Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2020.010406
ISSN 2790-0983 Vol. 1, Issue 4: 45-52

Research Information Security Technology Based on Machine Learning

Jiayao Ji*

The People's Procuratorate of Shanghai Hudong District, Hongkou, Shanghai, China

jijiayao1988@gmail.com

**corresponding author*

Keywords: Machine Learning, Network Security, Information Security Technology, Detection System

Abstract: With the rapid development and widespread popularity of the Internet, the types and quantities of abnormal traffic are also increasing day by day. Network intrusion detection has become an important fortress to ensure network information security. People put forward more and more urgent demands for the robustness and development of network security technology. Therefore, this paper studies information security technology based on machine learning. In this paper, the concepts of network security situation assessment and network intrusion detection are described in detail at first, then the machine based learning data preprocessing module, network vulnerability level and coarse and fine granularity hybrid detection system are designed, and finally the detection performance and defense performance of fine granularity hybrid detection system are analyzed in detail, and a conclusion is drawn.

1. Introduction

In recent years, information security technology has gradually matured and been widely used [1]. Network security situational awareness can help network security managers effectively grasp the security situation of the network, timely discover the security problems of the network system, and predict the future security situation of the network, providing a basis for making network security decisions [2-3]. From the current state of network security, information security technology needs to be developed more vigorously to ensure the security of the network, otherwise it will cause greater losses [4-5]. How to detect abnormal network traffic from the network intrusion data and stop the damage in time is the current difficulty to overcome [6]. For this kind of problem, a popular and effective approach is to combine machine learning to build a machine learning intrusion detection model based on network traffic data to efficiently identify network traffic data with attacks [7-8].

With the advent of the Internet era, the development of information security technology for network security is becoming increasingly serious, and a large number of researchers have conducted research on information security technology based on machine learning. For example, Dieudonne Tchuente et al. considered that the attack signal is unknown and used neural network techniques to approximate the attack signal eliminating the assumption that the attack signal has a known upper limit, and by combining state feedback with the estimated information of the attack can effectively compensate for the impact of the attack, and the results of the study showed that the theoretical results are valid [9]. Omer Faruk Beyca et al. proposed an active learning-based system that collaborates with experts to periodically classify incoming signatures and rank them through uncertainty sampling, showing that the system with MC decay performs best by passing more samples of "medium" importance to the experts and mitigating the imbalance in the training data imbalance in the training data set [10]. Machine learning-based information security technologies can contribute to the development of cyber security

At present, network intrusion attacks are on the rise, and if left unchecked can eventually lead to serious consequences, so this paper presents an in-depth study of information security technology based on machine learning. The research content of this paper includes three parts: the first part is an overview, including an overview of network security situation assessment and network intrusion detection; the second part is the system design, mainly divided into three parts: data pre-processing module design, network vulnerability level design and coarse and fine-grained hybrid detection system design; the third part is the system analysis part, mainly has two aspects of analysis, detection performance analysis and defence performance analysis.

2. Related Overview

2.1. Overview of Network Security Situation Assessment

Situation awareness is based on correct situation understanding. The feedback of situation understanding results is an accurate expression of the real state of things. The presentation of situation awareness results is the process of situation assessment. The focus of situation awareness is situation assessment [11]. The situation assessment system can timely find hidden dangers and threats in various situation elements, generate situation values that can accurately evaluate the network security status, and then predict the change trend of the network security status in a long period of time, so that network managers can timely master the network security situation [12]. For possible future security events, we should build protective measures and cut off attack sources as soon as possible, so as to ensure the normal operation of the network and reduce unnecessary economic losses [13].

2.2. Overview of Intrusion Detection

The event database is used to store data information in the detection process, which can store a large amount of data information [14]. The workflow of intrusion detection mainly includes the steps shown in Figure 1, and the workflow is shown in Figure 1.

(1) Information collection: The main purpose is to collect information generated in the computer network system, and the content of the collected data mainly includes log files, data streams and user activity status [15-16].

(2) Data processing: the collected data usually has a lot of noise, and in order to learn useful information from the data efficiently, the data needs to be serialized, normalized and feature-dimensionalized [17].

(3) Information analysis: This step is the core part of the process. The collected data needs to be

detected by a suitable and efficient model, so the optimization of the model is crucial in this step [18].

(4) Alarm and response: Based on the decision configuration information of this module, the detection results are responded to.

