ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20240223001345
Original scientific paper

Network Encryption Traffic Anomaly Detection Based on Integrated Machine Learning

Xiaoqing YANG*, Niwat ANGKAWISITTPAN

Abstract: This paper presents an anomaly detection method for encrypted network traffic using integrated machine learning. A stream feature extraction technique is employed to extract key features such as the median value of stream packets, median value of stream bytes, contrast stream, port growth rate, and source IP growth rate from the encrypted traffic. These features are then fed into an anomaly detection model that combines a collaborative neural network and a random forest classifier. An improved Bagging method is used to fuse and identify the anomalous characteristics of the encrypted traffic by weighted summation. Experimental results using the Trace dataset demonstrate that the proposed method achieves high precision and zero false positives in detecting various types of anomalies under different attack scenarios. The proposed approach offers a promising solution for ensuring network security and protecting against threats in encrypted communication channels.

Keywords: anomaly detection; flow characteristics; improved Bagging method; integrated; machine learning; network encryption traffic

1 INTRODUCTION

The rapid development of new technologies, such as cloud computing, Internet of Things and blockchain, has led to great changes in network traffic patterns and characteristics [1]. These new technologies are highly dependent on the network, which makes the network security issue more important [2]. The increasing popularity of cloud services has also brought new challenges to network security. As a new computing model, cloud computing has greatly promoted the process of enterprise and individual information construction by providing flexible and extensible computing resources. However, with the wide application of cloud services, cloud security issues have become increasingly prominent. Cloud service providers need to ensure the security and privacy of user data, and at the same time deal with threats from various network attacks. The surge of Internet of Things devices is a remarkable trend in recent years. With the improvement of network conditions and the popularization of Internet applications, the number of Internet of Things devices has experienced explosive growth. These devices are connected with each other through the Internet, forming a huge network, which makes data collection, transmission and processing more convenient. However, with the increasing number of Internet of Things devices, network security issues have become increasingly prominent. Because IOT devices usually have low computing power and storage capacity, and often lack professional security protection measures, they are vulnerable to various network attacks. The widespread use of mobile applications also brings new challenges to network security. Mobile applications have become an indispensable part of people's daily life, and they provide various convenient services and functions. However, with the increasing number of mobile applications, security issues have become increasingly prominent. Some malicious software or viruses will pretend to be normal applications and attack by stealing user information and destroying system stability. Therefore, the encryption traffic anomaly detection technology is constantly updated and developed to meet the challenges brought by new technologies and applications [3]. In order to protect the privacy of users, more and more network traffic is encrypted. This makes it difficult for traditional traffic analysis methods to effectively detect and identify abnormal traffic. Therefore, it is necessary to develop anomaly detection technology specifically for encrypted traffic.

It can be seen from relevant research materials that Cvitic et al. [4] studied a machine learning-based DDoS traffic detection method of the Internet of Things. By learning a large amount of network traffic data, we can automatically identify and classify DDoS attack traffic, and improve the accuracy and efficiency of detection. Compared with the traditional detection method based on feature matching, the detection method based on machine learning can better adapt to the high-speed and large-scale Internet environment and detect complex DDoS attack traffic in real time. But this method also has some shortcomings. Due to the diversity and complexity of network traffic, machine learning models may fail when facing unknown attack traffic. Brezolin et al. [5] studied an entropy based network vulnerability detection method, which infers whether there are abnormal behaviors or security threats in the network by analyzing the data entropy of network traffic and system state. Because entropy can measure the chaotic degree of traffic information, it can effectively detect covert attacks and unknown threat behavior traffic in the network. However, this method may be bypassed by an attacker. Some attackers may use some special means to specifically reduce the information entropy of network requests, which makes it difficult to find vulnerabilities and thus cannot effectively detect abnormal network traffic. Hubballi [6] is based on an efficient keyword matching network traffic classification method, which can quickly and accurately identify and classify network traffic, thus improving network performance and security. However, this method also has some shortcomings. First, the accuracy and coverage of keyword matching are limited. Due to the diversity and dynamic nature of network traffic, it is difficult to match all possible keywords completely and accurately, which may cause misjudgment and missing detection of some traffic. Secondly, this method has limited processing capacity for encrypted traffic. With more and more network applications using encrypted communication, classification methods based on keyword matching are difficult to effectively identify abnormal

problems in encrypted traffic. In addition, this method may also face performance bottlenecks and privacy leaks.

In order to solve the above problems, this paper proposes a network encryption traffic anomaly detection technology based on integrated machine learning. Ensemble learning is a machine learning method, and its basic idea is to combine multiple learners to produce a more powerful learning system. This method is sometimes called a multi-classifier system. The advantage of ensemble learning is that it can achieve better performance than a single learner by combining the results of multiple learners. This is mainly due to the diversity of different learners and the effective use of their relationships. Firstly, this paper introduces the basis of network encrypted traffic identification, that is, the object of network encrypted traffic identification; Then the interactive process of encrypted traffic is summarized; The feature extraction method of network encrypted traffic is studied. Finally, the network encryption traffic anomaly detection technology based on integrated machine learning is designed. The main contributions are as follows: under different attack conditions, the number of false alarms for abnormal detection of encrypted traffic is zero, which can improve the accuracy of abnormal detection of network encrypted traffic, thus reducing the error of abnormal detection, and has the ability of high-precision abnormal detection of network encrypted traffic.

