Innovative Design and Intelligent Manufacturing
L.C. Jain et al. (Eds.)

doi:10.3233/FAIA241097

A Study on the Construction of an Intelligent Prediction Model for Corporate Tax Risk

Zhen CHEN[a,1]
[a]*Software Engineering institute of Guangzhou, Guangzhou 510980, China*

Abstract. In order to solve the constraints on flexibility and scalability of traditional rule-based assessment methods in the comprehensive processing of massive, multi-dimensional and heterogeneous tax-related data for tax risk analysis of large enterprises, an intelligent prediction model for enterprise tax risk is proposed. Aiming at the typical scenarios of tax risk of related transactions of large enterprises, the characterization extraction method of tax data features is defined, and the tax risk analysis and prediction model based on artificial neural network multilayer perceptron is constructed. In order to evaluate the performance of the proposed model, a test dataset is constructed using real tax data and expert labeled data, and experiments are conducted under different positive and negative sample ratios and sample capacity sizes, and compared with several widely used machine learning models. The experimental results show that the model in this paper achieves better performance than the comparative models in both positive and negative sample balanced and unbalanced cases. Among them, when the ratio of positive and negative samples is 5:5, all the models achieve the optimal results relative to the data imbalance, while the model in this paper outperforms all the comparative models, and achieves the best precision rate, recall rate, F1 value and AUC value. Conclusion: The proposed method has good performance in terms of accuracy and effectiveness. With the continuous improvement of data labeling in tax authorities, the method of artificial neural network for tax risk assessment has a broad prospect in business use.

Keywords. Neural networks; data mining; tax risk; related transactions

1. Introduction

Risk warning has always been a major challenge. Since the 1930s, risk early warning methods have appeared in large numbers, but these methods can really predict the enterprise risk of very few, and even fewer of them are used in tax risk. To summarize, there are three main reasons. First, the early warning method mainly relies on the judgment of indicators, through the establishment of the indicator system for prediction. However, the indicators are often one-sided and subjective, and these defects of the indicators have led to the failure of early warning. Secondly, it is difficult to guarantee the quality of data. In the establishment of early warning indicator system, usually assuming that the data is true and reliable, but in reality, there are often missing data, anomalies and redundancy. Thirdly, the indicator system cannot be adjusted in real time, dynamically and individually [1]. Most of the indicators are added to the indicator system

[1] Corresponding Author: Zhen CHEN, huizi10000@126.com

after the occurrence of risk, and cannot be adjusted and revised in time with the type of risk and the specific situation of the enterprise, which often results in inaccurate prediction and even may be misleading.

Nowadays, big data has opened a major transformation of the times, changing the government, economy, humanities and other areas of society, and the tax field is no exception. Big data refers to a huge amount of information that can be acquired, stored, managed, analyzed, and organized to support daily decision-making based on hardware environment and software tools that are more than typical database tools [2]. In the era of big data, with the rapid development of the mobile Internet, the Internet of things, social networks and other media, effectively improve the transmission, storage, processing, mining and sharing of data and other aspects of the ability. At the same time, various internal operation data, financial data, and external correlation data of enterprises provide a rich source of data for enterprise risk early warning [3]. Therefore, how to quantify these massive, diverse, uncertain and instantly accessible data through big data analysis technology into the field of tax risk, so that the scientific nature of prediction, the effectiveness of the improvement and enhancement, will be the use of big data to crack the problem of early warning of tax risk is an important opportunity.

