EasyChair Preprint

№ 3405

A Method for Detecting Abnormal Traffic in Full-Stream Network Based on Machine Learning Technology

Yutong Han, Huaibin Wang and Jiongming Zhu

May 15, 2020

A method for detecting abnormal traffic in full-stream network based on machine learning technology

Han Yutong[1,2] Wang Huaibin[1,2] Zhu Jiongming[3]

1. Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Tianjin University of Technology, Tianjin 300384, PR China
2. Key Laboratory of Computer Vision and System (Ministry of Education), Tianjin University of Technology, Tianjin 300384, PR China
3. Jiayuan Huichuan (Tianjin) Technology Co., Ltd)

Abstract: For the network, each server computer, and even the terminal system, abnormal network traffic will cause a lot of CPU time slices and memory space occupation, and cannot respond to demand services normally. In order to solve these problems, it is necessary to build an analysis system of network traffic anomaly, which has good functions of early warning, alarm and traffic processing. This paper proposes a full-flow network abnormal traffic detection method based on machine learning technology, using machine learning technology as a classifier and interpreter to detect abnormal traffic data in the network and output a conclusion report. By importing the network traffic data intercepted from the network into the database, extracting relevant data from the database, constructing a data frame and data point collection, and designing a unique data conversion mechanism for the data, and finally detecting the data points in the data frame and classification and other operations, to obtain the analysis and explanation of normal data, abnormal data and abnormal behavior after classification, and output data analysis static report.

Keywords: machine learning; network abnormal traffic; full flow detection

Introduction

Network traffic is a kind of large data set which is common now. It is more and more difficult to analyze network traffic. Especially when abnormality occurs in the network, the detection of abnormal traffic is very challenging. Abnormal detection of network traffic can provide good information for network failures and security attacks to achieve monitoring and alarming. Nowadays, network security issues cannot be ignored. The traditional method is to use static rule matching network anomaly detection methods in a dynamic and complex network environment. It is difficult to detect unknown anomaly types and cannot meet the requirements of network security detection. Machine learning technology has the characteristics of self-learning and smoking. It can adapt to the complex and changing network environment and can detect unknown abnormal types to meet the needs of real-time accurate detection. Because machine learning system is composed of environment,

knowledge base and execution. The environment provides knowledge information to the learning system, and then the learning system updates the information of the knowledge base through some information. Finally, the executive part will complete the task according to the information of the knowledge base, and feed back the acquired experience to the knowledge base, so that the executive part has stronger execution ability and higher efficiency. For the network, each server computer and even the terminal system, the abnormal network traffic will lead to a large number of CPU time slices and memory space occupation, unable to respond to the demand service normally. In response to these problems, it is necessary to build a full-flow network traffic abnormality analysis system based on machine learning technology to effectively deal with network traffic.

Machine Learning Technology

Machine learning techniques can be divided into multiple categories, such as supervised learning and unsupervised learning. The former needs to provide samples for training. However, unsupervised learning is generally aimed at data without labels or difficult to label by manpower. For example, face recognition technology, there will be many similarities between people, such as the ontology and imitators of stars, because similarity is difficult to define, so unsupervised learning is required for clustering.

The main task of machine learning technology has two points: One is classification. Classification is the task of classifying things into different categories, or to classify labeled data. The classification is divided into single classification and multiple classification at the same time. The single classifier assumes that the training data belongs to only one class. During the training process, the learning objective is a certain function. This function is used to determine the internal points of the training data as positive and the external points as negative, and the single classification does not require labels. Multi classification includes many types, such as DOS, scan, Botnet, etc. The anomaly detection technology based on multi classification needs manually marked data sets, and can not identify the location attack. The second is clustering. Clustering is similar to classification, but the difference is that the class is unknown. By grouping things through the similarity between data, and anomaly detection based on clustering, it needs a larger cluster to be normal, and a smaller cluster to be attacked or intruded. Clustering technology is a common method in anomaly detection, including single chain path clustering algorithm, K-means algorithm and hierarchical clustering algorithm.

