Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2020.010203
ISSN 2790-0983 Vol. 1, Issue 2: 19-27

Malicious Network Attack and Intrusion Based on Decision Tree Algorithm

Yupeng Sang*

College of Information and Technology, Wenzhou Business College, Wenzhou 325035, China

sangyupeng@wzbc.edu.cn

**corresponding author*

Keywords: Decision Tree Algorithm, Malicious Network, Network Attack, Network Intrusion

Abstract: With the development of the network, the attack methods have become diverse. Malicious attackers can steal users' personal information by mining information. Therefore, this paper intends to study the role of decision tree algorithm in malicious network attacks and intrusions. The purpose is to improve the detection and defense of network attacks through this algorithm to ensure the information security of users. This paper mainly uses the method of experimental comparison and experimental construction to deeply explore the application of decision tree algorithm and PCA feature extraction in the system. The experimental results show that the intrusion accuracy based on decision tree algorithm can reach more than 90%, and the accuracy of PCA feature extraction can increase by 3%.

1. Introduction

Because there is a large amount of data information in the Internet, which is large in quantity, various in variety and complex. However, the maturity of hacker technology and the lack of management ability have led to many large-scale illegal website intrusions and spread to various fields. Illegal elements use these loopholes to cause serious harm to the vast number of Internet users, and wantonly destroy the network system platform. Computer network security technology involves two major parts: network security and user privacy protection. Decision tree algorithm is a new method in data mining. It deals with a large number of complex system problems by constructing a model.

There are many researches on the malicious network attack and intrusion of decision tree algorithm. For example, some people propose an intrusion detection system based on decision tree classification algorithm to improve detection efficiency and detection accuracy [1-2]. Others described anomaly detection technology, misuse detection technology, host based and

network-based intrusion detection systems in detail according to different classification standards of intrusion detection [3-4]. In addition, some scholars provide an intrusion detection system model based on Spark and data mining and design a prototype system [5-6]. Therefore, network intrusion is a social problem, and the research on network intrusion has the significance of the times and social value.

In this paper, network attacks and related modeling methods are briefly described. Secondly, the intrusion detection system is studied deeply. The characteristics and main problems of intrusion detection are proposed. Then the classification based on intrusion detection decision tree is studied, and the decision tree algorithm can play a role in intrusion network. Finally, the ability of decision tree algorithm is highlighted through experiments, and relevant conclusions are drawn.

2. Attack and Intrusion against Malicious Network Based on Decision Tree Algorithm

2.1. Network Attack Modeling Method

In the process of network attack, the attacker can choose to obtain information from a specific object or host to destroy system resources or control the entire network. Network attack refers to that an attacker obtains information available on other computer terminal devices through the Internet without authorization. The intruder captures some security mechanisms such as vulnerabilities and malicious code. If a hacker attempts to steal user data, illegally accesses sensitive computers, and returns the results to the hacker. On the contrary, the attack is triggered after sending a request to the host, making the attacker unable to normally respond to the execution process of these instructions. There are various types of network attacks, such as hackers and viruses. They can destroy or destroy the entire system by managing and controlling the data flow. In this process, it may be affected by some uncertain factors and cannot work normally. In order to reduce the probability of this situation, we need to establish a model to predict the computer operation status and network topology and other information security problems in the future, and propose corresponding solutions to improve the effectiveness and reliability of network attacks [7=8].

Model-based network attack is a typical complex intrusion prevention method. It sets a limit on hackers so that attackers cannot bypass their ports, steal or destroy other target computers. The system can be divided into completely irrelevant type (such as host) and local type according to the attack object, purpose and mode. The first type is mainly used to protect and detect external equipment. The non integrity or semi authenticity model is that when there is no information in the network, attackers cannot obtain internal files through their ports and cannot destroy other targets. Model based attack refers to hiding, extracting and filtering the target object or other information, so as to realize the attacker, and implement defensive strategies to obtain effective survival. When building a network attack model, first determine a tested domain. According to the data type captured by the host, select the same attribute values in different categories as candidate nodes. Secondly, one kind of information that can reflect all possible behavior patterns or states should be selected as a candidate node. Finally, the information is sorted and a new combination is formed to achieve the prediction goal. Model based network attack (FagedPressure) refers to that an attacker accesses a given data set without authorization to obtain the target object or impose information resources on the target computer. There are two main methods. First, obtain the required information directly from the user's location. The other is through connection with other nodes in the system [9-10].

