Modern Management based on Big Data II and Machine Learning and Intelligent Systems III
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210231

Tax Risk Prediction of Real Estate Based on Convolutional Neural Network

Meilin YIN[1] and Ning LUO
Guangdong University of Foreign Studies, China

Abstract. Risk management is an important link in tax administration. From China's taxation practice, risk identification has become the weakness of tax management. With the complexity of massive data and the secrecy of modern transactions, traditional tax risk identification can no longer adapt to the development of the times. In the past, most risk researches focused on the basic machine learning stage. There are gaps in the application of deep learning in tax risk management. Based on the tax risk management indicators, this paper took the real estate industry as an example. We used convolutional neural network (CNN) to construct a tax risk prediction model. The experiment shows that a tax risk prediction model based on CNN has higher accuracy in tax risk identification and has a stronger ability to process tax data. The model has a certain reference value for tax authorities to reduce tax risk and tax loss.

Keywords. Tax risk prediction, convolutional neural network

1. Introduction

Taxation plays an important role in organizing fiscal revenue and regulating economic operations. China has implemented tax system reforms many times. The reforms have improved the structure of tax system. But tax evasion has long existed in China. With the rapid development of big data, digital economy is currently important component of China's national economy [1]. The secrecy of those transactions affects the tax administration and risk identification. Tax risk identification is facing greater challenges, due to the diversified modern transactions, the complex financial accounting methods and the huge and complicated data. In the past, most of studies focused on the basic machine learning stage. They generally used the algorithms of random forest and BP neural network. There is a gap in the application of deep learning in tax risk management [2]. This paper takes the real estate industry as the research object. We build a tax risk prediction model of real estate industry based on convolutional neural network. The model provides scientific basis and technical support for the identification and prediction of taxation risks in China. It has certain reference value for tax risk management.

This paper presented below consists of four sections. The second section analyzes the status quo and causes of tax risk management in China and puts forward the tax risk management indicators adopted in this paper. The third section will briefly discuss the proposed methods. The fourth section will experiment and analyze the tax risk

[1] Corresponding Author: Meilin Yin, Network and Information Center, Experimental Teaching Center, Guangdong University of Foreign Studies, Guangzhou, China; E-mail: 383635698@qq.com.

prediction model adopted in this research. The fifth section elaborates the significance, the insufficiency and the direction of the follow-up research.

2. An overview of tax risk management

2.1. The current status and causes of tax risk management in China

The traditional forms of tax collection emphasize manual work. Calculation, collection and risk identification depend on the experience of tax personnel [3]. With the complexity of modern economy, the drawbacks of relying on manual greatly affect the efficiency of tax administration. The authorities have realized the importance of big data in administration. After more than 30 years, the construction of tax informatization has achieved considerable results. However, there has been a great contradiction between the advanced nature of big data and the backwardness of some grass-roots tax officials. They have not fully utilized the big data [4].

We take the real estate industry as an example. The real estate industry has played an important role in economic development. It has a high degree of relevance and a complex industrial chain [5]. Meanwhile, it involves many types of taxes. The calculation of funds is also more complicated. Therefore, the tax risk investigation of the real estate industry has become a very difficult part of the administration.

2.2. The main tax risk management indicators in China

According to the current tax risk management indicators of China's taxation department, and screening according to the availability of data, the tax risks are mainly divided into the following categories as the output of the research model in this paper.

- A-level risk indicators, are mainly used to monitor possible false transactions. The specific classification is as follows. ①Accounts receivable are greater than operating income. ②Accounts payable are greater than operating income. ③The total cost of the period is greater than 30% of the operating income.
- B-level risk indicators, are mainly to monitor possible tax evasion. The specific classification is as follows. ①Inventory is negative. ②Inventory is greater than 30% of operating income. ③The cost of sales is greater than the operating income. ④Accounts receivable are negative. ⑤The accounts payable are negative. ⑥The prepayment is negative.
- C-level risk indicators, are mainly used to monitor possible abnormal behaviors that are not directly related to tax items. The specific classification is as follows. ①Other receivables are greater than operating income. ②Other payables are greater than operating income.
- D-level risk indicators, are mainly used to monitor possible abnormal behavior of indirect tax-related items. The specific classification is as follows. ①Non-operating income exceeds 1% of operating income. ②Non-operating expense exceeds 1% of operating income. ③Other receivables are negative. ④Other payables are negative.

