Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2021.020303
ISSN 2790-0983 Vol. 2, Issue 3: 20-27

SSH Application Classification Based on Machine Learning

Jing Liu*

Philippine Christian University, Philippine

18636110600@163.com

**corresponding author*

Keywords: Machine Learning, Decision Tree Classification Algorithm, SSH Application Classification, SSH Tunnel

Abstract: The security feature of SSH protocol ensures the privacy and security of communication content or communication behavior. Often, APTs and malware also use SSH or a variant encryption protocol disguised as SSH to invade a computer or server. In order to solve the shortcomings of existing SSH application classification research, this paper discusses the SSH protocol framework, SSH tunnel and C4.5 decision tree classification algorithm, and briefly discusses the data collection and system development tools of SSH application classification system in this paper. Moreover, the SSH classification model based on machine learning is designed and discussed. Finally, the proposed C4.5 decision tree classification algorithm is tested on the classification results of protocol application. Experimental data show that the average recall and accuracy of C4.5 decision tree classification algorithm for the five protocols are more than 93.17%. Therefore, the C4.5 decision tree classification algorithm proposed in this paper has certain advantages for SSH application classification.

1. Introduction

With the continuous improvement of network bandwidth, more and more network applications are created, and more and more people begin to use SSH for communication. The protocol encrypts the transmitted data, so it is difficult for the administrator to supervise the behavior of accessing foreign illegal websites through SSH tunnel. Therefore, studying the application identification under SSH tunnel can prevent some illegal behaviors.

Nowadays, more and more scholars have conducted rich research in SSH application classification through various technologies and system tools, and have also achieved certain research results through practical research. Acharya describes a technology that uses Open

containment (OpenSSH) software to ensure secure, encrypted transmission of network traffic over mobile Lans. Whenever a mobile LAN implemented with a mobile IP LAN moves to a foreign network, its gateway (router) gets an IP address from the new network. IP tunnels are used to encapsulate IP, and then a "home proxy" is established from the gateway through the foreign network to its domestic network. These tunnels provide the mobile LAN with a virtual bidirectional connection to the home network as if the LAN were directly connected to its home network. Therefore, when the IP moves, the tunnel network traffic of the mobile LAN must traverse one or more foreign networks that may not be trusted. Such traffic may be eavesdropped and intercepted [1]. Donya mainly introduces the evolution of SSH, the necessity of SSH, the working principle of SSH, the main components and features of SSH. As the number of users on the Internet increases, so does the threat to data. The Secure Shell (SSH) protocol provides a secure way to perform remote logins and other secure network services over an insecure network. The SSH protocol is designed to support many functions as well as appropriate security. The architecture provides user authentication, integrity, and confidentiality with its mutually independent built-in layers, connection-oriented end-to-end delivery, reusing encrypted tunnels into several logical channels, providing datagram delivery across multiple networks, and optionally providing compression. The role of each layer of the architecture and connection establishment is also described in detail here. This paper also mentioned some threats, applications, advantages and disadvantages of SSH [2]. Saab believes that as an encrypted communication protocol, SSH not only provides security protection for remote login and other services, but also encapsulates some other unknown applications in tunnel applications, which brings potential impact on network security. Therefore, it is necessary to accurately identify these applications and take corresponding measures in time to protect network security. For the encryption features of SSH protocol, the methods based on traffic statistics can usually be used to identify these applications, and supervised machine learning methods are mostly adopted. By comparing supervised machine learning and unsupervised machine learning, five machine learning methods (C4.5, SVM, BayesNet, K-means and EM) were used to classify SSH applications. The results show that the unsupervised K-means method has the best classification effect and the highest accuracy in identifying HTTP in SSH tunnel [3]. Although the existing SSH application classification research is very rich, there are still some shortcomings in the SSH application classification research based on machine learning.

This paper first introduces the three main components of SSH protocol framework and related concepts of SSH tunnel, and briefly describes the mathematical model steps of constructing classification model in C4.5 decision tree classification algorithm, and then details the data collection of training and testing and the system development tools. Then, according to the characteristics of SSH protocol communication and the main characteristics of the classification algorithm in machine learning, the stream features to be used for classification are extracted. Finally, according to these extracted features, the SSH classification model based on decision tree classification algorithm is proposed, and the classification system based on machine learning is constructed.

2. SSH Application Classification Based on Machine Learning

2.1. SSH Framework

The SSH protocol framework consists of three parts. The first part is the transport layer protocol, which is at the lowest level of the SSH protocol framework and mainly provides various supports for server authentication, data confidentiality and information integrity [4]. The second part is the user authentication protocol, which can provide the client identity authentication to the server. The third part is the connection protocol part. The connection layer runs on top of the user

authentication protocol, which multiplexes many different concurrent encryption tunnels into the logical channel [5]. It allows login sessions and TCP forwarding and also provides flow control services [6].