2.2. Intrusion Detection System

The attack behavior in the network is an uncertain factor. Under different circumstances, intrusion detection may occur. From the perspective of current malware analysis, there are mainly the following categories: host log information, illegal intrusion attacks, host parameter settings, etc. to determine whether there are security vulnerabilities or types of security events, and the size of threats. However, these methods can not well identify the intrusion and theft methods of hackers or trojans, as well as the losses caused by their destructive consequences. And it is impossible to accurately assess the possible risks between the attacker and the protected. The task of network security can be summarized into three aspects: protecting the confidentiality and integrity of the transmitted communication content, preventing the disclosure of information in the transmission process, and ensuring that only legally authorized users can access or obtain it. Authentication of user's legal identity, establishment of security association between users, and maintenance of trust relationship between users and users, users and systems. While ensuring the normal operation of network transmission, protect the processing capacity of the client system from being affected [11-12].

Intrusion detection can timely detect intrusion behaviors in the network and make appropriate responses. As a dynamic monitoring technology, it usually first probes important information in the system, and then scans the system for illegal behaviors or traces of being attacked, so as to find out the process of attempting to attack, being attacked, or having completed the attack [13-14]. Although the rapid development of information security technology has driven the rapid change of intrusion detection technology, the current intrusion detection system still has the following problems:

The problem of false positives and false negatives. As a pair of technical indicators, how to balance the false negatives and false negatives to achieve the optimal value as much as possible has always been the primary concern of intrusion detection systems. User privacy and data security issues. The working intrusion detection system monitors and analyzes the data in the network and system and records any abnormal situation in a timely manner, so as to ensure the normal operation of IDS and may cause security problems such as privacy disclosure [15-16]. In large distributed networks, it takes a long time to detect some attacks, which tests the scalability of intrusion detection mechanisms in terms of time. In the intrusion detection system, the automatic response module usually has a higher priority. The problems of the information analysis module, the expansion of the network scale, the gradual accumulation of the amount of data generated in the network, and the requirements for the efficient analysis and processing capability of the intrusion detection system are also increasing [17-18].

Intrusion detection mainly analyzes, judges and identifies malicious data in the network. This includes feature extraction and classification technology. Feature extraction is to ensure that attacks can find objects accurately. We generally need to obtain some specific log information to infer user behavior habits or characteristics and provide corresponding strategies. For text files, it is required that they have a certain amount of validity to detect the correlation between different types and orders of magnitude of networks, so as to form a complete and error free detection system.

2.3. Classification Based on intrusion Detection Decision Tree

In the decision tree algorithm, hidden information is discovered through data mining, and then decisions are made according to the results obtained It can be used to deal with a large number of complex problems. Based on the attack and intrusion analysis of decision tree algorithm, the data

set is divided, and then all objects that may be marked as different types are represented by tags. Next, select a category in each category. Finally, other target user groups in each category are calculated according to the attribute values. Each attribute set consists of three subsets. Among them, the random part represents the characteristic value of some regular changes. The random allocation part represents that the state of each event is a discrete distribution feature. The final result is determined by continuously adjusting the attribute characteristics and output values of each training unit. Judge whether there is significant difference with other grouping sets. If a new combination is generated, it has been marked as a malicious attack event. If it has the same attribute characteristics, it will be treated as a subset, otherwise, it will be regarded as an intrusion.

