Financial Fraud Detection Using Deep Learning Based on Modified Tabular Learning

Meiying Huang and Wenxuan Li(✉)

School of Finance and Economics, Qinghai University, Xining, China
545118597@qq.com

Abstract. With the rapid development of Internet technology and the rapid progress of the financial industry, fraud is causing more and more damage, which not only brings huge losses to enterprises, but also has a significant impact on corporate image. Therefore, detecting fraud is an important topic. At present, there are roughly two methods to detect fraud. One is to establish corresponding standards in the financial field for manual detection. The defects of this method are slow detection speed, lagging update and high false positive rate. Another method is automatic recognition of the machine. However, the disadvantage of this method is that when the machine runs stably, too many will cause great pressure to the machine. Therefore, in recent years, with the application of artificial intelligence in the financial field, the application of artificial intelligence method in fraud detection has great potential. At present, the mainstream intelligent methods for fraud detection include convolutional neural network (CNN) and support vector regression (SVR). However, these methods are not interpretable in tabular data model, we proposed a feature-based deep learning regression model that can directly deal with tabular data. In order to verify the effectiveness of this model, we conducted an experiment on a real transfer record of a mobile payment company with the proposed method and mainstream method. The results show that the model has a good performance in detecting fraudulent behavior and verifies the feasibility of the model.

Keywords: Fraud Detection · Deep Learning · Neural Networks · Interpretability

1 Introduction

In recent years, with the rapid development of economy, financial fraud has emerged gradually. In 2020, as a large number of businesses moved online due to the COVID-19 pandemic, fraud has increased in the field as well. Especially in the financial field, a large number of businesses involving identity authentication and data audit, such as credit and transaction business, have also been transferred online. As the risk control system of financial institutions corresponding to online business is relatively weak, there is no good way to face financial fraud. Taking personal credit as an example, Tencent research [6] found that fraud caused about 40–70% of the total overdue; It can be found that about 75% of the overdue fraud cases have obvious behaviors such as data counterfeiting and

Z. Zeng et al. (Eds.): ECIT 2022, AHE 11, pp. 550–558, 2023.
https://doi.org/10.2991/978-94-6463-005-3_55

abnormal behavior in the process of overdue fraud. Companies have made efforts to detect fraud by choosing more secure service providers. Gartne's report [5] revealed that because of the high threshold of fraud effectiveness comparison verification, enterprises are difficult to choose traditional security service vendors, and need to rely on large Internet companies with a large amount of real user data.

As a sensitive topic in the financial field, fraud detection can also be analyzed by using artificial intelligence algorithms. As a result, advanced data analysis techniques to detect fraud have received a lot of attention in the past few years. For example, convolutional neural network (CNN) [4] in deep learning and support vector regression (SVR) [1] in machine learning. Convolutional neural network is a kind of feedforward neural network with deep structure including convolution computation, and is widely used in image recognition, video processing, natural language processing, machine learning, document analysis of deep learning algorithm. However, it also has the following disadvantages: It is easy to make the training result converge to the local minimum rather than the global minimum by using the gradient descent algorithm. In addition, the pooling layer will lose a lot of valuable information and ignore the correlation between the local and the whole. SVR is a supervised machine learning algorithm for classification, regression, and anomaly detection. In this algorithm, each data item will be plotted as a point in n-dimensional space, where N is the eigennumber and each eigenvalue is the value of a specific coordinate, and classification is carried out by looking for hyperplanes that distinguish the two classes. However, SVR algorithm is difficult to implement large-scale training samples and easy to be sensitive to missing data.

However, some methods such as convolution neural network (CNN) and support vector regression (SVR) can not explain the tabular data model, a feature-based fraud detection framework is proposed based on interpretability deep learning.

This work makes the following contributions:

1) Intelligent feature extraction. A method based on deep learning is adopted to automatically detect whether there is fraud in transfer records, eliminating the process of manual feature extraction;
2) Tabular data can be processed directly. Provides an end-to-end learning approach based on gradient descent with the benefit of semi-supervised learning and the ability to use information from another training model to learn to solve related tasks;
3) Diverse validation data sets. The same data set and different methods were used to verify the detection performance and universality. On the one hand, we conducted data type screening experiments and formed a small data set for transfer type. On the other hand, a monthly transfer log of a mobile payment company is taken as an example to verify the effectiveness of this method.

The rest of this paper is organized as follows. In Sect. 2, gives the description of the three methods and the mathematical expression of the algorithm used. In Sect. 3, the details of introduction to data sets, data processing and comparative experiments. And comparisons with other methods are shown in this section. Finally, Sect. 4 concludes this paper.

