A Novel Structure-based Feature Extraction Approach for Financial Fraud Detection

Hui Xia¹*, Hui Ma²

¹School of Accounting Chongqing University of Technology Chongqing, China
²Chongqing Rural Commercial Bank Chongqing, China

*Corresponding author e-mail: summertulip@cqut.edu.cn

Abstract. In finance and economic area, financial fraud detection plays an important role for both corporate management and capital market system. Feature extraction is one of the most important procedure in fraudulent firm detection. Current feature extraction approaches pay large amounts of attention on their financial attributes, which have explicitly limited the representation of 'normal' pictures of firms, and furthermore, reduced the financial fraud detection performance. Hence it is necessary to search for a better set of functions as features to represent firms more accurately. In this work, notice that the imitation behaviors among firms often happen in business management, while this structure patterns have not been utilized in financial fraud detection so far, we extract features under the constraint of both financial characters and structure patterns of firms. We also design three measurements to quantify the structure patterns. Experimental results have shown a great performance of the proposed approach.

Keywords: financial fraud, feature extraction, NMF, peer effects.

1. Introduction

Currently, most financial fraud detection systems can provide early warning signals, or red flags, which are especially useful in decision making. Moreover, the outputs can help creditors manage financial risks better, and investors transfer funds to cut loss in time. Generally speaking, financial fraud detection systems not only play an important role for both business operators and other interested parties such as stakeholders, investors and so on [1,2], but also are good for sustainable business development [3,4].

Prevalently, machine learning based approaches are the most effective ways for detection, such as clustering, matrix factorization, ensemble learning, neural networks and so on [5-12], which usually employ well-defined financial ratios as primitive inputs. Those ratios are defined and verified by domain experts and practices empirically. However, with the development of financial and accounting, the number of financial ratios increases rapidly, which aggregates the difficulties in artificial feature selection. Besides, the growing complexity of corporate business has made the well-defined financial ratios have more limitations in representation.

Hence, it is necessary to develop an effective feature selection strategy for financial fraud detection. In financial fraud detection, current feature extraction process focuses on the low rank
approximation[10-12] of each firm from financial data space while neglecting the structure patterns among firms. In the areas of economic and finance, this structure patterns between firms are described as peer effects, which represent how an individual's decision is influenced by her peers'[13]. Peer effects have drawn an increasing attention as firms often pay close attention to what their peer groups are doing, such as investment decision[14], marketing strategies making[15], financial policies shaping[16,17] and so on. As for the underlying financial illegal behavior of each firm, peer effects can not only lead to higher lawlessness of corporations, but also aggravate the infringement of legal interests of capital providers[18]. Though recognizing peer effects could create negative social multiplier effects[19], current feature extraction methods almost ignore this critical nature, resulting in a limited detection performance and a poor explanation on the derived results.

In our work, we illustrate a novel feature extraction model. We refine high-dimensional data considering both financial characters and structure patterns. The proposed model project the high-dimensional financial data into a new low-dimensional space under the constraints that matching known structure patterns of firms: peer effects (which have been neglected by current techniques), and represent each firm more accurately with a small number of functions searching from a potentially infinite set. Comparing with other feature extraction methods, the proposed model demonstrate more influences from peer groups, which are reasonable and more effective in describing the reality.

2. Background and Related Work

Current researches usually employ financial ratios as the input of financial fraud detection. Even though Bao et al.[9] have demonstrated the significance of raw financial data in financial fraud detection, large numbers of research pay attention on the financial ratios because the results are powerful and explainable in summarizing the financial data and indicating the health of a company[5]. In this work, we also employ the financial ratios. We gather financial ratios through widely accepted 4 aspects: solvency, profitability, growth ability and operating capacity. It is no doubt that a certain number of financial ratios has led the financial fraud detection problem to a high-dimensional space.