The spanning tree determines the nodes in the network by selecting a decision rule, and then integrates all possible and possible attack attribute sets into the tree structure diagram. The test log contains a lot of random data in each cluster. These data can be malicious information, abnormal information or error messages, and also include other types of events. All data packets from the original data set to the information flow sent by the end user may be triggered without authorization. If there are a lot of spam or incorrect file types in the network, the whole system will crash. Secondly, these data information should be classified and processed, and divided into different groups and assigned to corresponding locations, so as to mark the corresponding information list content according to specific addresses.

C4.5 algorithm uses information gain rate as the standard.

$$GR(X,Y) = \frac{G(X,Y)}{S(X,Y)} \tag{1}$$

Since Y is divided based on the value of type attribute X, S (X, Y) is the amount of information.

The attribute selection of C4.5 algorithm is based on the minimum information entropy. The entropy of set W is calculated as follows:

$$In(W) = -\sum_{i=1}^{1}((f(M_i,W)/|W| * \log_2 f(M_i,W)/|W| \tag{2}$$

Among them, $f(M_i,W)$ represents the number of samples belonging to the class (one of the possible classes) in the set W.

In the field of data mining, a very important aspect is to design experiments to evaluate actual mining algorithms. Only through extensive and systematic evaluation of various data mining algorithms on various data sets, can we have a comprehensive and profound understanding of a data mining algorithm, and can we choose an appropriate and effective mining algorithm for a specific problem.

3. Implementation of Decision Tree Algorithm in Network Intrusion Detection System

3.1. Experimental Environment

The setting of network security environment is mainly to ensure the normal and stable operation of the entire network system. Therefore, we need to classify attacks and make statistical analysis. Create a buffer between the host and the router to prevent malicious attackers from invading. There are protocol stack switches, firewalls and other different types of servers between the host and the router and the Internet. The experimental environment is Intel (R) Core (TM) 2 Duo CPU/4GB

120GB, and the operating system is Windows XP. The test environment is shown in Figure 1:

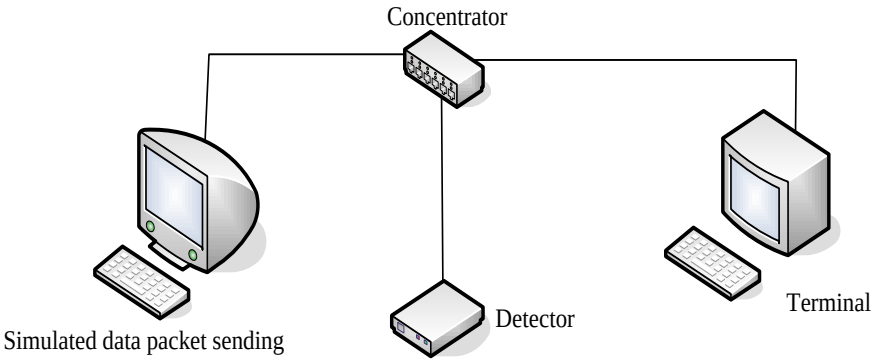

Figure 1. Experimental test environmental structure

Figure 1 is the simulation diagram of the test environment of the experiment. The sending of analog data packets is related experimental data, the detector is a decision tree classifier, and the terminal displays the classification results and test results.

3.2. Experimental Scheme

This experiment uses malicious programs to attack malicious data in the network, and uses intrusion tree algorithm and forward behavior technology to analyze whether different types of attacks exist. Two methods are mainly used in this experiment. One is predictive learning. The method is based on user behavior analysis, knowledge base construction and rule making to realize the reasoning calculation of training set. The second is to simulate anonymous verification strategy (BP neural network) and random scoring mechanism (ESO) to test network attacks and use them as a means of defense against malicious attackers, so as to finally achieve the purpose of intrusion prevention and control. This experiment is divided into three. The first is to use two test sets to judge the ability of C4.5 algorithm to detect unknown attacks. The second is to use C4. 5 algorithm as the decision tree classification algorithm. In addition, C4.5 algorithm is used to control PCA feature extraction variables to build a decision tree.

3.3. Test Indicators

The evaluation index of network attack refers to whether the attack is successful or how many times this event has occurred after analyzing the attacker and system. In this experiment, we use the number of detected attacks, detection rate and false alarm rate to evaluate the experimental results. Through the analysis of these three indicators, relevant conclusions are drawn.

