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.

© 2018 Elsevier Ltd. All rights reserved.

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:

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

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 $\sum_{i=1}^{|\mathbf{x}|}\left(x_i \log x'_i + (1 - x_i)\log(1 - x'_i)\right)$ 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 $\mathbf{x}$:

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

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

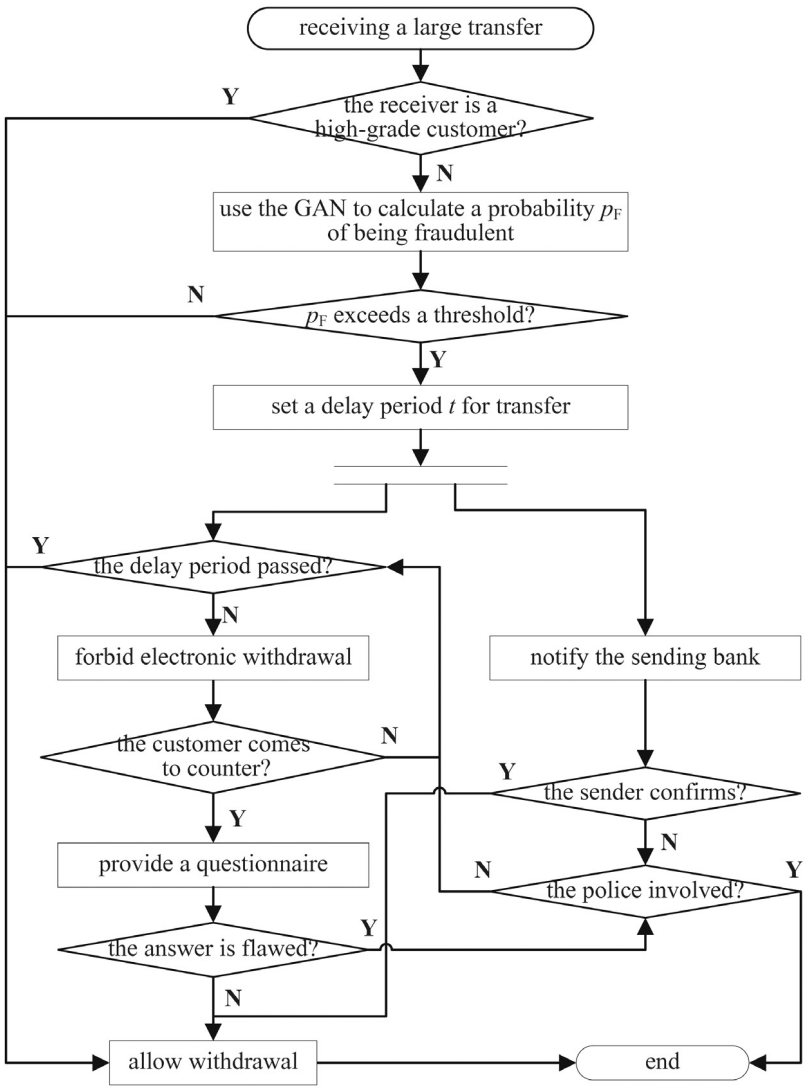

Fig. 1. The basic workflow of the receiving bank for fraud control.

probability that $\mathbf{x}$ in the data space is a sample from the data distribution (positive samples) that we are trying to model, and concurrently, the generator $G(\mathbf{z})$ generates samples from the prior $p(\mathbf{z})$ to the data space and tries to confuse the discriminator into believing that those samples come from the actual data distribution. D and G are simultaneously optimized through a two-player minimax game with the objective function:

$$\arg\min_G \left\{\arg\max_D \left[\mathbb{E}_{\mathbf{x}\sim\mathcal{X}} \log D(\mathbf{x}) + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})} \log\big(1 - D(G(\mathbf{z}))\big)\right]\right\}. \quad (3)$$

We construct a GAN based on a deep denoising autoencoder architecture, which consists of a two-hidden-layer encoder, a corresponding decoder, and two Gaussian mixture models denoted by GMM1 and GMM2, as shown in Fig. 2. A GMM calculates an output from the latent vector $\mathbf{z}$ of the denoising autoencoder as follows (Cardinaux, Sanderson, & Marcel, 2003; Gauvain & Lee, 1994):

$$\Phi(\mathbf{z}) = \frac{1}{|\mathbf{z}|} \sum_{i=1}^{|\mathbf{z}|} \log\Big(\sum_{j=1}^{N_G} w_j \mathcal{N}(z_i; \mu_j, \sigma_j)\Big) \quad (4)$$

where $\mathcal{N}(z_i; \mu_j, \sigma_j)$ is a high-dimensional Gaussian function with mean μ_j and diagonal covariance matrix σ_j, N_G is the number of Gaussians, and w_j is the weight for Gaussian j subject to $\left(\sum_{j=1}^{N_G} w_j\right) = 1$.