For the purpose of computation resource saving and the avoidance of overfitting, we need to process feature reduction first. One effective way for dimension reduction is the projection, such as PCA (Principal Component Analysis)[10,11] and NMF (Non-negative Matrix Analysis) [12], through transforming primitive data into a new space while retaining as much variations as possible presented in the entire data set. Generally, the extracted features can approximate the data very accurately in the low-dimensional space. However, the above approximation is totally based on data without considering the domain knowledge. This directly leads a poor explanation of extracted features and reduces the representation accuracy. Illustrated examples are NMF, which constructs features randomly and blindly. In financial management domain, we not only focus on extracting financial features of firms, but also pay attention on the structure patterns among them.

The structure patterns in finance and economy area are named as peer effects, which play an important role in shaping firm behaviors. Peer learning occurs within certain social contexts, taking place through observation, imitation and modeling [20]. Through observation, an individual can percept the frequency of peer groups perform actions in question. After obtaining the knowledge, the individual would repeat and practice through mental and physical rehearsal. Finally, the above imitation process helps to model their own behavioral patterns based on peer groups when facing more uncertainty. Peer effects are believed to be a critical, explanatory factor for research on finance and economic because of two reasons.

However, one key challenge is the measurement of firm structure patterns, peer effects. According to empirical studies, the connections between firms are usually measured by either in the same industry, or in the adjacent geographical positions. Generally, firms in the same industry and similar size are more likely to learn from each other, enabling them to avoid the potential risks and the possibilities of falling behind the rivals [15]. Besides that, the geographical agglomeration of firms can determine senior executives' social networks, following by the communication learning activities between them, and finally leading to the knowledge spillover [22]. Another significant method of forming peer groups
is the identification of the common sell-side analysts between firms. In this method, the senior executives are truly connected via analysts, and the choices of analysts directly reflect the relatedness of firms [23]. Both the specialization and the direct communication with senior executives can help analysts get firm conducts timely and accurately.

In this work, we do feature extraction through allowing a wider class of features, constrained by known behavior patterns of firms, peer effects, in order to avoid an explosion of possibilities and a more accurate representation on both financial features and structure patterns. In financial fraud detection procedure, we do not attempt here to do comparisons with all existing methods, which would be prohibitive given the size of the literature, we have deliberately chosen three exemplar approaches that cover a very wide range of ideas for financial fraud detection, and in doing so we aim to illustrate the general improvement that our approach obtains.

3. The Proposed Feature Extraction Model

3.1. Measurements of Structure Patterns

In order to quantify peer effects, we should have basic information of each firm, such as the industry, the geographical location (longitude and latitude) of office address. Note that most researchers regard imitation behaviors usually occurring in the same area and the same industry, we propose a structure pattern measurement of firms as:

$$\text{sim}_{ij} = \begin{cases} \exp(-\text{dis}(i_1, i_2)) & \text{if } \text{industry}^i_{i_1} = \text{industry}^i_{i_2} \\ 0 & \text{if } \text{industry}^i_{i_1} \neq \text{industry}^i_{i_2} \end{cases}$$

where \(\text{dis}(i_1, i_2)\) is the geographical distance between company \(i_1\) and \(i_2\), which is calculated by Haversine formula [21]. In time period \(\Delta T\), suppose there are \(T\) time unit, and \(j \in [1, J]\), then \(\text{industry}^i_j\) refers the industry of company \(i_1\) in the \(j\)th time unit. The whole equation means that if company \(i_1\) and \(i_2\) are in the same industry, and are close in geographical locations, they are more likely to imitate with each other, otherwise, there is no peer effects between \(i_1\) and \(i_2\).

3.2. A Structure based Feature Extraction Approach

The feature extraction problem can be expressed as searching for a set of functions which not only minimises the reconstruction error of financial data, but also highlights the structure patterns of peer groups.

$$F(U, V) = \text{argmin} \lVert R - UV \rVert^2_F + \lambda_1 \lVert S - (VV^TU^T) \rVert^2_F + \lambda_2 (\lVert U \rVert^2_F + \lVert V \rVert^2_F)$$

where \(R\) is the averaged financial ratio matrix of all firms in period \(\Delta T\), \(U\) and \(V\) can be seen as feature matrices of corporates and financial ratios respectively. \(K\) is the number of latent features; \(S\) is used for quantifying the structure similarity between firms; \(\lambda_1\) and \(\lambda_2\) are the parameters of the second, third and forth term in (2); \(\lVert \cdot \rVert_F\) represents the 2-norm of one matrix.