3. An overview of Convolutional Neural Network

3.1. The concept and application of convolutional neural network

Convolutional Neural Network (CNN) is a feed-forward neural network. It is a hierarchical model imitating human neural network and includes convolution calculation and deep structure. It has good performance in image recognition, classification and positioning [6-8]. In recent years, it has also been used in many fields such as financial currency and disease recognition. CNN is one of the representative algorithms of deep learning. It can be applied to diversified scenarios, and has good computational results in risk identification and target detection [9]. However, there is a gap in the application of CNN in tax risk management in China.

3.2. Convolutional neural network model

Convolutional Neural Network (CNN) consists of five basic structures, which includes input layer, convolution layer, pooling layer, fully connected layer and output layer. Each layer connects the same weighted neurons and maps them to different areas of the upper layer. We can obtain a neural network structure with translation invariance. Through the feature layers (convolutional layer and pooling layer), the original features are continuously extracted and compressed. More reliable high-level features are gradually obtained from low-level features. Then the last layer is used for tasks such as classification and regression. Its structure is shown in Figure 1.

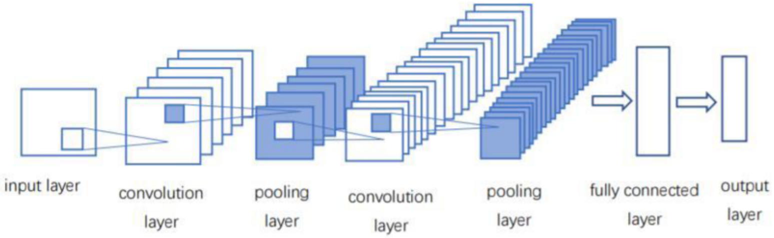

Figure 1. Structure of CNN.

The input layer is the training sample, which is the original raw data. In the convolutional layer, the input original data is divided into different regions by the convolution operation. In the actual operation process, in order to obtain higher-level features, we usually use multi-layer convolution. In simple terms, if we use the fewer convolution layers, we will get the lower feature extraction level. Multiple convolutions can make the low-level features gradually become high-level features. Pooling is mainly used to compress features. After the calculation of the fully connected layer, the result is passed to the output layer. This paper uses *Softmax* calculation to output the final classification results. The algorithm is as follows.

①The formula for calculating the *j* feature graph of the *l* layer of the convolution layer is:

$$X_j^l = f\ \left(\sum_{i \in M_j} (X_j^{l-1} \otimes K_{i,j}^l) + b_j^l\right) \tag{1}$$

In formula (1), $K_{i,j}^{l}$ represents the convolution kernel. ⊗ is the convolution operator. b_j^l represents a constant offset. X_j^{l-1} represents the feature graph of *(l-1)* layer, and is associated with X_j^l. ($M_j=1, ..., N^l$).

The convolution operator is an integral operation through convolution, which is used to calculate the area of the overlap region of two curves. It can be thought of as a weighted sum, replacing the pixel value of a point with a weighted average around it.

②The formula for the *(l+1)* pooling layer is:

$$X_j^{l+1} = p\ (X_j^l) \tag{2}$$

In formula (2), *p ()* is the down-sampling function. X_j^l represents the feature graphs of the *l* layer.

