Research on Network Intrusion Detection Method Based on Machine Learning

Guorong Xu*
Nanjing ZTE Software Co.Ltd, Nanjing, 210000, China
*Corresponding author’s e-mail: xu.guorong@zte.com.cn

Abstract. After years of development, the Internet has penetrated into all areas of society, and various network intrusions occur almost every day. In order to enhance the security of the network, intrusion detection technology has received more and more attention. Network intrusion detection is a branch of computer security. Its goal is to automatically and effectively detect intrusion traffic in the network and provide timely warnings. Network intrusion detection technology has become an important means to resist network intrusion in recent years, and network intrusion detection systems of various scales have been applied in the network. Due to the diversity of intrusion methods, traditional digital authentication and firewalls and other network security measures are difficult to meet the needs of network intrusion detection. The machine learning boom has gradually emerged, and machine learning methods have begun to be applied to the field of intrusion detection, which has become a research hotspot in this field. This article discusses machine learning methods, discusses adding machine learning algorithms to the network intrusion detection link to achieve effective detection of network intrusion traffic, so as to obtain a good network intrusion detection effect.

1. Introduction
Computer network not only brings us a lot of convenience, but also creates a series of security problems. With the continuous increase in the amount of data in the Internet and the continuous innovation of attack types, network security issues have become more and more prominent. Attacks by foreign hackers on websites in China are more common, and the consequences of data information leakage caused by attacks have become more serious. Intrusion activities mainly attempt to access or modify information, or damage the function of the system. When detecting intrusion activities, the main task is to monitor and analyze user, system and network activities, identify vulnerabilities, identify attack methods and analyze abnormal activities [1]. In response to the continuously deteriorating network environment, corresponding network protection technologies have emerged. The most commonly used network protection technologies include firewalls, data encryption, virus protection, identity verification, virtual private networks, authentication, and digital signatures. These technologies can play a role in security protection to a certain extent, but as static passive defensive network security tools, their service methods are relatively simple, and they can only protect against known attacks, and it is difficult to protect against unknown attacks. Therefore, it is difficult to protect against unknown attacks. Respond to complex and changeable network intrusions. In order to make up for the shortcomings of static protection technology, an intrusion detection system appeared, which is a new type of network defense technology for real-time monitoring of network transmission [2]. In recent years, machine learning methods have gradually been applied to network intrusion detection. Relevant research is based on machine learning and data mining theory, fusing different alarm information, mining the internal connections between compound...
attack sequences, presenting the scenarios corresponding to the attack process in a logical structure, and taking more efficient alarms for network managers the response plan provides help [3].

2. Network intrusion detection and machine learning

2.1. Network intrusion detection

Network intrusion refers to an attempt to access or manipulate information or damage the system without authorization, thereby causing the system to be unreliable or unusable. Intrusion detection refers to the process of collecting and analyzing key information from a network or computer system (including internal and external), so as to speculate whether there is an attack in the computer network. Network intrusion events originate from the monitored system or network, including network traffic information, mail entering and leaving the network, network packet capture information, and various log data [4].

Intrusion detection can be divided into host-based intrusion detection, network-based intrusion detection and hybrid intrusion detection. The host-based intrusion detection technology is mainly used in earlier intrusion detection systems, and is not affected by the exchange network and encrypted network information flow [5]. It is very effective in detecting attacks from within the system. Network-based intrusion detection obtains information by setting up traffic monitoring sites on important network segments, monitors and analyzes network data packets, and judges intrusions. Most intrusion detection systems are network-based, and they can successfully detect intrusions that the previous intrusion detection system cannot detect. The hybrid intrusion detection system combines these two types of intrusion detection systems to form a complete active defense system. The common network detection process is shown in Figure 1.

![Figure 1. Common network intrusion detection process](image)

2.2. Machine learning methods

Machine learning refers to the use of computers to simulate human learning activities, how to use computers to recognize existing knowledge, acquire new knowledge and continuously improve the level of cognition, so that the machine itself has the ability to continuously learn. Machine learning is interrelated with multiple disciplines such as artificial intelligence, mathematics, statistics, biology, image processing, and information theory, and is divided into supervised learning and unsupervised learning. Machine learning constructs a learning machine from existing experience, and further classifies or predicts unknown data through the learning machine is the goal of machine learning [6].

The supervised learning algorithm analyzes the training data and generates an inference function that can be used to draw new examples. The best solution will allow the algorithm to correctly determine the class labels of those invisible instances. In the network intrusion detection based on supervised learning, the learning result has the only test standard, which is to complete the learning task under the supervision of the mark [7]. The classification algorithms commonly used in network intrusion detection technology mainly include naive Bayes, decision trees, support vector machines, logistic regression, and K-nearest neighbor classification algorithms.