2.2. The SSH Tunnel

Tunnel technology is a new technology emerging in recent years. It originates from the idea of protocol disguise and is the embodiment of protocol disguise [7]. Protocol camouflage is to encapsulate one protocol into the appearance of another protocol through various methods, so that their protocol syntax and communication port are as consistent as possible [8]. In addition to the specific content of the communication, the more unified the other aspects, the more deception can be achieved. A simple example, such as via SSH protocol in the server and the client to open A port respectively, and then through the open ports, will forward requests to host A to host B, host response is returned to A, B finally returned to the requester, so that it can solve the problem of the requester cannot directly access the host B, and access to resources on the host B [9].

2.3. C4.5 Decision Tree Classification Algorithm

Machine learning methods mainly include supervised learning, unsupervised learning and semi-supervised learning. In this paper, the decision tree classification algorithm in supervised learning is used to study SSH application classification [10].

The classification pattern established by it is x tree architecture, and it is assumed that there are a type of $B_u = (u = 1,...x)$ SSH protocol training sample set [11]. The expected information (entropy of K) required for classification in the training set is:

$$G(K) = \sum_{u=1}^{x} f_u \log_2^{fu} \tag{1}$$

Where f_u is the probability that the SSH protocol in training set K belongs to category B_u [12]. Suppose that attribute S is chosen as the splitting node and attribute S has y different outputs according to the test of training SSH protocol data samples, which divides K into y subsets $K_v(v = 1,2,...y)$. The amount of information required for reclassification of the divided training samples is:

$$G_S(K) = \sum_{v=1}^{y} \frac{|K_v|}{K} * G(K_v) \tag{2}$$

Thus, according to equations (1) and (2), the information gain obtained by this division can be obtained:

$$Gain(S) = G(K) - G_S(K) \tag{3}$$

The algorithm actually uses the method of minimum entropy. The information gain rate must be considered for each split node determined by this method, but the calculation result of the information gain rate still depends on the probability distribution of the type in the training sample set [13]. This algorithm realizes the overfitting phenomenon of training through the pruning method [14]. Since the pruned classification tree has a smaller area and less difficulty, the pruned classification tree has a faster and better classification speed in the actual flow analysis [15].

3. Research on SSH Application Classification Based on Machine Learning

3.1. Data Collection

In the detection method based on machine learning, it is necessary to extract the feature attributes of the data stream and build the classification model according to the feature attributes, so the feature selection of the data stream plays a crucial role in the identification efficiency of the traffic [16]. We collected a large number of datasets, and made statistics on the size and direction of the first four packets with non-empty load after the successful establishment of the data flow in the dataset. The data flow in the dataset is shown in Table 1 [17].

Table 1. Data set distribution

Application project	Training data	The test data
GRTD	345	189
SKTY	2451	1452
QQ	2741	4124
SSH	5241	6524
WEB	1226	867

3.2. System Development Tools

The software module is mainly implemented under the Windows10 operating system, and the software master uses Python3.6 programming language. PyQt5 module set is mainly used in the design and implementation of the interface, and the system in this paper is mainly implemented based on the development tools in Table 2 [18].

Table 2. System Development Tools

Development projects	The development tools
The operating system	Windows10
CPU	Processor6136,16processors
CPU frequency	1x2400MHz
Memory	32G
The hard disk	250G
Development of language	Python3.6
Module sets	PyQt5
Compile environment and	GCC
Debugging environment	GDB

4. Research on SSH Application Classification Based on Machine Learning

4.1. Construction of SSH Application Classification Model Based on Machine Learning

Machine learning has been widely used in traffic identification systems. This method has good universality, but it needs to select specific attributes according to different environments. The design of machine learning SSH application classification method is shown in Figure 1. It is mainly divided into two parts: training module and testing module, and the structure of submodules in these two modules is also very similar. This paper focuses on the training module structure flow.