2 Methods

In order to show the structure of the work, our analysis includes the detailed descriptions of the mathematical expressions of CNN, SVR, the proposed method and performance evaluation metrics, which are listed as follows.

2.1 CNN Algorithm

CNN is known for its ability to automatically extract valid features. We use three Conv stages (Conv Stages) and two fully connected stages (FC stages) to complete training for the model, where each Conv Stage includes a Conv layer, a leaky ReLU function, and a Max-pooling layer [8]. The convolution layer uses the following formula to extract new features from the output of the previous layer:

$$O_m^n(i) = Q^n(i) * A_c^{n-1} + F^n(i) \tag{1}$$

where $O_m^n(i) \in R$ denotes the output of the e_{th} unit using the i_{th} filter in the n_{th} layer. $A_e^{n-1} \in R^{1\times t}$ is the e_{th} sub-vector of the input data in the previous layer $n-1$, t is the length filters and A_e^{n-1}.

Each element $O_m^n(i)$ in the extracted features is mapped by a non-linear activation function named Leaky ReLU with the following expression:

$$g_m^n(i) = O_m^n(i) \tag{2}$$

which helps to avoid gradient explosion and gradient disappearance problems. The activation output is denoted as $g^n(i) = \left|g_1^n(i), g_2^n(i), \ldots\ldots\ldots, g_m^n(i), \ldots\ldots, g_l^n(i)\right|$

The dimension reduced by the maximum pooling layer is used to reduce the number of training parameters, avoid overfitting and improve the robustness of the model [11]. However, some useful information can also be lost after the max-pooling process. In this experiment, we use a max-pooling layer in the first Conv phase, and the formula is as follows:

$$P_m^n(i) = \max_{\gamma=1,\ldots,s} g_\gamma^n + I_{pl(m-1)}(i) \tag{3}$$

where $s \in Z$ is the max-pooling size, and I_{pl} is the stride in the max-pooling layer.

Through stacking three Conv Stages above, a multi-stage structure is constructed for feature representation [10]. Then the output features of the multi-stage structure is flattened and pass to two FC Stages, where one is used for dimension reduction and another is used for the final prediction.

2.2 SVR Algorithm

SVR is famous for its stability and high prediction accuracy. Support vector regression (SVR) is a supervised learning algorithm used to predict discrete values. The goal of SVR is to reduce errors by determining the hyperplane and minimizing the range between predicted and observed values [3]. As shown in the figure:

$$\min \frac{1}{2}||w||^2 \tag{4}$$

$$s.t.\ |y_i - (wx_i + b)| \leq \zeta, \forall i \tag{5}$$

The minimization of w value is similar to the maximization of marginal value. In this study, the kernel function is designed to transform into another linear space by changing the inner product space to replace it with another kernel function space. Where the kernel of SVR is Gaussian kernel with a hyper parameter σ in this paper:

$$k(x_i, x_j) = \exp\left(-\|x_i - x_j\| / \left(2\delta^2\right)\right) \tag{6}$$

2.3 The Proposed Method

We adopt a tabular learning based domain adaptation method for transfer training. The tabular learning network consists of submodules of multiple steps, each of which focuses on features at a different level [9]. Each step is composed of Attentive transformer and Feature transformer and some auxiliary operations. Attentive transformer serves as a mask for output features and is used to learn the importance of each Feature. Feature transformer serves to extract features and generate a more effective representation of sample properties (Fig. 1).

Attentive Transformer outputs a Mask for feature selection as shown in Eqs. (6–8). i represent the step. a[i] represents the input characteristics, M[i] represents the Mask matrix, and P[i−1] represents the prior knowledge of the Mask. When the Mask M[i] of a feature output is high in the current step sub-module, the weight of the feature γ is reduced by subtracting the coefficient P[i] [2] (Fig. 2).