2 NETWORT ENCRYPTION TRAFFIC ANOMALY DETECTION TECHNOLOGY

2.1 Encrypted Traffic Anomaly Detection Target

Encrypted traffic identification objects mainly include stream level, host level, packet level and other modes [7]. Among them, stream-level and packet-level identification objects are the most widely used [8]. Tab. 1 shows the details of encrypted traffic anomaly detection targets. This paper mainly takes stream-level encrypted traffic as the detection object. There are two main reasons. On the one hand, the traffic characteristics are obvious. Stream-level encrypted traffic usually exists in the form of continuous data packets in network communication, and these data packets have clear correlation and time sequence. This feature enables the stream-level encrypted traffic to reflect the overall state and behavior pattern of network traffic more accurately, which is conducive to detecting abnormal traffic in the network. On the other hand, it is rich in information. Stream-level encrypted traffic contains rich information, such as source address, destination address, port number and transmission protocol, which is of great significance for identifying abnormal behavior in the network. Through the analysis of stream-level encrypted traffic, we can get the overall characteristics of network traffic and the characteristics of abnormal behavior, which provides strong support for anomaly detection. However, choosing stream-level encrypted traffic as the main detection target also has potential limitations: the detection of stream-level encrypted traffic depends on the data of network traffic, and if the data of network traffic is incomplete or has errors, the accuracy of detection results will be affected. However, the data of network traffic in this paper is complete, and the influence of limitations can be ignored.

Table 1 Details of encrypted traffic anomaly detection targets

Target type	Detection content	Core features
Flow level	Characteristics of flow and transmission process	Duration, number of bytes, etc.
Packet-level	The characteristics and transmission process of traffic grouping	Group size, transmission time, time interval
Host level	Correlation between hosts	Number of connections and ports
Session level	The characteristics and transmission process of conversations	Bytes, session time

2.2 Analysis of Interaction Process of Encrypted Traffic Based on TLS Encryption Protocol

When encrypting traffic anomaly detection, the anomaly problem mainly occurs in the interaction process of encrypted traffic, that is, the traffic transmission process. In the Internet environment, TLS encryption protocol mainly provides encryption services for any two communication applications during network traffic transmission. The protocol consists of TLS recording protocol and TLS handshake protocol. The former protocol is used to identify the type of encrypted network traffic and ensure the integrity of each encrypted network traffic. The latter protocol is responsible for establishing the encrypted network traffic transmission channel [9] before the client and server exchange encrypted network traffic data.

The TLS handshake behavior is when the client sends a request to the server to respond. The client and the server can exchange information only after the authentication is qualified [10]. The TLS handshake process is divided into four stages: Client Hello, Server Hello, Certificate & Key & Cipher Spec, and Change Cipher Spec. Fig. 1 is a schematic diagram of the TLS handshake process. The handshake process is the most vulnerable link to intrusion. If a hacker attacks encrypted traffic by taking advantage of a channel security vulnerability, it is very likely to cause an exception to encrypted traffic [11].

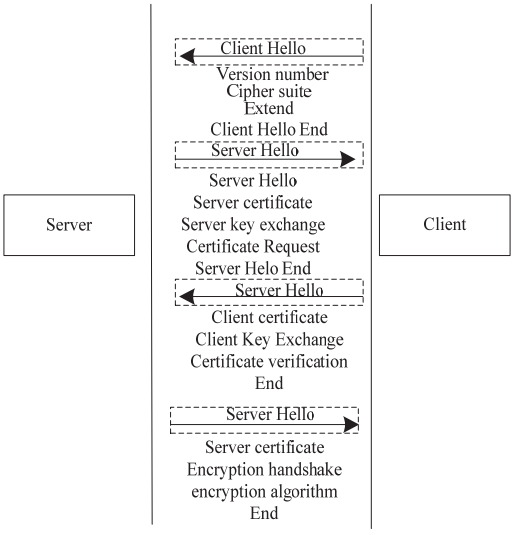

Figure 1 Schematic diagram of TLS handshake process

2.3 Network Encrypted Traffic Feature Mining Method Based on Stream Feature Extraction

As an important step of data classification detection [12], feature selection can effectively reduce the

redundancy of data, reduce the complexity of calculation, and eliminate potential noise [13]. In order to construct a good subset of network encrypted traffic features, it should be ensured that the selected network encrypted traffic features are highly correlated with the classified class tags, and the correlation between these network encrypted traffic features should be reduced as much as possible [14]. In other words, the characteristics of network encrypted traffic that are irrelevant and redundant to class tags should be eliminated. Therefore, this paper uses the network encrypted traffic feature mining method based on flow feature extraction to mine the network encrypted traffic feature for anomaly detection [15].

If the network encrypted traffic sample is $A=\left(A_1, A_2, .., A_j\right), j=1,2,...,q$. When the encrypted channel is attacked, the encrypted traffic will change in size and IP port access. Therefore, in order to detect the anomaly of encrypted traffic, this paper extracts the flow characteristics of encrypted traffic as a sample of anomaly detection [16]. The stream characteristics of encrypted traffic are mainly divided into the median number of stream packets X_1, the number of bytes in the stream X_2, contrast flow X_3, port growth rate X_4, source IP growth X_5. Because they can capture the abnormal patterns that may appear when encrypted traffic is attacked in network traffic analysis. These features are unique and play an important role in anomaly detection, as follows:

(1) Median value of stream packet number X_1:

Principle: The median value of the number of packets in a stream indicates the median number of packets in a traffic stream, which can help the detector identify those streams with abnormal increase or decrease in the number of packets, such as identifying signs that attackers are trying to hide their attack behavior or increase network congestion.

Contribution: By monitoring the change of the bit value in the number of flow packets, the detector can identify the traffic flow that deviates from the normal behavior pattern, thus triggering further anomaly detection.

$$X_1=\begin{cases}\dfrac{q_{(n+1)/2}+q_{(n-1)/2}}{2}, n \text{ is an odd number}\\ \dfrac{q_{\mathrm{n}/2}}{2}, n \text{ is an even number}\end{cases} \tag{1}$$

In the formula, n, $q_j \in q_n$ represent the number of encrypted traffic samples and the number of traffic packets in the traffic samples in ascending order, respectively, the j-th number of bit stream packets.

