Financial Risk Prediction of Listed Companies based on Text and Financial Data

Xu Wei and Yonghui Chen
Southwest University of Science and Technology School of Computer Science and Technology, Mianyang, China

Keywords: Deep Learning, CNN, Financial Risk, Annual Report Text.

Abstract: This paper uses the relevant financial data of 4348 A-share listed companies from 2010 to 2019 and the discussion and analysis of operation in the annual report as the research sample, and uses the Pytorch framework to build a neural network model to predict whether the listed companies fall into financial crisis. The experimental results show that when text data is combined with traditional financial index data, the prediction accuracy of the deep learning model can reach about 85%, which can significantly improve the accuracy of financial risk prediction compared with using only financial data.

1 INTRODUCTION

The financial crisis of listed companies is one of the main drivers of the financial risk and gains primary attention from creditors and investors. The financial risks of listed companies may further propagate recession and thus jeopardize the economy at large (Ding 2012). Therefore, it is of great significance to pay attention to the financial situation of listed companies.

The research of financial risk prediction is usually based on the company's stock market transaction information and the company's financial index data (Ye 2017, Song 2019, Hosaka 2019, Zhang 2021). However, as an unstructured and qualitative data form, the annual report text of listed companies also plays a very important role in how to convey information to the public. For example, the annual report submitted by a listed company to the regulator contains all the information about the company in the past year (Feng 2019). Among them, “Discussion and Analysis of Operating Conditions” mainly analyzes the operating conditions, operating models, and operating strategies of listed companies. It is a summary of the overall operating conditions of the past year and a generalization of future operating directions. Recent research has proven that qualitative company reports contain important information that can predict financial risks (Campbell 2014). Since it is difficult to obtain and quantify text data, it is still challenging to effectively combine financial text with financial data in a financial model (Lang 2015). In this research, we propose a new deep learning method that combines financial text and financial data to predict financial risks. We found that text data can supplement traditional accounting and market-based variables to predict financial risk, and the deep learning model that combines financial text and financial data has higher prediction accuracy than a model that uses a single type of input.

2 MATERIALS AND METHODS

2.1 Data

2.1.1 Financial Indicator Data

In this paper, financial indicator data of A-share listed companies from 2010 to 2019 were obtained from the Jukuan database. Since financial indicators are subject to change during the year, the data on financial indicators as of the date of disclosure of the company's annual report by the listed companies are taken.

Based on relevant studies, a total of 30 financial indicators are selected in this paper to analyze the solvency, operating capacity, development capacity, profitability, risk level, cash flow, and ratio structure indicators of enterprises. Based on general indicators, as many financial factors affecting the business conditions of enterprises are considered as possible,

Wei, X. and Chen, Y.
Financial Risk Prediction of Listed Companies based on Text and Financial Data.
DOI: 10.5220/0011172500003440
In *Proceedings of the International Conference on Big Data Economy and Digital Management (BDEDM 2022)*, pages 240-244
ISBN: 978-989-758-593-7

and rich indicators are selected for feature learning to avoid artificially removing features that have a significant impact on the prediction results. The selected financial indicators include current ratio, quick ratio, cash ratio, gearing ratio, cash flow ratio, accounts receivable turnover, inventory turnover, current asset turnover, total asset turnover, operating income growth rate, operating profit growth rate, net profit growth rate, net flow from operating activities growth rate, total assets growth rate, return on assets, total net asset margin, return on net assets, net operating margin, return on investment, financial leverage, operating leverage, consolidated leverage, net cash flow from operating activities per share, cash recovery rate, operating index, cash asset ratio, working capital ratio, fixed asset ratio, equity concentration index and z-index(Table 1).

Table 1: Financial index system of financial risk prediction.