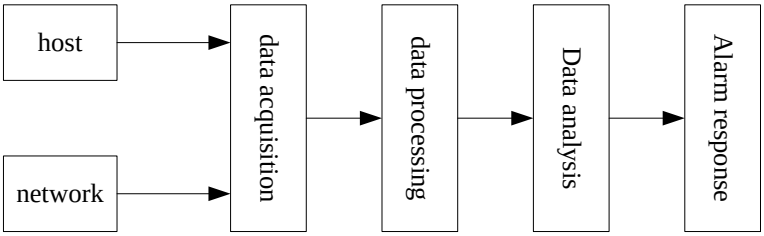

Figure 1. Intrusion detection workflow

3. System Design

3.1. Design of Data Preprocessing Module

In order to get accurate situation awareness results, it is necessary to comprehensively collect system data information, and the network system is constantly running, producing different data at all times. Multiple sensors are needed to obtain system data, but the data format obtained by different sensors is different, and the amount of data obtained is huge, not all of which are useful information, and can be used only after screening. Pre-processing the collected data can improve the analysis efficiency. Check whether the data is missing and in line with the actual situation, and determine whether the data needs to be reprocessed and reprocessed. Modify the data with errors, analyze the data that cannot be repaired, and select only the useful data. Data is transformed to form standardized data

3.2. Network Vulnerability Level Design

Vulnerability is the main way for hackers to attack the target network. In the network security situation assessment system, the value of network security situation largely depends on the level of network vulnerability. The calculation of network vulnerability level mainly depends on the configuration of the network and the types and numbers of vulnerabilities in the network. The specific quantitative formula is as follows:

$$R = \frac{\sum_{x=1}^{b} \sum_{y=1}^{a} T_{xy} D_x J_{xy}}{K} \tag{1}$$

R is the vulnerability level, b is the total number of hosts and servers, a is the total number of vulnerability categories, K is the total number of vulnerabilities, Txy is the level factor corresponding to the yth vulnerability in the xth server, the specific value can be found in the vulnerability scoring system, Jxy is the number of the yth vulnerability in the xth server, Dx is the importance of the xth server, and the specific formula of Dx is as follows:

$$D_x = \frac{H_x}{\sum_{x=1}^{b} H_x} \tag{2}$$

3.3. Design of Coarse and Fine Particle Mixed Detection System

The network anomaly detection system based on machine learning with coarse and fine granularity can effectively save computing cost and improve detection efficiency while ensuring accuracy. The fine particle size detection is used to avoid the masking effect and marsh abuse in the coarse particle size detection, and improve the detection accuracy. The system flow is shown in Figure 2.

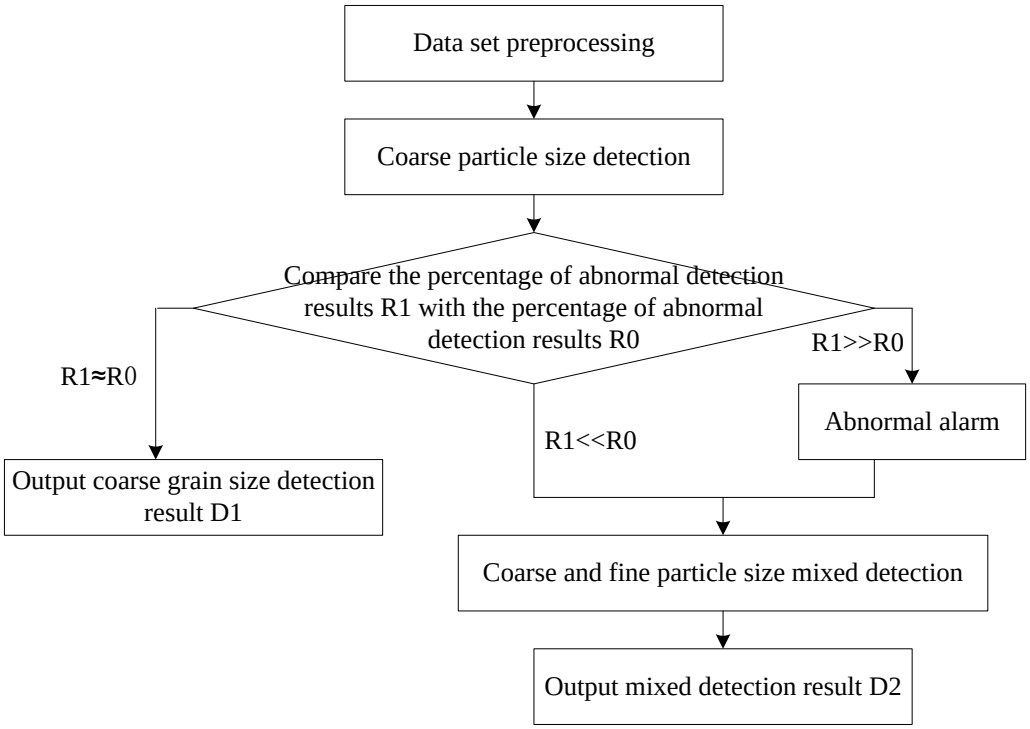

Figure 2. Network anomaly detection system flow chart

Step 1: preprocess the training set and test set of network traffic data to eliminate the features irrelevant to anomaly detection, so as to improve the detection speed.