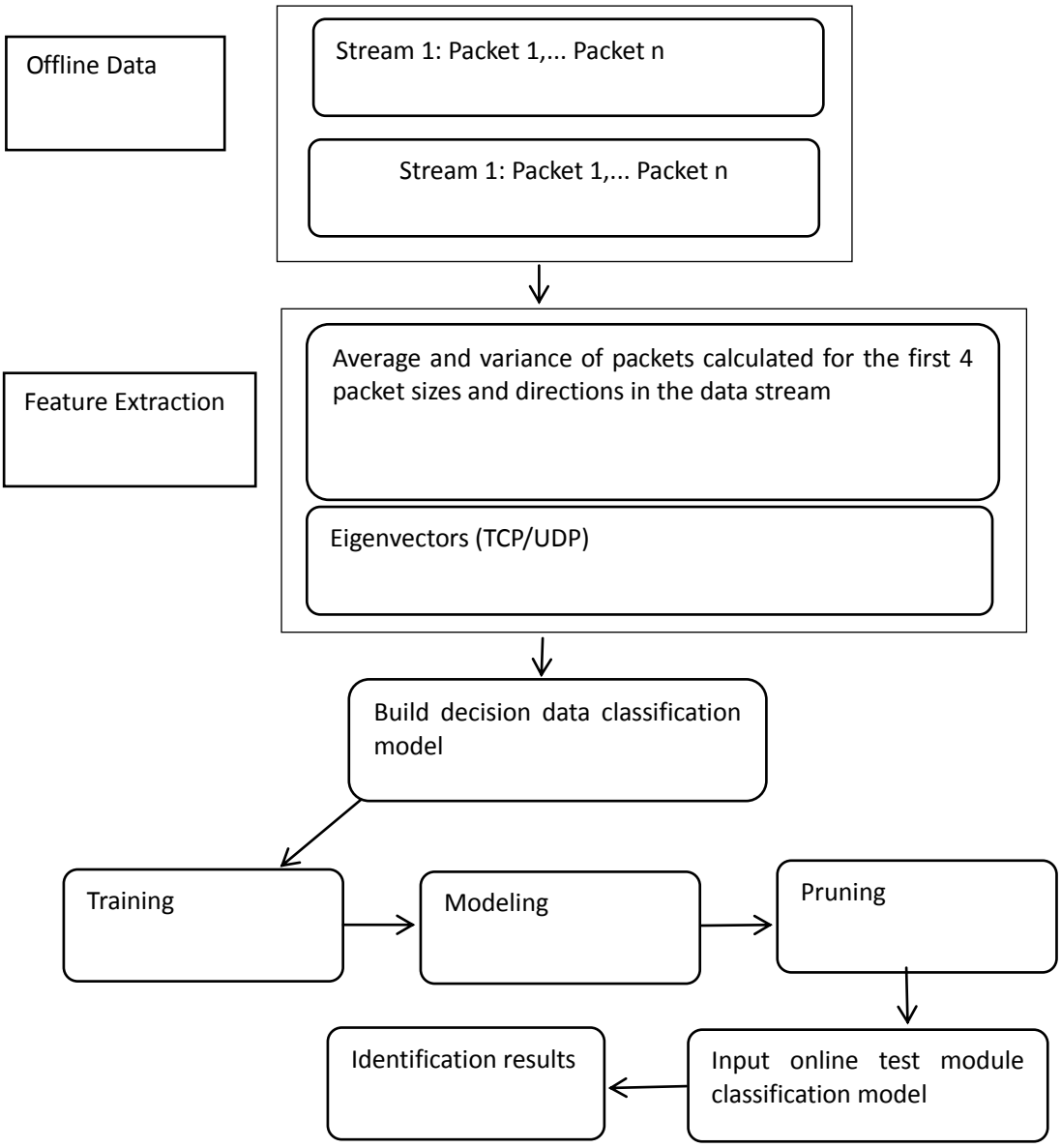

Figure 1. SSH application classification model based on machine learning

The training module is the foundation of the test module, which mainly includes three sub-modules: traffic collection, feature extraction and classification model establishment. It extracts feature attributes from labeled data sets and builds classification models according to feature vectors using classification algorithms. In this paper, the C4.5 decision tree algorithm is selected to complete the classification and identification of traffic, and the information of the first four packet sizes in the data stream (packet size, packet direction, packet size average and variance, TCP/UDP, etc.) is selected as the feature attributes of machine learning. Similar to the training module, the test module mainly includes three sub-modules, traffic collection, feature extraction and traffic classification. Traffic classification makes use of the classification model generated in the training module, and uses the feature vectors extracted from the online traffic to classify and identify the data stream, so as to achieve the purpose of online traffic identification.

4.2. SSH Application Classification Model Application Based on Machine Learning

After selecting the top four packet sizes and directions with non-empty load in the data stream, we calculate the average and variance of packet sizes, and take the packet size, packet direction, average packet size and variance of packet sizes as the feature vectors of the data stream. Since machine learning methods need training sets to construct classification models, we use C4.5 decision tree classification algorithm to construct a decision tree model. Meanwhile, we also use the generated decision tree model to evaluate and predict the test set. The final judgment results are shown in Table 3.

