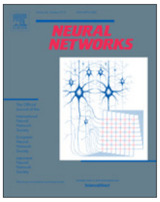
Neural Networks 102 (2018) 78–86

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Generative adversarial network based telecom fraud detection at the receiving bank

Yu-Jun Zheng[a,b,*], Xiao-Han Zhou[b], Wei-Guo Sheng[a], Yu Xue[c], Sheng-Yong Chen[b]

[a] *Institute of Service Engineering, Hangzhou Normal University, Hangzhou 311121, China*
[b] *College of Computer Science & Technology, Zhejiang University of Technology, Hangzhou 310023, China*
[c] *School of Engineering and Computer Science, Victoria University of Wellington, Wellington 6140, New Zealand*

ARTICLE INFO

Article history:
Received 29 December 2017
Received in revised form 22 February 2018
Accepted 26 February 2018
Available online 5 March 2018

Keywords:
Intelligent data analysis
Fraud detection
Generative adversarial network (GAN)
Denoising autoencoder
Deep learning

ABSTRACT

Recently telecom fraud has become a serious problem especially in developing countries such as China. At present, it can be very difficult to coordinate different agencies to prevent fraud completely. In this paper we study how to detect large transfers that are sent from victims deceived by fraudsters at the receiving bank. We propose a new generative adversarial network (GAN) based model to calculate for each large transfer a probability that it is fraudulent, such that the bank can take appropriate measures to prevent potential fraudsters to take the money if the probability exceeds a threshold. The inference model uses a deep denoising autoencoder to effectively learn the complex probabilistic relationship among the input features, and employs adversarial training that establishes a minimax game between a discriminator and a generator to accurately discriminate between positive samples and negative samples in the data distribution. We show that the model outperforms a set of well-known classification methods in experiments, and its applications in two commercial banks have reduced losses of about 10 million RMB in twelve weeks and significantly improved their business reputation.

1. Introduction

In Aug 2016, a telecom fraud case, following widespread media reports, became a hot topic in China: Xu Yuyu, an 18-year-old student who was just admitted to a national key university, received a phone call saying she would receive a scholarship, and then followed the caller's steps to use an ATM to pay 9900 RMB yuan into an account; She lost consciousness after realizing it was fraudulent, and died of a cardiac arrest. Two similar cases were reported in the same week. The cases have raised great concerns of the society on telecom fraud, which has increased dramatically in recent years and has caused the loss of tens of billions each year in China.

The high rate of telecom fraud in China is due to various reasons, including the absence of telecom supervision (e.g., the abuse of group message sending and caller ID spoofing), the lax regulation of banks (e.g., account identity theft), the lack of protection of personal information (e.g., in almost all cases the suspects can provide accurate information about the victims), and the small deterrent to criminals (e.g., low detection rate and light sentences). Therefore, full control of telecom fraud, which requires joint efforts of telecommunication providers, banks, lawmakers and law enforcers, and many other governmental and non-governmental agencies, remains a very challenging task at the present stage.

The aim of this paper is to develop an effective way for identifying transfers sent from victims to fraudsters just at the receiving bank. The basic idea is to identify for each large transfer (remittance) a probability of being sent from a victim to a criminal; if the probability exceeds a threshold, proactive measures can be taken to prevent the fraudsters to take the money.

Fraud detection is usually seen as a pattern classification problem of identifying abnormal patterns from the normality, for which classical statistical classification, data mining and machine learning methods have been widely used (Behdad, Barone, Bennamoun, & French, 2012; Bolton & Hand, 2002; El-Melegy, 2014; Ngai, Hu, Wong, Chen, & Sun, 2011; Raman, Somu, Kirthivasan, & Sriram, 2017; Song, Zheng, Xue, Sheng, & Zhao, 2017; Zheng, Ling, Xue, & Chen, 2014). In particular, artificial neural network (ANN) models, which are known for their capability of modeling highly nonlinear and complex functions from the ground up by simulating the properties of interacting neurons, have been successfully applied to various financial fraud detection problems including credit card fraud (Aleskerov, Freisleben, & Rao, 1997; Baesens, Setiono, Mues, & Vanthienen, 2003; Dorronsoro, Ginel, Sgnchez, & Cruz, 1997; Fu, Cheng, Tu, & Zhang, 2016; Ghosh, 1994; Syeda, Zhang, & Pan, 2002; Vlasselaer et al., 2015; Zakaryazad & Duman, 2016), telecom fraud (Mohamed et al., 2009; Sanver & Karahoca,