③The commonly used activation function is *ReLU*. When *x* is a negative number, the neuron will be 0. That is neuron necrosis. This paper uses the *ELU* function to avoid the death of some neurons to a certain extent. The calculation formula is:

$$y = \begin{cases} x, & x > 0 \\ \alpha\ (e^x - 1)\ , & x \leq 0 \end{cases} \tag{3}$$

In formula (3), α is an adjustable parameter. It controls when the negative portion of the *ELU* saturates. In the paper, $\alpha=1$.

④The formula of the fully connected layer is:

$$X_j^l = f\ (W_j^l \cdot X_j^{l-1} + b_j^l) \tag{4}$$

In formula (4), *f ()* is the ELU function. W_j^l represents the weight. b_j^l represents the offset.

⑤The formula of *Softmax* is:

$$y_i = \frac{e^{-W_i^L X^{L-1}}}{\sum_{j=1}^{M} e^{-W_i^L X^{L-1}}} \tag{5}$$

In formula (5), y_i. is a prediction vector ($y_i=y_1,...,y_M$). W_i^L represents the weight. X^{L-1} represents the multiple feature graphs X_j^l of the *l* layer. *M* represents the number of categories.

4. Tax risk prediction model for real estate based on convolutional neural network

4.1. Data collection and processing

This paper analyzes the subject of "Tax Risk Forecast Model for the Real Estate Industry". The data sample is the financial statement data of 142 listed companies in the real estate industry in the four quarters of 2019. The total amount of data is 8520, and many variables are obtained, such as operating income and accounts payable. The data covers most of the listed companies in the real estate industry. There are financial data reflecting the operating conditions of taxpayers. A large number of data conform to the selected indicators. The above data sources are true and reliable, and the information is sufficient. It is expected to achieve the desired goals of the research. The selection of data in this article is carried out according to the following procedures.

- The first step is to select the scope and content of the data according to the research and indicator requirements.
- The second step is to do a preliminary screening of the data. We will exclude data that are of little significance to the study.
- The third step is to eliminate missing value data.
- The fourth step is to classify and aggregate the scattered data items.

There are too many actual participation variables, which are limited by space and will not be listed one by one. The data comes from the CSMAR database. We took the data on December 31, 2019 as an example to screen and integrate the data, as shown in Table 1.

Table 1. Real estate listed companies fourth quarter financial data table.

Stock code	000002	000006	000011	Other：123
Operating income	3.67894E+11	3731330140	3961669942	…
Operating costs	2.3455E+11	2044318167	1433615885	…
Sales expense	9044496840	51736223.03	111553952.5	…
Non-operating income	714732128.7	2516753.07	23732348.28	…
Non-operating expense	788578652.7	8100823.59	4793503.85	…
…	…	…	…	…

According to this paper adopted by the tax risk level classification method, we will have visual chart data processing. Because there are too many data, we will not enumerate them one by one. We take the relationship between 2019 accounts receivable and operating revenue with the stock code of 000002 as an example. It is shown in Figure 2.

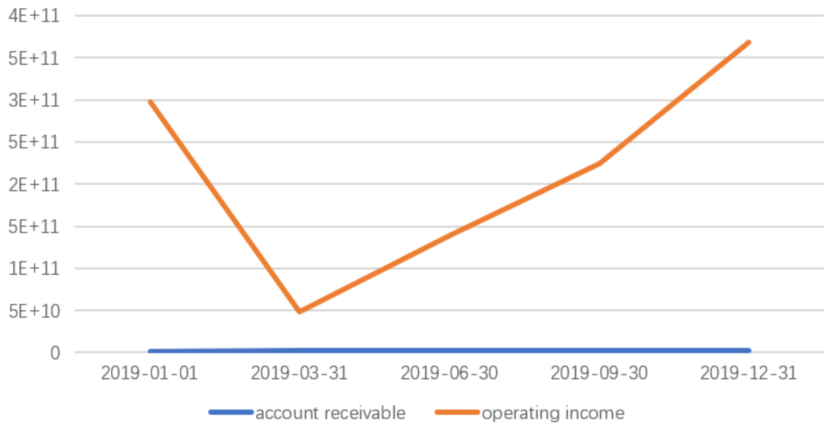

Figure 2. The relationship between the 000002's accounts receivable and operating income in 2019.

