Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 199 (2022) 262–268

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021)

# A risk detection framework of Chinese high-tech firms using wide & deep learning model based on text disclosure

Fang Da[a], Chao Peng[a], Haomin Wang[a], Tie Li[a]*

[a]School of Management and Economics, University of Electronic Science and Technology of China, Chengdu, 611731, P. R. China

**Abstract**

High-tech firms have been recognized as a major source of earnings in most of countries in the world, especially in China. A high degree of information asymmetry has led to the problem of difficult financing, high interest rates, and complicated audit procedures for high-tech firms. However, little works focus on an efficient automated framework to improve risk detection accuracy for high-tech firms. Our work proposes a new framework using wide & deep learning model with annual report text and financial data to detect risk of high-tech firms. We examine the effects of the proposed framework by comparing them with other mature classification methods utilizing real firms data in China. The results showed that the proposed framework with text improve classification performance compared to baseline methods with financial data. The rate of increase in accuracy, recall, AUC is 12.2%, 100% and 44.4%. Moreover, the results suggests that the importance of unstructured data and soft information of high-tech firms should be emphasized to improve risk detection accuracy.



*Keywords*: risk detection framework; wide & deep learning models; text disclosure; high-tech firms

## 1. Introduction

In China, high-tech firms develop business activities based on continuous research and development and the transformation of technological achievements. Taking a classical view, high-tech firms have been recognized as a major source of employment and income in the world and have made important contribution to global technological development and economic. Unfortunately, the finance support and development of high-tech firms are still in dilemma. From the perspective of high-tech firms, high interest rates and collateral requirements have greatly increased the difficulty of obtaining credit and finance. From the perspective of banks and financial institutions, high-tech firms represent high bankruptcy risk. This issue becomes even more complicated because

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
*E-mail address:* lteb2002@uestc.edu.cn.


10.1016/j.procs.2022.01.032

of the lack of data, difficulties in accessing the authentic information and the shifting landscape of high-tech firms.

Specifically, high-tech firms have peculiarities, such as high risk, high information asymmetry, no collateral, uncertain production and profit. Due to their higher risk than traditional industries and their importance to economic and technology, the Chinese government have provided tax relief and encouraged banks to provide funds support for high-tech firms. It appears to solve the problem of high-tech firms financing and development. The situations resemble that some high-tech firms receiving financing but bankruptcy has been repeated occurrence.

The fact is that high-tech firms are different from traditional firms [1]. Thus, the paradigm of risk detection needs to be changed. Due to high-tech peculiarities, it is impending for developing the suitable risk detection framework of high-tech firms to improve prediction accuracy rather than blindly granting funding. The popularity solution of credit and risk issues is to improve the risk assessment model of firms without considering the specificities [2]. Nevertheless, the differences and particularities between firms data, industries require that firms risk detection cannot be generalized.

In addition to the acknowledged financial data regarded as one of the risk detection indicators, non-financial variables can also predict firms risk [3]. And extant works have verified soft information is a crucial factor of business risk [4]. Further, according to the Schumpeterian theory statements [5], the uncertain of innovation can bring about tremendously risks. Therefore, the soft information, such as text, can influence the risk of high-tech firms. In order to implement the combination of structure data and unstructured data to detect high-tech firms risk, we utilize the wide & deep leaning model [6] to construct risk detection framework. The pressing concern for high risk of high-tech firms motivates us to contribute to improve risk detection accuracy from constructing the suitable framework perspective.

In this paper, we propose a risk detection framework of high-tech firms. First, we collect 965 Chinese high-tech firms data consisting of basic information, financial information, annual report text and status (risky or no risk). And then we design a risk detection framework employing the wide & deep learning model of high-tech firms to verify the effectiveness by comparing the performance of the basic classifiers using complementary text and traditional financial data. As a result, the prediction ability of risky high-tech firms is quite a bit improved, which is reflected in the recall metric.