Unsupervised learning algorithms need to automatically find the data structure and features in these unlabeled data. Unsupervised learning attempts to find hidden structures in unlabeled data. Since the examples provided to the learning machine are unlabeled, there is no test data set to evaluate potential solutions. Many methods used in unsupervised learning come from data analysis and data mining methods, and clustering methods are a typical representative of unsupervised learning.
2.3. Network intrusion detection based on machine learning

Machine learning intrusion detection is different from traditional network intrusion detection. It uses machine learning or expert knowledge to define the network attack behavior process in advance, and establishes a knowledge base of attack behavior association rules by mining the attack sequence, and uses rule association analysis technology to identify and attack instances with consistent attack knowledge defined in the basic alert, and then associate and match the new alert information in the network with the attack rules in the database, and finally construct an attack process between two [8].

The use of machine learning methods for network intrusion detection is an important application scenario of machine learning methods. It has a basic fixed process and can be divided into five stages: data cleaning, feature selection, imbalance processing, classifier training, and classification result evaluation. The network intrusion detection process based on machine learning is shown in Figure 2.

![Network intrusion detection process based on machine learning](image)

3. Intrusion detection based on single-step attack

At present, the intrusion detection algorithm researched by academia is basically based on the KDD99 data set. From a data point of view, the training and testing of machine learning classifiers requires a large amount of manually labeled data, that is, data with category labels. However, in the real network environment, a label that connects to the data cannot be obtained. The KDD99 data set is close to the real network environment through simulation. Normal and attack traffic are artificially added. Therefore, the connections in the KDD99 data set are labeled with labels. Some commercial data sets used in actual intrusion detection are not disclosed [9]. Capturing network traffic packets is the data source of the entire intrusion detection system. The open source traffic capture tools on the Internet are becoming more and more powerful. This article selects Win Pcap attacks for data collection in the traffic capture link.

3.1. Network packet capture

The network monitoring device has been set in the current capture parameters, then go directly to the next step. If it is not set, find all the network ports of the machine and decide which port to use for capturing. The listening device port can also be specified in the command mode. Initialize the set promiscuous mode of the monitoring device, capture all traffic data packets passing through the device, distinguish different monitoring devices, and then all operations are performed through the network monitoring device. Capture the network address number of the specified network device and the corresponding mask traffic packet. Create a set of filtering rules. The set of rules is saved in a string and needs to be converted into a format that can be recognized by the Win Pcap engine. You can also filter some required packets by defining the rules for capturing packets through the user interface. The capture function module enters the main loop to realize the capture of the data packet. At the same time, the captured data packet can be filtered through the kernel filter program by defining the filtering mechanism [10]. The previous step is to set the filtering conditions before the capture, and this filtering...
is Filter the captured data packets. After the data packet capture parameters are set, the captured data packets are decomposed. The process of capturing network traffic data packets is shown in Figure 3.

![Flowchart](image)

Figure 3. Process of capturing network traffic packets

3.2. Model building and data training

Construct an attack tree model to describe the behavior sequence of network sessions in the system. Attack behavior trees are divided into two types: AND nodes and OR nodes. Generally, the attack tree contains any combination of AND and OR, and a separate attack tree behavior sequence is generated in a depth-first traversal manner. After the feature reduction processing of the captured data set is completed, the input of the attack tree generation algorithm is the processed data set.

The idea of a single-step attack is to divide the link records of the network session into feature sets according to attributes, and add identification to the feature set according to the attack classification label, thereby establishing a training data set, and the unidentified data is used as the test data set. Use a supervised machine learning method to train the training set to establish a single-step attack intrusion detection model.

The data set of intrusion detection consists of training data with identification and test data without identification. Each connection record in the training data set contains 39 fixed feature attributes and 1 class identifier. Each feature provides a classification basis. The data features contained in each data record in the data set are within different value ranges. The data preprocessing stage discretizes and normalizes the continuous features of the data set. The intrusion detection performance of the decision tree, random forest, logistic regression, and naive Bayes algorithm in the supervised machine learning algorithm is compared and analyzed in the data set, and the algorithm with the best result is selected as the weak classifier of the integrated algorithm. The precision rate and recall rate are used as evaluation indicators, and F1-score is used as a comprehensive indicator to reconcile precision rate and recall rate to evaluate whether the Gradient Boost Decision Tree model algorithm helps to improve the accuracy of intrusion detection. The detection results of the machine learning algorithm are shown in Table 1.

| Table 1. Machine learning algorithm detection results |
|----------------|----------------|----------------|----------------|----------------|
| Results        | recall          | precision      | accuracy       | F1-score       | support        |
| Naive Bayes    | 0.82            | 0.88           | 0.8091         | 0.83           | 311029         |
| Logistic       | 0.91            | 0.93           | 0.9207         | 0.91           | 311029         |
| Regression     | Decision Tree   | 0.92            | 0.93           | 0.9223         | 0.92           | 311029         |
| Random Forest  | 0.93            | 0.93           | 0.9217         | 0.91           | 311029         |