In our GAN, the encoder together with GMM1 acts as the discriminator D. That is, the encoder accepts an input vector $\mathbf{x}$ representing a transfer (the input features of which are summarized in Table 1) and transforms it into a latent vector $\mathbf{z}$, and GMM1 calculates from $\mathbf{z}$ a possibility $\Phi_1(\mathbf{z})$ of the transfer being a real normal transfer from the data distribution, i.e., both fakes transfer from the generator and fraudulent transfers are regarded as negative samples. In this way, the encoder is trained for discovering some important implicit features indicating a fraudulent transfer and encoding them into latent vectors to facilitate final detection.

The decoder acts as the generator G, which accepts a latent vector $\mathbf{z}$ and an additional one-hot vector encoding of the class label (i.e., normal, fraudulent, or unseen), and outputs a set of features that constitute a (fake) transfer.

Finally, the encoder together with GMM2 acts as the classifier that outputs a possibility $\Phi_2(f(\mathbf{x}))$ of the input transfer $\mathbf{x}$ being a normal transfer rather than a fraudulent one (note that the denoising autoencoder only uses input corruption for initial training).

The discriminator D and the generator G are simultaneously trained using iterative gradient descent that alternates between K steps of optimizing D and one step of optimizing G according to Eq. (3) (Goodfellow et al., 2014), as shown in Algorithm 1.

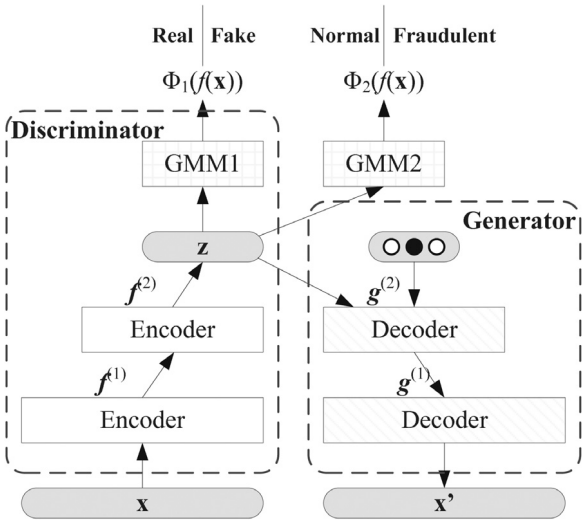

Fig. 2. The architecture of the GAN for fraud detection.

Once both D and G are trained, GMM2 is then trained with the expectation maximization algorithm (Dempster, Laird, & Rubin, 1977) to maximize the conditional log-likelihood for the training data:

$$\arg\max_{\Phi_2} \left[\mathbb{E}_{\mathbf{x}\sim\mathcal{X}^+} \log \Phi_2(f(\mathbf{x})) + \mathbb{E}_{\mathbf{x}\sim\mathcal{X}^-} \log(1-\Phi_2(f(\mathbf{x})))\right] \tag{5}$$

where $\mathcal{X}^+$ and $\mathcal{X}^-$ are the empirical distributions of positive samples (normal transfers) and negative samples (fraudulent transfers), respectively.

Algorithm 1 The algorithm for training the proposed adversarial deep denoising autoencoder.

while the stop criterion of generative adversarial learning is not satisfied **do**
  for $k = 1$ to K **do**
    Sample a minibatch Z of latent vectors;
    Generate adversarial examples X^A from the generator for Z;
    Sample a minibatch of normal transfers X^N;
    Sample a minibatch of fraudulent transfers X^F;
    Update the discriminator's weights θ_D by ascending along the gradient:

$$\nabla_{\theta_D}\left[\mathbb{E}_{\mathbf{x}\in X^F\cup X^A} \log D(\mathbf{x}) + \mathbb{E}_{\mathbf{x}\in X^N} \log\left(1-D(\mathbf{x})\right)\right]$$

  end for
  Sample a minibatch Z of latent vectors;
  Update the generator's weights θ_G by descending along the gradient:

$$\nabla_{\theta_G}\mathbb{E}_{\mathbf{z}\in Z} \log D(G(\mathbf{z}))$$

end while

4. Experiments

The data set has 330,000 large transfer samples (with amount larger than 5000 RMB yuan in this study), including 329,820 normal transfers and 180 real-world fraudulent transfers, all taking from real transactional data of two banks, Jiangxi Rural Commercial Bank and Shangrao Bank. We use a 5-fold cross-validation, that is, we partition the data set into 5 equal sized pieces and run validation for 5 times: at each time, we use four pieces to train both the generator and the discriminator, and then use the four pieces together with 6000 additional fraudulent transfers generated by the generator to train the final classifier; afterwards, we use the remaining piece to test the fraud detection ability of the whole model. Consequently, the combination of the validations is still the task of detecting 180 frauds among 330,000 transfers which is roughly in line with the actual data distribution, while the training set contains relatively enough fraudulent samples to address the curse of imbalanced training.