(2) The median value in the number of bytes of the stream X_2:

Principle: The median value in the number of bytes in a stream indicates the median number of bytes transmitted in a traffic stream, which plays an important role in detecting data leakage, DDoS attacks or other attacks that lead to a large amount of data transmission.

Contribution: When the median value of the number of bytes in the encrypted traffic suddenly increases, it indicates that abnormal data transmission activities are going on, such as data leakage or traffic amplification stage of DDoS attack.

$$X_2=\begin{cases}\dfrac{q'_{(n+1)/2}+q'_{(n-1)/2}}{2}, n \text{ is an odd number}\\ \dfrac{q'_{\mathrm{n}/2}}{2}, n \text{ is an even number}\end{cases} \tag{2}$$

In the formula, $q'_j \in q'_n$ represents the inflow samples, the number of bytes in the stream is arranged in ascending order, and the number of bytes in the stream, the j-th number of bit stream packets.

(3) Contrast flow X_3:

Principle: Contrast flow refers to other traffic flows that are similar or related to the current traffic flow in certain attributes (such as source IP, destination IP, port, etc.). By analyzing the behavior pattern of contrast flow, whether the current traffic flow is abnormal can be detected.

Contribution: If the contrast flow shows normal behavior and the current traffic flow is abnormal, it may be a sign of attack behavior. In addition, the contrast flow can also help the detector identify the attack traffic that tries to imitate the normal traffic to cover up its abnormal behavior.

In the encrypted channel, there is an interactive relationship between the access address and the destination address of the encrypted traffic [17]. If the traffic flows A is forward data stream, A' is the reverse data flow, then A, A' are interactive flow.

The contrast flow is:

$$X_3=2\times l/L \tag{3}$$

In the formula, l, L represent the logarithm and total number of interactive flows in encrypted traffic respectively.

(4) Port growth rate X_4:

Principle: Port growth rate indicates the growth rate of the number of new ports accessed by a source IP or destination IP address in a period of time, which plays an important role in detecting attacks such as port scanning and botnets.

Contribution: When the port growth rate increases abnormally, it indicates that an attacker is scanning the port of the target or trying to establish a large number of connections, thus triggering the anomaly detection mechanism.

When the encrypted channel is attacked, the attacker can attack the server through the randomly designed port number, and the port growth rate will also increase [18], so the increment of the port at a fixed time can be set as the flow characteristic of the encrypted traffic.

$$X_4=\frac{\widehat{x}}{T} \tag{4}$$

In the formula, $\widehat{x}$, T represents the port increment and time respectively.

(5) Growth rate of source IP X_5:

Principle: The growth rate of source IP indicates the growth rate of the number of new source IP addresses accessing a certain destination IP address in a period of

time, which is helpful to detect distributed attacks from multiple source IP addresses.

Contribution: When the growth rate of source IP increases abnormally, it indicates that multiple attackers or botnet nodes are attacking the target, thus triggering the anomaly detection mechanism.

When the encrypted channel is attacked, the attacker can attack the server through a randomly designed IP address. At this time, the source IP address of the attacked host increases [19]. Therefore, the source IP increment at a fixed time can be used to set the flow characteristics of encrypted traffic.

$$X_5 = \frac{IP_r}{T} \tag{5}$$

In the formula, IP_r represents the source IP increment.

2.4 Network Encryption Traffic Anomaly Detection Model Based on Integrated Machine Learning

Integrated machine learning can comprehensively consider the characteristics of network encrypted traffic and improve the accuracy of network encrypted traffic anomaly detection by integrating the advantages of multiple classification models [20]. It is helpful to identify and classify abnormal behaviors in network encrypted traffic more accurately. Fig. 2 is the technical flow of the network encryption traffic anomaly detection method based on integrated machine learning.

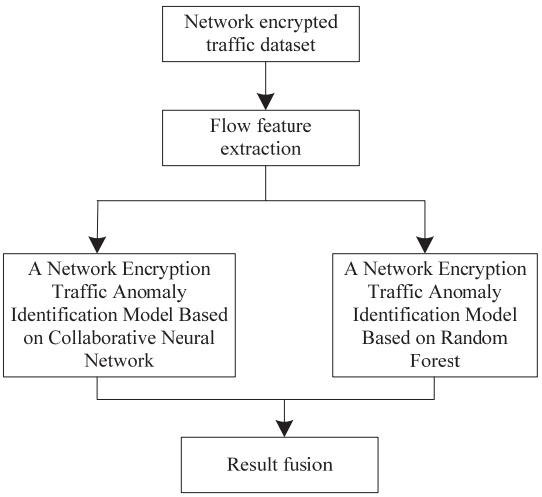

Figure 2 A network encryption traffic anomaly detection method based on integrated machine learning

2.4.1 Network Encryption Traffic Anomaly Identification Model Based on Collaborative Neural Network

Collaborative neural network performs well in dealing with complex data structures and pattern recognition tasks. In the recognition of network encrypted traffic anomalies, encrypted traffic data is usually highly complex and dynamic, and models are needed to capture these complex patterns and relationships. Collaborative neural network can automatically learn effective feature representation from encrypted traffic data, and identify abnormal traffic through its powerful learning ability. The advantages are as follows: on the one hand, feature learning ability: collaborative neural network can automatically extract features from the original encrypted traffic data, avoiding the tedious manual feature engineering process. On the other hand, nonlinear processing ability: the relationship in encrypted traffic data is often nonlinear, and collaborative neural network can handle this complex nonlinear relationship. The limitations are as follows: on the one hand, the demand for computing resources is high: the training and reasoning of collaborative neural networks usually require a lot of computing resources. On the other hand, the dependence on training data: the performance of collaborative neural network is highly dependent on the quality and quantity of training data.

