doi:10.3772/j.issn.1006-6748.2020.04.003

Fraudulent phone call recognition method based on convolutional neural network①

Xing Jian (邢 剑)[*,**,***], Wang Shupeng[*,**], Ding Yu②[*,**]
([*] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, P. R. China)
([**] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100093, P. R. China)
([***] Xinjiang Branch of National Computer Network Emergency Response Technical Team/Coordination Center of China, Urumqi 830000, P. R. China)

Abstract

With increasingly rampant telephone fraud activities, the social impact and economic losses caused to China have increased dramatically. Precise, convenient, and efficient fraudulent phone call recognition has become a challenge since telephone fraud became more varied and covert. To deal with this problem, many researchers have extracted some statistical features of telephone fraud behavior and proposed some machine learning algorithms on the field of fraudulent phone call recognition. In this paper, the calling detail records are utilized to construct a classifier for fraudulent phone call recognition. Meantime, a deep learning approach based on convolutional neural network (CNN) is proposed for better features learning and compared with the existing state-of-the-art machine learning algorithms. It learns phone number and call behavior features of telephone fraud, and improves the accuracy of classification. The evaluation results show that the proposed algorithm outperforms competitive algorithms.

Key words: fraudulent phone call recognition, convolutional neural network (CNN), calling detail records (CDR), deep learning (DL), telephone fraud

0 Introduction

The recognition of fraudulent phone call is an important task to guard against and combat telephone fraud. The traditional crowdsourcing methods of labeling fraudulent phone number have achieved good recognition results. In recent years, with the continuous transfer of telephone fraud to overseas countries and the widespread use of VoIP and phone number modification software, fraudulent phone numbers are constantly changing and becoming more covert[1]. The traditional methods based on blacklist are no longer effective as a result of these changes. With the behavior statistical features of fraudulent phone call, many machine learning algorithms are proposed on the field of fraudulent phone call recognition, such as random forest (RF)[2], support vector machine (SVM)[3], and so on. However, the accuracy of the algorithms is not high. Therefore, the precise recognition of fraudulent phone call has become a challenge.

Based on the simulation of a hierarchical structure existing in human brain, deep learning can establish the mapping between the low-level signals and the high-level semantics for achieving the hierarchical expression of data characteristic[4]. It has been widely used for image recognition and classification. In this paper, a new method based on convolutional neural network (CNN) is proposed for fraudulent phone call recognition. It has the ability of learning phone number and call behavior features of telephone fraud automatically and outperforms the state-of-the-art approaches. The phone number and call behavior features of telephone fraud are usually one-dimension vector. So, 1D-CNN (one-dimensional convolutional neural network) is used to process feature vectors to obtain abstract features.

The key contributions of this work are summarized as follows:

This paper designs and constructs a classifier that combines non-statistical features with statistical features for fraudulent phone call recognition. The classifier only utilizes calling detail records (CDR) in data processing, classification training and evaluation process-

① Supported by the National Natural Science Foundation of China (No. 61931019).
② To whom correspondence should be addressed. E-mail: dingyu@iie.ac.cn
Received on Sep. 16, 2019

ing, so it can be constructed easily, conveniently, and efficiently.

This study provides the first preliminary exploration of state-of-the-art deep learning (DL) algorithm applied to fraudulent phone call recognition, namely CNN. This paper designs, tunes and evaluates the model which is capable of automatically learning phone number and call behavior features of telephone fraud. It demonstrates that CNN-based fraudulent phone call recognition method achieves high accuracy, more than the state-of-the-art classification methods.

The rest of this paper is structured as follows. Section 1 describes the related work. Section 2 presents the proposed method in details. Section 3 displays the experimental result. Finally, Section 4 concludes with discussion.

1 Related work

Scam call activity regularity and behavior features analysis report 2016[5] released by 360 Internet Security Center shows that the survival period of fraudulen phone number is about 57.6 days, the continuous active period is 7.6 days, the average number of calls for a single number in a single day is 185, the average number of calls required for a successful fraud is 1 000, and the average time to complete a successful fraud is 5.4 days.

This indicates that there are some differences between fraudulent phone call and normal phone call in call frequency, call time, long-distance call rate and other behavior features [6]. At the same time, although fraudulent phone number has randomness and variability, the phone number itself also has certain regularity[7], such as non-standard number, international number, short number or fake number.