4.2. The model based on convolutional neural network

The obtained data is recorded as *X*, which is used as the input value of the model. We take the indicators, involved in the financial statement data of each enterprise, as the input of the model.

The tax risk prediction model of the real estate industry based on CNN is set as follows. Its structure is shown in Figure 3.

- Input layer: The indicators involved in the financial statement data of each enterprise are used as the input of the model.
- Convolutional layer: We use a 5 × 5 convolution kernel to obtain the corresponding feature map. Then we calculate it through the ELU function and add the deviation. Finally, we use it as the value of the neuron in the C1 layer.
- Pooling layer: Filter is set to 2 × 2. Then the corresponding feature map will be output. Among them, the scaling factor is 2 to control the compression speed.
- Output layer: Since we use tax risk management indicators to classify tax risk, the final output classification uses the *Softmax* function to achieve accurate tax risk classification. For a given test sample *X*, there are four possible outputs. If we input a vector element, it will output the probability value of [0-1]. This probability value is the probability of each tax risk classification.

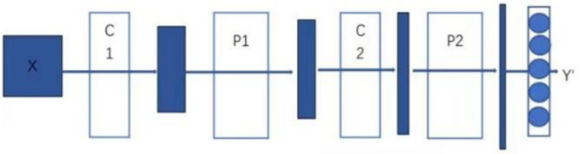

Figure 3. The model structure of tax risk prediction based on CNN.

5. Experimental results and analysis of tax risk prediction model based on convolutional neural network

We adopt the *ELU* function in the paper. Although the calculation amount is larger than that of the *ReLU* function, it can avoid neuron necrosis to a certain extent. When the value is negative, ELU has the characteristic of soft saturation. It improves the robustness to input changes. Moreover, since the output value of *ReLU* has no negative value, the mean value of the output is greater than zero. It makes the activation unit of the next layer have bias shift. The average value of ELU output is close to zero, which reduces the computational complexity. Therefore, compared with the traditional *ReLU* function algorithm, the convergence speed of ELU function is greatly improved. The function comparison is shown in Figure 4.

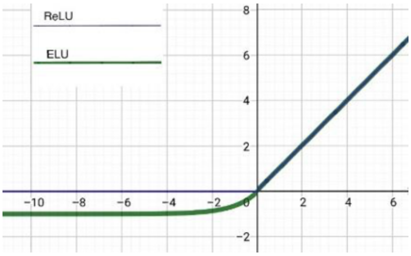

Figure 4. Functional image contrast between ReLU and ELU.

In this paper, we build a tax risk prediction model based on CNN. 80% of the data is set as the train set, while 20% of the data is used as the test set. Of the final 7260 data sets, 6000 are training sets and the remaining 1260 are test sets. In 1260 test sets, there are 1230 data consistent with the actual results, with an accuracy of 97.96%. We take the listed company with the stock code of 000002 as an example. The output result shows that the possible tax risk levels are A, B and C. Among them, the possibility with higher probability is B. We conduct data analysis on this company, as shown in Figure 5. In 2019, the company's annual inventory is more than 30% of the operating income, so there may be B-level tax risk. In the second and third quarters, accounts payable are greater than operating income, so there may be A-level tax risk. Other payables are greater than operating income, so there may be a C-level tax risk. The predicted results are consistent with the actual results.

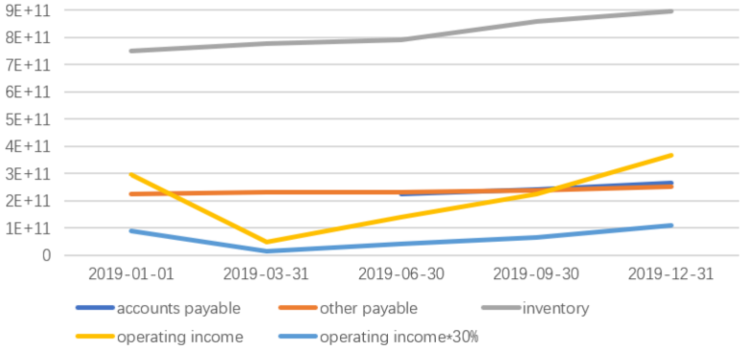

Figure 5. The relationship between the variables of the 000002.