2. Literature Review

Didimo, W. et al. believed that risk assessment is an important step of risk management by studying the risk management of the construction industry. It also provides a reasonable systematic optimal model under each selection standard [4]. He, G. et al. studied the relationship between the actual tax rate and the relative scale of the underground economy by using the data and non-parametric regression analysis method [5]. Dhawan, A. et al. took the tax related projects in each link of a real estate company as the research object, analyzed the tax related matters of the real estate company in the preliminary preparation stage, construction stage, sales stage and retention stage, and used questionnaires and Delphi method to analyze the management's attitude to tax risks, legal awareness, employee quality, business status, tax related documents In terms of information, this aspect studied the tax risk of the enterprise, and analyzed the reasons for its abnormal changes with the business tax income ratio, tax profit ratio, stamp tax rate change rate, and sales profit ratio as auxiliary indicators, and put forward relevant suggestions on the tax risk control of the enterprise [6]. Kim, J., et al. took the VAT burden of 15 listed companies in the retail industry as the research object, constructed the enterprise VAT tax risk assessment model and assessment index system. This paper believed that the change of VAT tax burden rate was mainly affected by the purchase and sales ratio, gross profit rate of sales, and applicable VAT tax rate. Using the tax burden comparison method, the research results were obtained through analysis, The proportion of purchase and sales is inversely proportional to the VAT burden. The gross profit rate of sales is an important factor affecting the VAT burden rate of enterprises. The farther an enterprise's gross profit rate of sales deviates from the industry's gross profit rate of sales, the greater its VAT tax risk [7].

Savi ć, M. proposed to process the secondary response variables in Logistic regression analysis through quadratic programming, so as to carry out risk early warning, which greatly improved the discrimination ability of Logit model [8]. Chyz, J. A. and others constructed a new risk early warning model for structural risk minimization to predict local government debt risk [9]. A feasible and effective risk early warning model

is obtained by constructing training samples with TOPSIS method and DeFeer method, and training samples with support vector machine. Rahman, R. A. and others applied the risk early warning model to the doctor-patient relationship. It is proposed to build a risk early warning model of doctor-patient relationship through particle swarm optimization of back-propagation neural network, so that the model can converge faster and predict more accurately [10]. Abernathy, J. L. and others have established a risk early warning model based on the systematic risks in China's banking industry, combined with principal component analysis and regression analysis, and through empirical methods, so as to better implement macro supervision on the banking industry [11]. The research on the risk early-warning model of real estate enterprises is more focused on the financial risks of enterprises.

This paper explores a tax risk assessment model based on multi-layer perceptron of artificial neural network for large enterprise tax risk analysis scenarios. Taking the risk identification of typical related party transactions as a case study, the characteristics closely related to related party transactions are extracted. Based on these efficient characteristics, artificial neural networks are trained to assess tax risk.

3. The Research Methodology

3.1 Research Questions

For the following scenario: related enterprise Co_1 and enterprise Co_2 have transfer pricing tax avoidance, the transaction between the two is T_{12}, including commodity $\{item_1, item_2, \ldots, item_k\}$. If there is a tax burden difference between enterprise Co_1 and Co_2, determine whether there is transfer pricing tax avoidance between Co_1 and Co_2 by solving the following three problems [12–14].

A) Whether the enterprise Co_1 and Co_2 are related parties;

B) whether there are key commodities with abnormal pricing in T_{12} $\{item_1, item_2, \ldots, item_k\}$;

And c) Whether there is a tax difference between Co_1 and Co_2.

The whole problem can be formally described as: the given data set C, T and the pending test set X train a model $moudle(C,T)$ to realize the prediction and recognition of X.

a) Given the data set C, $moudle\_1(C)$ is realized through graph traversal to find out the association relationship that satisfies the specific graph topology path.

b) Given dataset T, $moudle\_2(T)$ is implemented by normalization to achieve classification T_key for attributes P_1, P_2, P_3.

c) given the training set R, T_key, D, B, and the tested set X, train the classifier $moudle\_3(R,T,T\_key,D,B)$.

3.2 Model Description

Order $Co_m = \{Co_1, Co_2, \ldots, Co_n\}$ is the set of all enterprise Co_i, where $Co_i = (\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_m)$, the value space of A j is the set $\{a_1, a_2, \cdots, a_{s_j}\}$. s_j is the number of available values of A j. In addition, $T_{ab} = \{\text{item}_1, \text{item}_2, \cdots, \text{item}_n\}$ is the transaction between Co_a and Co_b; $T = \{T_{ij}\}, i,j = 1,2,\cdots,n,\ i \neq j$ is the transaction of Co_m.