For our GAN-based model (denoted by GAN-DAE), the input dimension is 168 (as we can conclude from Table 1, but it can be extended when more information is available), and we empirically set the number of hidden neurons in the first and second layers to 96 and 48, respectively. For the GAN training algorithm, we set $k = 2$ and the maximum number of iterations to 1000. The threshold of GMM is set to 0.5.

To validate the performance of our model, we compare it with the following six models:

- A Bayesian belief network (BBN).
- A three-layer feed-forward artificial neural network (ANN), where the number of hidden neurons is tuned to 13.
- A Takagi–Sugeno type fuzzy inference system (FIS) which uses fuzzy IF-THEN rules to map input variables into class labels and trains the model parameters based on rank adjusted transductive similarity and incremental learning (Tencer, Reznakova, & Cheriet, 2015).
- A basic deep autoencoder (DeepAE) (Vincent, Larochelle, Lajoie, Bengio, & Manzagol, 2010), which is trained for minimizing the loss function (1) and uses a GMM as the final classifier.
- A basic adversarial autoencoder model (denoted by GAN-AE) (Makhzani et al., 2015), which takes adversarial training to minimize the objective function (3), but uses the basic autoencoder rather than the denoising autoencoder for the adversarial networks, and does not use another GMM classifier as our model.
- Another GAN-based model (denoted by GAN-AE2) that uses the similar architecture (with two GMM classifiers) of our model, but uses the basic autoencoder rather than the denoising autoencoder for the adversarial networks.

We use the following three measures (Webb & Copsey, 2011) to evaluate the fraud detection performance results of the models (where *TP*, *FP*, *TN* and *FN* refer to true positives, false positives, true negatives and false negatives, respectively):

- *Precision* that denotes what percentage of transfers identified as fraudulent are actually such:

$$precision = \frac{TP}{TP+FP} \tag{6}$$

- *Recall* that denotes what percentage of fraudulent transfers are identified as such:

$$recall = \frac{TP}{TP+FN} \tag{7}$$

- *Fallout* that denotes what percentage of normal transfers are misclassified as frauds:

$$fallout = \frac{FP}{FP+TN}. \tag{8}$$

Fig. 3 compares the experimental results of the seven models. As we can see from the results, the performance of ANN is the worst: it detects less than 60% frauds and misclassifies nearly 30% normal transfers, mainly because the calibration method of ANN is not very suitable for the considered fraud detection problem which requires a variety of complex and implicit classification rules. BBN achieves higher recall and lower fallout than ANN, showing its higher accuracy in estimating both prior probability and joint probability distributions of the related features. In comparison with

Table 1
Summary of the input features to our GAN for fraud detection.

Type	Information
Basic information of the transfer	Amount, Launch time
Basic information of the sending account	Sending bank[a], Sending region[a]
Transfer records from the sending account to the receiving bank	Frequency of transfers$^{\times 3}$, Average amount of transfers$^{\times 3}$, Frequency of large transfers$^{\times 3}$, Average amount of large transfers$^{\times 3}$
Transfer records from the sending account to the receiver	Frequency of transfers$^{\times 6}$, Average amount of transfers$^{\times 6}$, Frequency of large transfers$^{\times 6}$, Average amount of large transfers$^{\times 6}$
Basic information of the receiving account	Receiving region[a], Balance of the account, Duration of the account, Region where the account is opened, Prefix of the holder's phone number[a], Age of the holder[a], Gender of the holder[a], Birthplace of the holder[a], Number of other accounts with the holder, Total amount of the other accounts, Number of different opening regions of the other accounts, Number of different phone numbers with the other accounts, Average duration of the other accounts, Standard deviation of the durations of the other accounts
Transactional records of the receiver	Average balance$^{\times 6}$, Frequency of deposits$^{\times 6}$, Average amount of deposits$^{\times 6}$, Frequency of large deposits$^{\times 6}$, Average amount of large deposits$^{\times 6}$, Frequency of counter deposits$^{\times 6}$, Average amount of counter deposits$^{\times 6}$, Frequency of large counter deposits$^{\times 6}$, Average amount of large counter deposits$^{\times 6}$, Frequency of withdrawals$^{\times 6}$, Average amount of withdrawals$^{\times 6}$, Frequency of large withdrawals$^{\times 6}$, Average amount of large withdrawals$^{\times 6}$, Frequency of counter withdrawals$^{\times 6}$, Average amount of counter withdrawals$^{\times 6}$, Frequency of large counter withdrawals$^{\times 6}$, Average amount of large counter withdrawals$^{\times 6}$, Rate of change of withdrawal locations$^{\times 6}$, Rate of change of large withdrawal locations$^{\times 6}$