In the identification of network encryption traffic abnormality in the collaborative neural network, the mode u_N number of the traffic characteristic sample X of the network encryption traffic that needs to be identified is set to N, and the dimension of the mode vector u_N is set to M, then the identification dynamic equation is:

$$\dot{p} = \sum_N \beta_N u_N \left(u_N^+ p\right) - \sum_N \left(u_N^+ p\right)^2 \left(u_N^+ p\right) u_N - \left(p^+ p\right) p + O \tag{6}$$

In the formula, p represents the input initial state of the network encrypted traffic characteristic sample X; $\dot{p}$ represents the identification result of encrypted traffic anomaly; β_N, O respectively represent the fluctuation amplitude of attention parameters and the transmission rate of network encrypted traffic; u_N^+ represent the adjoint vector of u_N.

$$\left(u_N^+, u_N\right) = u_N^+ u_N \tag{7}$$

By two ordered parameters $\delta_N = \left(u_N^+, p\right) = u_N^+ p$, E, the dynamic evolution equation of network encryption traffic anomaly detection is constructed:

$$\dot{\delta}_N = \delta_N \left(\beta_N - E + \delta_N^2\right) \tag{8}$$

Using the order parameter equation, a collaborative neural network is designed to identify the abnormal network encryption traffic. The network structure is shown in Fig. 3.

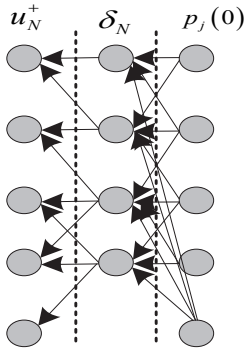

Figure 3 Collaborative neural network for identifying abnormal encrypted traffic in networks

In the cooperative neural network structure used to identify the abnormal network encryption traffic, the component amount $p_j(0)$ is set by the initial value of the network encrypted traffic characteristic sample pattern vector input by the input layer element *i*. The middle layer forms neurons from order parameters, which can be multiplied and summed by each $p_j(0)$ and u_N^+. There are order parameters δ_N, and each neuron can classify whether there is abnormality in the characteristic samples of network encrypted traffic. Considering the dynamic nature of the network encrypted traffic feature samples, this network uses the dynamic equation to dynamically classify the network encrypted traffic feature samples, and the classification result of the encrypted traffic feature output by the output layer is the anomaly detection result:

$$p_j(t) = \sum_N \delta_N(t) u_N \tag{9}$$

2.4.2 Network Encryption Traffic Anomaly Identification Model Based on Random Forest

Random forest is an integrated learning method based on decision tree, which improves the accuracy of classification by constructing multiple decision trees and integrating their prediction results. In the identification of network encrypted traffic anomalies, random forest can process high-dimensional data and evaluate the importance of features, and has good generalization ability and robustness. The advantages are as follows: on the one hand, high accuracy: Random forest can usually achieve high classification accuracy by integrating the prediction results of multiple decision trees. On the other hand, it is robust: random forest has good tolerance to noise and outliers, and it is not easy to over-fit. The limitations are as follows: On the one hand, the demand for computing resources: Although the demand for computing resources of random forest is lower than that of neural network, it still needs some computing resources when dealing with large-scale data sets. On the other hand, the processing of high-dimensional data: when the feature dimension is very high, the performance of random forest may decline because it needs feature selection and division in each decision tree.

In the random forest-based network encrypted traffic anomaly identification model, consider the sum ϖ of the weight coefficient of the *m* network encrypted traffic characteristics as a whole, then:

$$\varpi = \left[\varpi_{1,1}, \varpi_{1,2}, \ldots, \varpi_{1,m}\right] \tag{10}$$

In the process of running the random forest-based network encrypted traffic anomaly identification model, it is necessary to design the decision tree for identifying the anomalous traffic and update the encrypted traffic anomaly detection sample weight coefficients. Then *X* is randomly sampled to construct a training set of traffic features for encrypted traffic anomaly detection. When a decision tree is used for detection of anomalies in encrypted traffic features at time of *t*, the method of updating the weight coefficients is expressed as follows:

$$\varpi(t+1) = \left[\varpi_{1,1}(t+1), \varpi_{1,2}(t+1), \ldots, \varpi_{1,m}(t+1)\right] \tag{11}$$

In the process of encrypted traffic anomaly detection, when the encrypted traffic anomaly detection results are accurate, the weight coefficients of such traffic feature samples need to become smaller when updating, this operation can ensure that the number of repeated detections will be reduced in the next classification detection. For this reason, this paper uses the confusion factor *z*, which is used to control the weight coefficients and values of all feature samples as a whole after each weight coefficient update. Then:

$$\begin{cases} \varpi_{m+1,1} = \left(\varpi_{m+1,1} / z\right) \times \varepsilon \times \theta^{\pm\gamma} \\ z = \left[\sum_{m=1}^{\varpi} \varpi_{m+1,1}\right] \times \varepsilon \cdot \times \theta^{\pm\gamma} \\ \gamma = \ln\left(\left(1-\theta_m\right)/\theta_m\right) \times \dfrac{1}{2} \end{cases} \tag{12}$$

In the formula, ε, γ represent general parameters, weight balance coefficients, respectively. θ represents the number of feature samples accurately identified.

Random forest-based network encrypted traffic anomaly identification model is trained, and the classifier weights of each decision tree are calculated:

$$\varpi = \frac{2}{\phi^{-1} + \varphi^{-1}} \tag{13}$$

In the formula, ϕ, φ represent the ratio of θ to the total number of encrypted traffic features, and the mean value of the cost of an anomaly detection for a particular encrypted traffic feature sample as a percentage of the total anomaly detection process.