Taking Co_m, Co_b and T_{ab} as inputs, the tax assessment model of related party transactions can be described as $\{Pr_{ab}, AT_{ab}\} = F(Co_a, Co_b, T_{ab})$, where Pr_{ab} is defined as the probability of Co_m, Co_b occurring as a key transaction AT_{ab} for the purpose of tax avoidance. The model is schematic shown in Figure 1.

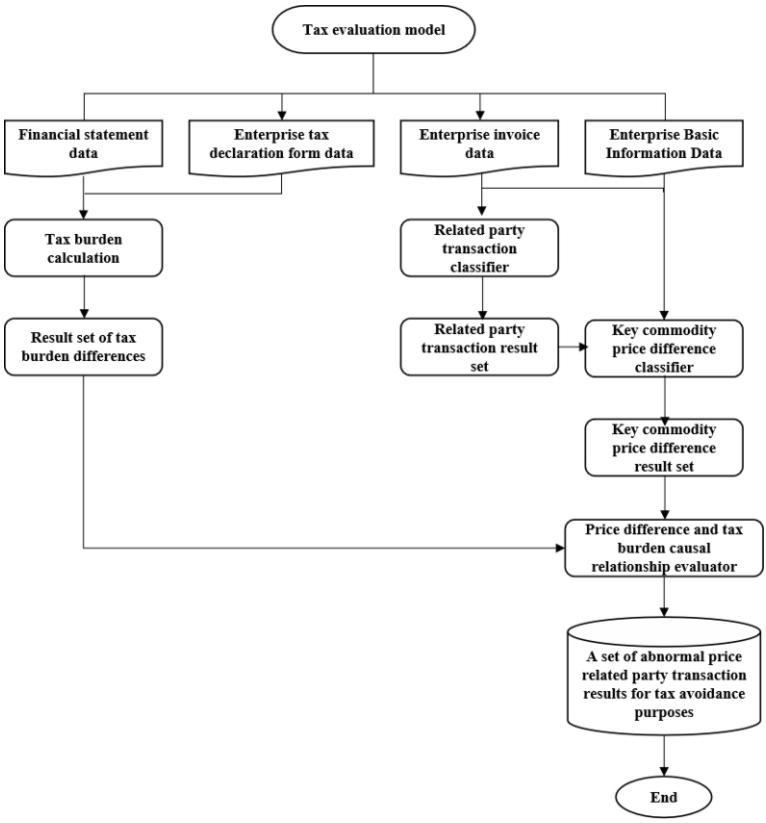

Figure 1. Tax assessment model.

In this paper, we go about solving the problem in three steps$F(Co_a, Co_b, T_{ab})$, and the process of which is shown in Figure 2.

4. Experimental Validation

4.1. Description of the Experimental Data-Set

The experimental data consists of the database of a district tax bureau of the State Administration of Taxation, the public information of enterprises and the query results of third-party databases, including the taxpayers' information form, the annual tax return of enterprise income tax, and the invoice data. The taxpayer information table is filled in by the enterprise when it first registers for tax, and consists of 43 fields, including registration serial number, state and local tax type code, tax file number, taxpayer identification number, taxpayer name, registration type code of taxable body, registration

type code, name of legal representative, and type code of legal representative's identity document [15].

4.2. Data Preprocessing and Baseline Test Dataset

Through pre-processing, the shareholding information of 236 enterprises under the jurisdiction of three large enterprises is obtained, and 63,000 invoices corresponding to the enterprise information, including ordinary invoices, special invoices and electronic invoices, as well as 330,000 commodities corresponding to the invoices, are obtained by querying the electronic ledger system.

However, the manual audit results of the tax authorities in these 236 enterprises show that only 5 enterprises have related-party transaction tax avoidance behaviors, and the valid samples are too small, so the a priori information is used to construct the baseline test dataset in order to validate the validity of the method [16].

4.3. Experimental Steps

Aiming at the insufficient number of samples of enterprises with related-party transaction tax avoidance behaviors in practice, the baseline test data set with different positive and negative sample ratios is set as the training set, and the positive and negative sample ratios are 5:5, 4:6, 3:7, 2:8, and 1:9, respectively, and the sample capacity is 100,000, which is used to test the validity of the proposed model and compare it with other common machine learning models.