$\times 3$ For which we calculate the corresponding values over the last year, the last month, and the last week as three input components, respectively.
$\times 6$ For which we calculate the corresponding values of the current recipient account and the other accounts of the same holder over the last year, the last month, and the last week as six input components, respectively.
[a] For which we calculate the rate of the number of fraudulent transfers related with the current feature value to the total number of cases (according to data released by the police) as an input component. E.g., for "sending bank" of the transfer, the rate of the number of fraudulent transfers from the bank to the total number of cases is used as an input component.

BBN and ANN, the detection accuracy of FIS improves greatly, but its misclassification rate of 20.5% is not much better than BBN, and thus is still only very suitable for this very unbalanced classification problem.

The DeepAE model exhibits much better performance than both BBN and ANN in terms of all three measures; its detection accuracy is similar to that of FIS, but its misclassification rate is much lower, showing the advantage of the deep neural network in modeling the unknown probabilistic relationship among the input features for identifying abnormality. Nevertheless, the DeepAE's misclassification rate of 15.3% is still unacceptable. For example, the number of larger transfers received per month by a first-level branch of the investigated bank is around 150,000 (the branches at lower levels do not have permissions or abilities to access and analyze the data), and such a misclassification rate indicates that more than 750 normal transfers would be wrongly accused per day, the in-depth investigation of which could be very expensive. Moreover, since a majority of large transfer receivers are (potentially) high quality customers, delay transferring might create high complaints and push them to rivals, which would be unaffordable for the bank.

Using the same model architecture but different objective of learning, GAN-AE detects 1.66% more fraudulent transfers than DeepAE, and more importantly, reduces more than 86% of misclassification cases. As a consequence, the number of suspected transfers to be investigated per day reduces to around 100, which would be a feasible solution for most banks. This demonstrates that, in comparison with the normal loss-minimization learning, our adversarial learning approach is much more effective in training the classifier to differentiate samples *fitting* and *not fitting* the normal data distribution.

In comparison with GAN-AE using a single GMM classifier, GAN-AE2 uses two GMM classifiers, one for discriminating real transfers and artificially generated transfers, and the other for classifying normal transfers and fraudulent transfers. The results show that

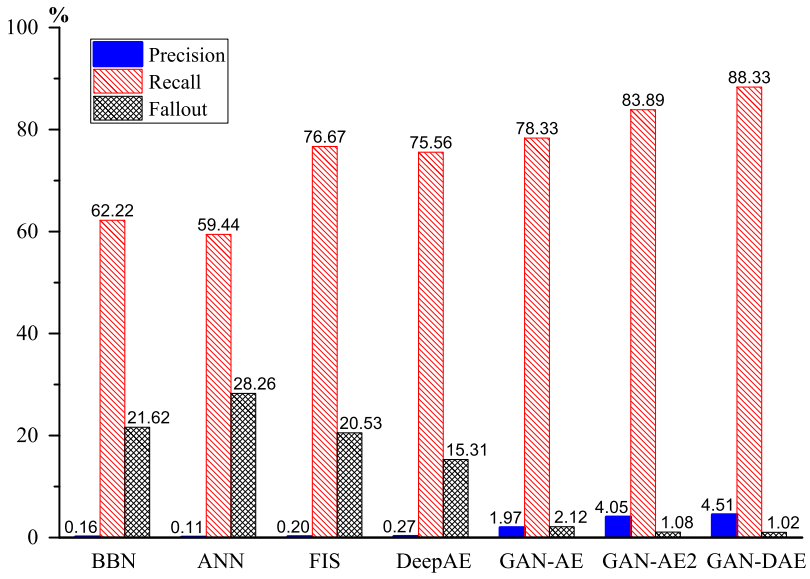

Fig. 3. The experimental results of the seven models on the test data.

