Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 4, No. 1, 2022

Research on Network Traffic Anomaly Detection Method Based on Python

Mingfeng Cai

International Department of Fujian Quanzhou Shenghu Experimental Middle School, Fujian, China

Abstract: Abnormal traffic is the traffic that differs from the normal range of network services. Objective social and natural phenomena, network equipment failures on hardware, and man-made malicious attacks can all lead to abnormal network traffic. Python is a computer programming language that can realize cross-platform interaction, and it is also an object-oriented explanatory and interactive scripting language. Based on this, this paper studies the network traffic anomaly detection method based on Python. By sampling the data sets divided by each layer with different strategies, multiple balanced sub-data sets are obtained, and the feature selection fusion method proposed in the previous section is applied to each sub-data set to obtain the corresponding optimal feature subset, which is used to train multiple base classifiers to perform anomaly detection in this layer. The results show that Python-based network traffic anomaly detection method is superior to the traditional algorithm in accuracy and F1-Score.

Keywords: Python, Abnormal traffic, Traffic detection.

1. Introduction

Anomaly detection of massive network traffic is an important technology of new network attack detection, but the processing efficiency can't meet the real-time requirement because of the massive data [1-2]. Information technology is affecting people's life style and working mode. From government agencies to various enterprises, network information technology is needed to provide work convenience. The security and reliability of the network are related to the business operation all the time. Therefore, it is increasingly difficult to detect abnormal network traffic, and it has become one of the hot issues to study new methods of abnormal network traffic, deal with endless types of abnormal network traffic, and improve the efficiency and accuracy of detection [3]. Python is a computer programming language that can realize cross-platform interaction, and it is also an object-oriented explanatory and interactive scripting language. Based on this, this paper studies the network traffic anomaly detection method based on Python.

2. Research Method

2.1. Overall design of model

The types of abnormal network traffic mainly include Alpha Anomaly, DDos, Port Scan, Network Scan, Worms and Flash Crowd [4-5]. Common network attacks can be divided into three categories: first, reconnaissance and tracking attacks: stealing all kinds of information of the target computer; Second, access attack: using system vulnerabilities to gain host control rights; The third is to refuse service. Network traffic anomaly detection is to analyze network traffic by applying various anomaly detection techniques to find network attacks in advance.

Network traffic collection is a complete tool system, which includes the acquisition, transmission, storage, analysis and display of hardware network traffic data. In the network traffic collection system, the method of collecting raw traffic data is its core technology. In most cases, it is used in LAN. Because the network probe is not routed, the installation of the network probe has little impact on the overall bandwidth of the network. The transmission rate of the connection interface where the network probe is installed, the cache and data processing capability of the monitoring host will all affect the accuracy and efficiency of traffic collection. Therefore, the applicability of this method to the network with large data transmission traffic is relatively poor.

The detection methods that directly take the dimensions of network traffic data header as data attributes mainly include unsupervised learning, supervised learning and semi-supervised learning. Because of the huge volume of network traffic data and its real-time updating, it is necessary to use sliding window and other methods to improve the detection efficiency, so as to realize online detection. At present, researchers have proposed a large number of clustering learning algorithms in unsupervised learning environment [6]. In practical applications, the choice of clustering algorithm depends on the type of data and the purpose of clustering. At present, Python technology has been widely used in computer website development, artificial intelligence, platform back-end management and other fields [7-8].

Python-based network traffic anomaly detection system mainly includes four modules: data acquisition module, information statistics module, anomaly detection module and anomaly warning information module. The design structure of the system detection model is shown in Figure 1.