$$M[i] = sparc\max(P[i-1] \cdot h_i(a[i-1] \tag{7}$$

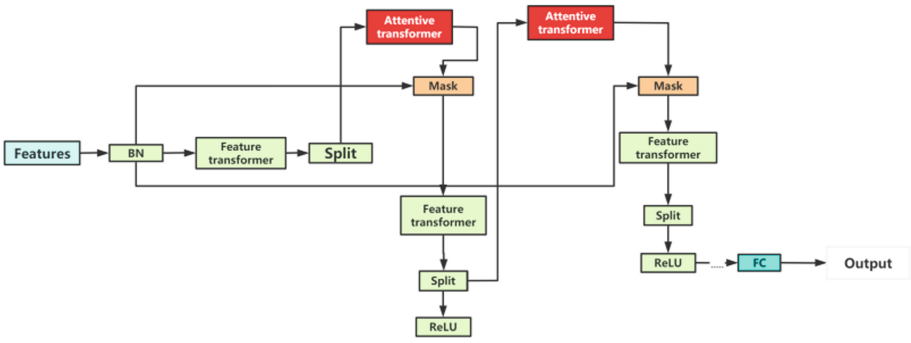

Fig. 1. Tabnet Encoder Architecture

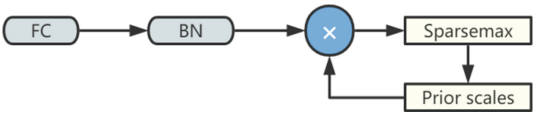

Fig. 2. Attentive transformer

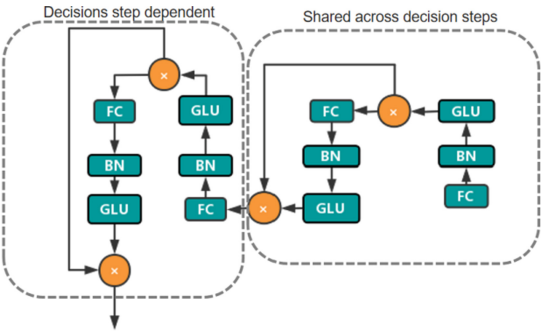

Fig. 3. Feature transformer

$$P[i] = \prod_{j=1}^{i} |\gamma - M[j]| \tag{8}$$

$$\sum_{j=1}^{D} M[i]_{b,j} = 1 \tag{9}$$

In order to control the sparsity of feature mask selection, a regular term L_{sparse} is added to the loss function, where N_{steps} is step number, B is batch size, and D is feature dimension.

$$L_{sparse} = \sum_{i=1}^{N_{seps}} \sum_{b=1}^{B} \sum_{j=1}^{D} \frac{-M_{b,j}[i]}{N_{steps} \cdot B} \log\left(M_{b,j}[i] + \xi\right) \tag{10}$$

It can be seen from the Fig. 3 that the inner product of Mask and original feature is split after feature transformer, part of which is output and part of which is input of the next step. Feature transformer includes a shared parameter layer and an independent decision layer. The parameter sharing layer shares step parameters, and the parameters of the independent decision layer are only obtained by the step training. At the same time, in order to realize the learning of large batch size, except the original BN of the first input feature, the other BNS are all ghost BN. Finally, the global embedding d_{out} is obtained, and the final output is obtained through a linear mapping $W_{final} d_{out}$.

$$[d[i], a[i]] = f_i(M[i] \cdot f) \tag{11}$$

$$d_{out} = \sum_{i=1}^{N_{steps}} \mathrm{Re}LU(d[i]) \tag{12}$$

2.4 Evaluation Metrics

We adopt one metric to evaluate the performance of fraud detection for each transfer record [7]: The root mean square error (RMSE) described in Eq. (13) is used to calculate the deviation between CNN and the proposed Method for each transfer record.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2} \tag{13}$$

where y_i and $\hat{y}_i$ are CNN's training value and the training value of the proposed method of one transfer record, respectively. n is the number of cycles from initial status to failure status of one transfer record. $Error_i = \hat{y}_i - y_i$ and $\overline{Error}$ is the average value of $Error$, while $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

3 Description of the Fraud Detection Data Sets

In this experiment, we chose a data set of monthly transfer records of mobile payment companies to predict corporate fraud. The data includes 11 indicators, including time, transfer type and transfer amount. The description of the data is shown in Table 1.

3.1 Data preparation

There are many different types of data, often with different dimensions and units, which affect the results of data analysis. The existence of incomplete sample data will increase the experimental training time and error. Therefore, the standardized and normalized data are processed before the experiment, so that the range of the processed data set is reduced, that is, restricted within a certain range. If the data has missing values, the following processing is usually required:

1) To make fillna and dropna work, fill 0 with None;
2) Delete harmful rows;
3) Fill missing values;
4) Show variable correlations (Fig. 4).

Then the data is divided into training set and test set, part of the data is selected for model training, and the data of test set is used for model testing. Specifically, we conduct fraud detection on data of transfer and cash out type.