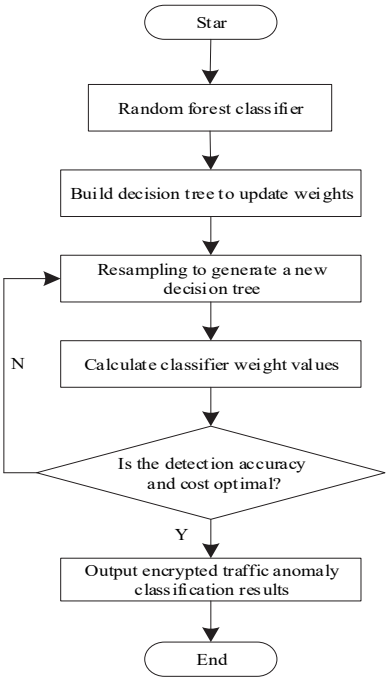

Figure 4 The operation process of a network encryption traffic anomaly recognition model based on random forest

The classifier weight of each decision tree can reflect the detection accuracy of encrypted traffic anomaly. If the recognition accuracy of a decision tree is significant, the number of correctly identified encrypted traffic feature samples is large. At this time, the value ϕ is large. If the recognition accuracy of a decision tree is significant and its cost is low, and the recognition performance is optimal at this time, the corresponding weight coefficient is large, then the output of this decision tree is the result of encrypted traffic feature anomaly recognition $p_j(t)$. Fig. 4 shows the operational flow of the random forest based network encrypted traffic anomaly identification model.

2.4.3 Encryption Traffic Anomaly Detection Based on Improved Bagging Method

Bagging method, also called bootstrap aggregation method, is one of the commonly used methods in integrated learning technology. In this method, the samples of the detection results of network encrypted traffic anomaly in sections 2.4.1 and 2.4.2 can be sampled in the way of playback, and reorganize the detection results of the two anomaly detection models into V data sets of encrypted traffic anomaly detection results. The detection results of the network encryption traffic anomaly recognition model based on synergetic neural network and the network encryption traffic anomaly recognition model based on random forest are the classification results of the weak classifier in the improved Bagging method. In the field of encryption traffic anomaly detection, due to the complexity of network environment and the variability of data, it is often difficult for a single classifier to achieve the ideal detection effect. Therefore, ensemble learning technology, especially Bagging method, has become an important means to improve the performance of anomaly detection system because it can effectively integrate the advantages of multiple classifiers.

The function of the improved Bagging method is to improve the accuracy and robustness of the whole anomaly detection system by combining the prediction results of several weak classifiers. Weak classifiers refer to the network encryption traffic anomaly identification model based on collaborative neural network and the network encryption traffic anomaly identification model based on random forest respectively. Each weak classifier has its own unique advantages and limitations. The improved Bagging method can achieve complementary advantages by reasonably combining the outputs of these classifiers, thus obtaining more accurate and comprehensive anomaly detection results.

The specific implementation details are as follows:

Step 1: Data sampling: A plurality of sample subsets are extracted from the original data set according to the method of putting back, and the size of each subset is usually the same as or similar to the original data set.

Step 2: Training weak classifiers: using each sample subset to train two weak classifiers (cooperative neural network and random forest) respectively. Because there is put-back sampling, the same sample may appear in multiple subsets, which is helpful to increase the difference and diversity between weak classifiers.

Step 3: Prediction and combination: For the encrypted traffic data to be detected, each weak classifier will give its own prediction result. The improved Bagging method fuses these prediction results into a final detection result by voting.

Step 4: Output the result: output the final anomaly detection result according to the combined prediction result.

However, in the process of "sampling in the way of putting back", if the distribution of data samples of the encrypted traffic anomaly detection results is poor, it will lead to errors in the final network encrypted traffic anomaly identification results. For this reason, this paper improves the Bagging method. First, according to a certain proportion, perform the operation of "sampling in the way of putting back". If the total number of samples in the network encryption traffic anomaly identification result is K, the number of samples of a certain class is ρ, the number of training sample sets needed is μ, then the number of samples with put-back sampling for encrypted traffic anomaly detection is:

$$\partial = \frac{\rho}{\mu} \times K \tag{14}$$

The anomaly detection results of 2 network encrypted traffic anomaly recognition models for many times, are set into the following matrix pattern:

$$p(X) = \begin{Bmatrix} p_1(X) \\ p_2(X) \\ \vdots \\ p_n(X) \end{Bmatrix} = \begin{Bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{Bmatrix} \tag{15}$$

In the formula, $p_{mn} \in \{0,1\}$, $\sum_{i=1}^{K} p_{ji} = 1$ represents the detection result of the j-th network encrypted traffic anomaly recognition model is the i-th traffic data, when the final decision result of the j-th detection model, the formula is:

$$P_j = \underset{K}{\operatorname{argmax}}\, p_{ji} \tag{16}$$

Finally, using the voting method, the outputs of the 2 detection models were integrated and analyzed, and the number of votes obtained in the integrated analysis was Ω_j, then the category with the highest number of votes is the final network encryption traffic anomaly detection result:

$$P''' = \underset{K}{\operatorname{argmax}} \sum_{i=1}^{n} \pi\left(\Omega_j = \rho\right) \tag{17}$$

In the formula, π represents an exponential function, if π is 1, then the encrypted traffic is anomalous at this time, and vice versa, when π is 0, the encrypted traffic is normal. Fig. 5 shows the technical architecture of encrypted traffic anomaly detection based on the improved Bagging method.