The neural network has 4 layers, 24 and 12 nodes in the hidden layer, 2 nodes in the output layer and 12 nodes in the input layer. The Adam solver is used to train the neural network, and the batch size is set to 200, and the target value of training accuracy is 0 001, the maximum training times is 10000, the momentum factor is 0.9, and the activation function is ReLU function.

4.4. Experimental Results

This section builds a machine learning environment based on the above ideas. The experimental environment is: Intel XeonTM Gold 6140 CPU 2.30 GHz, memory 128 GB, operating system Ubuntu, and programming language Python 3.7. The implementation of the model can be roughly divided into three parts: environment building, data processing and model implementation. The Python libraries related to machine learning that the above environment depends on include Pandas, NumPy, Scipy, Matplotlib, Sklearn, guppy, psutil, etc.

For the price difference and tax burden causal relationship evaluator, the proposed model is compared with other common machine learning models. For different models, training was conducted on the baseline test data set with a sample size of 100000 and a positive negative sample ratio of 1:9, 2:8, 3:7, 4:6, and 5:5. After training each model in a training set with a sample size of 100000 and a positive negative sample ratio of 5:5, the ROC curve obtained on the test set is shown in Figure 3.

The P-R curve and ROC curve results with different ratios of positive and negative samples on the test dataset show that the model in this paper achieves better performance than the comparison models in both positive and negative sample balanced and unbalanced cases. Among them, when the ratio of positive and negative samples is 5:5, all the models achieve the optimal results relative to the unbalanced data, while the model

in this paper outperforms all the comparative models, and achieves the best precision rate, recall rate, F1 value and AUC value.

5. Conclusion

On the real dataset, this paper screens out 143 companies with the status of surviving, of which 32 have related party transactions and 7 are considered to have tax risks, among which 5 enterprises are labeled as having tax risks by manual audit of tax authorities.

There is still room for improving the quality of the dataset used in the current model. In fact, the search for related parties and the calculation of average prices in the commodity industry are complex issues in themselves, such as data integrity, missing fields, insufficient labeled data, and inconsistencies in the caliber of some statistics. The supplementation of the empirical dataset, the improvement of the labeled data, and the validation of the baseline test dataset are being carried out continuously. With the continuous accumulation of empirical cases of tax department audit, the tagged sample dataset will be further improved; at the same time, combined with the optimization of the parameters of the feature selection model, the effect of this paper's tax risk assessment method for large-scale enterprises based on artificial neural network multilayer perceptron in enterprise tax risk assessment is expected to be further improved.

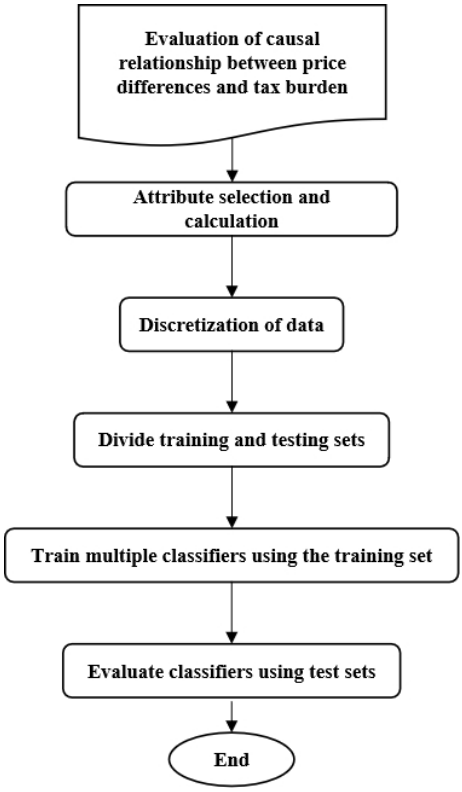

Figure 2. Assessment of the causal relationship between price differences and tax burden.