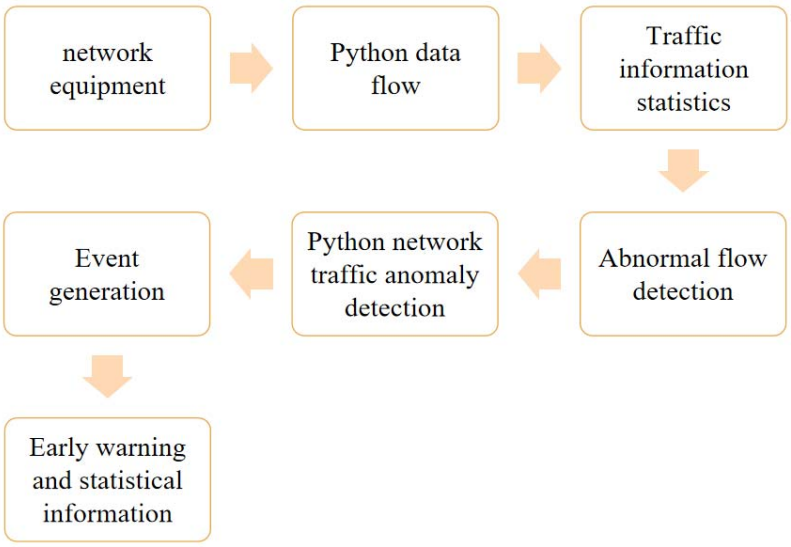

Figure 1. Design structure of system detection model

Through the single-stream detection results, the abnormal events are forewarned and the collected information is stored in the database. Then, the data is generated into statistical analysis data information, and the information concerned by users is presented for network managers to manage the network security. When abnormal traffic occurs in the network and is detected by the detection system, the system will automatically enter the network abnormal data traffic detection module.

If the network anomaly detection system can't detect the abnormal traffic in the network well, it is necessary to modify and adjust the relevant similarity threshold, so that the system can detect the abnormal traffic information in the network well; And the relevant data information is stored in the database of the system, which is convenient for early warning of the same network traffic anomaly in the future.

Analyze the feature selection methods of network traffic, build a basic feature selector, and get the optimal feature subset by fusing the results of feature selection [9]. Then, aiming at the data imbalance phenomenon in multi-layer network traffic anomaly detection, a method of data division and data resampling is proposed. The idea of ensemble learning is introduced to construct different sub-classifiers. Feature selection engineering can reduce over-fitting, make the generalization ability of classification model stronger, and enhance the understanding between features and feature values.

2.2. Python-based network traffic anomaly detection method

Abnormal traffic is the traffic that differs from the normal range of network services. Objective social and natural phenomena, network equipment failures on hardware, and man-made malicious attacks can all lead to abnormal network traffic. The situation that leads to network abnormality is not only the malicious attacks of hackers, but also the abnormal network operation and network configuration errors of operators, the congestion of hardware flash memory, etc. If network traffic anomaly detection is to be carried out, the traffic data generated on the network is the model of detection. In serious cases, the network facilities will be overwhelmed or even close to partial paralysis, and the network of operators will be extremely congested and the loan bandwidth resources will be occupied in large quantities. Therefore, now we attach great importance to the detection technology of Abnormal traffic.

Traffic collection refers to monitoring network traffic data, collecting traffic data from the outside of the traffic analysis system, and inputting it into the internal interface of the system to obtain the real-time status of the network, which is the foundation of the whole network traffic analysis system [10]. In computer networks, abnormal network traffic patterns mean that computer information has been leaked and transmitted to unauthorized terminals.For example, abnormal behavior in credit card transactions may indicate the existence of fraudulent banks. In recent years, network attacks have become more and more rampant. According to McAfee Labs threat report, the types of attacks in the past six months include malware account hijacking, DDoS, etc. Many new types of attacks such as malicious scripts, commercial emails, misconfiguration, etc. In contrast, denial of service/distributed denial of service attacks, network scanning attacks, worm attacks, and non-malicious behavior attacks, etc., have become the network anomalies that anomaly detection really pays attention to because of their strong aggressiveness and great damage, and they are also the types of attacks that this paper focuses on. This section will focus on the causes and phenomena of these attacks.

In order to analyze and classify network traffic more effectively and accurately, this paper proposes an anomaly detection method of network traffic based on Python. By sampling the data sets divided by each layer with different strategies, multiple balanced sub-data sets are obtained, and the feature selection fusion method proposed in the previous section is applied to each sub-data set to obtain the corresponding optimal feature subset, which is used to train multiple base classifiers to perform anomaly detection in this layer.

Error correction is carried out in the python statement of the candidate network application layer. According to the data classification standard, on the basis of capturing sensitive data in the whole DHCP protocol, the correlation between this statement and captured information is evaluated [11]. δ is assumed to be a random variable in DHCP protocol, and the process of judging this variable can be expressed as formula

(1).

$$P_{\delta}(s|0)=\frac{1}{Z_{0_{c\in(s,0)}}\int_{z}(s_c,o_c)} \tag{1}$$

s is a random scene model for DHCP protocol; c represents the random probability of vulnerability exception caused by the change of sensitive data; o is expressed as a machine learning algorithm function; z is the validity parameter of identification information; P is the domain name of sensitive data in the network application layer.