Logistic Regression

Logistic regression algorithm is a special classification algorithm, it has positive class and negative class, namely: $y \in \{0,1\}$, where 0 represents negative class and 1 represents positive class. When faced with a classification problem: y = 0 or 1, the possible situations are:

$$h_\theta(x) > 1 \text{ or } h_\theta(x) < 0$$

It is impossible to generalize the results. At this time, we need to use logical regression, and the results can meet the following requirements:

$$0 \leq h_\theta(x) \leq 1$$

Since logistic regression requires our output value to be between 0 and 1, we need to have a hypothesis function that satisfies $0 \leq h_\theta(x) \leq 1$:

$$h_\theta(x) = \mathrm{g}(\theta^T X)$$

Where, X is the eigenvector, G is the logical function, also known as the Sigmoid Function, which is specifically as follows:

$$\mathrm{g}(z) = \frac{1}{1 + e^{-z}}$$

The specific representation of the logic function on the image is as shown in Figure 1:

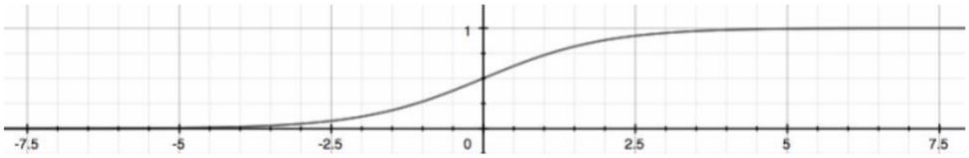

Fig.1. Sigmoid Function

The left side of the logic function approaches 0 infinitely, and the right side approaches 1 infinitely. The output value of the model that meets the needs is between 0 and 1. The function of $h_\theta(x)$ is to input the given variable according to the set parameters, and calculate the possibility that the value of the output variable is 1, namely:

$$h_\theta(x) = P(y = 1|x; \theta)$$

Framework

In order to process the flow data obtained from the data set, the data processing tool USTC-TL2016 is used to process the data in the experimental preprocessing link, including: traffic split, traffic clean, image generation, IDX conversion. At the same time, the final idx3 file formed in the preprocessing process is composed of raw traffic bytes, not only the flow characteristics or packet characteristics. The data of the input model is the original flow data, rather than the characteristics of manual extraction, which can save labor. End-to-end framework overview is shown in Figure 2.

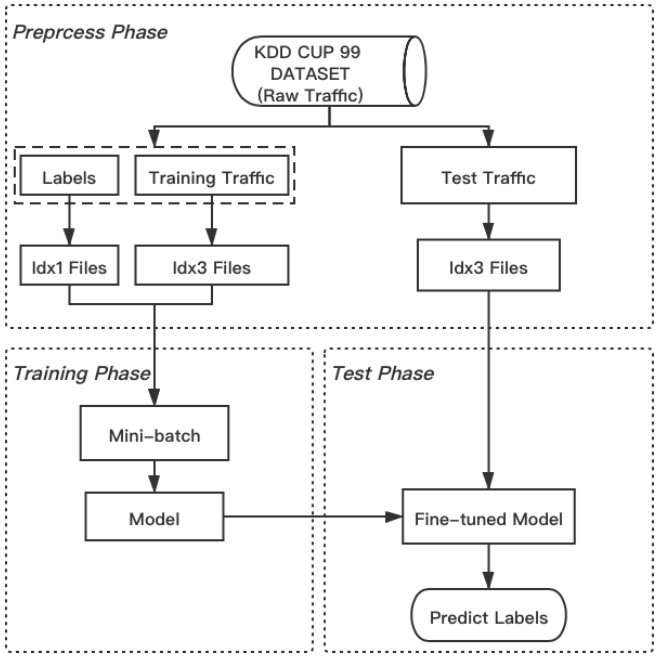

Fig.2. End-to-end Framework Overview

Full Flow Network Anomaly Detection

In order to provide response analysis of dynamic data sources, the machine can use a stream-based data analysis method to filter, highlight, and summarize data, and filter and summarize data before it reaches the user. Since end users do not have the ability to manually analyze each result in large-scale data, they can maximize the effectiveness of each result by using computing resources, thereby helping end users to analyze. In other words, large-scale data needs a full flow data analysis method based on machine learning to help identify data and data trends.

Full-flow Network Anomaly Detection Process Model

First of all, the whole traffic data intercepted from the network is stored in the set database, the connection with the database is established, and the interface is called to extract the data flow from the database for analysis. It designs a column based data framework for the incoming data and processing, and then constructs a collection of data points to be processed. Each data point contains two parts: Measurement and attribute value, in which measurement can be used to detect abnormal traffic and attribute can be used to explain abnormal behavior. Secondly, we transform the extracted data into features, classify the transformed data, use machine learning technology to identify the abnormal items in the network traffic data, and analyze and explain the normal data and abnormal data combined with predicate classifier. The full flow network anomaly detection process model is shown in Figure 3.