GAN-AE2 detects 7% more fraudulent transfers and misclassifies 50% less normal transfers than GAN-AE. This demonstrates that the combination of loss-minimization learning and adversarial learning can lead to further classification performance improvement.

The difference between GAN-AE2 and GAN-DAE is that the latter uses denoising pre-training. The reason behind our choice is that the data set often contains incomplete and noisy information. For example, due to identity theft and loose control of the real-name system, the statistical data on the bank accounts can be inaccurate and misleading. The results show that, by incorporating the denoising mechanism, GAN-DAE detects 5% more fraudulent transfers and misclassifies 6% less normal transfers.

In summary, the experimental results show that the classification performance of our adversarial learning model is much better than the traditional Bayesian-based classifier, the shallow ANN model, the fuzzy inference model, as well as the basic deep learning model. Moreover, our architecture using the denoising autoencoder and two separated GMM classifiers can further improve the detection rate and reduce the misclassification rate.

5. Real-world applications

The proposed GAN-DAE model has been integrated with the business transaction systems of two banks and applied for fraudulent transfer detection at five first-branches during September 5 to November 27 (12 weeks), 2016. Since the third week, at each weekend we add the data tuples of confirmed normal transfers and fraudulent transfers during the week before the previous week to the data set (it typically requires about 7–15 days to confirm a fraudulent case), and re-train the model for classifying new transfers.

Table 2 presents the total number N of large transfers, the number N_I of transfers identified as fraudulent, the number N_{IF} of fraudulent transfers identified, and the number N_{NIF} of fraudulent transfers not identified in the twelve weeks. As we can see, the system successfully detects 321 of 367 true fraudulent cases, i.e., the total detection rate is more than 87%, which contributes greatly to reducing customer losses. On the other hand, the misclassification rate is about 0.68%, and the average number of cases to be investigated in a branch is about 30–50 per day, which is generally acceptable.

The misclassification rate of the GAN-DAE model in the twelve weeks are shown in Fig. 4, and the N_{IF} and N_{NIF} values are shown in Fig. 5. As we can see, by incremental learning, the misclassification rate generally decreases during the first six weeks and keeps around 0.6% during the last six weeks. Similarly, the detection rate generally increases during the first six weeks, and keeps around 99.4% during the last six weeks. Moreover, since the third week, the number of fraudulent cases decreases obviously, because the fraud detection and the subsequent active measures effectively deter the (potential) criminals. This demonstrates the effect of using our GAN based approach in fraud control in practice.

Table 2
Application results of GAN-DAE in the two commercial banks.

	N	N_I	N_{IF}	N_{NIF}
W1	176 052	1638	45	7
W2	200 870	1717	53	10
W3	221 036	1672	42	6
W4	168 757	1359	27	3
W5	195 056	1320	29	4
W6	181 335	1113	26	3
W7	191 557	1152	20	2
W8	207 364	1312	23	3
W9	185 071	1108	15	2
W10	197 059	1196	16	3
W11	200 215	1199	12	1
W12	190 448	1150	13	2
SUM	2 314 820	15 936	321	46

We also test the other six classification models used in Section 4 *on the data of first two weeks* and compare their results with our GAN-DAE model in Fig. 6. As we can see, GAN-DAE has significant performance advantage over the other models, which is similar to the experimental results described in Section 4. Moreover, since the other models have lower detection accuracy and higher misclassification rate, if they were employed in practice, much fewer samples could be confirmed and added to the data set for incremental training, and their performance would be even worse than our GAN-DAE model since the third week.

Furthermore, according to the analysis of the banks, the average cost for performing in-depth investigation on a transfer is around 220 RMB, and thus the total investigation cost in the twelve weeks is around 3.5 million. The average amount of a fraudulent transfer is around 20,000 RMB. By using our approach, the banks successfully detect and defeat 321 fraudulent cases. Moreover, in comparison with the previous period, the incidence rate decreases around