Step 2: Use the coarse particle size detection method. The pre-processed network traffic data training set is trained in an unsupervised manner. The adopted coarse-grained detection method does not calculate the specific distance between data samples, and uses a single threshold as the identification basis to obtain coarse-grained detection data sets D1 and R1. R1 refers to the proportion of abnormal samples detected in the total number of samples. Coarse grained detection method saves a lot of computing cost caused by distance computing.

Step 3: Make statistics in advance of the proportion of exceptions in the network environment under general conditions, which is recorded as R0. Step 2 After coarse particle size detection, the percentage of abnormal detection is obtained, which is recorded as R1. If R1 $\approx$ R0, go to step 4; If R1>>R0, it indicates that the abnormal proportion is high, turn to step 5 and alarm; When R1<<R0, the network quality seems to be better than the normal value, but this also indicates that the network may not be able to detect exceptions, which requires mixed coarse and fine granularity detection.

Step 4: Compare R1 and R0. If the result is R1 $\approx$ R0, it indicates that the current network condition is normal. No fine particle detection is required, and the coarse particle detection result D1 is directly output

Step 5: This step is triggered. After comparing R1 with R0, R1 is greater than R0, which indicates that the current network state is detected as abnormal. The system sends an abnormal

alarm to isolate the abnormal device, and stores the network traffic data for further mixed coarse and fine granularity detection.

Step 6: First, use coarse particle size detection to distinguish a large number of positive samples with high normality. After reducing the number of positive samples, the gap between the number of positive and negative samples can be narrowed, and at the same time, the outliers are isolated, making the distribution difference of data sets smaller; The fine particle detection can effectively improve the detection effect.

Step 7: After the coarse and fine particle mixed detection in step 6, output the coarse and fine particle mixed detection result D2.

4. System Analysis

4.1. Test Performance Analysis

In the experiment, the isolated forest algorithm is selected as the coarse granularity detection algorithm, and the Kmeans algorithm is selected as the fine granularity detection algorithm. In the pre experiment, three isolated forest models are established according to the training samples. Anomaly detection is essentially a classification problem, and accuracy, accuracy and recall are the basic indicators to judge whether the classifier is good or not. See Table 1 for the detection results of coarse and fine grain mixing anomaly.

Table 1. Detection results of coarse and fine grain mixed anomaly

	Precision(%)	Accuracy(%)	Recall(%)
Test Set 1	97.6	96.8	92.7
Test Set 3	99.4	98.7	94.9

It can be seen from Table 1 that the accuracy rates of test set 1 and test set 3 are both above 95%, which indicates that the traffic samples predicted as normal in the test set are basically accurate, and the recall rate is also above 90%, which proves that more than 90% of all normal samples are predicted correctly. The overall accuracy rate is more than 95%, which proves that the anomaly detection scheme can achieve more than 95% accuracy in both isolating abnormal traffic and detecting normal data, and the accuracy rate is high. It proves that the coarse and fine grain hybrid anomaly detection algorithm is basically correct in judging each sample in anomaly detection, and there is almost no misjudgment. The low accuracy and recall rate indicates that there is some missing judgment. Test set 1 and test set 3 simulate the sampling of normal network conditions. The proportion of exceptions is low, and the advantages of coarse and fine grain mixed detection are not obvious. Therefore, three groups of data sets are randomly selected to simulate the attacked network conditions. See Table 2 for the test results of the mixed coarse and fine particle size detection scheme and Table 3 for the test results of the single coarse particle size detection scheme.

Table 2 Test results of coarse and fine particle size mixed testing scheme

	Test Set 2	Test Set 5	Test Set 7
Precision (%)	95.4	97.9	97.4
Accuracy (%)	96.1	95.3	96.8
Recall (%)	92.7	93.8	95.3

It can be seen from the data in Table 2 that although the accuracy rate of coarse and fine grain mixed detection is lower than that of test sets 1 and 3, it still remains at more than 93%, and the accuracy is consistent with the original, without much decline. It can be seen from Table 3 that the results of anomaly detection using coarse grain size alone for the three test data simulating abnormal conditions are not ideal. The average accuracy of the three test sets is 88.1%, and the

accuracy is 86.53%. The main reason is that when using a single threshold to judge, it is impossible to accurately distinguish the data results in the middle, which is affected by the swamping and masking effects.

