**Multidisciplinary Project Report**

Study on Identification of Botnet based on Machine Learning

**NAME:** Wu You

**MAJOR:** Avionics

**Student ID No:** 2017122063

**Tutor:**  Professeur Gu Zhaojun

SINO-EUROPEAN INSTITUTE OF AVIATION ENGINEERING

**ABSTRACT**

The current network environment grows fast, it has a enormous amount of traffic and stores lots of valuable information, the modern civil aviation is dependent on the Internet, so the cyber security becomes important. But there exist quite mass of threat with various types. A most threatening attack is the botnet, botnet is a cluster of hosts controlled by attacker, and can be used to conduct enormous attack activities. Recently, there has been many attack affairs using botnet.

The traditional method to differentiate traffic load is based on deep packet inspection(DPI), in which we extract substring or expressions directly from the packet, but it relies on human work and cannot adapt to the fast changing environment. Using machine learning can solve these problems. But potential issues for machine learning is the selection of algorithm and features, even the classification mechanism.

In this paper, we start by constructing a simulating network environment of botnet. The environment will be built using several servers with virtual machine in it, communicating with each other. The traffic load will be fully collected, and used for subsequent analysis. Then we used python to implement the data process module and machine learning module. We extracted useful features from the traffic load, then we made a preliminary comparison of different types of classifiers. Then we optimized the feature set of the selected algorithm and proposed an improvement of the classification method. Finally we evaluated the advantages and disadvantages of the proposed method.

**Keywords:** botnet; machine learning; cyber security; alogrithm

**RÉSUMÉ**

L’environnement réseau actuel se développe rapidement, il y a énormément de trafic et stocke beaucoup d’informations précieuses, l’aviation civile moderne se dépend d’Internet, la cyber-sécurité devient donc importante. Mais il existe une masse de menace avec différents types. Le botnet est l’une des attaques les plus menaçantes. Botnet est un groupe d’hôtes contrôlé par un aggresseur et peut être utilisé pour mener d’énormes activités d’attaque. Récemment, il y a eu de nombreuses attaques utilisant le botnet.

La méthode traditionnelle de différenciation de la charge de trafic repose sur l’inspection approfondie des paquets (DPI), dans laquelle nous extrayons des sous-chaînes ou des expressions directement à partir du paquet, mais elle repose sur un travail humain et ne peut pas s'adapter à un environnement en évolution rapide. L'apprentissage automatique peut résoudre ces problèmes. Mais les problèmes potentiels pour l'apprentissage automatique sont la sélection de l'algorithme et des datas, aussi le mécanisme de classification.

Dans cet article, nous commençons par construire un environnement réseau simulé le botnet. L’environnement sera construit à l’aide de plusieurs serveurs contenant une machine virtuelle et communiquant entre eux. Le flux de trafic sera entièrement collectée et utilisée pour une analyse ultérieure. Nous avons ensuite utilisé python pour implémenter le module de processus de données et le module d'apprentissage automatique. Nous avons extrait des caractéristiques utiles de la charge de trafic, puis nous avons procédé à une comparaison préliminaire de différents types de classificateurs. Nous avons ensuite optimisé l’ensemble des caractère de l’algorithme sélectionné et proposé une amélioration de la méthode de classification. Enfin, nous avons évalué les avantages et les inconvénients de la méthode proposée.

**Mots clés:** botnet; apprentissage automatique; cyber-sécurité;algorithme

**摘要**

. 当前的网络环境发展迅速，拥有大量的流量并存储大量有价值的信息，现代民航依赖于互联网，因此网络安全变得十分重要。但是存在着各种类型的大量威胁。其中最具威胁性的攻击方法是僵尸网络，僵尸网络是由攻击者控制的主机群集，可用于进行大规模的攻击活动。近几年，发生了多起使用僵尸网络的攻击事件，并造成了很大损失。

