A Network Traffic anomaly Detection method based on CNN and XGBoost

Dan Niu
Key Laboratory of Measurement and Control of CSE, Ministry of Education School of Automation, Southeast University
Nanjing, China
danniu1@163.com

Jin Zhang
School of Automation, Southeast University
Nanjing, China

Li Wang
Elex Cybersecurity INC
Nanjing, China

Kaihong Yan
Elex Cybersecurity INC
Nanjing, China

Tao Fu
Elex Cybersecurity INC
Nanjing, China

Xisong Chen
School of Automation Southeast University
Nanjing, China

2020 Chinese Automation Congress (CAC) | 978-1-7281-7687-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/CAC51589.2020.9327030

***Abstract*—With the rapid development of information network, network security is becoming more and more important. Intrusion detection is an important component of the network security system. The traditional signature-based matching detection method is difficult to cope with the increasingly complex network environment. On the contrary, anomaly detection which is based on network traffic pattern analysis has obvious advantages in dealing with encryption attacks, zero-day attacks and other new attacks. This paper studies the network traffic anomaly detection, and proposes a traffic anomaly detection model which combines convolution neural network and eXtreme Gradient Boosting algorithm. First, the collected traffic data is preprocessed into appropriate format that meets the input requirements of the model. Then the improved LeNet-5 convolution neural network is used for feature learning, and finally XGBoost algorithm is used to classify the learning features. Experimental results show that the proposed network traffic anomaly detection method based on CNN and XGBoost has a high accuracy, and good experimental results have been achieved in both two-classification and multi-classification.**

Keywords—Network anomaly detection, convolution neural network, eXtreme Gradient Boosting.

I. INTRODUCTION

With the rapid development of network technology, the 5G era is coming. The level of network communication technology has become a precursor of the development level of a country. The network makes people's life more convenient and fast, but at the same time, the security risk is also huge. The frequent occurrence of network security events has brought huge losses to the social economy, so the realization of network security is quite important.

Cyberspace security system can be divided into network security system and computer host security system, including firewall, anti-virus software and intrusion detection system. The network anomaly detection studied in this paper is an important way to realize intrusion detection. The traditional intrusion detection method is misuse detection, which is based on the known attack signature library, and determining whether the intrusion is based on the matching degree between the object to be detected and the library. Although the effect of anomaly detection is not as good as misuse detection in the detection of known attack types, but with the increasing complexity of network structure and environment, the unique advantage of anomaly detection is highlighted. It does not[1] need to analyze the payload of data packet and it analyzes the behavior pattern of traffic data.

In related research, network traffic anomaly detection methods generally include four kinds of methods, based on classification (like SVM, Bayesian network, neural network), based on clustering, based on statistics and based on information theory [1]. The method studied in this paper is the network anomaly detection based on classification method. General network traffic classification methods include port-based identification, deep packet detection, statistics-based and behavior-based methods [2]. The first two methods are static matching rules, while the traffic classification methods based on statistics and behavior belong to the category of machine learning methods.

Wang Wei [3] combines the two classical network structures of convolutional neural network and recurrent neural network to learn the spatial and sequential series characteristics of traffic data respectively, and realizes an end-to-end anomaly detection process. Feature learning is completed by CNN and RNN to achieve accurate detection of abnormal traffic, and the detection accuracy and false alarm rate on DAPRA1998 and ISCX2012 data sets are greatly improved compared with the general method. In [4], an intrusion detection method is introduced based on deep autoencoder (AE), which uses greedy layer-wise pre-training to avoid over-fitting, and has high accuracy in both binary and multi-classification of KddCup99 dataset. Al-Qatf [5] proposed an effective deep learning method STL-IDS for feature learning and dimensionality reduction. This model adopts sparse automatic encoder mechanism and is an effective feature learning algorithm for unsupervised reconstruction. After the pre-training stage, the new features are input into the support vector machine algorithm to improve its intrusion detection ability and classification accuracy. The results show that this method accelerates the training and testing time of SVM, and performs better than most previous methods in the performance measurement of second-class and multi-class classification. In [6], a new type of unsupervised