Table 3. Test results of using coarse grain inspection scheme alone

	Test Set 2	Test Set 5	Test Set 7
Precision (%)	88.7	85.8	89.8
Accuracy (%)	86.5	87.7	85.4
Recall (%)	83.4	84.2	81.6

4.2. Defense Performance Analysis

The experimental results of the defense performance test are shown in Figure 3. Analyzing the contents of the figure, the horizontal axis corresponds to the packet number obtained for the data collection module, the vertical axis corresponds to the destination IP entropy, the red line corresponds to the change of the destination IP entropy when there is a DDoS attack on the network, and the blue line corresponds to the change of the source IP entropy after the location defense module is opened when there is an attack on the network.

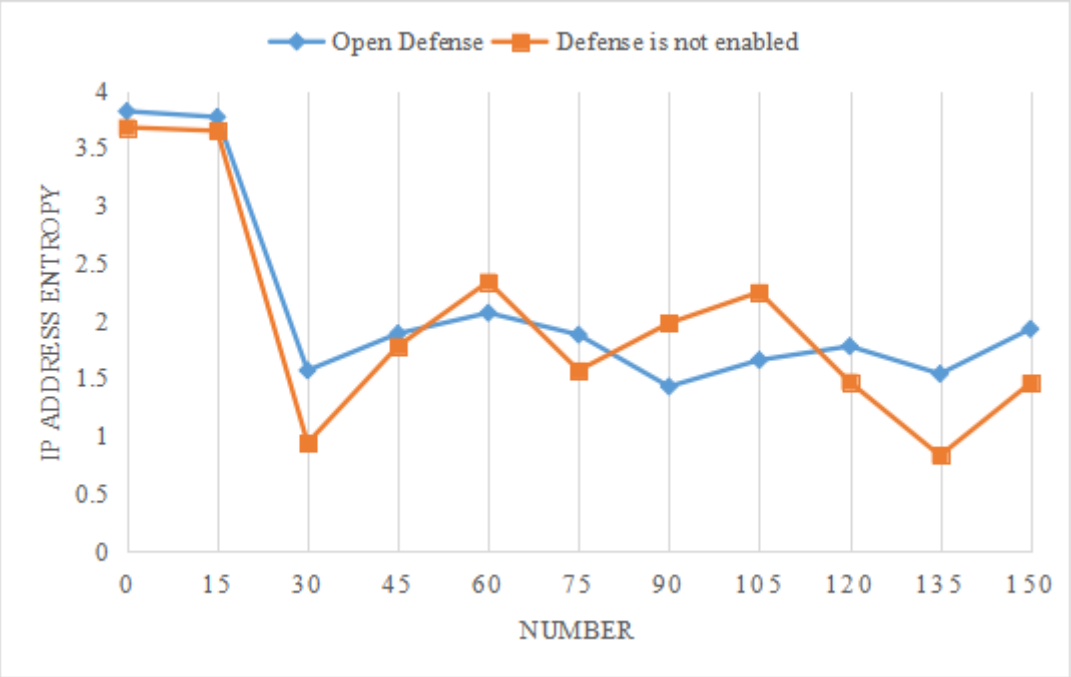

Figure 3. Effect of the defense module on the destination IP address entropy

The blue line in the figure represents the change of the entropy of the destination IP address when the defense module is turned on, and the red line represents the change of the entropy of the destination IP address when the defense module is not turned on. It can be seen from Figure 3 that the No. 30 packet starts to suffer from DDoS attacks. At this time, the destination IP gradually decreases and finally reaches the minimum entropy value. Comparing it with the normal network state, the entropy value of the destination IP decreases by 2.25. From the perspective of the entropy value change of the destination IP, the network has obvious anomalies. At this time, the initial detection module detects abnormal traffic and wakes up the coarse and fine grain mixed detection module. After the coarse and fine grain mixed detection module determines the attack traffic, it starts the defense module.

5. Conclusion

With the development of the Internet, network intrusions occur frequently. It is necessary to study the information security technology to maintain network security. Therefore, this paper studies information security technology based on machine learning to maintain network security. Through the analysis of detection performance, it is found that the coarse and fine granularity hybrid network anomaly detection system based on machine learning has high detection accuracy and precision, and good detection performance, but the accuracy of using coarse granularity detection scheme alone needs to be further improved. Through the analysis of the defense performance results, it is found that the system defense performance is good. After the coarse and fine granularity mixed detection module determines the attack traffic, start the defense module. There are many aspects to be improved in this paper.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Ons Aouedi , Kandaraj Piamrat, Salima Hamma, Menuka Perera Jayasuriya Kuranage: Network traffic analysis using machine learning: an unsupervised approach to understand and slice your network. Ann. Des Telecommunications 77(5-6): 297-309 (2020).