[*] Corresponding author at: Institute of Service Engineering, Hangzhou Normal University, Hangzhou 311121, China.
E-mail address: yujun.zheng@computer.org (Y.-J. Zheng).

https://doi.org/10.1016/j.neunet.2018.02.015
0893-6080/© 2018 Elsevier Ltd. All rights reserved.

2009), insurance fraud (He, Wang, Graco, & Hawkins, 1997; Viaene, Dedene, & Derrig, 2005; Xu, Wang, Zhang, & Yang, 2011), etc. There are also studies using explicit entity-relation networks to infer possible fraudulent activities (Subelj, Stefan, & Bajec, 2011; Vlasselaer, Eliassi-Rad, Akoglu, Snoeck, & Baesens, 2016). However, most data mining and machine learning methods heavily rely on vast quantities of transactional or operational data to discover abnormality. For example, credit card fraud may be detected by comparing suspicious transactions with customers' previous usage patterns mined from long-term history data (Srivastava, Kundu, Sural, & Majumdar, 2008), and tools for telecom fraud detection often utilize information such as average call duration, number of calls, and location of the caller from operational database of the telecommunication provider (Bolton & Hand, 2002). However, in the application scenario of our research such data is often unavailable because:

- For most cross-bank transfers, the receiving bank cannot access detailed information about the sending accounts.
- The receiving bank also cannot obtain call records of the recipients of transfers from the telecommunication provider.

That is, the receiving bank has to rely mainly on its own transactional data to infer whether the recipients of transfers are fraudsters, which significantly increases the difficulty of supervised learning. Moreover, given that normal transfers constitute a much larger fraction, a small imperfection in classifying them will result in a large number of false positives due to the base rate fallacy (Axelsson, 2000; Du, Vong, Pun, Wong, & Ip, 2017; Fu et al., 2016; Pérez-Ortiz, Gutiérrez, Tino, & Hervás-Martínez, 2016; Zheng, Chen, Xue, & Xue, 2017; Zheng, Sheng, Sun, & Chen, 2017). For example, assuming a bank needs to identify 10 fraudulent cases from ten thousand transfers in one day, for which a detection method with an accuracy of 90% might be regarded as highly effective, i.e., nine fraud cases of ten could be correctly identified; however, there would also be one thousand normal transfers being wrongly accused, which would be extremely costly to take corresponding measures and would be very detrimental to customer relations.

In this paper, we propose a new approach, called adversarial deep denoising autoencoder, for telecom fraud detection at the receiving bank based on generative adversarial network (GAN) (Goodfellow et al., 2014), which establishes an adversarial game between a discriminator model for distinguishing between generated and real data and a generative model for generating data to fool the discriminator. Compared with the basic adversarial autoencoder model (Makhzani, Shlens, Jaitly, Goodfellow, & Frey, 2015), our approach uses a deep denoising autoencoder (Vincent, Larochelle, Bengio, & Manzagol, 2008) to handle noisy inputs, and employs two top-level classifiers, one for discrimination and the other for classification, to enhance the learning effectiveness. The main contributions of this paper are two-fold:

- We propose a novel adversarial learning structure which achieves not only high accuracy and but also low misclassification rate for telecom fraud detection, and we believe it will also be useful for many other anomaly detection problems where the training set is limited.
- Our approach has been successfully applied to two commercial banks, significantly reducing the customer losses and improving the business reputation of the banks.

The rest of the paper is structured as follows. Section 2 describes our basic workflow for fraud control in the receiving bank. Section 3 presents the proposed adversarial learning approach for fraud detection, Section 4 presents the computational experiments, and Section 5 concludes with discussion.