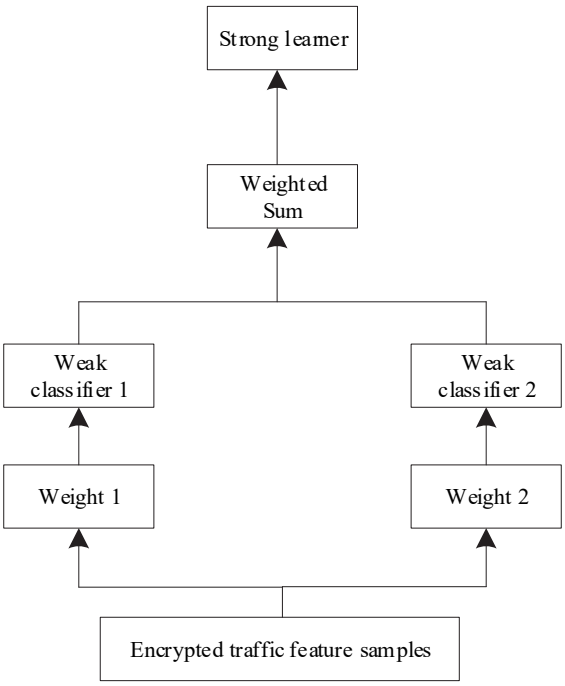

Figure 5 Technical architecture of encrypted traffic anomaly detection based on improved Bagging method

Because each weak classifier has its unique advantages and limitations, combining them by improved Bagging method can realize complementary advantages, thus improving the accuracy of the whole anomaly detection system. The improved Bagging method can effectively reduce the possible over-fitting risk of a single classifier and improve the robustness and generalization ability of the system by sampling and combining the prediction results of multiple weak classifiers. In practical application, the abnormal data of encrypted traffic is often far less than the normal data, resulting in unbalanced data distribution. The improved Bagging method can deal with this imbalance problem by adjusting the proportion of different types of samples in the sampling process, and further improve the effect of anomaly detection.

3 EXPERIMENTAL ANALYSIS

To test the use effect of this technology, Trace dataset is used as the encrypted traffic anomaly detection data source of this technology. The 24-hour encrypted traffic in an encrypted channel is randomly extracted from this dataset. Pre-processing the original traffic data, including data cleaning (removing invalid or duplicate data), feature extraction (extracting meaningful features from traffic data, such as number of packets, number of bytes, port speed increase, etc.) and feature coding (converting the extracted features into a format suitable for machine learning model processing). Because the data set is continuous flow data for 24 hours, these data are divided in time order. Taking the traffic data of the first 20h as the training set and the traffic data of the last 4h as the test set ensures that the training set and the test set are continuous in time, thus more accurately reflecting the actual changes of network traffic. In integrated machine learning, some parameters are set for the model to control the complexity and performance of the model, such as the number of decision trees, the maximum depth of each decision tree, and the minimum number of samples required for each node to split. In order to obtain better model performance, cross-validation and grid search techniques are used to optimize model parameters. Because the integrated machine learning model usually involves a lot of calculations, it needs the server hardware resources of computing server, cloud computing resources, multiple CPU cores and GPU accelerators to support the training and reasoning of the model, and Scikit-learn, T machine learning library and framework are installed to support the implementation and training of the model. See Tab. 2 for details.

Table 2 Experimental data details

Data stream encoding	Number of flow records/bit	Continuous time consumption /s
1	24864	1821.9
2	23802	1696.8
3	22933	1724.2
4	22286	1784.2
5	21649	1658.6
6	55836	1666.8
7	55495	1665.6
8	65036	1614.4

Fig. 6 shows the data flow statistics of encrypted traffic, the green line represents the instantaneous bandwidth demand of the encrypted channel on a particular day, and the black area indicates the encrypted data block.

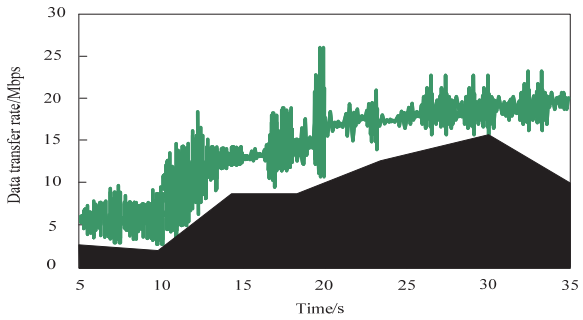

Figure 6 Statistical chart of encrypted channel data flow

In the encrypted traffic dataset, several anomalous samples shown in Tab. 3 are mixed.

Table 3 Details of abnormal data introduced by network encrypted traffic data

Data stream encoding	Abnormal data type	Number of flow records /bit
1	Analyze attacks	837
2	Backdoor attack	584
3	Vulnerability exploitation	4195
4	Pan attack	25807
5	Reconnaissance attack	19225
6	Out of buffer	7307
7	Mite attack	1455
8	Denial of service	165

The anomaly detection results of network encrypted traffic under different attack modes by the techniques in this paper are shown in Tab. 4.

As the data in Tab. 4 shows, this paper's technique for the introduction of anomalous data flow record samples, the detection of abnormal traffic samples is accurate, and can accurately detect network encryption traffic anomaly samples.

Table 4 Results of abnormal detection of network encrypted traffic using this article's technology

Data stream encoding	Number of encrypted traffic stream records / bit	Number of stream records after introducing abnormal data / bit	The number of abnormal data detected by the technology in this article / bit
1	24864	25701	837
2	23802	24386	584
3	22933	27128	4195
4	22286	48093	25807
5	21649	40874	19225
6	55836	63143	7307
7	55495	56950	1455
8	65036	65201	165

DNS spoofing is also called domain name spoofing. DNS spoofing methods include cache poisoning and information hijacking. The schematic diagram of two kinds of deception is shown in Fig. 7 and Fig. 8.