[2] Abigail Goldsteen, Gilad Ezov, Ron Shmelkin, Micha Moffie, Ariel Farkash: Data minimization for GDPR compliance in machine learning models. Al Ethics 2(3): 477-491 (2020).

[3] Suresh Dara, Swetha Dhamercherla, Surender Singh Jadav, Ch Madhu Babu, Mohamed Jawed Ahsan : Machine Learning in Drug Discovery: A Review. Artif. Intell. Rev. 55(3): 1947-1999 (2020).

[4] Leandro Miranda, Jose Viterbo, Flavia Bernardini: A survey on the use of machine learning methods in context-aware middlewares for human activity recognition. Artif. Intell. Rev. 55(4): 3369-3400 (2020).

[5] Muhammad Waqas, Shanshan Tu, Zahid Halim, Sadaqat ur Rehman, Ghulam Abbas, Ziaul Haq Abbas: The role of artificial intelligence and machine learning in wireless networks security: principle, practice and challenges. Artif. Intell. Rev. 55(7): 5215-5261 (2020).

[6] Simon Penny: Review of Art in the Age of Machine Learning by Sofian Audry. Artif. Life 28(1): 167-169 (2020). https://doi.org/10.1162/artl_r_00352

[7] Zied Ftiti, Kais Tissaoui, Sahbi Boubaker: On the relationship between oil and gas markets: a new forecasting framework based on a machine learning approach. Ann. Oper. Res.313(2): 915-943 (2020). https://doi.org/10.1007/s10479-020-03652-2

[8] Dieudonne Tchuente, Serge Nyawa: Real estate price estimation in French cities using geocoding and machine learning. Ann. Oper. Res. 308(1): 571-608 (2020).

[9] Omer Faruk Beyca, Ibrahim Yazici, Omer Faruk Gurcan, Halil Zaim, Dursun Delen,Selim Zaim: A comparative analysis of machine learning techniques and fuzzy analytic hierarchy process to determine the tacit knowledge criteria. Ann. Oper. Res.308(1): 753-776 (2020). https://doi.org/10.1007/s10479-020-03697-3

[10] Koushiki Dasgypta Chaudhuri, Bugra Alkan: A hybrid extreme learning machine model with harris hawks optimisation algorithm: an optimised model for product demand forecasting applications. Appl.Intell. 52(10): 11489-11505 (2020).

[11] Pradip Dhal, Chandrashekhar Azad: A comprehensive survey on feature selection in the various fields of machine learning. Appl. Intell. 52(4): 4543-4581 (2020).

[12] Marta Fernandes , Juan Manuel Corchado, Goreti Marreiros: Machine learning techniques applied to mechanical fault diagnosis and fault prognosis in the context of real industrial manufacturing use-cases: a systematic literature review. Appl. Intell. 52(12): 14246-14280 (2020).

[13] Ekaterina Gurina, Nikita Klyuchnikov, Ksenia Antipova, Dmitry A. Koroteev: Forecasting the abnormal events at well drilling with machine learning. Appl. Intell. 52(9): 9980-9995 (2020).

[14] Elena Hernandez-Pereira, Oscar Fontenla-Romero, Veronica Bolon-Canedo, Brais Cancela-Barizo, Bertha Guijarro-Berdinas, Amparo Alonso-Betanzos: Machine learning techniques to predict different levels of hospital care of CoVid-19. Appl. Intell. 52(6): 6413-6431 (2020).

[15] luri Krak, Olexander Barmak , Eduard Manziuk : Using visual analytics to develop human and machine-centric models: A review of approaches and proposed information technology. Comput. Intell. 38(3): 921-946 (2020). https://doi.org/10.1111/coin.12289

[16] Atika Qazi o, Najmul Hasan, Olusola Abayomi-Alli, Glenn Hardaker o, Ronny Scherer, Yeahia Sarker, Sanjoy Kumar Paul, Jaafar Zubairu Maitama: Gender differences in information and communication technology use & skills: a systematic review and meta-analysis. Educ. Inf. Technol. 27(3): 4225-4258 (2020).

[17] Ichiro Ide, Huynh Thi Thanh Binh: Special issue on "The Eighth International Symposium on Information and Communication Technology- SolCT 2017". J. Heuristics 28(2): 147-148 (2020).