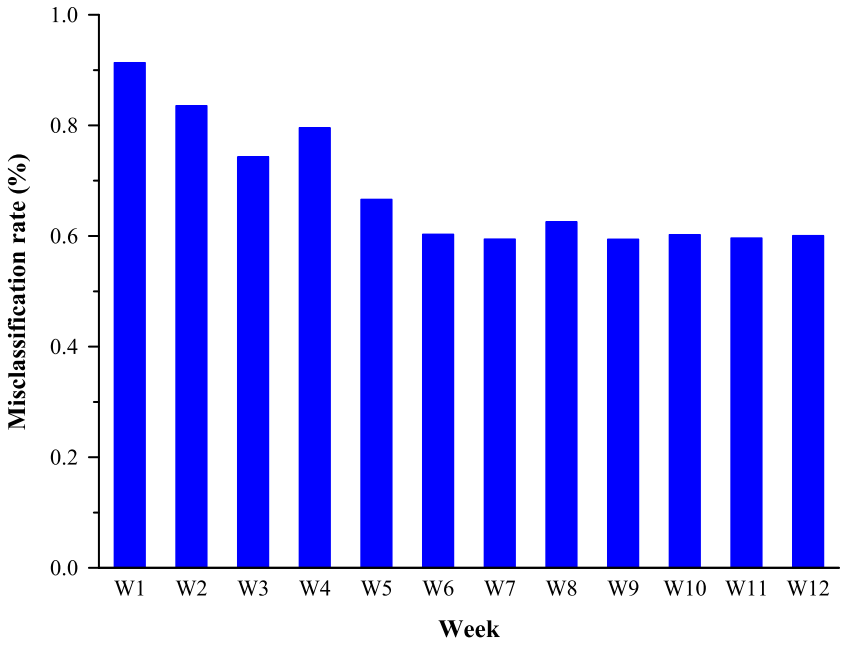

Fig. 4. The misclassification rates of the GAN-DAE model in the twelve weeks of application.

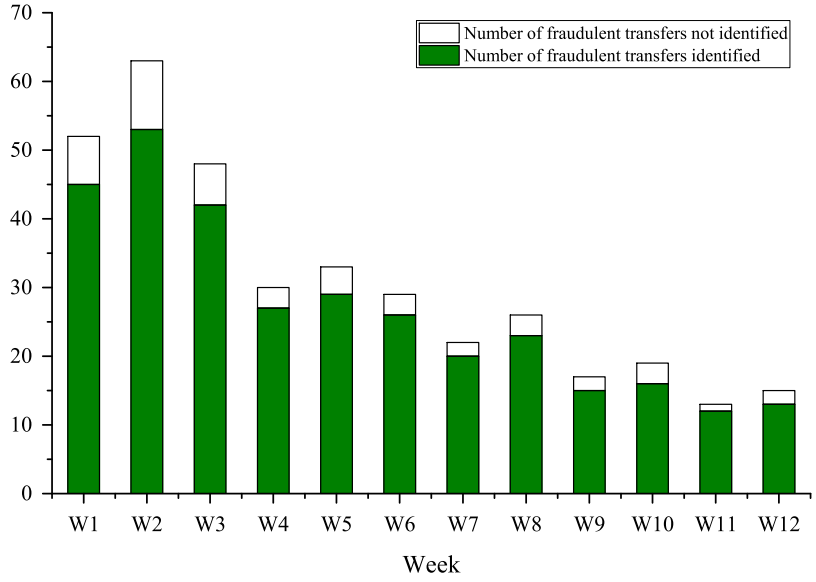

Fig. 5. The numbers of fraudulent transfers identified and not identified by the GAN-DAE model in the twelve weeks of application.

36%, i.e., there are around 180 fraudulent cases being prevented or avoided. Therefore, our approach helps the banks to reduce customer losses of nearly 10 million RMB, which is much higher than the total investigation cost. Although the banks pay for the cost and benefit the customers, they have successfully promoted their reputation which would be far beyond the cost.

6. Conclusion

Telecom fraud has become a serious problem in recent years in China. It deserves to be noted that, very recently, the People's Bank of China has published a new regulation which states that, since the end of 2016, any large transfer conducted on ATM and between two accounts of different holders will be automatically locked for 24 h. We think that using such an indiscriminate measure is not a good idea, and a more effective way is to infer for each large transfer a probability that it is fraudulent, and take appropriate measures based on the probability.

This paper proposes such an inference approach for the receiving bank to detect transfers that are induced by fraudsters. The approach is based on adversarial machine learning of a generative model and a discriminator model, which can effectively differentiate samples fitting and not fitting the normal data distribution. The model performance is further improved by the combination with loss-minimization learning and the incorporation of denoising mechanism. Experimental results show that our model has considerable performance advantages, in terms of both classification accuracy and misclassification rate, over other well-known classification models. The proposed approach has also been applied to two commercial banks and successfully detected and defeated 321 fraudulent cases, and thus effectively reduced customer losses and improved the reputation of the banks. We are currently