[1] *This work was supported by National key R&D Program of China (No. 2018YFC1506900), Zhishan Youth Scholar Program of SEU, National NSF of China (No. 61504027), the Key R&D Program of Jiangsu Province (No. BE2017076, BE2019052), the Key R&D industrialization Program of Suzhou (No. SGC201733, SGC201854).

asymmetric stacked autoencoder is put forwarded for feature learning, and then random forest algorithm is used to classify the acquired sample features. The author performed experiments on KDDCup99 and NSL-KDD data sets respectively. The experimental results show that the accuracy and recall rate of the model are very good, and the detection time is greatly shortened. Compared with the DBN (deep belive network) method, the accuracy is improved by 5%. The training time has been shortened by as much as 98.81%. Ren-Hung Hwang [7] proposed an effective anomaly traffic detection mechanism DmurPack, which consists of a convolutional neural network and an unsupervised autoencoder, which is used to automatically analyze traffic patterns and filter abnormal traffic. An important innovation is that D-PACK only examines the first l bytes of the first n packets in each flow, thus reducing the amount of preprocessed data so that early detection can be performed. The experimental results show that on the USTC-TFC2016 dataset [8], by examining only the first two packets in each stream, D-PACK still performs with nearly 100% accuracy and has a very low false alarm rate, such as 0.83%.

Most of the research work is to use machine learning algorithms to build models for a set of artificially defined feature data sets (such as protocol type, port number, data length, connection time, etc.). Since the features are artificially selected, it is difficult to avoid the influence of the producer's own subjective consciousness, and the selected features are not necessarily representative. The general idea of these anomaly detection models is as follows: first, analyze the original traffic data according to the rule protocol, extract features, filter features, obtain packet features, flow features, then use classifiers (such as SVM, Softmax, etc.) to classify and obtain tags.

This paper proposes a network traffic abnormal detection method based on CNN and XGBoost. As shown in Fig. 1, firstly, the read network traffic data is preprocessed, then use convolution neural network algorithm is employed to learn traffic characteristics automatically. Finally XGBoost algorithm is utilized to classify and realize network traffic abnormal detection.

The next section presents the dataset and preprocessing. The third section is the proposed model architecture. Finally the conclusions are given.

II. DATA PREPROCESSING

A. Dataset

Up to now, the abnormal behavior detection of network traffic based on machine learning method cannot be applied to the ground. One of the important reasons is the lack of good available public datasets. The datasets used in many related research are private datasets collected and made by themselves. Some classic public data sets also have problems: for example, some datasets are too old to adapt to the current network environment, or some datasets are generated by laboratory simulation and whether they can represent reality is a big problem.

At present, some public data sets used for traffic anomaly detection and traffic classification research can be divided into the following two types: the original traffic data file and the predefined characteristic dataset. The dataset used in this paper is USTC-TFC2016. The data format is pcap data stream file and the size is 3.71GB, including 10 kinds of normal traffic and 10 kinds of malicious traffic, as shown in Fig. 2.

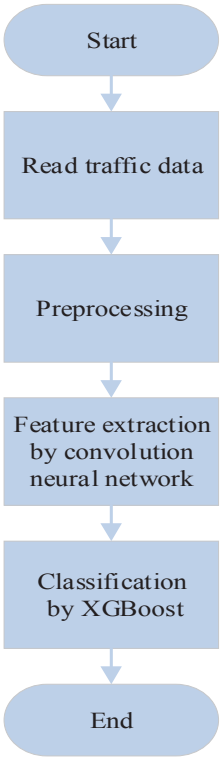

Fig. 1 Flow anomaly detection process based on CNN+XGBoost

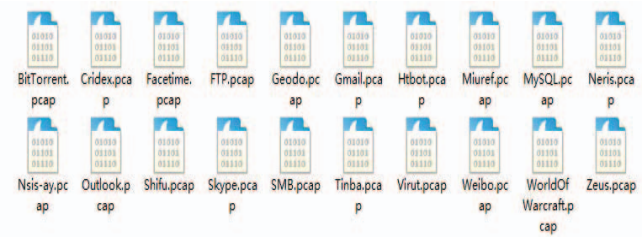

Fig. 2 USTC-TFC2016 dataset

B. Preprocessing

The proposed method is to use convolution neural network to extract the characteristics of traffic data. The data is the original traffic data packet, so it is necessary to preprocess the data to meet the input requirements of CNN network. As shown in Fig. 3, the data preprocessing work mainly includes: flow splitting, data cleaning, uniform length and flow grayscale image generation and other steps to convert the pcap original data stream into flow gray image.