2. The basic workflow for fraud control

First we introduce the basic workflow of our approach for fraud control at a receiving bank. Periodically, the bank conducts customer classification based on a set of predefined rules. For each account not belonging to a high-grade customer, whenever it receives a large transfer, we use the GAN to calculate a probability that the transfer is fraudulent, i.e., the receiving account is manipulated by a fraudster. If the probability exceeds a threshold, a delay period is set for the transfer, i.e., the receiver could not take the money by electronic means such as ATM and e-bank until the delay period is over. If the customer complains about this, the bank will suggest him to take the money from the counter. If the customer does come to the counter, the teller will ask him to fill out a questionnaire, and can block the transfer or even call the police when the answer has obvious flaws (but the teller could not prevent the customer to take the money if there is no obvious flaws, otherwise the bank would assume the risk of default).

Moreover, at the beginning of the delay period, the bank notifies the sending bank about the suspicion of fraud. The sending bank may (but is not obligated to) contact the sender to reconfirm the transfer: if the sender reconfirms, the delay period will be terminated; if the sender realizes or suspects it is a fraud, the bank will suggest him to call the police to block the transfer; if there is no response, the transfer will be accepted when the delay period is over.

Fig. 1 summarizes the basic workflow for fraud control, the efficiency of which depends primarily on the classification accuracy of the GAN.

3. An adversarial deep denoising autoencoder for fraud detection

Our basic idea is to use a deep neural network to extract latent representations that can support much more effective classification than raw input features, and employs adversarial learning to further improve the accuracy of discriminating between positive samples and negative samples in the data distribution.

We take autoencoder (Bengio, Lamblin, Popovici, & Larochelle, 2007) as the building block of our model. An autoencoder consists of an encoder that encodes an input vector $\mathbf{x}$ to a hidden (latent) representation $\mathbf{z} = f_\theta(\mathbf{x})$ and a decoder that decodes $\mathbf{z}$ to a reconstructed vector $\mathbf{x}' = g_{\theta'}(\mathbf{z})$, where f and g are affine mappings that can be sigmoid functions, and θ and θ' are vectors of weight and bias parameters of the encoder and the decoder, respectively. Autoencoder training consists in minimizing the reconstruction error:

$$\arg\min_{\theta,\theta'} \mathbb{E}_{\mathbf{x}\sim\mathcal{X}}\big[L\big(\mathbf{x}, g_{\theta'}(f_\theta(\mathbf{x}))\big)\big] \tag{1}$$

where $\mathcal{X}$ is the empirical distribution defined by the training set D, and L is the loss function. Typical choices for $L(\mathbf{x}, \mathbf{x}')$ include the squared error $\|\mathbf{x} - \mathbf{x}'\|^2$ for real-valued vectors and the negative log-likelihood $\sum_{i=1}^{|\mathbf{x}|}\big(x_i \log x'_i + (1 - x_i)\log(1 - x'_i)\big)$ for vectors of bits or bit probabilities (Bernoullis).

A denoising autoencoder (Vincent et al., 2008) is a simple variant of the basic autoencoder where the encoder accepts a noised input $\widetilde{\mathbf{x}} = (\mathbf{x}, \epsilon)$ and transforms it to the latent $\mathbf{z} = f_{\boldsymbol{\theta}}(\widetilde{\mathbf{x}})$. Denoising autoencoder training still consists in minimizing the average reconstruction error, but the key difference is that the latent $\mathbf{z}$ is a function of $\widetilde{\mathbf{x}}$ rather than $\mathbf{x}$ and thus the result of a stochastic mapping of $\mathbf{x}$:

$$\arg\min_{\theta,\theta'} \mathbb{E}_{\mathbf{x}\sim\mathcal{X}}\big[L\big(\mathbf{x}, g_{\theta'}(f_\theta(\widetilde{\mathbf{x}}))\big)\big]. \tag{2}$$

GAN (Goodfellow et al., 2014) is a pair of generator and discriminator networks, where the discriminator $D(\mathbf{x})$ computes the