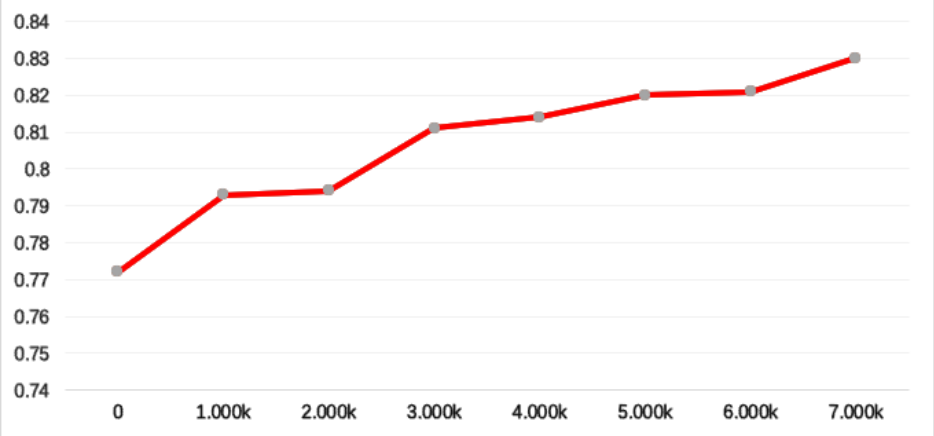

Figure 2. Changes of ACC during training

To sum up, it can be seen from the above figure that the accuracy of single-layer training approaches 82% and the loss value approaches 0.4 when using the training set. It can be clearly seen that the accuracy rises and the loss value decreases in the training process of deep learning network. In the process of training, the accuracy increased by about 10%, while the loss decreased by about 0.1. With the increase of the training set, for the deep neural network algorithm, the operation efficiency, accuracy and loss value will be significantly improved.

5. Conclusions

This paper mainly studies the classification of bank customers based on decision tree CART algorithm and deep neural network algorithm, and builds a bank credit evaluation model based on CART decision tree and deep neural network. Different from previous similar studies, this paper demonstrates the necessity of machine learning algorithm interpretability in the practical application of financial institutions, and avoids the problem that the evaluation index of the algorithm is too single. By adjusting the training ratio of data sets, pruning the decision tree and adjusting neurons, different application scenarios of the models are constructed.

Generally speaking, the decision tree is more effective in the case of small amount of data and discrete features, while the deep learning algorithm is more suitable for data sets with huge amount of data, and it can also better deal with continuous problems. According to the data characteristics of financial institutions, the application of deep learning algorithm is more extensive.

Due to the limitations of objective conditions such as preparation time and personal ability, there are still some shortcomings in this paper: the data set used in the experiment is not large enough to make a clear contrast between the decision tree algorithm and the deep learning algorithm, and the interpretable display and interpretation of machine learning are not perfect enough.

References

[1] OECD (2011). New Sources of Growth: Intangible Assets, Preliminary evidence and policy issues [EB/OL]. www.oecd.org/innovation/strategy.

[2] Report of the Appellate Body, Argentina-Financial Services.

[3] Zhang M..(2018) Prosperity or Conflict: Outlook of World Trade Situation in 2018. Tsinghua Financial Review, 02.

[4] MALEKIPIRBAZARI M,Aksakalli V. (2015) Risk Assessment Insocial Lending via Random Forests. Expert System with Applications,42(10):4621-4631.

[5] Durand D .(1941) Risk elements in consumer instalment financing. National Bureau of Economic Research.

[6] Wu Qiong. Research and application of decision tree algorithm in bank personal credit user classification [D]. Nanjing University, 2012

[7] WANG Hongge. Research on customer credit card default prediction based on AHP-Logistic model [D].Liaoning University,2019.

[8] Butaru F,Chen Q Q,Clark B, et al. (2016) Risk and Risk Mangement in the Credit Card Industry. Journal of Banking &Fiance,72:218-329.

[9] LECUNY, BENGIOY, HINTONG.(2015)Deeplearning.Nature,521(7553):436-444.

[10] FU Yufei, WANG Mingyan. (2022) Review of the application of deep learning in the financial field. Software Engineering, 25(03):1-4.