## 2. Literature review

The firms risk is the impact of future uncertainty on the realization of business objectives. Most of the previous research about firms risk, risk warning system, bankruptcy prediction and default prediction focus on the firms evaluation models and frameworks establishment, especially classification methods [7]. A famous z-score model [8] is extensively applied in firms risk, default area. And Wiginton [9] proposed the logistic regression model for firms credit risk prediction. With the development of computer and computing power, some scholars show that the machine learning methods and data mining have better performance than statistical methods [10]. The most frequently used techniques in data mining are random forest (RF) [11], support vector machine (SVM) [10], neural networks (NN) [12] involving deep learning methods [13].

In recent years, deep learning methods are epidemic because of their ability of effectively learning complex and nonlinear relationships from data. Google scientists developed and applied a wide & deep learning algorithm in mobile application recommendation on the Google Play store [6]. The main contribution of wide & deep learning is to integrate linear models and neural networks to achieve memory and generalization ability.

## 3. A risk detection framework of high-tech firms

This section is divided into two parts, the first subsection introduces the wide & deep learning model foundation and the second subsection explain the proposed risk detection framework of high-tech firms using wide & deep learning model.

### *3.1. Wide & deep learning model*

Generalized linear models with non-linear feature transformation are widely used in regression and classification problems. Wide & deep learning model is a hybrid model composed of a single-layer wide part (logistic regression) and a multi-layer deep part (deep neural network). The main function of the wide part is to have strong memory ability. The main function of the deep part is to have generalization ability. The model has both the advantages of logistic regression and deep neural network, which can quickly process and memorize a large number of historical behavior characteristics, and have powerful expression capabilities.

Fig.1 illustrates the structure of wide & deep learning models. The wide & deep learning framework jointly trains feed-forward neural networks with embeddings and linear model with feature transformations, which have been used for recommender systems with sparse inputs. The wide models on the left of the Fig.1 are generalized linear models, and the deep models are feed-forward neural networks on the right. During training period, joint training simultaneously optimizes all parameters by weighted summing the parameters of the wide models and the deep models.

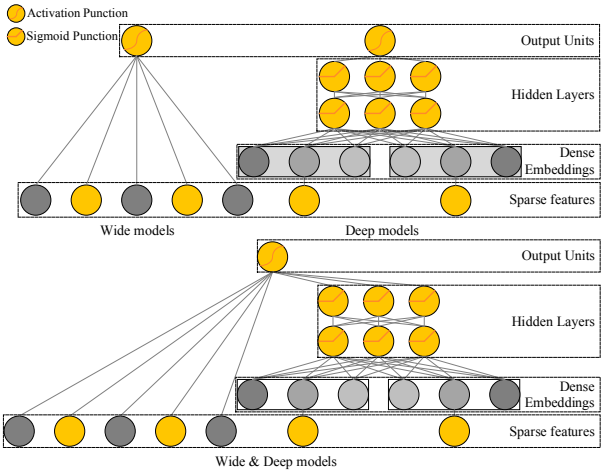
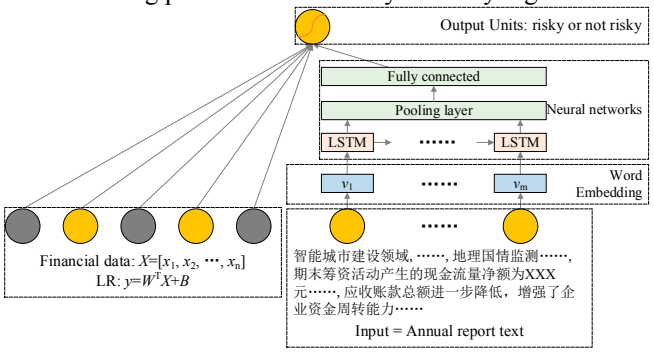



Fig. 1. The structure of wide & deep models [6]

### *3.2. The proposed risk detection model of high-tech firms*

Due to the importance of the soft information of enterprises, especially for high-tech firms, the traditional risk detection models are appropriate. The risk detection framework that combines structured data processing and unstructured data processing is imperative since the soft information largely contains unstructured data and financial data is structured data. The structure of the wide & deep learning model is applicable to process financial data and annual report text.