	Selected indicators	Abbreviation	Calculation description
Solvency	X1 Current ratio	RA	Current assets / current liabilities
	X2 Quick ratio	RAT	Quick assets / current liabilities
	X3 Cash ratio	CR	Monetary capital / current liabilities
	X4 Asset liability ratio	LEV	Total liabilities / total assets
	X5 Cash flow ratio	CASHCL	Net operating cash flow / current liabilities
Operating capacity	X6 Accounts receivable turnover	ARTURNOV	Average balance of operating income / accounts receivable
	X7 Inventory turnover	INTURNOV	Operating income / inventory
	X8 Turnover rate of current assets	VOL	Net income from main business / average balance of current assets
	X9 Total asset turnover	TATO	Operating income / total assets
Development capacity	X10 Growth rate of operating revenue	RG	Increase in operating income / operating income of the previous period
	X11 Operating profit growth rate	PGR	Increase in operating profit / operating profit of the previous period
	X12 Net profit growth rate	EG	Increase in net profit / net profit of the previous period
	X13 Growth rate of net flow from operating activities	GNFOA	Increase in net operating cash flow / net operating cash flow of the previous period
	X14 Growth rate of total assets	TAGR	Increase in total assets / total assets at the beginning of the period
Profitability	X15 Return on assets	ROA	Profit without financial expenses / average total assets
	X16 Net interest rate of total assets	ROT	Net profit / total assets
	X17 Return on net assets	ROE	After tax profit / net assets
	X18 Net operating interest rate	NPM	Net profit / operating income
	X19 Return on investment	ROI	Average total profit / total investment
Risk level	X20 financial leverage	DFL	Change rate of earnings per share of common stock / change rate of EBIT
	X21 Operating leverage	DOL	Change rate of EBIT / change rate of production and sales volume
	X22 Integrated lever	DTL	Change rate of net profit / change rate of main business income
Cash flow analysis	X23 Net cash flow from operating activities per share	NCFOPS	Net operating cash flow / number of common shares outstanding
	X24 Cash recovery rate	CRA	Net operating cash flow / total assets at the end of the period
	X25 Operating index	OI	Net operating cash flow / gross operating cash flow
Ratio structure	X26 Cash asset ratio	CAR	Cash assets / total assets
	X27 Working capital ratio	WCR	Total working capital / assets
	X28 Fixed assets ratio	LTCR	Fixed assets / total assets
Internal governance	X29 Equity concentration index	HERF	Number of shares of the largest shareholder / total number of shares of the company
	X30 Z index	Z	Number of shares of the first largest shareholder / number of shares of the second largest shareholder

In the actual data, the numerical ranges of different features are different, which may appear in the feature space. Individual features with large values have a dominant impact on the sample. To make all features on the same scale, it is necessary to map them to the same scale, to improve the accuracy of the model and speed up the fitting speed, so the samples should be normalized in this experiment. You can use the MinMaxScaler provided in the preprocessing class in Sklearn, the specific formula is (1). Where min represents the minimum value of each feature in the data, and max is the maximum value of each feature.

$$X_{\text{scale}} = \frac{x-\min}{\max-\min} \quad (1)$$

2.1.2 Text Data

We obtained the annual reports of all listed companies from 2010 to 2019 in PDF format from Eastmoney.

Python provides many class libraries for parsing PDF files, among which PDFMiner and PDFPlumber are some of the most common parsing methods. This paper finally uses PDFMiner to analyze the annual report in PDF format. After obtaining the text, clean it first, and use the regular formula to extract the chapter of "Discussion and Analysis of Operating Conditions" in the annual report. It is worth mentioning that, according to the requirements of the CSRC, "Discussion and Analysis of Operating Conditions" was not used as a separate chapter before 2016, but inserted into other chapters. Therefore, different methods should be taken to extract the annual reports in different periods. In this study, we exclude the blank samples in the section of "Discussion and Analysis of Operating Conditions" and finally obtain 28549 text data.

2.2 Methods

2.2.1 Text Representation

Before combining text with financial indicator data into the model, we need to consider how to represent the text. The usual approach is to convert text into vectors. That is, the text is mapped into a new space and represented by multi-dimensional continuous real number vectors.

We use the word2vec model based on Skip-gram (Deoras 2013). The goal of the Skip-gram model is to make a word predict the words around it. For a sequence of words W1, W2, W3, ..., Wn, we maximize log p.

$$\max \frac{l}{n} \sum_{t=1}^{n} \sum_{j=-c, j\neq 0}^{c} \log\, p\left(w_{t+j} \mid w_t\right) \quad (2)$$

In formula (2), c represents the number of words before and after the word is considered. In this paper, according to the size of the text data training set, the dimension of the word vector is set at 100 and the fixed-length is set at 5. Negative sampling is used to calculate log p, and the sub-sampling of words is proportional to their inverse frequency. In word2vec, words with similar semantics have high cosine similarity and allow vector calculation of words.

This paper uses Jieba word segmentation to segment our financial text data set to obtain the financial text corpus. We use the general vocabulary training set to train a general word vector model, and then use it for the second training. The final word vector contains financial background information, which is more suitable for our task. We calculate the average of all word vectors in each text, i.e., each financial text is represented by a 100-dimensional vector.

2.2.2 Model Building

We use Pytorch to combine financial index data with text data to build our model. The model structure is shown in Figure 1:

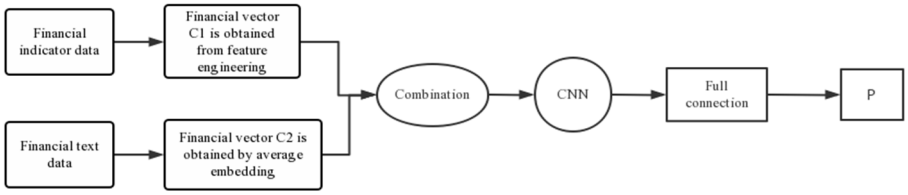

Figure 1 Schematic diagram of financial indexes and financial text combination model.