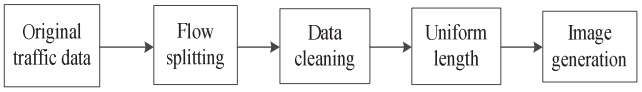

Fig. 3 Data preprocessing flow chart

The purpose of traffic splitting is to divide large pcap files into small pcap files according to certain rules. Pcap files are commonly used Datagram storage formats, as shown in Fig. 4. Pcap files include 24-byte headers and packets. Each packet is 16 bytes of packet headers and specific data composition.

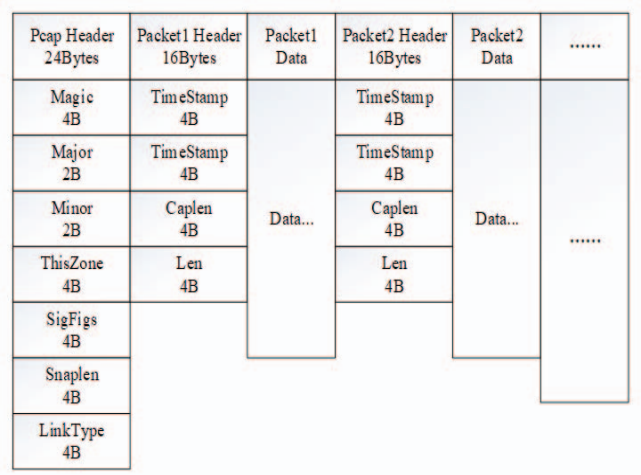

Fig. 4 Pcap file structure

In this paper, the segmentation preprocessing uses splitcap tool and the speed is very fast. The original packets of the dataset are processed in a 5-tuple way, in which the packets with the same source IP, destination IP, source port, destination port and protocol number are grouped together.

Data cleaning is mainly aimed at some redundant data and some streams without application layer data. In addition, it is found that there are too many samples regardless of category after processing, so downsampling is carried out to balance the data in order to prevent serious data imbalance during training. Figure 5 shows the results of the sample distribution after processing. For too many classes, 10240 samples are selected as the experimental data set.

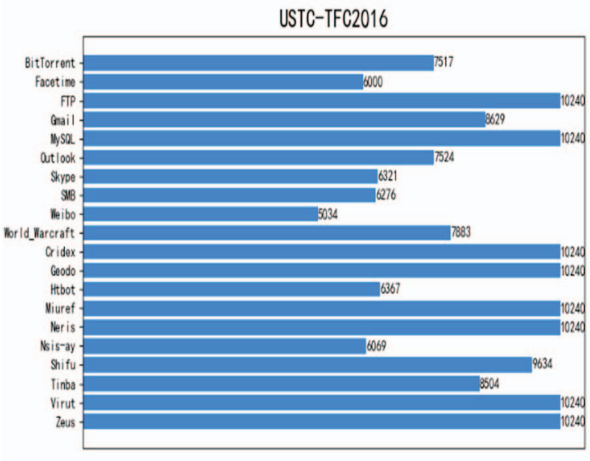

Fig. 5 Data distribution after preprocessing

After traffic segmentation and data cleaning, the number and length of packets of these flows is inconsistent. They cannot meet the input requirements of the neural network. Here, the first n bytes of each flow is intercepted as a result. If the flow length is less than n, padding with 0. After the length is unified, each byte of the flow is treated as a pixel value, just from 0 to 255, and then resized to get a grayscale image. As shown in Fig. 6, the texture features of flow grayscale images are different after the 20 types of traffic data in the USTC-TFC2016 dataset are converted into flow grayscale images.