区分流量负载的传统方法是基于深度包检测（DPI），我们直接从数据包中提取子串或表达式，并基于此对数据包进行匹配。但这种方法有诸多缺点，它依赖于人工从而不能适应快速变化的环境，同时还需要对大量的特征库进行维护，随着互联网的发展，将耗费大量资源。而使用机器学习得方法可以有效解决这些问题。但机器学习的潜在问题是算法和特征的选择，甚至还包括分类机制的设计。

在本文中，我们首先构建一个僵尸网络的模拟网络环境。使用多个相互通信的服务器并在其中配置虚拟机用于构建环境，我们将对流量进行收集，并用于后续分析。然后我们使用python来实现数据处理模块和机器学习模块。本文从流量负载中提取有用的特征，然后对不同类型的分类器进行了初步比较。然后，我们优化了所使用的特征集，并提出了一种改进的检测方法。最后，我们评估了所提方法的优缺点。

**关键词：** 僵尸网络；机器学习；网络安全；算法

**Table of contents**

[I. Introduction 1](#_Toc493532394)

[1.1 Background and Meaning 1](#_Toc493532395)

[1.2 Domestic and foreign development 2](#_Toc493532396)

[II. Research Program 4](#_Toc493532397)

[2.1 Research objectives 4](#_Toc493532398)

[2.2 Research Content 4](#_Toc493532399)

[2.3 Key Issues 4](#_Toc493532400)

[2.4 Technical Route 4](#_Toc493532401)

[2.5 Possible Innovation 6](#_Toc493532402)

[2.6 Work Plan 6](#_Toc493532403)

[2.7 Expected Purpose and Result 7](#_Toc493532404)

[References 8](#_Toc493532405)

# Introduction

## 1.1 Background and Meaning

Due to the rapid development of Internet technology and the enormous economic benefits for national development, the scale of China's Internet will continue to expand over a period of time. With the continuous popularization of the Internet, the number of users has been increasing, and a variety of new types of network applications have emerged. The network has covered all aspects of production and life, and the dependence of social development on the network has gradually increased. The development of network scale also brings certain problems. First, some new network applications such as P2P or VoIP-based software have higher network occupancy rate and affect the operation of other network applications. Secondly, with the rise of e-commerce, the network The continuous increase in the value of information has led to an increasingly prominent network security problem. Finally, Internet information has spread rapidly, providing conditions for the dissemination of bad information. In recent years, the aviation industry has become increasingly dependent on the Internet, and the supervision of the network environment needs to be strengthened. [1]

The current network environment has the characteristics of large traffic, various types, and rapid development. Whether it is network supervision and management or security protection, it is undoubtedly half the effort on the basis of full traffic. It is urgent to classify traffic for targeted work. Therefore, the classification technology of network traffic needs to be continuously upgraded to cope with the increasingly complex network environment. Traffic classification and identification is the first step in the network analysis and management. At the same time, the research results in this direction can be partially migrated to the subsequent analysis steps. Therefore, the research of traffic classification technology is of great significance to the current network security. [2]

The traditional methods of traffic classification include port-based analysis methods and deep packet analysis. Port analysis directly classifies traffic based on port information. However, in the current network environment, there are a large number of variable port protocols and applications, plus port concealment, The extensive use of hopping technology, port information can no longer be used as an independent classification basis, can only be used as auxiliary information for classification. Deep packet analysis can directly obtain a certain pattern feature by extracting the content of the collected data packets, and then use these features to classify the traffic. This method can obtain higher accuracy, but its defect determines that it is not applicable to the current Web environment. First of all, in the current network, the use of encryption technology is more and more extensive, and due to the randomness of the contents of the encrypted data packet, the method is almost ineffective for the encrypted traffic. Secondly, the current network develops rapidly, the traffic type is increasingly multi-functional, and the load is gradually increased. The method needs to maintain a large feature library, and needs to be updated frequently, while being too dependent on labor, the running cost is too large. [3]