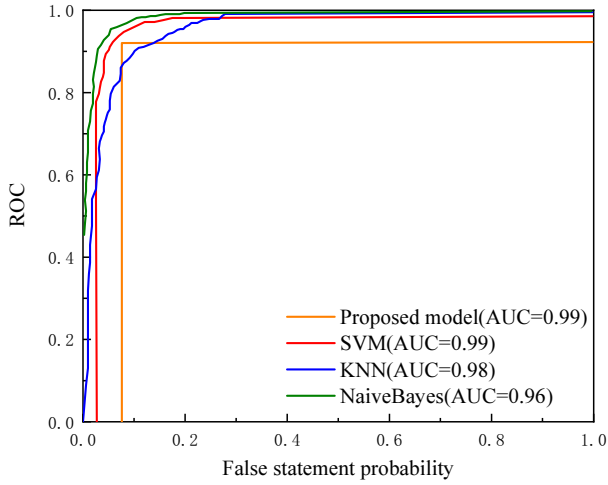

Figure 3. ROC Curve Obtained by Each Model

References

[1] Beasley, M. S., Goldman, N. C., Lewellen, C. M., & McAllister, M. (2021). Board risk oversight and corporate tax-planning practices. Journal of Management Accounting Research, 33(1), 7-32.
[2] Abedin, M. Z., Chi, G., Uddin, M. M., Satu, M. S., Khan, M. I., & Hajek, P. (2020). Tax default prediction using feature transformation-based machine learning. IEEE Access, 9, 19864-19881.
[3] Choi, J., & Park, H. (2022). Tax avoidance, tax risk, and corporate governance: evidence from Korea. Sustainability, 14(1), 469.
[4] Didimo, W., Grilli, L., Liotta, G., Menconi, L., Montecchiani, F., & Pagliuca, D. (2020). Combining network visualization and data mining for tax risk assessment. Ieee Access, 8, 16073-16086.
[5] He, G., Ren, H. M., & Taffler, R. (2020). The impact of corporate tax avoidance on analyst coverage and forecasts. Review of Quantitative Finance and Accounting, 54(2), 447-477.
[6] Dhawan, A., Ma, L., & Kim, M. H. (2020). Effect of corporate tax avoidance activities on firm bankruptcy risk. Journal of Contemporary Accounting & Economics, 16(2), 100187.
[7] Kim, J., McGuire, S., Savoy, S., & Wilson, R. (2022). Expected economic growth and investment in corporate tax planning. Review of accounting studies, 27(2), 745-778.
[8] Savić, M., Atanasijević, J., Jakovetić, D., & Krejić, N. (2022). Tax evasion risk management using a Hybrid Unsupervised Outlier Detection method. Expert Systems with Applications, 193, 116409.
[9] Chyz, J. A., Gal-Or, R., Naiker, V., & Sharma, D. S. (2021). The association between auditor provided tax planning and tax compliance services and tax avoidance and tax risk. The journal of the american taxation association, 43(2), 7-36.
[10] Rahman, R. A., Masrom, S., Omar, N., & Zakaria, M. (2020). An application of machine learning on corporate tax avoidance detection model. IAES International Journal of Artificial Intelligence, 9(4), 721.
[11] Abernathy, J. L., Finley, A. R., Rapley, E. T., & Stekelberg, J. (2021). External auditor responses to tax risk. Journal of Accounting, Auditing & Finance, 36(3), 489-516.
[12] Chen, X., Cheng, Q., Chow, T., & Liu, Y. (2021). Corporate in-house tax departments. Contemporary Accounting Research, 38(1), 443-482.
[13] Cao, Y., Feng, Z., Lu, M., & Shan, Y. (2021). Tax avoidance and firm risk: evidence from China. Accounting & Finance, 61(3), 4967-5000.
[14] Jacob, M. (2022). Real effects of corporate taxation: A review. European Accounting Review, 31(1), 269-296.
[15] Zwick, E. (2021). The costs of corporate tax complexity. American Economic Journal: Economic Policy, 13(2), 467-500.
[16] Eberhartinger, E., & Zieser, M. (2021). The effects of cooperative compliance on firms' tax risk, tax risk management and compliance costs. Schmalenbach Journal of Business Research, 73(1), 125-178.