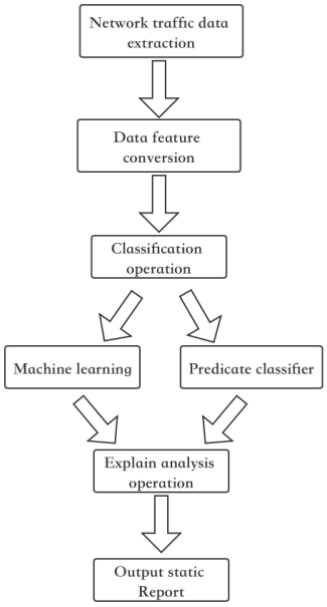

Fig.3. Full-flow network anomaly detection process model

Data Feature Conversion

After the completion of data extraction, we design a feature transformation mechanism to transform the data in the data framework, so that users can analyze various types of data. For example, the IP address is converted from string type to numerical type, and the probability of the combination of source and destination IP addresses of each network traffic in the statistics data set is stored in the Times column. Setting the data feature conversion function allows users to encode and analyze data sets in specific fields without modifying subsequent classifiers and interpreters, enhancing the practicality of abnormal flow detection methods.

Classification Operation

Predicate Classifier

The predicate classifier is used to identify the data traffic with the same attribute and specific value in the network traffic as outliers. Specific implementation steps: first, select a single column as a set of metrics, and secondly select predicates, set thresholds, and determine the metrics of data items one by one. Data items that match the selected predicates and thresholds are identified as outliers and set to 1.0. Mismatched data items are identified as normal values and set to 0.0. Finally, the classification indicators are integrated into one column and added to the last column of the data frame to form a new data frame, which is saved and displayed in the report.

Data Analysis

The adaptive damping reservoir algorithm is used to realize flow based data analysis. First of all, by setting a certain size of storage, to keep the entry inserted into the storage, second, when inserting a data entry, if the storage space is sufficient, then the entry is increased by 1; The entries are put into the storage at a certain ratio, and the existing entries are randomly expelled from the storage.

Explain Operation

Interpretation operation is used to classify multiple network traffic data points, group and summarize the data points, analyze and explain the normal and abnormal behaviors of each attribute group. Firstly, find all the abnormal value entries and normal value entries in the classified network traffic data. Secondly, find the attribute combination with the minimum abnormal support in the traffic data, record the minimum abnormal support, calculate the risk ratio of the single attribute value of the traffic data, and the minimum risk ratio, so as to find the attribute combination meeting the following conditions. The member attribute's abnormal support degree is greater than or equal to the minimum abnormal support degree, and the risk ratio is greater than or equal to the minimum risk ratio. The attribute combination is used to build a prefix tree on the abnormal value entry, where the prefix tree is presented in the way of attribute decreasing, and the attribute combination with the risk ratio less than the minimum risk ratio in the superior is filtered out. Finally, the risk ratio of each attribute combination of network traffic data is obtained.

Experiment

Dataset-KDD CUP 99

Data Preprocessing

The data set collected 9 weeks of TCPdump network connection and system audit data. The original data contains two parts:

1) 7 weeks of training data, including more than 5000000 network connection records;
2) The test data for the remaining 2 weeks contains approximately 2,000,000 network connection records.

By identifying the data set KDD CUP 99, the data set can be divided into five categories: Normal, DOS, Probe, R2L, and U2R, as shown in Table 1-1. After completing the identification work, data cleaning steps are required. There are very few non-compliant data in the KDD99 data set, which need to be checked and then deleted.

TABLE 1-1 KDD CUP 99 DATASET

Identification Type	Implication	Content
Normal	Normal record	Normal
DOS	Denial of service attack	back、land、neptune、pod、smuf、teardrop
Probing	Surveillance and other detection activities	ipsweep、nmap、portsweep、satan
R2L	Illegal access from remote machine	ftp_write、guess_passwd、imap、multihop、phf、spy、warezclient、waremaster
U2R	Ordinary users' illegal access to local super user privileges	buffer_overflow、loadmodule、perl、rootkit

Feature Conversion

For the discretized data, we need to make a brief analysis and filter the attributes. If the attribute has no obvious discrimination after discretization, it can be judged that it is less helpful for the classification algorithm. In this step, num_outbound_cmd, is_host_login, urgent, su_attempted, num_shell, num_failed_login, num_filecreation, these attributes show a tendency to be excessively concentrated in a certain range, so they are excluded. A total of 35 attributes were left for classification.