Previous studies have shown that fraudulent phone call can be effectively recognized through cognitive learning of the above phone number and call behavior features. Zhou et al. [6] made a statistical analysis of the call behavior of users and found that the call time frequency, call time interval, call frequency of the same object, call cycle and call interval had obvious regularity. However, due to the limited number of samples, it failed to extract the behavior features of fraudulent phone call. Wang et al. [8] proposed a recognition method of nuisance calls based on the random forest. It preliminarily found that phone numbers had features that could be used to identify them. However, the accuracy of the algorithm was only 84.30%. Ji et al. [3] proposed a recognition method of fraudulent phone call based on SVM. It only constructed a classifier for the call behavior feature of fraudulent phone call, but did not analyze the phone number features of fraudulent phone call and the accuracy of the algorithm was only 76%. Other researchers[9-11] chose to use the decision tree and naive Bayesian models to classify and analyze call behavior features. This paper constructs a classifier for fraudulent phone call recognition, explores one deep learning method and finally achieves a higher classification accuracy than the state-of-the-art machine learning algorithms.

2 Proposed method

2.1 The constructed classifier

In this paper, an easy, convenient, and efficient classifier is designed. Fig. 1 shows the overview of constructed classifier. It consists of feature extraction & data preprocessing phase and training & evaluation phase.

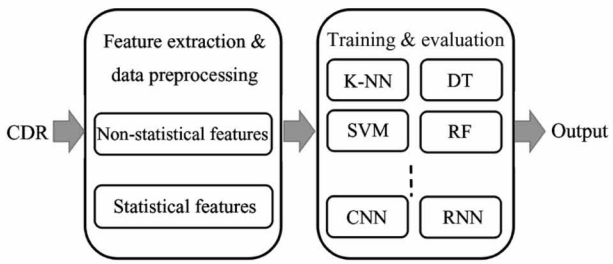

Fig. 1 Overview of the classifier

In the first phase, it extracts non-statistical features and statistical features from CDR, and then preprocesses the above data for next stage. In the second phase, it utilizes special algorithms to train the model and complete the evaluation task.

(1) Feature extraction and data preprocessing

Seven features are extracted from 6 fields of CDR, which result in 176 dimensions. The 6 fields are START _ TIME, END _ TIME, CALLING _ NUMBER, CALLED _ NUMBER, CALL _ DURATION, and CALLED _ LOCATION.

Non-statistical feature: CALLING _ NUMBER is extracted from all records as non-statistical feature. Meanwhile, duplicate CALLING _ NUMBER in one day is removed.

The main operation of data preprocessing is to complete the length of the CALLING _ NUMBER to 17 digits with zero and use One-Hot Encoding for digital conversion. Finally, a length-170 array is constructed, which represents a non-statistical feature, namely, CALLING _ NUMBER.

Statistical feature: Based on the above CALLING

_ NUMBER, six features are extracted from all records as statistical features. They are the number of CALLED _ NUMBER, the number of CALLED _ NUMBER (deduplication), the maximum similarity of CALLED _ NUMBER, the average similarity of CALLED _ NUMBER, the average CALL _ DURATION, and the number of CALLED _ LOCATION. The statistical period is one day.

The main operation of data preprocessing is to Min-Max Normalization of all statistical features except the maximum similarity of the CALLED _ NUMBER and the average similarity of the CALLED _ NUMBER. The Min-Max Normalization is calculated according to Eq. (1). Finally, all statistical features are converted in the range of [0,1], and a length-6 array is constructed, which represents the 6 statistical features.

$$x' = (x - X\_min) / (X\_max - X\_min) \quad (1)$$

(2) Training and evaluation

Two types of features are concatenated together and used as input to this layer. They jointly optimize the same category target.

Multiple supervised machine learning algorithms, such as k-nearest neighbors (KNN), decision tree (DT), RF, SVM and some DL algorithms, are utilized to train the model and evaluate whether the phone call is fraudulent.

2.2 The proposed algorithm

Convolutional neural network is extension of traditional multi-layer perception, based on local receive fields, shared weights and spatial or temporal sub-sampling. A CNN consists of an input layer and an output layer, as well as multiple hidden layers. The hidden layers include convolutional layers, pooling layers, fully connected layers and normalized layers. Convolutional layers apply a convolution operation to the input, passing the result of the next layer. Pooling layers combine the outputs of neural clusters at one layer into a single neuron in the next layer. Fully connected layers connect every neuron in one layer to every neuron in another layer. The above 7 feature vectors can be reshaped into one-dimensional structure. They are reshaped into 1×176. A 1D-CNN is used to process these feature vectors.

The architecture of network is imported from Alexnet, and summarized in Fig. 2. The net contains 8 layers. The first 2 layers are convolutional, the third layer is pooling, the fourth and fifth layers are convolutional, the sixth layer is pooling, and the remaining 2 are fully-connected. The output of the last fully-connected layer is fed to a 2-way softmax which produces a distribution over the 2 class labels.