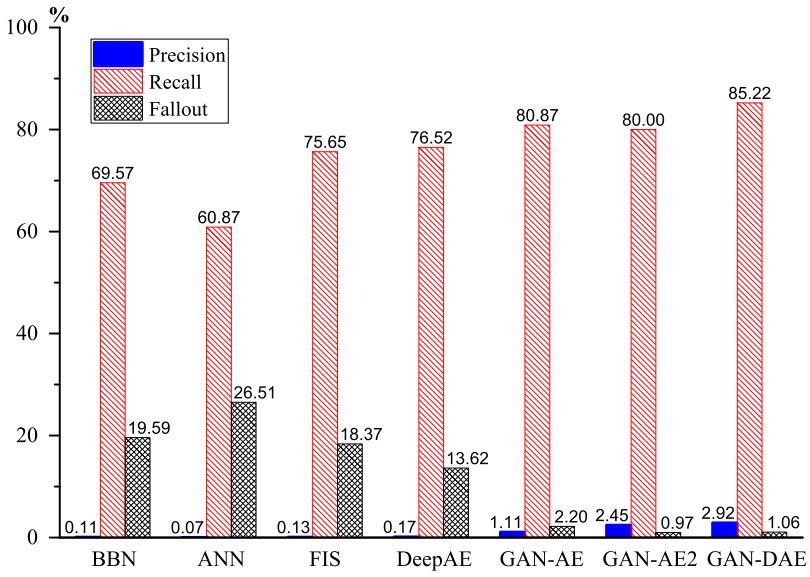

Fig. 6. The results of the seven models on the application data of the first two weeks.

applying and adapting the proposed model to some similar problems such as terrorist identification, and we believe that the emerging adversarial learning provides a new powerful method for complex classification problems. Another ongoing work is to study the use of evolutionary algorithms to improve the training efficiency for such complex deep networks with a large number of parameters (Bengio, 2014; David & Greental, 2014; Zheng, Ling, Chen, & Xue, 2015).

Funding

This work was supported by National Natural Science Foundation of China under Grant Nos. 61325019, 61473263, and U1509207.

References

Aleskerov, E., Freisleben, B., & Rao, B. (1997). CARDWATCH: A neural network based database mining system for credit card fraud detection. In *Proc. IEEE/IAFE 1997* (pp. 220–226).

Axelsson, S. (2000). The base-rate fallacy and the difficulty of intrusion detection. *ACM Transactions on Information Systems Security*, *3*(3), 186–205.

Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science*, *49*(3), 312–329.

Behdad, M., Barone, L., Bennamoun, M., & French, T. (2012). Nature-inspired techniques in the context of fraud detection. *IEEE Transactions on Systems, Man, and Cybernetics Part C: Applications and Reviews*, *42*(6), 1273–1290.

Bengio, Y. (2014). Deep learning and cultural evolution. In *Proceedings of the companion publication of the 2014 annual conference on genetic and evolutionary computation* (pp. 1–2). New York, NY, USA: ACM.

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. In J. P. Bernhard Schölkopf, & T. Hoffman (Eds.), *Advances in neural information processing systems, Vol. 19* (pp. 153–160). MIT Press.

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistics Science*, *17*(3), 235–255.

Cardinaux, F., Sanderson, C., & Marcel, S. (2003). Comparison of MLP and GMM classifiers for face verification on XM2VTS. In J. Kittler, & M. S. Nixon (Eds.), *Audio- and video-based biometric person authentication* (pp. 911–920). Springer Berlin Heidelberg.

David, O. E., & Greental, I. (2014). Genetic algorithms for evolving deep neural networks. In *Proc. GECCO* (pp. 1451–1452). Vancouver, Canada: ACM.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B. Statistical Methodology*, *39*(1), 1–38.

Dorronsoro, J. R., Ginel, F., Sgnchez, C., & Cruz, C. S. (1997). Neural fraud detection in credit card operations. *IEEE Transactions on Neural Networks*, *8*(4), 827–834.

Du, J., Vong, C.-M., Pun, C.-M., Wong, P.-K., & Ip, W.-F. (2017). Post-boosting of classification boundary for imbalanced data using geometric mean. *Neural Networks*, *96*(Supplement C), 101–114.

El-Melegy, M. T. (2014). Model-wise and point-wise random sample consensus for robust regression and outlier detection. *Neural Networks*, *59*(Supplement C), 23–35.

Fu, K., Cheng, D., Tu, Y., & Zhang, L. (2016). Credit card fraud detection using convolutional neural networks. In A. Hirose, S. Ozawa, K. Doya, K. Ikeda, M. Lee, & D. Liu (Eds.), *Neural information processing, Vol. Part III* (pp. 483–490). Cham: Springer International Publishing.

Gauvain, J. L., & Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, *2*(2), 291–298.

Ghosh, Reilly (1994). Credit card fraud detection with a neural-network. In *27th Hawaii int'l conf. syst. sci., Vol. 3* (pp. 621–630).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, Vol. 27* (pp. 2672–2680). Curran Associates, Inc..