## 1.2 Domestic and foreign development

Roughan et al. began the study of traffic classification based on machine learning, which used k-NN algorithm and linear classification, and analyzed the validity of individual data features. [4] In 2004, Antony et al. of the University of Waikato attempted to classify using an unsupervised clustering algorithm. [5] Moore et al. of Cambridge University conducted traffic collection in 2005 and proposed 248 traffic characteristics that can be applied to traffic classification. Finally, using the data obtained to extract corresponding features and using the improved kernel function of Naoba Baye The algorithm is combined with the feature optimization algorithm to classify traffic. After that, various algorithms were applied to traffic classification, and research problems became more targeted. [6] In 2009, Giuseppe et al. used deep learning combined with feature optimization to study the classification of encrypted mobile device traffic. [7] In 2010, Yuan et al. used the Support Vector Machine (SVM) for more accurate traffic classification. [8] Tomasz et al. proposed a traffic classification method based on C5.0 decision tree algorithm in 2012. [9] In 2014, Peng et al. used RBF networks to study the classification techniques at the beginning of traffic generation. [10] Arash et al. studied the method of identifying VPN traffic in 2016[12] and studied the classification of Tor traffic in 2017. [12] In recent years, based on the previous foundation, more research work on cyber threat detection has emerged. In 2017, Qiu et al. studied the application of semi-supervised learning algorithms in botnet detection. [13] Wang et al. combined the convolutional neural network to study the detection of malicious traffic. [14]

Due to the development of the domestic Internet and the late start of machine learning research, the research results in traffic classification work are slightly behind. In 2009, Xu Peng et al. studied the traffic classification technology based on support vector machine. [15] In 2012, Zhang Zhen et al. studied traffic classification methods based on propagation learning and semi-supervised learning. [16] Pan Wubin et al. studied feature selection techniques in traffic classification in 2014.[17] In 2017, Cao Jie et al. further studied the SVM-based network feature method and studied the feature dimension reduction method. [18] Research Program

## 2.1 Research objectives

The increasingly complex Internet environment has brought many challenges to network supervision. In order to solve the regulatory problems, it is first necessary to distinguish network traffic, thus reducing the difficulty of supervision. The goal of this project is to use Python to realize the preprocessing and feature optimization of network traffic data, and propose a reasonable scheme to achieve better traffic classification effect and realize the recognition of botnet.

## 2.2 Research Content

a. Security issues in the current network environment

b. Principles and characteristics of various machine learning algorithms, and data analysis theory

c. The working principle of the network protocol and the composition of the traffic data packet

d. Compare and analyze the classification effect of different algorithms on the data, and propose a scheme to classify the network traffic better and realize the identification of botnet traffic.

## 2.3 Key Issues

a. Feature extraction and optimization of network traffic data

b. Optimization of machine learning algorithms

c. Effective identification of botnets

## 2.4 Technical Route

The research topics to be adopted in this Question are as shown in Flowchart 2-1:

a. The first step is to pre-process the traffic data, including the removal of invalid data, redundant data, and the processing of missing data values.

b. After the data processing, firstly analyze the composition and distribution of the data samples, and determine the reasonable evaluation indicators for evaluating the subsequent experimental results. Then analyze the characteristics of various machine learning algorithms and compare the experiments. Choose the right algorithm.

c. After determining the algorithm used, perform the feature selection of the data, use Python to implement different feature selection methods, and perform feature screening separately, compare the screening results of different methods, and select the optimal according to the previously determined indicators. Feature set.

d. After determining the data characteristics used, perform the tuning of the algorithm model, first perform a rough search within a larger parameter range, then narrow the search range, and narrow the search step size. Finally, based on the cross-validation method, the most Excellent algorithm model.

e. By observing the data and combining the characteristics of the algorithm, an improved method is proposed to achieve a better data classification effect, which can be improved by data processing, hybrid method, training mechanism or algorithm itself.

The choice and improvement of machine learning algorithms is the core of this topic. At this stage, the way to use machine learning for classification is mainly supervised learning, and then other methods will be studied. Supervised learning means that the category of the data has been pre-calibrated. The algorithm model will use the data label as the basic fact, compare it with the classification result, and set an optimized target. Through continuous iterative update, a satisfactory result will be obtained. model.