Fig.2 below describes the proposed framework for risk detection of high-tech firms utilizing wide &deep learning model. The inputs of the experiment are financial data and annual report text. In this experiment, the high-tech firms are in China and the annual report text is Chinese. The annual report text translation is located below the framework in Fig.2. The outputs are high-tech firms' status, risky or no risky. The left side of Fig.2 (wide model) is a logistic regression (LR) model, which used to process financial data. The right side of Fig.2 (deep model) is long short-term memory (LSTM) part handling the annual report text. The difference between the proposed model and the original model is the deep part. The procedure annual report text of high-tech firms has three steps. Firstly, the text words should be transformed into vectors by word embedding. Secondly, the vectorized text data is sent into the neural network module, which has LSTM, pooling layer, and fully connected layer. Thirdly, the outputs of neural networks are put into the sigmoid functions with wide part outputs. The training process of the proposed model optimizes the parameters of LR and LSTM synchronously. In the testing period, the test data exploits the training parameters to classify the risky high-tech firms and no risky firms.




| Annual report text translation: Chinese-English |
|---|
| Smart city construction field, ……, Geographical national conditions monitoring……, The net cash flow from financing activities at the end of the period was RMB XXX……, The total amount of accounts receivable was further reduced, which enhanced the company's capital turnover capacity. …… |

Fig. 2. The proposed risk detection framework of high-tech firms using the wide & deep learning model

## 4. Experiments

This section contains the data description and experimental results. The first part states the sample data and experimental settings. The second part is experimental results of the proposed risk detection framework.

### *4.1. Data and experimental settings*

we collected experimental data on 965 Chinese high-tech firms, which were listed in the new over the counter market in 2018. The data includes financial data and annual report text. Table 1 show the sample data distribution. The original data of high-tech firms has basic information, financial information, and annual report text. The features in this experiment contains financial data and annual report text, which belong to numerical data and unstructured data. The financial data has 9 variables and is not displayed in this paper for privacy protection.

The experiment employs the jieba tokenization toolkit and python 3.6.5 with TensorFlow-GPU 1.8 backend at a personal computer of i5- 7300HQ, 8GB RAM, with Windows 1064x. The disk is 1TB+128GB SSD and GPU is NVIDIA Geforce GTX 1050Ti.

We selected four frequently used classification methods to validate the risk classification performance of the proposed framework. The classification methods used in the experiment are logistic regression (LR), support

vector machine (SVM), random forest (RF), and neural network (NN). We selected five widely used classification evaluation metrics: accuracy, precision, recall, G-mean, and area under the receiver operating characteristic (ROC) curve (AUC).

Table 1. Sample data

| Samples | Training set | Test set | Total |
|---|---|---|---|
| Total | 965 | 73 | 1038 |
| Risky firms | 770 | 50 | 820 |
| No risky firms | 195 | 23 | 218 |

### *4.2. Results*

Table 2 shows the classification results of the five approaches. The proposed model handles the 2 kinds of data and obtain the best performance among these approaches in all metrics. All classifiers achieved good performance in accuracy. Recall is an important metric in risk prediction because it reflects how many risky high-tech firms have been classified correctly. The proposed model is more effective than other approaches in correctly classifying the risky high-tech firms. Fig. 3 is the classification ROC of the experiment, the ROC curve of the proposed model based on wide & deep learning is on the top, which means the proposed model has the best classification results.

The result of the experiment reveals that the features combing financial data and text with wide & deep learning approach has pretty classification effect. It illustrates that unstructured data, especially soft information, is helpful for risk detection of high-tech firms. Moreover, since the diversity of data require the advanced framework, the deep learning-related methods are useful and valuable.