4. Analysis of Experimental Results

4.1. Decision Tree's Ability to Detect Unknown Attacks

In this paper, letters are used to represent different types of data. D represents DoS, R represents R2L, U represents U2R, P represents PROBING, and N represents normal. The total number of samples in test set 1 is 3000, and the total detection accuracy is 90%. The detection results are as shown in 1:

Table 1. Results of testing the decision tree with test set 1

	Number	False report rate	Detection rate
D	895	2.56%	97.44%
R	415	19.78%	80.22%
U	145	27.96%	72.04%
P	290	3.45%	96.55%
N	1255	8.89%	91.11%

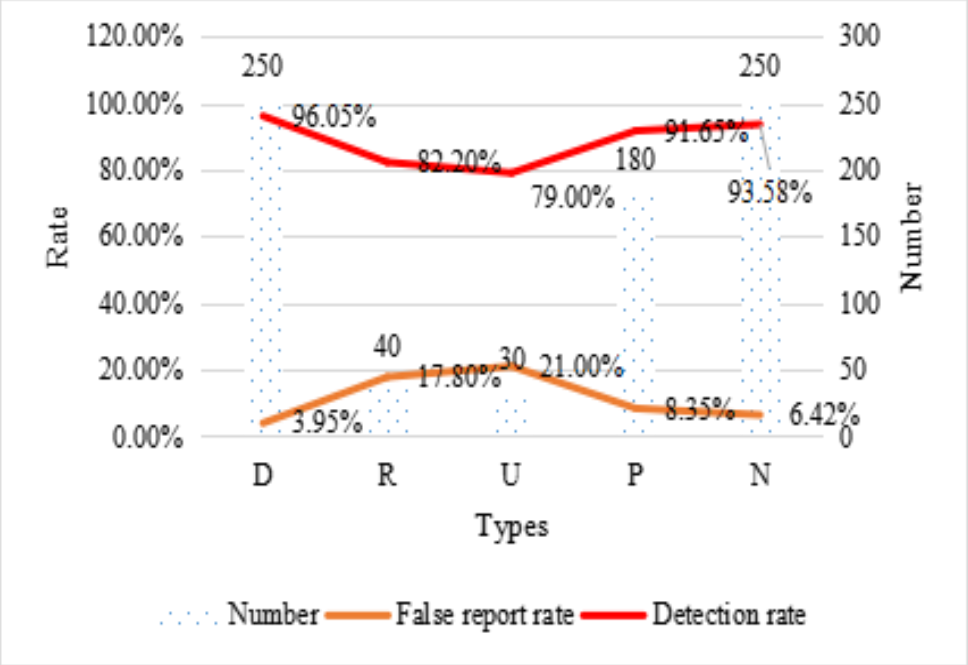

Figure 2. Results of testing the decision tree with test set 2

As shown in Figure 2, we can see that the total number of samples in test set 2 is 750, and the total detection accuracy is 93%. The accuracy of this test set is higher. The error rate of R2L and U2R attacks is higher. In the detection of unknown attacks, it is also necessary to improve the relevant capabilities.

4.2. Effect of PCA Feature Extraction on Decision Tree Performance

PCA feature extraction for test set 1. The detection results of different types of data sets are different, but the general trend is consistent. The total detection accuracy is 92%, and the detection results are shown in Table 2:

Table 2. Test results of test set 1 after pca feature extraction

	Number	False report rate	Detection rate
D	895	1.35%	98.65%
R	415	18.01%	81.99%
U	145	20.26%	79.74%
P	290	4.22%	95.78%
N	1255	6.57%	93.43%