Whenever a sample in F is input C, C calculates the output H according to the existing parameters, compares the result with the target value O of the sample, and corrects the parameter according to the deviation of the predicted result to obtain a new classifier C'. And use C' for the next training, until the end of the training, the resulting classifier can be used for the classification of the new sample.

Among them, the integrated machine learning method is a kind of effective supervised learning algorithm. The Adaboost algorithm and the random forest algorithm are relatively classic integrated machine learning algorithms.

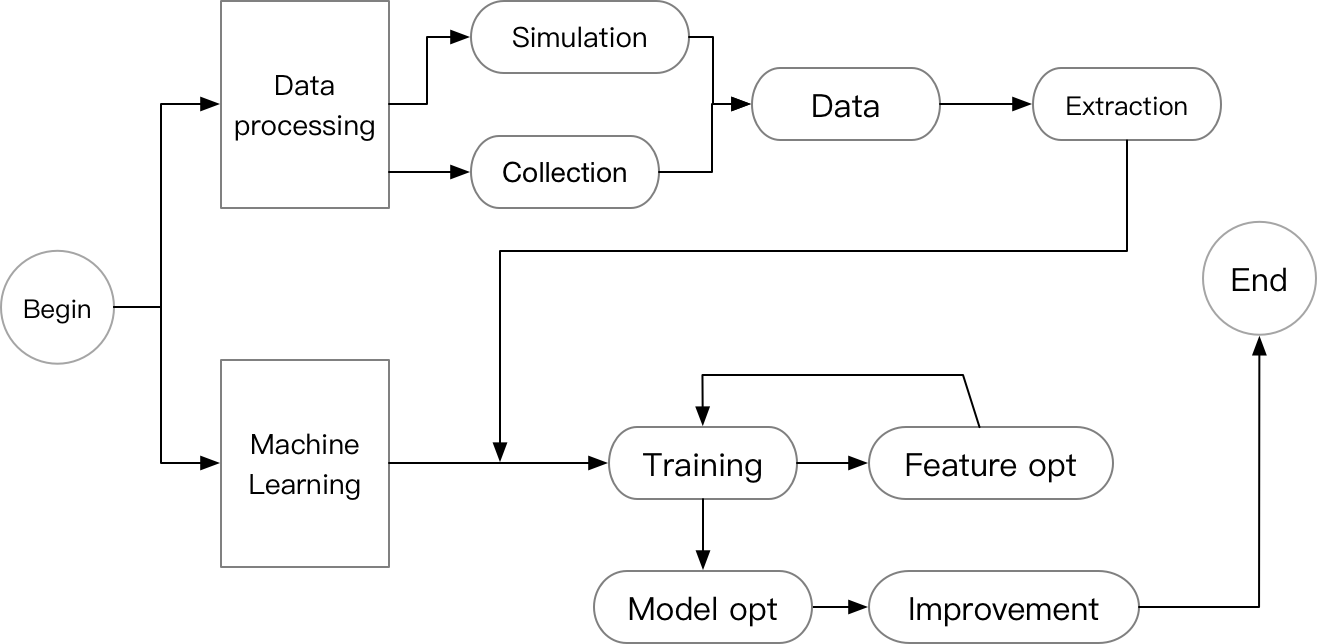


Figure 2-1 Research Process

## 

Figure 2-2 The process of machine learning

## 2.5 Possible Innovation

A new improved method for identifying botnet

## 2.6 Work Plan

(1) September-November 2018

Find and read the literature and materials related to the topic, learn how to use Python's various APIs, implement machine learning algorithms and perform data analysis.

(2) December 2018 - February 2018

    Study the main workings of botnets and build a simulation environment in the lab to mimic the behavior of botnets and collect network traffic to get the available data.

(3) March-April 2019

    Complete the preprocessing of the data, and conduct a preliminary analysis of the data, determine the evaluation indicators of the algorithm, and compare the effects of different algorithms. The feature optimization work is carried out, different feature selection strategies are applied, the selection results are analyzed, and the feature set finally used is determined, and then the feature set is used to perform the tuning of the algorithm model.