For the model, this experiment uses four evaluation indicators: accuracy, detection accuracy, recall, and F1 value to evaluate the effect of the model. The predicted results are shown in Table 2.

As shown in Table 2, the use of CNN to build a model for tax risk prediction has a high accuracy rate of up to 97.35%. It also has high overall detection and good generalization performance. This result shows that CNN has a good performance in predicting tax risks. To a certain extent, it helps the authorities to manage the tax compliance of enterprises.

Table 2. Test results of CNN.

	Train set	**Test set**
Accuracy	0.9856	0.9735
Detection	0.9825	0.9751
Recall	0.9801	0.9706
F1 value	0.9865	0.9743

6. Conclusion

This paper takes real estate industry as the research object, and uses convolutional neural network algorithms to build tax risk prediction models. The experiment proves that CNN can be effectively used for tax risk prediction. Compared with traditional models, CNN can effectively and accurately identify the tax risks. The model can prevent and reduce tax risks, and provides an important reference for authorities. It can also help to improve corporate tax compliance of real estate companies and other industries. It has important theoretical and practical significance for preventing and reducing tax losses, and increasing the country's fiscal revenue.

The research of this paper has two deficiencies. First, we only predict the tax risk of the real estate industry of listed companies, with limited sample data. Second, only the deep learning method of CNN is used for model prediction, and there is a lack of more research and comparison of other deep learning models. Based on the above shortcomings, the research work of this article can be expanded from the following aspects. The first one is to improve and enrich machine learning algorithms. For example, deep learning algorithms such as Recurrent Neural Network (RNN) can be included in the comparative experiment. The second one is to expand the types and scope of data, so that the research is not only aimed at listed companies, and it can be applied to a wider range of different industries. Third, other researchers use various tax risk indicators, and we can increase our research on indicator selection in subsequent studies. In future studies, improvements and further studies will be made on the above aspects.

References

[1] Xingyi Song, Yongsheng Song. Path selection of tax risk management under big data environment. Taxation Research. 2020,(03).

[2] Weiling Wang, Jing Wang. Research on the Development Trend and Promotion Policy of my country's Digital Economy. Economic Review Journal. 2019 (01), 69-75.

[3] Zhijuan Lin. Study on the Tax Management Path of Colleges from the Perspective of Internal Control. 2018 International Conference on Economics,Finance,Business,and Development (ICEFBD 2018). 2018 Oct.

[4] Yuanzhao Gao, Xingyuan Chen, Xuehui Du. A Big Data Provenance Model for Data Security Supervision Based on PROV-DM Model. IEEE Access. 2020 Feb. PP(99):1-1.

[5] Cunyi Sun. Tax risk management from the perspective of big data. Taxation Research. 2019 Jul. 107-111.

[6] Fengmei Cui. Deployment and integration of smart sensors with IoT devices detecting fire disasters in huge forest environment. Computer Communications. 150 (2020), 818–827.

[7] M. Anbarasan, BalaAnand Muthu. Detection of flood disaster system based on IoT, big data and convolutional deep neural network. Computer Communications. 150 (2020), 150-157.

[8] S. Jothi Shri, S. Jothilakshmi. Crowd Video Event Classification using Convolutional Neural Network. Computer Communications. 147 (2019) 35–39.

[9] Min Chen, Yixue Hao, Kai Hwang, Lu Wang. Disease Prediction by Machine Learning Over Big Data From Healthcare Communities. IEEE Access. 2017 Apr. PP(99):1-1.