The first term tries to get a low-rank approximation \(UV\) of financial data. The second and forth terms are regular terms in avoid of overfitting problem. As for the second term, it means that the low-rank approximation of each firm should also keep the structure patterns as much as possible, that is, minimizing the structure pattern reconstruction error.

Similar to NMF algorithm, we apply the multiplicative iteration method to optimize \(U\) and \(V\), as shown in (3) and (4).

$$U = U \cdot \frac{RV^T + \lambda_2 U + 2\lambda_1 S U V V^T}{U V V^T + 2\lambda_2 U V V^T U V V^T}$$

$$V = V \cdot \frac{\lambda_2 V + U^T R + 2\lambda_1 U^T S V}{U^T U V + 2\lambda_2 U^T U V V^T U V V^T}$$
3.3. A Structure based Feature Extraction Algorithm

Based on (3) and (4), we could get a iterative feature extraction approach, as shown in below.

Algorithm 1: The Structure based feature extraction algorithm

Inputs: Financial ratio matrix, $R$; Similarity matrix, $S$;
Parameters $\lambda_1$ and $\lambda_2$.

Outputs: Feature matrix of users, $U$; Feature matrix of financial ratios, $V$.

1. Initialize the feature matrices of users $U$ and financial ratios $V$ randomly;
2. While not convergence:
3. Update $U$ according to (3);
4. Update $V$ according to (4);
5. Calculate reconstruction error $\epsilon = \frac{1}{2} \sum_{i,j} (R_{i,j} - U_{i,j}V_{j,i})^2$
6. End of while

Algorithm 1 introduces the feature extraction process in detail. We check the convergence through either the residuals $\epsilon$ or the iteration times $b$. The preset bounds of $\epsilon$ and $b$ are $\varepsilon$ and $\beta$ respectively. If $\epsilon$ reduces to $\varepsilon$ in $b$ ($b \leq \beta$) iteration times, it means the algorithm converges.

4. Experimental Results

In this section, we apply several feature extraction approaches to the financial data of publicly traded China firms, and measure the performance through financial detection approach. The financial detection approach we use is error-based statistical approach. Since the low-rank approximation matrix is considered as 'normal' picture of firms, then the large error deviation will be classified as fraudulent ones. The compared feature extraction approaches are NMF, PCA and SDA (Statistical approach without Feature Extraction).

4.1. Metrics

We label the ground truth of financial fraud firms as a set $GF$, while real non-fraud firms are $GN$. Similarly, the detected set of financial fraud firms is $D^r$, where the applied method is r. We conclude a pair of metrics for performance evaluation: detection rate ($DR$) and false alarm rate ($FAR$). $DR$ refers to the detected financial fraud firms divided by the total financial fraud firms, $FAR$ is the number of normal firms misclassified as illegal ones divided by all normal firms as seen in (5) and (6).

$$DR = \frac{|GF \cap D^r|}{|GF|}$$

(5)

$$FAR = \frac{|GN \cap D^r|}{|GN|}$$

(6)

where $|$ means the size of a set.

4.2. Data description

Our samples consist of all the publicly traded construction companies listed on the Shenzhen Stock Exchange and Shanghai Stock Exchange in China. All data is derived from the China Stock Market and Accounting Research (CSMAR) database. According to CSMAR database, there are totally 217 financial ratios. This work focuses on firms over period between 2009 and 2017. Through learning the financial attributes and structure patterns in the past several years, we try to extract the low space features. All experiments are repeated 100 times to get statistical and convincing results. The sample
ratio we used here is 2:1, that is, for each financial fraud firm, we select two ‘normal’ firms with the same industry and similar business size.