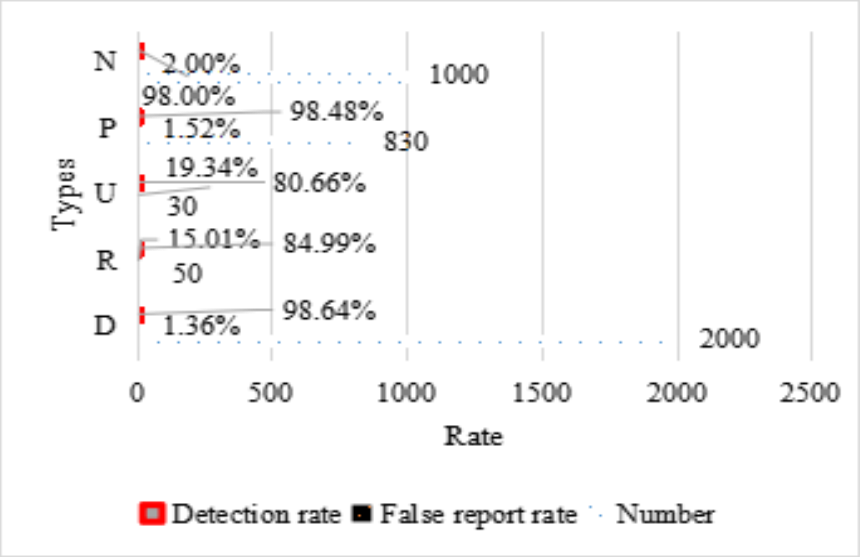

Figure 3. Training results of the training set after PCA feature extraction

As shown in Figure 3, compared with other test sets, we can see that after PCA feature extraction, the detection rate of type D is 98.65% that of type P is 95.78%, that of type N is 93.43%, and that of type R is 91.99%. These data are enough to prove that PCA feature extraction can improve the accuracy of the system.

5. Conclusion

On the Internet, network attack is a common phenomenon. The rapid development of the Internet has brought people a lot of convenience. However, network attacks are accompanied by a series of risks. Computer viruses, hackers and malicious programs will all have a certain threat and impact on human beings. This paper proposes an intrusion detection method based on decision tree algorithm by comparing and analyzing various types of network attacks and their characteristics and combining the current development of the Internet industry. The experiment in this paper shows that PCA feature extraction can improve the system speed in network intrusion detection. The data collection in this paper is not enough to bear the growing demand of reality, so it is necessary to strengthen the data processing capacity of the system and improve relevant technologies.

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] *Maman Abdurohman, Aji Gautama Putrada, Mustafa Mat Deris: A Robust Internet of Things-Based Aquarium Control System Using Decision Tree Regression Algorithm. IEEE Access 10: 56937-56951 (2020).*

[2] *Hadeer Mahmoud, Mostafa Thabet, Mohamed Helmy Khafagy, Fatma A. Omara: Multiobjective Task Scheduling in Cloud Environment Using Decision Tree Algorithm. IEEE Access 10: 36140-36151 (2020).*

[3] *Marziye Narangifard, Hooman Tahayori, Hamid Reza Ghaedsharaf, Mehrdad Tirandazian: Early Diagnosis of Coronary Artery Disease by SVM, Decision Tree Algorithms and Ensemble Methods. Int. J. Medical Eng. Informatics 14(4): 295-305 (2020).*

[4] *Chandrashekhar Azad, Bharat Bhushan, Rohit Sharma, Achyut Shankar, Krishna Kant Singh, Aditya Khamparia: Prediction Model Using SMOTE, Genetic Algorithm and Decision Tree (PMSGD) for Classification of Diabetes Mellitus. Multim. Syst. 28(4): 1289-1307 (2020).*

[5] *Leidiane C. M. M. Fontoura, Hertz Wilton De Castro Lins, Arthur S. Bertuleza, Adaildo G. D'Assunção, Alfredo Gomes Neto: Synthesis of Multiband Frequency Selective Surfaces Using Machine Learning with the Decision Tree Algorithm. IEEE Access 9: 85785-85794 (2020).*