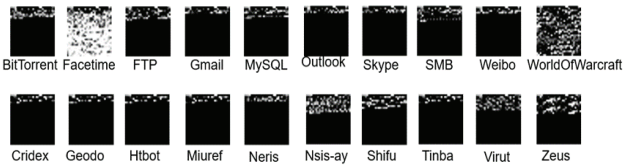

Fig. 6 Flow grayscale image of USTC-TFC2016 dataset

III. THE ANOMALY DETECTION MODEL BASED ON CNN AND XGBOOST

The proposed abnormal traffic detection method in this paper is combing CNN and XGBoost. The convolution neural network for feature extraction of traffic data adopts improved LeNet-5 structure, which not only ensures the detection accuracy, but also has a simple model structure. It greatly reduces the training time and improves the real-time performance of model detection, which is very important in traffic anomaly detection. Figure 7 shows the CNN structure.

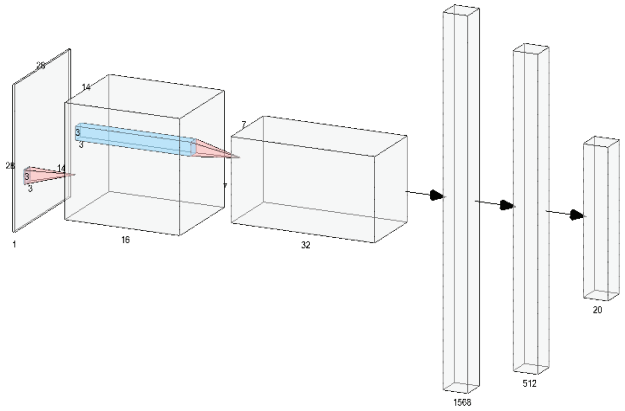

Fig. 7 Convolution neural network structure diagram

Table 1 lists the specific parameters of CNN: two layers of convolution-batch normalization-pooling, plus two layers of full connectivity. After the CNN training, 512 features are calculated and saved as the input of the XGBoost classification algorithm. The XGBoost is an integrated boosting algorithm and an additive model. The prediction process is to add the scores of all the base models to get the final prediction score. The training process is a process of iterative learning of new sub-models, that is, learning the fitting error of the new function according to the current prediction residual. XGBoost has achieved good results in some data mining competitions.

It is well known that there is a significant problem in network traffic anomaly detection based on machine learning method: the data is unbalanced, that is, the number of normal samples is much more than abnormal sample data. Therefore, some optimization is performed for this problem, including sampling-based and weighting-based methods. Sampling includes downsampling and oversampling is to equalize the data through the number of samples. The weighting method is to assign different weights to different categories of samples in the loss function. In this work, we first try to balance the data by oversampling methods such as smote and adasyn, but the effect is not obvious. Then focal loss function instead of mean square error or cross-entropy loss is employed to train CNN, and the classification effect is better.

TABLE I. SPECIFIC STRUCTURAL PARAMETERS OF CNN

layer	type	Filters/neurons	Stride	Padding
1	2DConV+Relu +Batch Normalization	16 (kernel size=3)	1	2
2	Maxpooling	Kernel size=2	2	-
3	2DConV+Relu +Batch Normalization	32 (kernel size=3)	1	2
4	Maxpooling	kernel size=2	2	-
5	Dense	512	-	-
6	Dense	20	-	-

Focal loss is a dynamic cross-entropy loss function. Eq. (1) is that focal loss in binary classification. γ is a weight hyper parameter, which is greater than 1 and it is used to adjust the attenuation rate. $\hat{y}$ is the predicted value. As shown in Fig. 8, the loss of simple samples attenuates is more obviously. Focal loss is more sensitive to samples that are difficult to classify, so it is a kind of difficult data mining.

$$L_{fl} = \begin{cases} -(1-\hat{y})^g \log \hat{y}, & y = 1 \\ -\hat{y}^g \log(1-\hat{y}), & y = 0 \end{cases} \tag{1}$$