The input of the model is a 130-dimensional vector, which consists of two parts:

- Financial indicator data: including 30 financial indicators, each of which is normalized;
- Financial text: the section of "business discussion and analysis" in the annual report.

The model splices the financial index data and text data through the intersection of pd.merge function of pandas library, and send them to the convolutional neural network (CNN). The model parameters of convolution neural network include the number of convolution cores, the size of convolution cores, the size of the pooling layer, and so on. To select the best parameters to fit the model in this paper, we reset the value range of parameters. For example, the CNN convolution kernel size d∈{2,3,4,5}, the number of CNN convolution cores h∈{64,100,128,256}, the pool layer size c∈{5,6,7,8}, and the learning rate λ∈{0.01, 0.001, 0.0001}, epoch∈{5, 10, 15}, the weight value of cross-entropy loss function f∈{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 3, 4}.

3 RESULTS & DISCUSSION

We compare the model proposed in this paper with other models. These models are:

S-CNN: Feature vectors are constructed based on financial data, and then the CNN model is used to extract features and realize classification.

S-SVM: The model based on financial data uses SVM to classify.

S-XGB: The model based on financial data uses XGBoost to classify.

The evaluation results of each model are shown in Table 2:

Table 2: Experiment summary table.

	Accuracy	True Positive Rate	True Negative Rate	F1 Value
S-CNN	78.00%	89.02%	54.57%	0.676623
T&S-CNN	85.00%	93.38%	77.67%	0.848035
S-SVM	70.83%	75.60%	63.38%	0.689527
S-XGB	77.12%	89.02%	54.57%	0.676623

It can be seen from the table that the prediction effect of the CNN deep learning model based on financial data is not significantly better than the traditional machine learning model based on financial data. After the combination of financial data and financial text, the CNN model is higher than other models' inaccuracy, true positive rate, true negative rate, and F1 value. There may be two main reasons:

- The convolutional neural network model pays more attention to information, which leads to insufficient attention to important information. After adding the financial text features, although there is still a lot of information, with the help of the financial text features, important features are highlighted.
- From the perspective of the financial text, the more information combined with the data, the better. In this way, after the combination of important information and data, after the screening of multi-layer neural networks, the more important information can be selected.

4 CONCLUSIONS

As more and more financial documents appear in the stock market, investors, regulators, and researchers need more deep learning models to process and analyze the information disclosures of listed companies. Taking all A-share listed companies in the recent ten years as samples, this paper builds a financial risk prediction model based on financial text and financial data. The experimental results show that compared with using only financial data, the F1 value of the financial risk prediction model based on the combination of text and financial data is significantly improved, indicating that the latest

progress of neural network can extract useful information from financial text, and the financial risk prediction combined with traditional financial data can further improve the prediction effect.

REFERENCES

Campbell, J.L., Chen, H., Dhaliwal, D.S., et al. The Information Content of Mandatory Risk Factor Disclosures in Corporate Filings [J]. Social Science Electronic Publishing.

Zhang, Chunmei, Zhao, Mingqing, Guan, Junqi. Financial risk portfolio early warning model of manufacturing listed companies based on lasso + SVM [J] Practice and understanding of mathematics, 2021, 51 (5): 12.

Deoras A, Tomá? Mikolov, Kombrink S, et al. Approximate inference: A sampling based modeling technique to capture complex dependencies in a language model[J]. Speech Communication, 2013, 55(1):162-177.

Ding, A.A., Tian, S., Yu, Y., et al. A Class of Discrete Transformation Survival Models with Application to Default Probability Prediction.

Hosaka, T. Bankruptcy prediction using imaged financial ratios and convolutional neural networks[J]. Expert Systems with Application, 2019, 117(MAR.):287-299.

Lang, M., Stice-Lawrence, L., Textual analysis and international financial reporting: Large sample evidence[J]. Journal of Accounting and Economics, 2015.

Ye, Lanzhou, Zhou J.Y., Management, S.O., Application of Support Vector Machine in Risk Management of Listed Companies[J]. Science Technology and Industry, 2017.

Mai, F., Tian, S., Lee, C., et al. Deep learning models for bankruptcy prediction using textual disclosures[J]. EUROPEAN JOURNAL OF OPERATIONAL RESEARCH, 2019.

Song, Y., Peng, Y., A MCDM-Based Evaluation Approach for Imbalanced Classification Methods in Financial Risk Prediction[J]. IEEE Access, 2019, PP (99): 1-1.