International Journal of Pure and Applied Mathematics
Volume 118 No. 24 2018
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue

http://www.acadpubl.eu/hub/

Network Traffic Analysis Using Weka Tool

Rishabh Tulsyan [1], Asha. S [2],
School of Computing Science and Engineering,
VIT University Chennai Campus,
Chennai,
Tamilnadu, India.600127
rishabh.tulsyan2017@vitstudent.ac.in
asha.s@vit.ac.in

May 26, 2018

Abstract

: One of the foremost and heavy threats on the web nowadays is malicious software package, typically noted as a malware. Cyber Security has been a major threat to the world as the digitalization continues to grow and expand. The networks which enables this connections are been hacked, important information's are leaked and most of all there is a major economic loss. This research deals with the analysis of such networks, which is not secure and is a threat to the system. Wireshark a powerful tool that collects and displays information about each network connected to it. It creates a dataset for these collected networks. Weka use is to classify these collected datasets from the Wireshark tool and classify it based on the various algorithms.

***Key Words*:**Cyber Security, Weka, Wireshark, Network Traffic.

1 INTRODUCTION

Network Traffic Classification has become an important topic in every field of digitalization. It is has become an important and an essential part for ISPs (Internet Service Providers) for managing the performance of the network. The first step includes the classification of the unknown networks. This step is an essential part of network classification, the network security. Traffic classification is the first step to identity and classify unknown network classes. Network Traffic Classification plays a very vital role in network security and management, such as Intrusion Detection, Quality of Service (QoS). Through this technique, network operators can take some actions such as to block some flows and manage resources. They can also find the growth of network applications.

In the last two decades, numerous network traffic classification techniques [3] [4] have been proposed to classify unknown classes. The first one is Port Based Technique. It is a great technique for network traffic classification / identification. This technique includes a port, which is firstly registered in Internet Assign Number Authority (1ANA) [2]. However, this technique failed due to increase of Peer-to-Peer applications (P2P) in [5], which use dynamic port numbers. Dynamic port number means unregistered number with Internet Assign Number Authority (1ANA). Then second, one is Payload Based technique. This technique gives accurate results in network traffic classification. This technique is Deep Packet Inspection (DPI) technique. However, there is a problem in this technique. The problem is that it cannot be used for encrypted data network applications as numerous network applications use encrypted techniques to protect data from detection. Therefore, this technique also failed due to use of encrypted flow of applications. Thereafter, the researchers proposed another method called Machine Learning Technique (ML) to classify internet traffic as well as to know what type of applications flow in the network. Machine Learning Technique gives very promising accuracy results in network traffic classification. This technique is based on training and testing data sets to classify unknown classes. Contribution: In this paper, we

discuss Network traffic classification techniques and discuss. Then we discuss Comparative analysis of four machine-learning classifiers. We first capture network traffic using packet-capturing application Wireshark [15]. After that using Net Mate tool, we extract features [21] from the capture traffic and then we apply four machine-learning classifiers to classify WWW, DNS, FTP, P2P and Telnet applications. The experimental result shows that C4.5 classifier give high accuracy as compare to other machine learning classifiers, which are 78.91% the rest of the paper is structured as follows:

Section 2 introduces the basic introductory information about developed techniques.

Section 3 Demonstrates Internet Traffic Classification Model.

Finally, we draw conclusion in Section 4.

2 NetworkTraffic Classification Techniques

The technique is to classify the collected Network Traffic according to the network protocol and the IP address it belongs. Here we used the port-based Technique. Payload based technique and Machine Learning Technique.

2.1 Port Based Technique

This technique is the traditionally used technique. A classification of networks done using well-known port numbers. The network applications registered in their port as assigned by the Internet Assigned Number Authority (AINA). This helps in identifying the corresponding port number in accordance with the port number registered at the AINA. Below is the list of port numbers assigned by AINA.

Table 1. AINA assigned port Number

Assigned Port	Application
20	FTP Data
21	FTP
22	SSH
23	Telnet
25	SMTP
53	DNS
80	HTTP
110	POP3
123	NIP
161	SNMP
3724	WoW

2.1.1 Algorithm

Connect to the Server that is connected. Capturing the IP address. Identifying the Port numbers. Looking for the assigned ports.