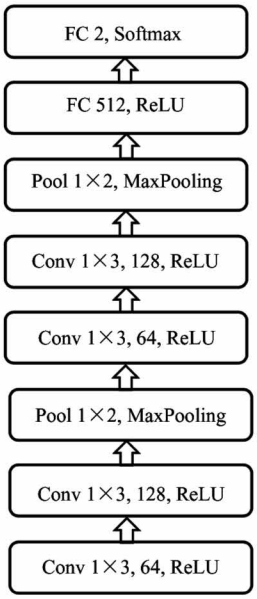

Fig. 2 The structure of convolutional neural network model

The ReLU (rectified linear) is applied to the output of every convolutional and fully-connected layer. The first convolutional layer filters the 1×176 input data with 64 kernels of size 1×3 with a stride of 1 step. The second convolutional layer takes the output of the first convolutional as input and filters it with 128 kernels of size 1×3. The maximum pooling function is used in the third layers, and its pool size is 1×2. The fourth to sixth layers are the same as the first 3 layers. The first fully-connected layer has 512 neurons. The cross entropy is selected as the loss function. Dropout is deployed to reduce the over-fitting. After the third and sixth layers, 25% of the neuron information is discarded. After the first fully-connected layer, 50% of the neuron information is discarded. All neurons are used in the test.

3 Experiments

3.1 Datasets

The data collected includes all CDR (from BICC/ISUP and SIP signaling) for 6 months from September 2018 to February 2019. There are more than 6 million normal phone calls and 8 284 fraudulent phone calls. In real-world environments, the proportion of fraudulent phone call samples to normal phone call samples is very small.

The experiment is conducted in 4 datasets, which are summarized in Table 1. All samples are randomly divided into 2 parts: training set and test set. The training set consists of 5 000 normal phone call samples and 5 000 fraudulent phone call samples. The test set consists of 3 000 normal phone call samples and 3 000

fraudulent phone call samples. The proportion of normal phone call samples to fraudulent phone call samples is 1:1. In the remainder of the text, this dataset is referred to as SC_1. Similarly, for datasets that the proportion of normal phone call samples to fraudulent phone call samples is 10:1, 100:1, and 200:1 are referred to as SC_{10}, SC_{100}, and SC_{200} accordingly.

Table 1 The datasets overview

Dataset		Number of normal phone call	Number of fraudulent phone call
SC_1	Training set	5 000	5 000
	Test set	3 000	3 000
SC_{10}	Training set	50 000	5 000
	Test set	3 000	3 000
SC_{100}	Training set	500 000	5 000
	Test set	3 000	3 000
SC_{200}	Training set	1 000 000	5 000
	Test set	3 000	3 000

3.2 Experimental setting

Several classic and popularly used machine learning algorithms are used for comparison including KNN, SVM (linear kernel) and SVM (RBF kernel).

In KNN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

Support vector machine is a kind of generalized linear classifier which classifies data by supervised learning. Its decision boundary is the maximum-margin hyperplane for solving learning samples. The selection of SVM kernel plays a vital role in its performance.

The implementation of the CNN classifier uses Keras[12] with Tensorflow[13] back-end. The experimental environment is a server with an Intel i9-9900k, 64 GB DDR4 memory and one Nvidia RTX2080Ti GPU.

3.3 Evaluation

Accuracy is the criterion of evaluation. The accuracy rate is the proportion of all the correct sample sizes to the training data during the iterative training, which is calculated according to Eq. (2). To ensure the reliability of experiments, the models' performance is estimated by conducting a 10-fold cross-validation on each dataset.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Table 2 shows the accuracy results of 4 algorithms under 4 datasets. CNN has the highest accuracy and its effect remains stable. It is 1.6%, 3.7%, 8%, 2.43% higher than the second best algorithm using 4 datasets, respectively. It achieves the highest accuracy of 98.67% in SC_{10}.

Table 2 The accuracy of classification under different algorithms and different datasets

Algorithm	SC_1	SC_{10}	SC_{100}	SC_{200}
KNN	95.30%	94.97%	89.48%	87.00%
SVM (linear kernel)	91.08%	84.28%	50.23%	50.03%
SVM (RBF kernel)	93.95%	91.10%	84.40%	78.48%
CNN	96.90%	98.67%	97.48%	89.43%

The results are depicted in Fig. 3 for 4 algorithms. All algorithms achieve high accuracy in the first 2 datasets. With the change of sample equilibrium, namely, like real-world environments, the number of normal phone call sample in the training set is far more than fraudulent phone call sample, KNN, SVM (RBF kernel), and CNN are getting less accurate but still effective in the last 2 datasets. However, the accuracy of SVM (linear kernel) has decreased dramatically to about 50%. For binary classification, this means that the algorithm fails. One possible reason for the performance drop is that the classifier trained and evaluated in small data size might learn the partial or error features instead.