Output Report

After training, 1 normal identification type was identified as normal and 22 training attack types in the training data set. In addition, 14 types of attacks only appeared in the test data set. Finally, it is concluded that each connection record in the KDD cup 99 training data set contains 41 fixed feature attributes and one class identifier, which is used to indicate whether the connection record is normal or a specific attack type. Of the 41 fixed feature attributes, 9 are of the symbolic type and the others are of the continuous type.

Summary

Nowadays, network security issues cannot be ignored. Generally speaking, there are some aspects to generate abnormal traffic that causes major network failures: first, denial of service attack, which is a very harmful and common attack mode, called DOS, distributed denial of service attack, also known as DDoS. The second is network worm traffic and other abnormal traffic. The abnormal traffic of these networks can cause the slowdown and paralysis of the backbone network, which has great harm and destructive power. The main forms are bandwidth occupation, network blocking, and frequent packet loss caused by failure to send normal data. In this paper, a machine learning based anomaly detection method for full flow network can effectively handle traffic. The next step will focus on the research of blockchain network anomaly detection based on machine learning, and strive to integrate machine learning technology and blockchain to realize the network traffic anomaly detection technology on the blockchain.

Reference

[1] 王伟. 基于深度学习的网络流量分类及异常检测方法研究[D]. 2018.

[2] 陈冠衡, 苏金树. 基于深度神经网络的异常流量检测算法[J]. 信息网络安全, 2019(6):68-75.

[3] 张松清, 刘智国. 一种基于半监督学习的工控网络入侵检测方法[J]. 信息技术与网络安全, 2018(1).

[4] 袁华兵. 一种基于机器学习的 P2P 网络流量识别算法研究[J]. 计算机与数字工程, 2019(10):2387-2391.

[5] 崔丹丹. 机器学习在网络入侵检测中的应用[J]. 信息与电脑, 2018(1):30-32.

[6] 柯宗贵, 王凤娇, 江纬,等. 基于机器学习的恶意文档检测与对抗性学习研究[C]// 第 33 次全国计算机安全学术交流会. 0.

[7] 刘岩, 巨汉基, 丁恒春, et al. 基于机器学习决策树的计量设备异常分析[J]. 自动化与仪器仪表, 2018(5).

[8] 夏景明, 李冲, 谈玲,等. 改进的随机森林分类器网络入侵检测方法[J]. 计算机工程与设计, 2019(8).

[9] 陈胜, 朱国胜, 祁小云,等. 基于机器学习的网络异常流量检测研究[J]. 信息通信, 2017, No.180(12):44-47.

[10] 张晓艳. 基于机器学习的网络异常流量检测方法[J]. 现代电子技术, 2015, v.38;No.454(23):84-87.

[11] 童行行. 基于机器学习的网络流量分析研究[D]. 清华大学.

[12] 左昌盛, 宋歌. 基于机器学习的动态基线及其在银行网络流量数据监测中的应用[J]. 金融经济, 2016, 446(20):94-96.

[13] 左申正. 基于机器学习的网络异常分析及响应研究[D]. 北京邮电大学.

[14] 王伟. 基于深度学习的网络流量分类及异常检测方法研究[D]. 2018.

[15] 杜天宇. 基于机器学习的网络通信流量分析技术研究[J]. 黑龙江科技信息, 2015(15).

[16] 陈飞宇. 基于集成学习算法的异常检测研究[D]. 2015.

[17] 刘垚磊, 杨瑞, 杨艺. 基于 iForest 的虚拟机异常检测机制[C]// 全国计算机安全学术交流会. 0.

[18] 贾伟峰. 网络入侵检测中机器学习方法的应用研究[D]. 电子科技大学.

[19] 程恩. 基于机器学习的入侵检测系统研究[D]. 华中科技大学.

[20] 杨斌. 基于聚类的异常检测技术的研究[D]. 中南大学.