He, H., Wang, J., Graco, W., & Hawkins, S. (1997). Application of neural networks to detection of medical fraud. *Expert Systems with Applications*, *13*(4), 329–336.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2015). Adversarial autoencoders. Co mputing Research Repository (CoRR), abs/1511.05644.

Mohamed, A., Bandi, A. F. M., Tamrin, A. R., Jaafar, M. D., Hasan, S., & Jusof, F. (2009). Telecommunication fraud prediction using backpropagation neural network. In *Proc. SOCPAR*, Dec (pp. 259–265).

Ngai, E., Hu, Y., Wong, Y., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, *50*(3), 559–569.

Pérez-Ortiz, M., Gutiérrez, P. A., Tino, P., & Hervás-Martínez, C. (2016). Oversampling the minority class in the feature space. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(9), 1947–1961.

Raman, M. G., Somu, N., Kirthivasan, K., & Sriram, V. S. (2017). A hypergraph and arithmetic residue-based probabilistic neural network for classification in intrusion detection systems. *Neural Networks*, *92*(Supplement C), 89–97.

Sanver, M., & Karahoca, A. (2009). Fraud detection using an adaptive neuro-fuzzy inference system in mobile telecommunication networks. *Multiple-Valued Logic and Soft Computing*, *15*(2–3), 155–179.

Song, Q., Zheng, Y.-J., Xue, Y., Sheng, W.-G., & Zhao, M.-R. (2017). An evolutionary deep neural network for predicting morbidity of gastrointestinal infections by food contamination. *Neurocomputing*, *226*, 16–22.

Srivastava, A., Kundu, A., Sural, S., & Majumdar, A. (2008). Credit card fraud detection using hidden markov model. *IEEE Transactions on Dependable Secure Computing*, *5*(1), 37–48.

Šubelj, L., Furlan, Štefan, & Bajec, M. (2011). An expert system for detecting automobile insurance fraud using social network analysis. *Expert Systems with Applications*, *38*(1), 1039–1052.

Syeda, M., Zhang, Y.-Q., & Pan, Y. (2002). Parallel granular neural networks for fast credit card fraud detection. In *Proc. FUZZ-IEEE'02. Vol. 1* (pp. 572–577).

Tencer, L., Reznakova, M., & Cheriet, M. (2015). TITS-FM: Transductive incremental takagi-sugeno fuzzy models. *Applied Soft Computing*, *26*, 531–544.

Viaene, S., Dedene, G., & Derrig, R. (2005). Auto claim fraud detection using bayesian learning neural networks. *Expert Systems with Applications*, *29*(3), 653–666.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proc. 25th int'l conf. machine learning* (pp. 1096–1103). New York, NY, USA: ACM.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, *11*, 3371–3408.

Vlasselaer, V. V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., et al. (2015). APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems*, *75*, 38–48.

Vlasselaer, V. V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2016). GOTCHA! network-based fraud detection for social security fraud. *Management Science*, *63*(9), 3090–3110.

Webb, A., & Copsey, K. (2011). *Statistical pattern recognition* (3rd ed.). New York: John Wiley & Sons.

Xu, W., Wang, S., Zhang, D., & Yang, B. (2011). Random rough subspace based neural network ensemble for insurance fraud detection. In *4th Int'l joint conf. computational sciences and optimization* (pp. 1276–1280).

Zakaryazad, A., & Duman, E. (2016). A profit-driven artificial neural network (ANN) with applications to fraud detection and direct marketing. *Neurocomputing*, *175*(Part A), 121–131.

Zheng, Y.-J., Chen, S.-Y., Xue, Y., & Xue, J.-Y. (2017). A Pythagorean-type fuzzy deep denoising autoencoder for industrial accident early warning. *IEEE Transactions on Fuzzy Systems*, *25*(6), 1561–1575.

Zheng, Y.-J., Ling, H.-F., Chen, S.-Y., & Xue, J.-Y. (2015). A hybrid neuro-fuzzy network based on differential biogeography-based optimization for online population classification in earthquakes. *IEEE Transactions on Fuzzy Systems*, *23*(4), 1070–1083.

Zheng, Y., Ling, H., Xue, J., & Chen, S. (2014). Population classification in fire evacuation: A multiobjective particle swarm optimization approach. *IEEE Transactions on Evolutionary Computation*, *18*(1), 70–81.

Zheng, Y. J., Sheng, W. G., Sun, X. M., & Chen, S. Y. (2017). Airline passenger profiling based on fuzzy deep machine learning. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(12), 2911–2923.