Table 2. Classification performance of different data and approaches

| Approaches | Data | Accuracy | Precision | Recall | G-mean | AUC |
|---|---|---|---|---|---|---|
| LR | Financial data | 0.880 | 0.292 | 0.778 | 0.830 | 0.845 |
| SVM | Financial data | 0.931 | 0.438 | 0.778 | 0.855 | 0.867 |
| RF | Financial data | 0.924 | 0.364 | 0.444 | 0.651 | 0.904 |
| NN | Financial data | 0.950 | 0.533 | 0.831 | 0.921 | 0.914 |
| The proposed model based on wide & deep learning | Financial data + Text | **0.987** | **0.889** | **0.889** | **0.940** | **0.945** |

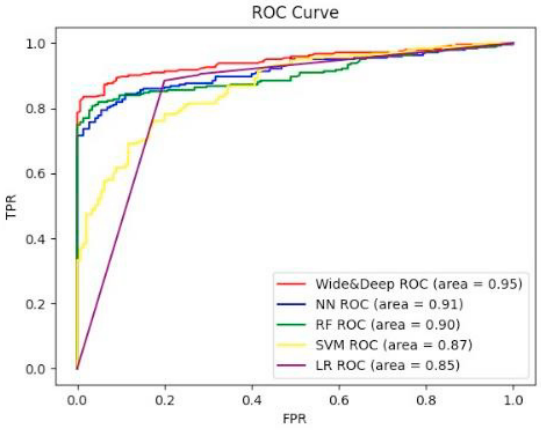



Fig. 3. The risk classification ROC of different approaches

## 5. Conclusion

Extant studies on risk detection of high-tech firms are rare, in spite of the topics of firms' risk evaluation are prevalent. Our study complements the risk detection framework of high-tech firms. We propose the risk detection model based on the wide & deep learning model with financial data and annual report text. The classification results of the proposed model obtain the outstanding performance comparing with the baseline models with sole financial data. Furthermore, the experimental result suggests that financial institutions should pay more attention to soft information and unstructured data to detect risks of high-tech firms.

## Acknowledgements

This work was supported in part by grants from the Ministry of Education Project of Humanities and Social Science (#20YJC630064).

## References

[1] Hong J, Feng B, Wu Y, et al. Do government grants promote innovation efficiency in China's high-tech industries?[J]. Technovation, 2016, 57: 4-13.
[2] Ammann M. Credit risk valuation: methods, models, and applications[M]. Springer Science & Business Media, 2002.
[3] Altman E I, Sabato G, Wilson N. The value of non-financial information in SME risk management[J]. Available at SSRN 1320612, 2008.
[4] Cornée S. The relevance of soft information for predicting small business credit default: Evidence from a social bank[J]. Journal of Small Business Management, 2019, 57(3): 699-719.
[5] Aghion P, Akcigit U, Howitt P. What do we learn from Schumpeterian growth theory?[M]//Handbook of economic growth. Elsevier, 2014, 2: 515-563.
[6] Cheng H T, Koc L, Harmsen J, et al. Wide & deep learning for recommender systems[C]//Proceedings of the 1st workshop on deep learning for recommender systems. 2016: 7-10.
[7] Wang H, Kou G, Peng Y. Multi-class misclassification cost matrix for credit ratings in peer-to-peer lending[J]. Journal of the Operational Research Society, 2020: 1-12.

[8] Altman E I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy[J]. The journal of finance, 1968, 23(4): 589-609.
[9] Wiginton J C. A note on the comparison of logit and discriminant models of consumer credit behavior[J]. Journal of Financial and Quantitative Analysis, 1980: 757-770.
[10] Kim H S, Sohn S Y. Support vector machines for default prediction of SMEs based on technology credit[J]. European Journal of Operational Research, 2010, 201(3): 838-846.
[11] Shi T, Horvath S. Unsupervised learning with random forest predictors[J]. Journal of Computational and Graphical Statistics, 2006, 15(1): 118-138.
[12] Huang X, Liu X, Ren Y. Enterprise credit risk evaluation based on neural network algorithm[J]. Cognitive Systems Research, 2018, 52: 317-324.
[13] Mai F, Tian S, Lee C, et al. Deep learning models for bankruptcy prediction using textual disclosures[J]. European journal of operational research, 2019, 274(2): 743-758.