(4) May-July 2019

    Algorithm optimization, combining comprehensive consideration of the characteristics of data and algorithms, make effective theoretical analysis, and improvement of existing classification methods, and implement experimental comparison of improved methods and classification methods of the original method to verify the method

Effectiveness.

(5) August-December 2020

## Work on a summary and write a dissertation.

## 2.7 Expected Purpose and Result

Expected Purpose:

Accurate identification of botnet traffic and classification of general application traffic

Expected Result:

# Complete the writing of a graduate thesis and publish a Chinese core journal paper References

[1]. Bai Jun, Xia Jingbo, Wu Jixiang, et al. Overview of Real-Time Network Traffic Classification Research[J]. Computer Science, 2013, 40(9): 8-15.

[2]. Pan Wubin, Cheng Guang, Guo Xiaojun, et al. Review and prospect of network encryption traffic identification research[J]. Transactions of Communications, 2016, 37(9): 154-167.

[3].Boutaba, Raouf, et al. "A comprehensive survey on machine learning for networking: evolution, applications and research opportunities." Journal of Internet Services and Applications 9.1 (2018): 16.

[4].Roughan M, Sen S, Spatscheck O, et al. Class-of-service mapping for QoS:a statistical signature-based approach to IP traffic classification[C]// ACM SIGCOMM Conference on Internet Measurement. ACM, 2004:135-148.

[5].Mcgregor A, Hall M, Lorier P, et al. Flow Clustering Using Machine Learning Techniques[C]// International Passive and Active Network Measurement International Workshop, PAM 2004, April 19-20, 2004, Proceedings. DBLP, 2004:205-214.

[6].Moore, Andrew, Denis Zuev, and Michael Crogan. Discriminators for use in flow-based classification. 2013.

[7].ACETO, Giuseppe, CIUONZO, Domenico, MONTIERI, Antonio, et al. Mobile Encrypted Traffic Classification Using Deep Learning.

[8].Yuan, Ruixi, et al. "An SVM-based machine learning method for accurate internet traffic classification." Information Systems Frontiers 12.2 (2010): 149-156.

[9].Bujlow, Tomasz, Tahir Riaz, and Jens Myrup Pedersen. "A method for classification of network traffic based on C5. 0 Machine Learning Algorithm." Computing, Networking and Communications (ICNC), 2012 International Conference on. IEEE, 2012.

[10.]Peng, Lizhi, Bo Yang, and Yuehui Chen. "Hierarchical RBF neural network using for early stage internet traffic identification." Computational Science and Engineering (CSE), 2014 IEEE 17th International Conference on. IEEE, 2014.

[11.]Draper-Gil, Gerard, et al. "Characterization of Encrypted and VPN Traffic using Time-related." Proceedings of the 2nd international conference on information systems security and privacy (ICISSP). 2016.

[12.]Lashkari, Arash Habibi, et al. "Characterization of Tor Traffic using Time based Features." ICISSP. 2017.

[13.]Qiu, Zhicong, David J. Miller, and George Kesidis. "Flow based botnet detection through semi-supervised active learning." Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017.

[14.]Wang, Wei, et al. "Malware traffic classification using convolutional neural network for representation learning." Information Networking (ICOIN), 2017 International Conference on. IEEE, 2017.

[15.] Xu Peng, Liu Qiong, Lin Sen. Research on Internet Traffic Classification Based on Support Vector Machine[J]. Journal of Computer Research and Development, 2009, 46(3): 407-414.

[16.] ZHANG Zhen, WANG Bin-qiang, LI Xiang-tao, et al. Semi-supervised traffic classification method based on neighbor propagation learning[J]. Acta Automatica Sinica, 2013, 39(7):1100-1109.

[17.] PAN Wu-Bin, CHENG Guang, GUO Xiao-Jun, WANG Yan.Feature Selection of Embedded Network Flow Based on Selective Integration Strategy[J].Chinese Journal of Computers,2014,37(10):2128-2138.

[18.] Cao Jie. Research on feature reduction and classification of network traffic based on SVM[D]. Jilin University, 2017.