SVM (Support Vector Machines) is a supervised machine learning technology, which can classify different types of data from different research fields through training. Compared with other algorithms, SVM has less chance to generate model over-fitting, and the classification result is more accurate. In this paper, the radial basis function kernel is used to construct the SVM-based classifier of the proposed multi-layer network traffic anomaly detection method. The radial basis function is shown in Formula (2):

$$K(x_i,x_j)=e^{-\gamma\left\|x_i-x_j\right\|^2},\quad \gamma>0 \tag{2}$$

The kernel function calculates the Euclidean distance between two numerical vectors, and maps the input data into a high-dimensional space, so as to best classify the given data into their respective network traffic categories. Radial basis is especially effective when separating data sets sharing complex boundaries.

For a class k problem, the number of all possible structures is N_k, and the specific calculation formula is shown in Formula (3).

$$N_k=\prod_{i=1}^{k-1}2*i-1 \tag{3}$$

The difference of construction methods lies in the uncertainty of classification order, and the problem of uncertainty of classification order lies in the lack of appropriate strategies to determine the classified categories of each layer, which leads to the lack of optimal classification order.

Before the abnormal detection of network data flow, network data is collected for a period of time and preprocessed, and a characteristic attribute value of the current network is obtained by analysis. In this paper, the linear transformation of data is carried out by using the minimum-maximum normalization pair, and the original data is converted into a unified [0,1] range.

$$x^*=\frac{x-\min_a}{\max_a-\min_a}(new\_\max_a-new\_\min_a)+new\_\min_a \tag{4}$$

Map the value x of a to x^* in the interval $[new\_\min_a, new\_\max_a]$. Where $x_{\min}$ is the minimum value of this attribute value and $x_{\max}$ is the maximum value of this attribute value; If $x_{\min}=x_{\max}$, then $x^*=1$.

3. Experimental Analysis

The experimental environment for programming and data set verification of the algorithm is mainly Windows operating system, and the algorithm is mainly realized by python code, and PyCharm integrated development environment is used. The computer system of is configured as Inter Core i5 processor, 4G memory, specifically 64-bit Windows 10 educational operating system.

Through the design and implementation of each module, we have built an experimental environment that can be used to verify the effectiveness of the prototype system. The experiment environment is tested in the laboratory room of the laboratory building of the school computer center. Now, test the data of network characteristics and attributes in the normal state of the network and when Dos attacks are in progress. Generally, we choose the time period from 9: 00 to 10: 00 a.m. to observe the historical traffic during this time period. As shown in Figure 2.

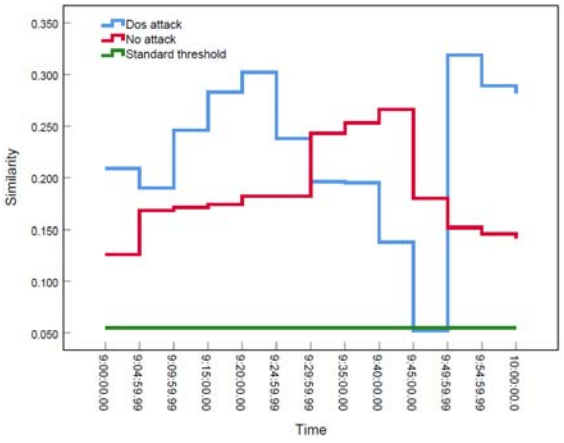

Figure 2. Historical flow curve

From the above figure, we can clearly see that the traffic curves of the three subnets are relatively smooth, with few peaks, during the period from 9: 00 to 10: 00 in the morning. It shows that the network traffic is stable from 9: 00 to 10: 00. Then the Dos attack is tested, and the similarity calculated here is improved by unified dimension.

From the similarity curve, we found that from 9: 10 to 9: 50, the similarity curve is relatively stable in this interval, and the value is relatively small compared with the standard threshold. After 9: 55, the similarity began to be higher than the standard threshold. From this point, we can see that the attack has stopped, and then we can conclude that there was an abnormality in the network from 9: 15 to 9: 45.