Table 1. The experimental results

Layer	type	amount	nameorig	oldbalanceorg	newbalanceorig	is Fraud	is Flagged Fraud
0	payment	9839.64	C1231006815	170136	160296.36	0	0
1	payment	1864.28	C1666544295	21249	19384.72	0	0
2	transfer	181	C1305486145	181	0	1	0
3	Cash out	181	C840083671	181	0	1	0
4	payment	11668.14	C2048537720	41554	29883.86	0	0

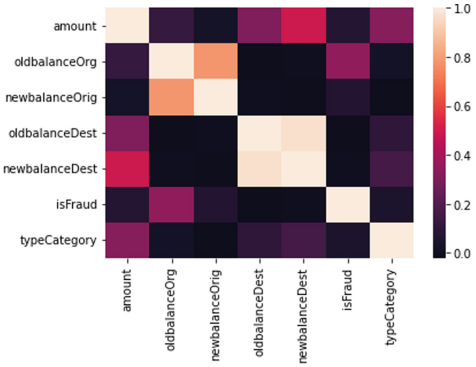

Fig. 4. Variable correlations

Table 2. The experimental results

Method	loss	accuracy
The proposed method	**0.0449**	**98.9044**
CNN	0.0458	95.2526
SVR	0.0747	85.2404

3.2 Fraud Detection

The fraud detection results of the method proposed in this paper for transfer records are shown in Table 2. The first column shows the size of the loss value, and the second column shows the accuracy of detection.

We use RMSE metric described in Eq. 13 to evaluate the detection performance of the proposed method.

3.3 Comparison and Discussion

In order to verify the effectiveness of the proposed method, we compared the proposed method with two mainstream artificial intelligence methods as CNN and SVR using the same transfer record. The results are shown in the Fig. 5. In the figures, we can clearly find that the detection error of our proposed method is smaller and the accuracy is higher.

The overall comparison results of different methods are shown in the figure. CNN and SVR generate the largest RMAE for all records, and we believe that using SVR it is difficult to find a common space for the large number of source and target data. Although the proposed method shows relatively small RMSE compared with CNN, it is not satisfactory for the universality of all transfer records. The method presented in this paper has obvious advantages in precision and versatility.

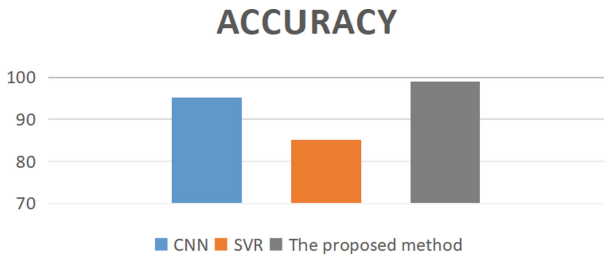

Fig. 5. The accuracy of the three methods for training the same transfer data set.

4 Conclusions

In this paper, we propose a method for detecting fraud based on deep learning, which consists of Attentive Transformer and Feature Transformer and some auxiliary operations. This method can directly process table data and automatically extract data features. To validate the effectiveness of the proposed method, we conducted experiments on fraud detection using different methods on real data sets. The results show that the detection performance of this method is significantly improved compared with several mainstream artificial intelligence methods. In the future, we want to detect fraud using more accurate model.

References

1. Al-magsoosi AAD, Mohammed GN, Ramadhan ZA (2021) Comparison and analysis of supervised machine learning algorithms. Period Eng Nat Sci 9(4):1102–1109
2. Arık SO, Pfister T (2021) TabNet: attentive interpretable tabular learning. In: AAAI, vol 35, pp 6679–6687
3. Armaghani DJ, Koopialipoor M, Bahri M, Hasanipanah M, Tahir MM (2020) A SVR-GWO technique to minimize flyrock distance resulting from blasting. Bull Eng Geol Env 79(8):4369–4385. https://doi.org/10.1007/s10064-020-01834-7
4. He HX, Zheng JC, Liao LC, Chen YJ (2021) Damage identification based on convolutional neural network and recurrence graph for beam bridge. Struct Health Monit 20(4):1392–1408
5. Hobart M (2020) The 'dark data' conundrum. Comput Fraud Secur 2020(7):13–16
6. Holmes C, King R (2019) The evolution of business-to-business FinTech: what the future holds. J Paym Strategy Syst 13(3):217–225
7. Janiesch C, Zschech P, Heinrich K (2021) Machine learning and deep learning. Electron Mark 31(3):685–695. https://doi.org/10.1007/s12525-021-00475-2
8. Jung JW, Heo HS, Kim JH, Shim HJ, Yu HJ (2019) RawNet: advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification. arXiv preprint arXiv:1904.08104
9. Tao R, Wei Y, Jiang X, Li H, Qin H, Wang J, Ma Y, Zhang L, Liu X (2021) Towards real-world X-ray security inspection: a high-quality benchmark and lateral inhibition module for prohibited items detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10923–10932

10. Yu K, Lin TR, Ma H, Li X, Li X (2021) A multi-stage semi-supervised learning approach for intelligent fault diagnosis of rolling bearing using data augmentation and metric learning. Mech Syst Signal Process 146:107043
11. Zhang S, Zhang S, Zhang C, Wang X, Shi Y (2019) Cucumber leaf disease identification with global pooling dilated convolutional neural network. Comput Electron Agric 162:422–430