[6] *Kumpol Saengtabtim, Natt Leelawat, Jing Tang, Wanit Treeranurat, Narunporn Wisittiwong, Anawat Suppasri, Kwanchai Pakoksung, Fumihiko Imamura, Noriyuki Takahashi, Ingrid Charvet: Predictive Analysis of the Building Damage From the 2011 Great East Japan Tsunami Using Decision Tree Classification Related Algorithms. IEEE Access 9: 31065-31077 (2020).*

[7] *Ferdinand Bollwein, Stephan Westphal: A Branch & Bound Algorithm to Determine Optimal Bivariate Splits for Oblique Decision Tree Induction. Appl. Intell. 51(10): 7552-7572 (2020).*

[8] *Firoozeh Karimi, Selima Sultana, Ali Shirzadi Babakan, Shan Suthaharan: Urban Expansion Modeling Using an Enhanced Decision Tree Algorithm. GeoInformatica 25(4): 715-731 (2019). https://doi.org/10.1007/s10707-019-00377-8*

[9] *Muhamad Hasbullah Mohd Razali, Rizauddin Saian, Yap Bee Wah, Ku Ruhana Ku-Mahamud: An Improved ACO-based Decision Tree Algorithm for Imbalanced Datasets. Int. J. Math. Model. Numer. Optimisation 11(4): 412-427 (2020).*

[10] *Rajendra Mahla, Baseem Khan, Om Prakash Mahela, Anup Singh: Recognition of Complex and Multiple Power Quality Disturbances Using Wavelet Packet-Based Fast Kurtogram and Ruled Decision Tree Algorithm. Int. J. Model. Simul. Sci. Comput. 12(5): 2150032:1-2150032:23 (2020). https://doi.org/10.1142/S179396232150032X*

[11] *Matheus Guedes Vilas Boas, Haroldo Gambini Santos, Luiz Henrique de Campos Merschmann, Greet Vanden Berghe: Optimal Decision Trees for the Algorithm Selection Problem: Integer Programming Based Approaches. Int. Trans. Oper. Res. 28(5): 2759-2781 (2020). https://doi.org/10.1111/itor.12724*

[12] *Vinay Arora, Rohan Singh Leekha, Inderveer Chana: An Efficacy of Spectral Features with Boosted Decision Tree Algorithm for Automatic Heart Sound Classification. J. Medical Imaging Health Informatics 11(2): 513-528 (2020).*

[13] *Neda Mehdizadeh, Nazbanoo Farzaneh: An Evidence Theory based Approach in Detecting Malicious Controller in the Multi-Controller Software-defined Internet of Things Network. Ad Hoc Sens. Wirel. Networks 51(4): 235-260 (2020).*

[14] *Chiara Ravazzi, Francesco Malandrino, Fabrizio Dabbene: Towards Proactive Moderation of Malicious Content via Bot Detection in Fringe Social Networks. IEEE Control. Syst. Lett. 6:*

2960-2965 (2020).

[15] Scaria Alex, T. Dhiliphan Rajkumar: An Approach for Malicious JavaScript Detection Using Adaptive Taylor Harris Hawks Optimization-Based Deep Convolutional Neural Network. Int. J. Distributed Syst. Technol. 13(5): 1-20 (2020). https://doi.org/10.4018/IJDST.300354

[16] Sivaraman Eswaran, Vakula Rani, Daniel Dominic, Jayabrabu Ramakrishnan, Sadhana Selvakumar: An Enhanced Network Intrusion Detection System for Malicious Crawler Detection and Security Event Correlations in Ubiquitous Banking Infrastructure. Int. J. Pervasive Comput. Commun. 18(1): 59-78 (2020).

[17] Mohit Kumar, Priya Mukherjee, Kavita Verma, Sahil Verma, Danda B. Rawat: Improved Deep Convolutional Neural Network Based Malicious Node Detection and Energy-Efficient Data Transmission in Wireless Sensor Networks. IEEE Trans. Netw. Sci. Eng. 9(5): 3272-3281 (2020).

[18] V. Brinda, M. Bhuvaneshwari: Identifying Malicious Secondary User Presence within Primary User Range in Cognitive Radio Networks. Wirel. Pers. Commun. 122(3): 2687-2699 (2020).