In this experiment, the program written in python language is used to preprocess the data, and then the standardized data is used in the experiment. In this experiment, accuracy, detection rate and F1-Score are used as evaluation measures. Finally, the network traffic anomaly detection model proposed in this paper is compared with other traditional classification methods KNN (k-Nearest Neighbor) and SVM, and the results are shown in Figure 3.

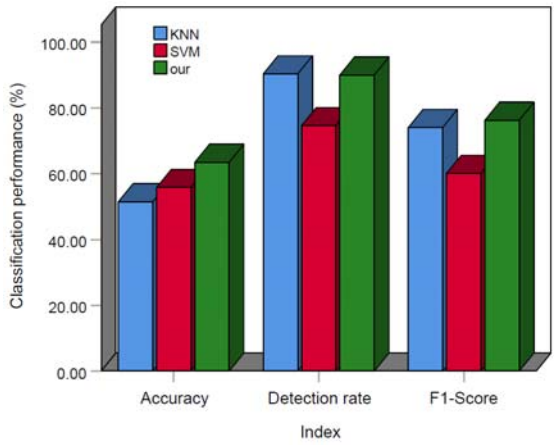

Figure 3. Comparison between this method and other classification methods

It can be seen that Python-based network traffic anomaly detection method is superior to KNN and SVM four classification algorithms in accuracy and F1-Score. The detection rate performance is lower than KNN classification method and higher than SVM four classification method.

4. Conclusions

Information technology is affecting people's life style and working mode. From government agencies to various enterprises, network information technology is needed to provide work convenience. The security and reliability of the network are related to the business operation all the time. This paper focuses on Python-based network traffic anomaly detection method. In the experiment, the program written in python language is used to preprocess the data, and then the standardized data is used in the experiment. In this experiment, accuracy, detection rate and F1-Score are used as evaluation measures. The results show that Python-based network traffic anomaly detection method is superior to the traditional algorithm in accuracy and F1-Score.

References

[1] Monshizadeh, M. , Khatri, V. , Gamdou, M. , Kantola, R. , & Zheng, Y. . (2021). Improving data generalization with variational autoencoders for network traffic anomaly detection. IEEE Access, 2021(99), 1-1.

[2] Pei, J. , Zhong, K. , Jan, M. A. , & Li, J. . (2022). Personalized federated learning framework for network traffic anomaly detection. Computer networks,2022(22), 209.

[3] Zhang, S. T. , Lin, X. B. , Wu, L. , Song, Y. Q. , & Liang, Z. H. . (2020). Network traffic anomaly detection based on ml-esn for power metering system. Mathematical Problems in Engineering, 2020(1), 1-21.

[4] Xia, H. , Fang, B. , Roughan, M. , Cho, K. , & Tune, P. . (2018). A basisevolution framework for network traffic anomaly detection. Computer Networks, 135(22), 15-31.

[5] Hosseinpour, M. , Yaghmaee, M. H. , Seno, S. A. H. , Roshkhari, H. K. , & Asadi, M. . (2018). Anomaly‐based dos detection and prevention in sip networks by modeling sip normal traffic. International Journal of Communication Systems, 31(18), 25-26.

[6] Murugan, K. , & Suresh, P. . (2018). Efficient anomaly intrusion detection using hybrid probabilistic techniques in wireless ad hoc network. International Journal of Network Security, 20(4), 730-737.

[7] Dutta, V. , Chora, M. , Pawlicki, M. , & Kozik, R. . (2020). A deep learning ensemble for network anomaly and cyber-attack detection. Sensors, 20(16), 4583.

[8] Al-Badawi, A. . (2021). Attack-aware iot network traffic routing leveraging ensemble learning. Sensors, 22(41), 37.

[9] Carvalho, L. F. , Abrao, T. , Mendes, L. , & Proenca, M. L. J. . (2018). An ecosystem for anomaly detection and mitigation in software-defined networking. Expert Systems with Applications, 104(10), 121-133.

[10] Wang, J. , Jia, S. , Zhao, H. , Xu, J. , & Lin, C. . (2018). Internet anomaly detection based on complex network path. IEICE Transactions on Communications, 101(12), 2397-2408.

[11] Wang, Y. N. , Wang, J. , Fan, X. , & Song, Y. . (2020). Network traffic anomaly detection algorithm based on intuitionistic fuzzy time series graph mining. IEEE Access, 2020(99), 1-1.