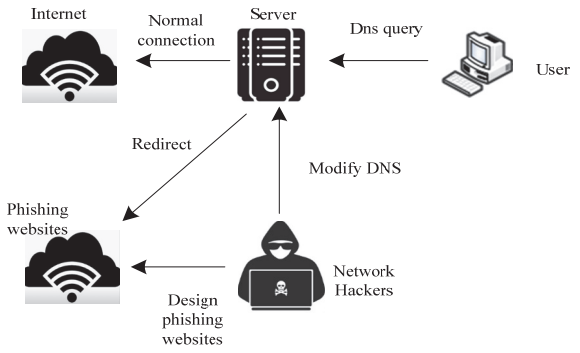

Figure 7 Cache poisoning

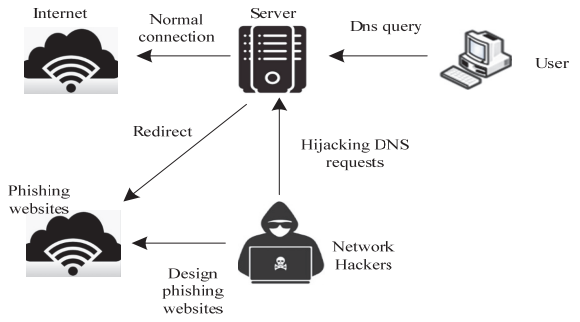

Figure 8 Information hijacking

In these two conditions, the number of false alarms of the encrypted flow anomaly detection is tested before and after the use of the technique in this paper, and the results are shown in Fig. 9 and Fig. 10.

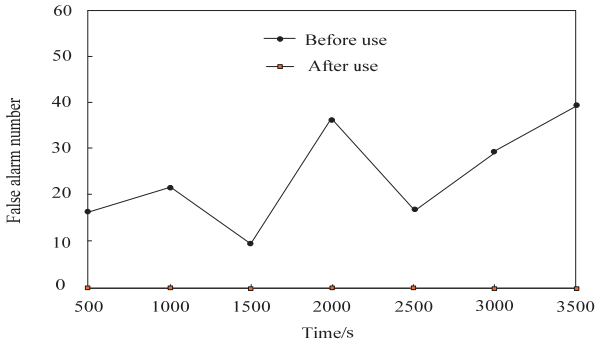

Figure 9 The number of false positives in detecting encrypted traffic anomalies during cache poisoning conditions

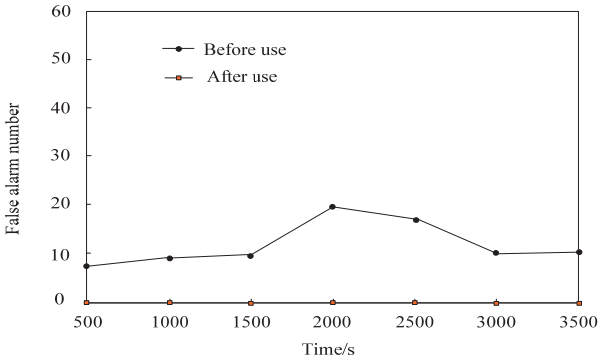

Figure 10 The number of false alarms in detecting abnormal encrypted traffic during information hijacking conditions

According to Fig. 9 and Fig. 10, it can be seen that in the cache poisoning, information hijacking conditions, before and after the use of this paper's technology, the number of false alarms of encrypted traffic anomaly detection has obvious changes; before the use of this paper's technology, the number of false alarms of encrypted traffic anomaly detection are more than 5 times, while after the use of this paper's technology, the number of false alarms of encrypted traffic anomaly detection are 0 times. Comparing the results of this test, it can be seen that the use of this paper's technology can improve the accuracy of network encrypted traffic anomaly detection, thus reducing the anomaly detection error, so the number of false alarms of anomaly detection is zero.

4 CONCLUSION

In this paper, we study a network encrypted traffic anomaly detection technique based on integrated machine learning, which has high value in network encrypted traffic anomaly detection problem and has a broad future outlook. The use of this technique can improve the accuracy of network encrypted traffic anomaly detection and thus reduce the anomaly detection error.

The network encryption traffic anomaly detection technology based on integrated machine learning proposed in this paper has shown remarkable advantages and wide application potential in real-world scenarios. Through deep learning and intelligent analysis of encrypted traffic, this technology can effectively identify abnormal behaviors in the network and provide a strong guarantee for network security.

Firstly, by comparing with other latest anomaly detection methods for encrypted traffic, the superiority of this method and the rationality of model selection are confirmed, which shows high performance and proves the effectiveness of integrated machine learning in dealing with complex network encrypted traffic data.

Secondly, through the sensitivity analysis of key parameters of cooperative neural network and random forest model, the influence of these parameters on detection performance is deeply understood, and the potential areas for further optimization are determined, which provides important guidance for selecting the best parameter settings according to the network environment and data characteristics in practical application.

In addition, the test results on larger and more diverse data sets show that this method has good scalability and

computational efficiency, can cope with the increasing encrypted traffic and increasingly complex abnormal behavior in the real network environment, and provide real-time and efficient anomaly detection services for network administrators and security analysts.

Finally, when analyzing the importance of different features and the decision-making process of the model, the potential anomaly detection mechanism is found, which provides valuable insights for understanding the anomaly detection process and helps to explain and convey the detection results to network administrators or security analysts.

In the future, integrated machine learning techniques will continue to play an important role in network encrypted traffic anomaly detection. As network attacks evolve and encrypted traffic increases, more efficient and accurate anomaly detection methods are needed to ensure network security. Integrated machine learning technology can improve the accuracy and efficiency of anomaly detection by continuously improving and optimizing the model, and become one of the important means to deal with network threats.

In order to further improve the performance of integrated machine learning techniques for anomaly detection of encrypted network traffic, future research efforts can be carried out in the following areas:

(1) To study how to improve the real-time and response speed of the model in order to better cope with high-speed and large-scale network traffic.

(2) Consider how to protect user privacy and data security, and ensure that the anomaly detection process does not disclose the user's personal information and data.

Acknowledgements

This work was supported by Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi (STIP, Research on Encrypted Traffic Identification and Anomaly Detection Based on Machine Learning, No. 2023L422).