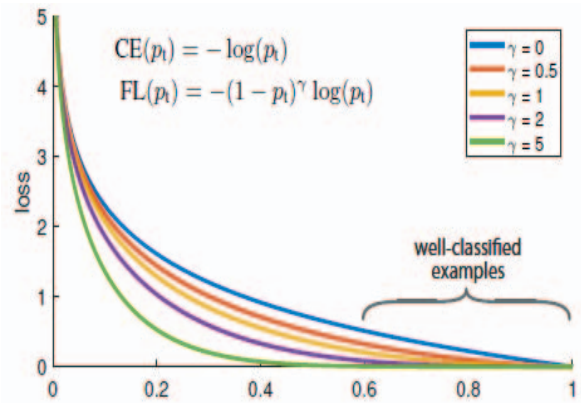

Fig. 8 Focal loss

IV. EXPERIMENT AND ANALYSIS

The deep learning framework used in this experiment is that the TensorFlow1.15, processor is configured with Intel (R) Core (TM) i7-5500U CPU @ 2.40GHz 8.00GB memory. CNN training is a training set with a test set ratio of 4 to 1, and the loss function is focal loss, back propagation optimization algorithm is AdamOptimizer. The XGBoost classifier uses the XGBClassifier class of the xgboost library.

A. Evaluation indicator

The definitions of true positive, false positive, true negative and false negative in abnormal traffic detection are given below: True positive(TP): abnormal flow, and the predicted result is also abnormal. False positive(FP): normal flow, but the predicted result is abnormal. True negative(TN): normal flow, and the predicted result is also normal. False negative(FN): abnormal flow, but the predicted result is normal.

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

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$recall = \frac{TP}{TP + FN} \tag{4}$$

$$f_1 = \frac{2TP}{2TP + FP + FN} = \frac{2\,precision * recall}{precision + recall} \tag{5}$$

$$False\ Alarm\ Rate = \frac{FP}{TN + FP} \tag{6}$$

$$False\ Negative\ Rate = \frac{FN}{TP + FN} \tag{7}$$

The accuracy in Eq. (2) measures the proportion of the total number of correct classifications. Eq. (3), Eq. (4), and Eq. (5) define accuracy, recall rate, and F1 metric, respectively. The first two indicators reflect the impact of incorrect classification on accuracy, and the last index is a comprehensive measure of accuracy and recall rate. The FAR, in Eq. (6), also known as FPR, is used to measure benign flow rates that are misclassified as malicious. The FNR in Eq. (7) measures the rate of malicious traffic that is misclassified as benign.

B. Analysis of experimental results

Fig. 9 shows the confusion matrix of the classification results of the CNN-XGBoost model on the USTC-TFC data set. The horizontal axis of the confusion matrix is the prediction label, and the vertical axis is the real label. The value of $C_{i,j}$ in each position of the matrix indicates the ratio that the sample with the real label i is predicted to be label j. As can be seen from the figure, the darker the color block in the confusion matrix is, the closer the value is to 1. The color on the main diagonal is almost 1, and the color in other positions is very light. And most of the values are less than 0.01. Thus it can be seen that the proposed CNN-XGBoost model in this paper achieves a good classification effect on the data set USTC-TFC, and the traffic type can be classified with nearly 100% accuracy. This is a 20-classification problem. In practice, abnormal traffic only needs to be detected, and there is no need for specific classification. So if it is two-classification, the effect will be better.

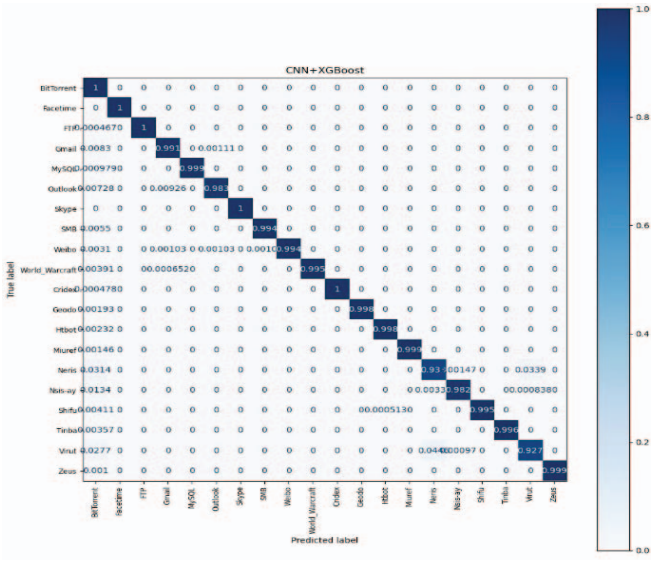

Fig. 9: Confusion matrix

Table 2 gives the specific indicators of the classification effect of each category of traffic, including accuracy, recall rate and f1 value.