Table 3. Classification result data of decision tree model

Application Protocols	Accuracy	Search completion rate
GRTD	98.54%	93.5%
SKTY	95.67%	91.42%
QQ	89.98%	94.12%
SSH	94.17%	88.56%
WEB	99.34%	86.49%

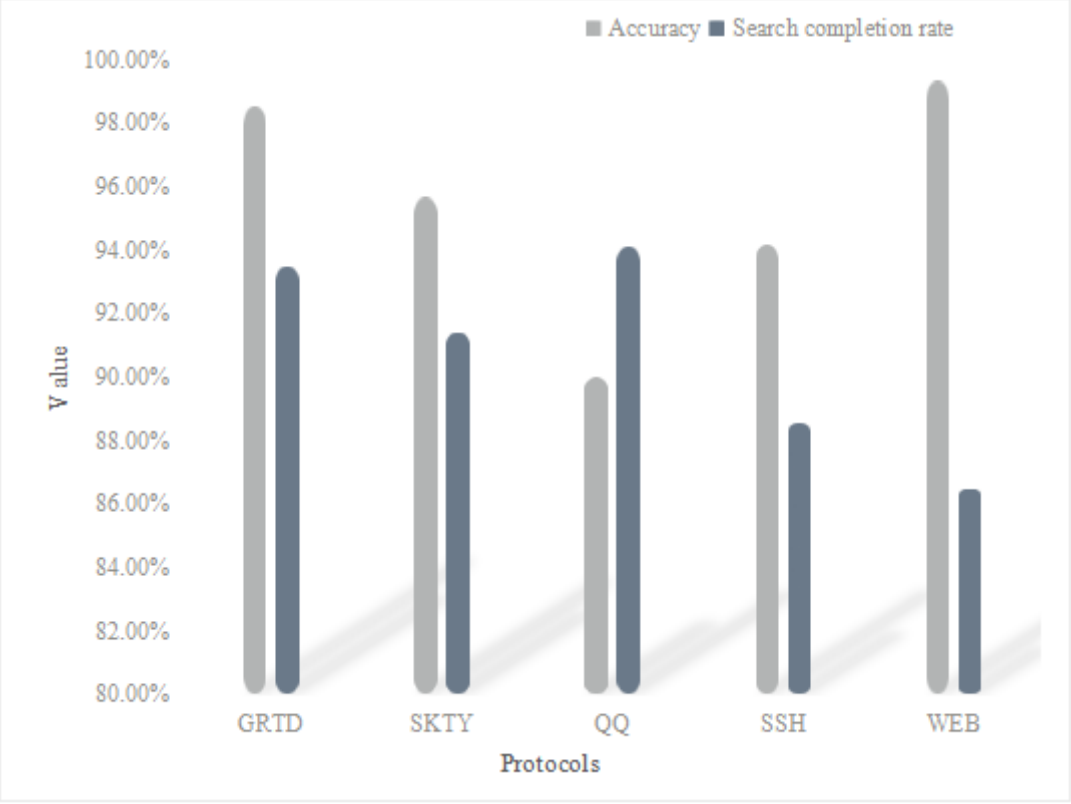

Figure 2. Comparison of classification results of decision tree model

It can be seen from the data in Figure 2 that the accuracy of C4.5 decision tree classification algorithm for GRTD application protocol reaches 98.54% and the recall rate reaches 93.5%. The accuracy of the application protocol SKTY is 95.67%, and the recall rate is 91.42%. The accuracy and recall of the application protocol QQ reached 89.98% and 94.12%, respectively. The accuracy and recall of SSH are 94.17% and 88.56% respectively. The accuracy of SSH is 99.34%, and the recall rate is 86.49%. In conclusion, the classification accuracy and recall rate of C4.5 decision tree classification algorithm are more than 85%. It has a good classification effect.

5. Conclusion

This paper first points out the SSH protocol framework, SSH tunnel and C4.5 decision tree classification algorithm, and then makes an in-depth investigation and analysis of the data collection and system development tools of SSH application classification model based on machine learning, and discusses the classification of the data sets of each application protocol. The randomness of SSH packet load is detected by using C4.5 decision tree classification in machine learning. The relatively simple chi-square detection method is selected, and the chi-square calculation method suitable for data stream is explained in detail. This paper also gives the implementation process of C4.5 algorithm in traffic identification environment. The detection method based on payload is easy to implement, but the load characteristics of application protocols need to be summarized first. Therefore, we summarized the payload of SKTY protocol, GRTD protocol and SSH protocol. Based on packet size distribution, the traffic identification method is relatively simple and easy to implement. Moreover, this method works on the network layer and is effective for both encrypted and unencrypted data traffic.