The main conclusion here is that the CNN-based classifier is capable of extracting accurate identifying

information from the phone number and call behavior features of telephone fraud. It works very well and outperforms other competing methods.

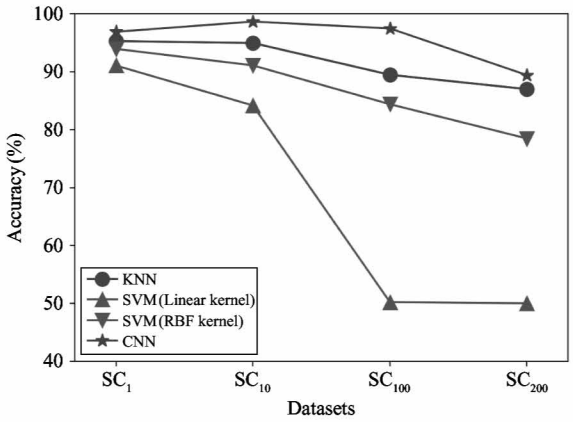

Fig. 3 The accuracy of 4 models for 4 datasets

4 Conclusions

Fraudulent phone call recognition represents an essential task for both preventing and curbing fraud effectively. In this study, a new method is proposed for fraudulent phone call recognition which is evaluated on the real-world datasets. The experimental results show that the proposed novel method has the ability of learning phone number and call behavior features of telephone fraud automatically and outperforms other competing methods. The obtained success rate exceeds 98% in the datasets evaluation. The method has 3.9% more classification accuracy than the state-of-the-art method on average. In conclusion, the application of deep learning algorithm makes fraudulent phone call recognition accurate and effective.

References

[1] 360 Internet Security Center. Telecomm fraud activity pattern and behavioral characteristics report[EB/OL]. http://zt.360.cn/1101061855.php? dtid=1101062366&did=490106344: 360 Safety Technology Co. Ltd., 2016

[2] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Bogota, Columbia, 2014:1725-1732

[3] Ji Z H, Ma Y C, Li S, et al. SVM based telecom fraud behavior identification method[J]. *Computer Engineering & Software*, 2017, 12 (38):104-109

[4] Huang T D, Yu C M, Kao H Y. Data-driven and deep learning methodology for deceptive advertising and phone scams detection[C]// Technologies and Applications of Artificial Intelligence, Taibei, China, 2017:166-171

[5] 360 Internet Security Center. China telecom fraud situation analysis report [EB/OL]. http://zt.360.cn/1101061855.php? did=490024605&dtid=1101061451: 360 Safety Technology Co. Ltd., 2016

[6] Zhou G M, Chen G X, Zhou Y Z. User behavior in telecommunication fraud based on CDR analysis[J]. *Information Security and Communications Privacy*, 2015, 11: 114-118

[7] Li L K, Ma Z X, Chen Q N, et al. Research of technology solutions and operation countermeasures to telephone fraud prevention and control[J]. *Telecom Science*, 2014, 11:166-171

[8] Wang Y Q, Wang H C. Research on a combining algorithm for harassing calls to identify[J]. *Telecom Science*, 2017, 7:71-78

[9] Xu T. The Design and Implementation of Visualization Character Relationship Analysis System Based on Mining of Call Records[D]. Harbin: Software Institute, Harbin Institute of Technology, 2014:1-80 (In Chinese)

[10] Tseng V S, Ying J C, Huang C W, et al. FrauDetector: a graph-mining-based framework for fraudulent phone call detection[C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 2015: 2157-2166

[11] Ying J C, Zhang J, Huang C W, et al. PFrauDetector: a parallelized graph mining approach for efficient fraudulent phone call detection[C]// International Conference on Parallel and Distributed Systems, Wuhan, China, 2016: 1059-1066

[12] Chollet F. Keras [EB/OL]. https://github.com/fchollet/keras: Keras, 2015

[13] Abadiet M, A. Agarwal, P. Barham, et al. TensorFlow: Large-scale machine learning on heterogeneous systems [EB/OL]. https://www.tensorflow.org/: TensorFlow.org, 2015

Xing Jian, born in 1983. He is pursuing his Ph.D degree in Institute of Information Engineering, Chinese Academy of Sciences. He received his B.S. and M.S. degrees from Xinjiang University in 2004 and 2008 respectively. His research interests include big data management and analysis, network security.