4.3. Experiment on varying numbers of latent features $K$

In this work, the number of latent features determines how we know about the financial features of each firm. Hence, there is a trade-off of choosing a best $K$, as shown below.

| Metrics | The values of K |
|---------|-----------------|
| 3       | 40              | 70          | 100         | 130         | 160         | 190         |
| DR      | 0.6984          | 0.8230      | 0.7733      | 0.7891      | 0.7601      | 0.7543      |
| FAR     | 0.1421          | 0.0780      | 0.0626      | 0.0721      | 0.0801      | 0.0887      | 0.0594      |

As Table I shows, we compare the detection performance through selecting $K$ in the range [3, 40, 70, 100, 130, 160, 190]. Before $K$ increases to 70, the detection rates of all three curves are increasing, while the false alarm rates are decreasing. However, with the continuing growth of $K$, the detection performance declines, as the detection rate decreases and the false alarm rate increases. Hence, it can be concluded that the best trade-off of $K$ should be 70.

4.4. Experiment of the proposed approach

The detection performance of all approaches has been illustrated on Table II. Obviously the proposed structure based feature extraction approach outperforms the rest. Among all, the proposed approach has the best detection precision (with DR value as high as 0.8747, FAR value as low as 0.0626). This is reasonable since the geographical location based similarity measurement takes the view that companies of the same industry and in the same region usually confront the similar business environment, and are more likely to imitate peer groups.

Generally, comparing with the rest PCA, NMF and SDA approaches, the proposed structure based approach has explicitly better detection rates and lower false alarm rates. This demonstrates the effectiveness of structure pattern in normal picture profiling.

| Metrics | Methodologies |
|---------|---------------|
|         | The proposed approach | NMF | PCA | SDA |
| DR      | 0.8747        | 0.6779 | 0.4716 | 0.4722 |
| FAR     | 0.0626        | 0.1611 | 0.2642 | 0.2639 |

4.5. Experiment on samples without sampling

The above experiments are all on the sampled data, in which each financial fraud firm is matched with at least one non-fraud firms. This is not appropriate in reality. In this work, we validate the effectiveness of the proposed feature extraction approach comparing with the rest three methods using all publicly traded China firms without any sampling strategies. According to Table III, the performance of all four methods are not better than the sampled data scenario as shown in Table II. This is because the models can not search for the distinctive features without pairwise sampling. In real detection scenario, there are too many disturbances in feature extraction, since all industries are included in experiment, and even in the same industry, the management characteristics of companies are different. This directly weaken the detection performance. In this condition, the proposed feature extraction approach still shows the largest detection rate and the smallest false alarm rate.
Table 3. The Detection Performance of Four Methods on Real Data without Sampling

| Metrics | Methodologies |
|---------|---------------|
|         | The proposed approach | NMF | PCA | SDA |
| DR      | 0.7534         | 0.5947 | 0.4523 | 0.4415 |
| FAR     | 0.2765         | 0.2843 | 0.2917 | 0.2989 |

5. Conclusion

Traditional feature extraction approaches in finance and economic area normally focus on the attributes of data, and ignore the domain characters, which has limit its wide and effective application. In this paper, we address this problem by producing a set of functions as features constrained by structure patterns, which has not been used in feature extraction as far as we know. We then evaluate the proposed feature extraction model using data of publicly traded China firms. Comparing with the other three approaches, the proposed model can describe 'normal' picture of firms more accurately, with higher detection rate and lower false alarm rate. However, there are many issues left, such as the issue of generalization and timeliness, which require large amounts of further working on the perfection of feature extraction model.

Acknowledgments

This work is supported by the Science and Technology Research Project of Chongqing Education Commission (KJQN201801103) and the Humanities and Social Science Research Project of Chongqing Education Commission (20SKGH176). We also thank to the kindly supported from the Science and Technology Research Project of Chongqing Education Commission (KJQN202001140).