It can be seen from Table 1 that the decision tree algorithm has the highest comprehensive index compared to other classic algorithms in the intrusion detection results. Therefore, the decision tree algorithm is selected as the weak classifier of the integrated algorithm to establish the intrusion model. The maximum number of weak learners in the framework and the weight reduction coefficient of each weak learner are jointly adjusted to find the best fit effect. For the same result of the training set prediction effect, the reduction of the weight reduction coefficient means that the basis is required. More iterations of the learner.
4. SVM network intrusion detection based on clustering algorithm

The intrusion detection algorithm based on machine learning is basically to design a classifier. For a network connection, after counting its characteristics, the trained classifier calculates the category of the connection (normal connection or attack connection) to realize the attack Behavioral alarm. SVM is a machine learning method with good performance, and it has also achieved good results when applied to network intrusion detection. In order to further improve the network intrusion detection effect of the intrusion detection system, you can try to combine SVM with other technologies to improve the existing SVM-based network intrusion detection method.

As a classic clustering algorithm, K-means can effectively cluster large-scale data, but the traditional K-means method needs to set the number of target categories for clustering in advance. Most of the time, it is difficult to set the value based on experience. Get it properly. In order to improve the performance of the algorithm, setting the number of initial cluster centers and target categories has become the research focus of improving the method. In this paper, based on the improved K-means method of distance threshold, we study SVM network intrusion detection technology.

The maximum allowable distance between the cluster center and the sample points in the data set is called the distance threshold. Randomly select a point in the sample set as the initial cluster center point. If the distance between the sample point and the cluster center is less than the distance threshold, it is determined that the sample point belongs to this cluster. If the distance between the two is greater than the distance threshold, then create a new cluster with this sample point as the cluster center. According to this process, a complete traversal of the samples in the data set is achieved. The biggest advantage of this improved method is that it can automatically obtain the number of clustering categories. By setting the distance threshold, the training set samples are clustered into multiple new sets, and the cluster center of each set is used as a new training sample. After the clustering is completed, a new training set will be generated. After removing the redundant data in the training set, a small amount of key data is extracted as the new training set data for subsequent training of the classifier.

We use the new training set formed after clustering to train the SVM classifier, and compare the final detection results with the results before the introduction of the clustering method. After the introduction of the clustering method, the SVM classifier is effective against Probe, U2R, and R2L attacks. The detection rate has increased. The detection rate of each category before and after clustering is shown in Figure 4.

![Figure 4. Comparison of detection rates before and after clustering](image)
5. Conclusion
The network intrusion detection technology that emerged at the end of the last century has become an important means to resist network intrusion in recent years. The network intrusion detection system based on machine learning is a subversion of the traditional network intrusion detection system. By learning a large amount of training data containing attack samples, an intrusion detection model that can distinguish between normal traffic and attack traffic is established, which greatly improves intrusion detection system performance. In the event collection stage of network intrusion detection, this paper adds feature engineering in machine learning to research. The application of machine learning methods in the field of intrusion detection is introduced, including attack type identification, evaluation criteria, and traditional single-step attack analysis algorithms. Use the improved K-means algorithm to filter the data in the training set, remove redundant data and reduce the complexity of model training. Combining the improved K-means algorithm with support vector machines is a better classification method, which can improve the efficiency of network intrusion detection.

References
[1] Zhang X 2018 Network intrusion detection based on machine learning algorithm[J]. Modern Electronic Technology, 41(03): 124-127.
[2] Luo F C, Wu F and Chen Q, et al. 2020 Intrusion detection algorithm for wireless sensor networks based on machine learning. Journal of Harbin Engineering University, 41(03):433-440.
[3] Liu Z Y 2019 Application of support vector machine algorithm in ship network intrusion detection. Ship Science and Technology, 41(18): 169-171.
[4] Wang Z Y 2019 Network intrusion detection based on k-nearest neighbor density peak clustering hybrid algorithm. Automation Technology and Application, 38(12): 48-52.
[5] Zhu P Z 2019 Network intrusion detection method based on machine learning. Journal of Hunan Vocational and Technical College of Industry, 19(06): 8-12.
[6] Yu H H, Zhou F Y, Chen M M 2019 Analysis of KDD-CUP99 network intrusion detection data set based on machine learning. Computer Engineering and Science, 41(S1): 91-97.
[7] Zhou F F 2020 Experimental research on network intrusion detection based on support vector machine method. Journal of Southwest China Normal University (Natural Science Edition), 45(01): 57-61.
[8] Zhu J and Wu P 2020 Research on network intrusion detection of shipbuilding industry based on deep learning. Ship Science and Technology, 42(09):181-183.
[9] Su M and Ma L 2020 Network intrusion detection based on ant colony algorithm selecting neural network parameters[J]. Modern Electronic Technology, 43(22): 114-117.
[10] Song Y, Hou B N and Cai Z P 2021 Network intrusion detection method based on deep learning feature extraction [J/OL]. Journal of Huazhong University of Science and Technology (Natural Science Edition): 1-7[2021-02-15]. https://doi.org/10.13245/j.hust.210219.