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] Acharya, Vishwanath, Bora, et al. Classification of SDSS photometric data using machine learning on a cloud. Current Science: A Fortnightly Journal of Research, 2018, 115(2):249-257. https://doi.org/10.18520/cs/v115/i2/249-257

[2] Donya, Dezfooli, Seyed-Mohammad, et al. Classification of water quality status based on minimum quality parameters: application of machine learning techniques. Modeling Earth Systems and Environment, 2018, 4(1):311-324. https://doi.org/10.1007/s40808-017-0406-9

[3] Saab, Fadi, Jaff, et al. Chronic Total Occlusion Crossing Approach Based on Plaque Cap Morphology: The CTOP Classification. Journal of endovascular therapy: an official journal of the International Society of Endovascular Specialists, 2018, 25(3):284-291. https://doi.org/10.1177/1526602818759333

[4] Bae S Y , Shin J S , Kim Y S , et al. Decision tree analysis on the performance of zeolite-based SCR catalysts. IFAC-PapersOnLine, 2021, 54(3):55-60. https://doi.org/10.1016/j.ifacol.2021.08.218

[5] Baranauskas, Jose, Augusto, et al. A tree-based algorithm for attribute selection. Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies, 2018, 48(4):821-833. https://doi.org/10.1007/s10489-017-1008-y

[6] Kantavat P , Kijsirikul B , Songsiri P , et al. Efficient decision trees for multi-class support vector machines using entropy and generalization error estimation. International Journal of Applied Mathematics & Computer Science, 2018, 28(4):705-717. https://doi.org/10.2478/amcs-2018-0054

[7] Acharya, Vishwanath, Bora, et al. Classification of SDSS photometric data using machine learning on a cloud. Current Science: A Fortnightly Journal of Research, 2018, 115(2):249-257. https://doi.org/10.18520/cs/v115/i2/249-257

[8] Al B , Spm B , Fhc D , et al. Predicting the surfactant-polymer flooding performance in chemical enhanced oil recovery: Cascade neural network and gradient boosting decision tree. Alexandria Engineering Journal, 2021, 61(10):7715-7731.

[9] Laiou A , Malliou C M , Lenas S A , et al. Autonomous Fault Detection and Diagnosis in Wireless Sensor Networks Using Decision Trees. Journal of Communications, 2019, 14(7):544-552. https://doi.org/10.12720/jcm.14.7.544-552

[10] Rozzini R . Patients' preferences and "paternalistic approach" in elderly patients with atrial fibrillation. BMJ, 2021(7246):1380-1384.

[11] Alavian S M , Sharafi H , Borba H H , et al. Economic evaluation of pan-genotypic generic direct-acting antiviral regimens for treatment of chronic hepatitis C in Iran: a cost-effectiveness study. BMJ Open, 2021, 12(6):161-176.

[12] Sd A , It B . Interpretable machine learning approach in estimating traffic volume on low-volume roadways - ScienceDirect. International Journal of Transportation Science and Technology, 2020, 9(1):76-88. https://doi.org/10.1016/j.ijtst.2019.09.004

[13] Shetty C , Sowmya B J , Seema S , et al. Air pollution control model using machine learning and IoT techniques - ScienceDirect. Advances in Computers, 2020, 117(1):187-218. https://doi.org/10.1016/bs.adcom.2019.10.006

[14] Sombolestan S M , Rasooli A , Khodaygan S . Optimal path-planning for mobile robots to find a hidden target in an unknown environment based on machine learning. Journal of ambient intelligence and humanized computing, 2019, 10(5):1841-1850. https://doi.org/10.1007/s12652-018-0777-4

[15] Chittora D . How al and machine learning helps in up shilling to better career opportunities. Pc Quest, 2019, 32(3):20-21.

[16] Baumhauer, Judith, Mitten, et al. Using PROs and machine learning to identify "at risk" patients for musculoskeletal injury. Quality of life research: An international journal of quality of life aspects of treatment, care and rehabilitation, 2018, 27(Suppl.1):S9-S9.

[17] Paiva F D , Cardoso R N , Hanaoka G P , et al. Decision-making for financial trading: A fusion approach of machine learning and portfolio selection. Expert Systems with Application, 2019, 115(JAN.):635-655. https://doi.org/10.1016/j.eswa.2018.08.003

[18] Arslan, Atakan, Kardas, et al. Are RNGs Achilles' Heel of RFID Security and Privacy Protocols? Wireless personal communications: An Internaional Journal, 2018, 100(4):1355-1375. https://doi.org/10.1007/s11277-018-5643-3