5 REFERENCES

[1] Tahira, M., Jae, W., Lim, S. T., & Shah, M. Y. C. (2022). A Novel Deep-Learning-Enabled QoS Management Scheme for Encrypted Traffic in Software-Defined Cellular Networks. *IEEE Systems Journal*, *16*(2), 2844-2855. https://doi.org/10.1109/JSYST.2021.3089175

[2] Gracy, T. W., Prakash, M., & Betina. J. (2021). Multicast on-route cluster propagation using to identify the network intrusion detection system in mobile ad hoc network. *International journal of communication systems*, *34*(11), e4850. https://doi.org/10.1002/dac.4850

[3] Nakashima, M., Sim, A., & Kim, Y. (2021). Automated Feature Selection for Anomaly Detection in Network Traffic Data. ACM*Transactions on Management Information Systems*, *12*(3),1-28. https://doi.org/10.1145/3446636

[4] Cvitic, I., Perakovic, D., Perisa, M., & Botica, M. (2021). Novel approach for detection of IoT generated DDoS traffic. *Wireless networks*, *27*(3), 1573-1586. https://doi.org/10.1007/s11276-019-02043-1

[5] Brezolin, U., Vergutz, A., & Nogueira, M. (2023). A method for vulnerability detection by IoT network traffic analytics. *Ad hoc networks*, *149*(10), 1-10. https://doi.org/10.1016/j.adhoc.2023.103247

[6] Hubballi, N. & Khandait, P. (2022). KeyClass: Efficient keyword matching for network traffic classification. *Computer communications*, *185*(3), 79-91. https://doi.org/10.1016/j.comcom.2021.12.021

[7] Khandait, P., Hubballi, N., & Mazumdar, B. (2021). IoTHunter: IoT network traffic classification using device specific keywords. *IET Networks*, *10*(2),59-75. https://doi.org/10.1049/ntw2.12007

[8] Iman, A., Mohammad, A., Salahuddin, L. V., Noura, L., Raouf, B., Bertrand, M., Stephanie, M., & Stephane, T. (2021). A Look Behind the Curtain: Traffic Classification in an Increasingly Encrypted Web. *Performance Evaluation Review*, *49*(1),23-24. https://doi.org/10.1145/3447382

[9] Michael, L., Wall, P. T., Gregory, Q., & Kaden, R. A. H. (2022). Tensor-network discriminator architecture for classification of quantum data on quantum computers. *Physical Review*, *105*(6),1-17. https://doi.org/10.48550/arXiv.2202.10911

[10] Roy, S., Shapira, T., & Shavitt, Y. (2022). Fast and lean encrypted Internet traffic classification. *Computer communications*, *186*(3), 166-173. https://doi.org/10.1016/j.comcom.2022.02.003

[11] Lateef, I. & Akansu, A. N. (2021). Machine learning in eigen subspace for network path identification and flow forecast. *IET communications*, *15*(15), 1997-2006. https://doi.org/10.1049/cmu2.12230

[12] Canard, S. & Li, C. (2021). Towards practical intrusion detection system over encrypted traffic. *IET information security*, *15*(3), 231-246. https://doi.org/10.1049/ise2.12017

[13] Guannan, H. & Kensuke, F. (2023). Characterizing Privacy Leakage in Encrypted DNS Traffic. *IEICE Transactions on communications*, *106*(2), 156-165. https://doi.org/10.1587/transcom.2022EBP3014

[14] Montieri, A., Bovenzi, G., Aceto, G., Ciuonzo, D., Persico, V., & Pescape, A. (2021). Packet-level prediction of mobile-app traffic using multitask Deep Learning. *Computer networks*, *200*(9), 1-23. https://doi.org/10.1016/j.comnet.2021.108529

[15] Kiana, D. & Masaki, B. (2021). Bandwidth Efficient IoT Traffic Shaping Technique for Protecting Smart Home Privacy from Data Breaches in Wireless LAN. IEICE *Transactions on communications*, *104*(8), 961-973. https://doi.org/10.1587/transcom.2020EBP3182

[16] Dinh, D. L., Nguyen, H. N., Thai, H. T., & Le, K. H. (2021). Towards AI-Based Traffic Counting System with Edge Computing. *Journal of advanced transportation*, *2021*(5), 5551976.1-5551976.15. https://doi.org/10.1155/2021/5551976

[17] Thoma, M. & Hadjicostis, C. N. (2021). Detection of collaborative misbehaviour in distributed cyber-attacks. *Computer communications*, *174*(6), 28-41. https://doi.org/10.1016/j.comcom.2021.04.013

[18] Verma, J., Bhandari, A., & Singh, G. (2022). INIDS: SWOT Analysis and TOWS Inferences of State-of-the-Art NIDS solutions for the development of Intelligent Network Intrusion Detection System. *Computer communications*, *195*(11), 227-247. https://doi.org/10.1016/j.comcom.2022.08.022

[19] Andresini, G., Appice, A., & Malerba, D. (2021). Nearest cluster-based intrusion detection through convolutional neural networks. *Knowledge-based systems*, *216*(15), 106798. https://doi.org/10.1016/j.knosys.2021.106798

[20] Li, B., Nie, Z. L., & Chang, J. Y. (2021). Embedded Heterogeneous Internet of Things Ciphertext Data Dynamic Capture Method. *Computer Simulation*, *38*(2), 282-286. https://doi.org/10.3969/j.issn.1006-9348.2021.02.061

Contact information:

Xiaoqing YANG,
(Corresponding author)
Faculty of Computer Engineering,
Shanxi Vocational University of Engineering Science and Technology,
No. 369, Wenhua Street, Yuci District, Jinzhong City, Shanxi Province,
030619, China
E-mail: yangxiaoqing@sxgkd.edu.cn

Niwat ANGKAWISITTPAN,
Research Unit for Electrical and Computer Engineering Technology (RECENT),
Mahasarakham University,
No. 41/20, Kantarawichai District, Maha Sarakham, 44150, Thailand
E-mail: niwat.a@msu.ac.th