References

[1] Khanh Nguyen. Financial statement fraud: Motives, methods, cases and detection. Master’s thesis, Queensland University of Technology, 2008.
[2] Yuri Mishina, Bemadine J Dykes, Emily S. Block, and Timothy G. Pollock. Why ’good’ firms do bad things: The effects of high aspirations, high expectations and prominence on the incidence of corporate illegality. The Academy of Management Journal, 53(4):701–722, 2010.
[3] Saeed Nosratabadi, Amir Mosavi, Shahaboddin Shamshirband, Edmundas Kazimieras Zavadskas, and Kwok wing Chau. Sustainable business models: a review. Sustainability, 11, 2019.
[4] Weisheng Lu, Meng Ye, K.W. Chau, and Roger Flanagan. The paradoxical nexus between corporate social responsibility and sustainable financial performance: Evidence from the international construction business. Corporate Social Responsibility and Environmental Management, 25(5), 2018.
[5] Ravisankar P, Ravi V, Raghava Rao G, and Bose I. Detection of financial statement fraud and feature selection using data mining techniques. Decision Support Systems, 50, 2011.
[6] Chi-Chen Lin, An-An Chiu, Shaio Yan Huang, and David C Yen. Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts judgments. KnowledgeBased Systems, 89, 2015.
[7] Chengwei Liu, Yixiang Chan, Syed Hasnain Alam Kazmi, and Hao Fu. Financial fraud detection model: Based on random forest. International Journal of Economics and Finance, 7(7), 2015.
[8] Ran Wang, Vahid Asghari, Shu-Chien Hsu, Chia-Jung Lee, and JiehHaur Chen. Detecting corporate misconduct through random forest in china’s construction industry. Journal of Cleaner Production, 268, 2020.
[9] Yang Bao, Bin Ke, Bin Li, Y. Julia Yu, and Jie Zhang. Detecting accounting fraud in publicly
traded U.S. firms using a machine learning approach. Journal of Accounting Research, 58(1), 2020.

[10] Shin-Ying Huang, Rua-Huan Tsaih, and Fang Yu. Topological pattern discovery and feature extraction for fraudulent financial reporting. Expert Systems with Application, 41:4360–4372, 2014.

[11] A tutorial on principal component analysis: https://www.cs.princeton.edu/picasso/mats/pca-tutorial-intuition.jp.pdf.

[12] Lee DD and Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 1999.

[13] Huagang Tong and Jianjun Zhu. New peer effect-based approach for service matching in cloud manufacturing under uncertain preferences. Applied Soft Computing Journal, 94, 2020.

[14] Cen Wu, Xin Yu, and Ying Zheng. The spillover effect of financial information in mergers and acquisitions. The British Accounting Review, 52, 2020.

[15] Shenglan Chen and Hui Ma. Peer effects in decision-making: Evidence from corporate investment. China Journal of Accounting Research, 10, 2017.

[16] Ethan M.J.Lieber and William Skimmyhorn. Peer effects in financial decision-making. Journal of Public Economics, 163, 2018.

[17] Brunner Markus and Ostermaier Andreas. Peer influence on managerial honesty: The role of transparency and expectations. Journal of Business Ethics, 154(2), 2019.

[18] Rong Lu and Wei Chang. Peer effect in corporate fraud. Journal of Financial Research, 8(458), 2018.

[19] Jonathan M. Karpoff. The future of financial fraud. Journal of Corporate Finance, 2020.

[20] Nelwati, Khatijah Lim Abdullah, Mei Chan Chong, and Lisa McKenna. The effect of peer learning on professional competence development among Indonesian undergraduate nursing students: A quasi-experimental study. Journal of Professional Nursing, 2020.

[21] The haversine formula: https://www.longitudestore.com/haversine-formula.html.

[22] Stijn Kelchtermans, Daniel Neicu, and Peter Teirlinck. The role of peer effects in firms usage of R&D tax exemptions. Journal of Business Research, 108, 2020.

[23] Markku Kausti and Ville Rantala. Social learning and corporate peer effects. Journal of Financial Economics, 117, 2015.