TABLE II. ACCURACY, RECALL AND F1 VALUES OF 20 CATEGORIES IN USTC-TFC DATASETS

flow type	precision	recall	f_1-score
BitTorrent	1.00	1.00	1.00
Facetime	1.00	1.00	1.00
FTP	1.00	1.00	1.00
Gmail	0.99	0.99	0.99
MySQL	1.00	1.00	1.00
Outlook	0.99	0.99	0.99
Skype	1.00	1.00	1.00
SMB	1.00	1.00	1.00
Weibo	1.00	1.00	1.00
World_Warcraft	1.00	1.00	1.00
Cridex	1.00	1.00	1.00
Geodo	1.00	1.00	1.00
Htbot	1.00	1.00	1.00
Miuref	1.00	1.00	1.00
Neris	0.95	0.93	0.94
Nsisay-ay	0.99	0.98	0.99
Shifu	1.00	1.00	1.00
Tinba	1.00	1.00	1.00
Virut	0.95	0.94	0.94
Zeus	1.00	1.00	1.00

Table 3 shows the results of the evaluation of the overall classification, which is a comprehensive evaluation index.

TABLE III. OVERALL CLASSIFICATION EFFECT

Comprehensive evaluation	precision	recall	f1-score
micro avg	0.99	0.99	0.99
macro avg	0.99	0.99	0.99
weighted avg	0.99	0.99	0.99
samples avg	0.99	0.99	0.99

V. CONCLUSIONS

In this paper, the network anomaly detection based on network traffic behavior pattern analysis is studied, and a anomaly detection model based on CNN and XGBoost is proposed. Firstly, the pcap data stream files are processed, and then the CNN with improved LeNet-5 structure is employed to extract traffic features. After the completion of CNN network training, XGBoost classification is used to classify the features to achieve traffic classification and anomaly detection. In this process, data enhancement and batch normalization are utilized. Good experimental results on the USTC-TFC2016 data set are achieved. The accuracy is improved, and the false alarm rate is also reduced. At the same time, due to the low computational complexity and good real-time performance of the model, it can meet the practical application requirements.

REFERENCES

[1] Ahmed M, Mahmood A N, Hu J. A Survey of Network Anomaly Detection Techniques. Journal of Network and Computer Applications, 2015, 60: 19-31.

[2] Biersack E, Callegari C, Matijasevic M. Data Traffic Monitoring and Analysis. Springer Berlin Heidelberg, 2013.

[3] Wei wang, Yiqiang Sheng, Jinlin Wang, Xuewen Zeng and Ming Zhu. HAST-IDS learning hierarchical spatial-temporal feature using deep neural networks to improve intrusion detection, IEEE access, 2018, 6:1792-1806.

[4] Farahnakian F, Heikkonen J. A deep auto-encoder based approach for intrusion detection system, 2018 the 20th Inter-national Conference on Advanced Communication Technology (ICACT). New York: IEEE, 2018: 178-183.

[5] Al-Qatf M, Lasheng Y, Al-Habib M, et al. Deep Learning Approach Combining Sparse Autoencoder With SVM for Network Intrusion Detection, IEEE Access, 2018, 6: 52843-52856.

[6] Javaid A, Niyaz Q, Sun W, et al. A deep learning approach for network intrusion detection system, Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies(Formerly BIONETICS). New York: IEEE, 2016: 21-26.

[7] Hwang RH, Peng MC, Huang CW, et al. An unsupervised deep learning model for early network traffic anomaly detection, IEEE Access, 2020, 8: 30387-30399.

[8] Wang W, Zhu M, Zeng X, et al. Malware traffic classification using convolutional neural network for representation learning, 2017 International Conference on Information Networking (ICOIN), Da Nang, 2017: 712-717.