Thus, this traditionally used technique is quite helpful in classification of the network traffic. However, these classifications are not so accurate, fails due to a dynamic port number of application to, and hence not detected.

2.2 Payload Based Technique

This method is a deep packet inspection. (DPI). In this technique, the packet contents analyzed for signatures and authentications of network applications in the traffic. It is a used to propose Peer to peer (P2P) application. This solves the problem of the dynamic port number problem.

Below is the table illustrates the example of Karagiannis et al. in [5].

TABLE 2. Karagiannis is describe strings at the beginning P2P protocol payload

PIP Protocol	String	'Ira es.
Edankey. 2000	0xe319010000	TCP/UDP
	0xe53f010000	
Fast-track	"Get /.hash"	TCP
	0x2700000002980	UDP
BitTorrent	lx13Bit"	TCP
Gnutella	"GNUT" "GIV"	TCP
Aress	"GET hash"	UDP
	"Get Shal"	

However, this technique also has some major issues the first one being a complex hardware for pattern searching in a payload. The second is that, this technique does not work in the encrypted network application traffic. Lastly, this technique requires a continuous updating of signature patterns of new applications.

2.3 Machine Learning Technique

Machine Learning technique based on the data set. In this technique, the machine learning classifiers are trained as an input. Then using this trained sample, the unknown classes are classified. There are two types of machine learning technique supervised and unsupervised.

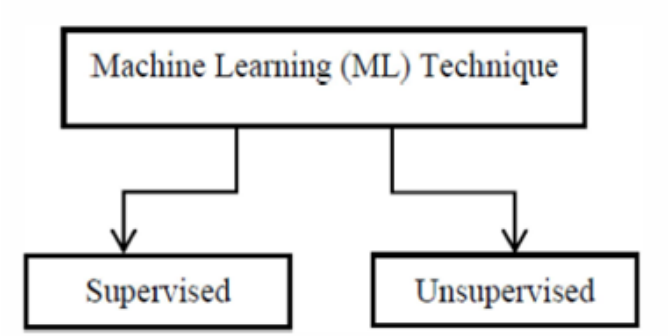

Figure 1. Kinds of machine learning

2.3.1 Supervised Learning

Supervised learning technique is a machine learning technique [11]. This technique is also called classification methods. This technique needs a complete labeled data set to classify unknown classes. Below are the two figures. Which are discussed in details in [12]. It means that the supervised learning technique trains the model with some labeled data set and then it will produce prediction output in new data samples.

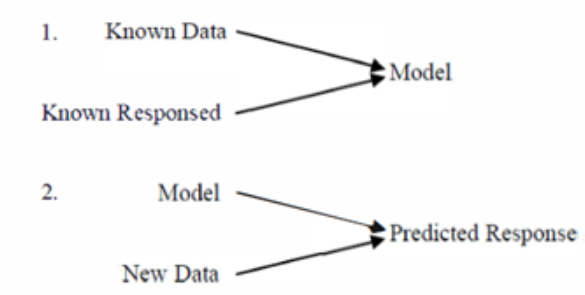

Figure 2. Method description by mathworks

This method infers function from labeled training data set. This method starts with a training dataset.

Training Dataset (TS) = <x_1,y_1>, <x_2,y_2> <x_N,y_M>.

Such that x is the feature vector which belongs to ith and y_i is its output predicted value.

2.3.2 Unsupervised Learning

Unsupervised technique is also called a cluster technique. In this method, there is no need of complete labeled data sets. Unsupervised is a type of machine learning. Thus, the result output of machine learning training does not identity or classify instances in predefined classes.

3 NETWORK TRAFFIC CLASSIFICATION MODEL

In this section, we explain the network traffic classification structure and model. Which includes systematic process as shown in Fig. 3. This systematic process method will show you how to use network traffic classification technique to identity / classify unknown network traffic classes using machine-learning technique.

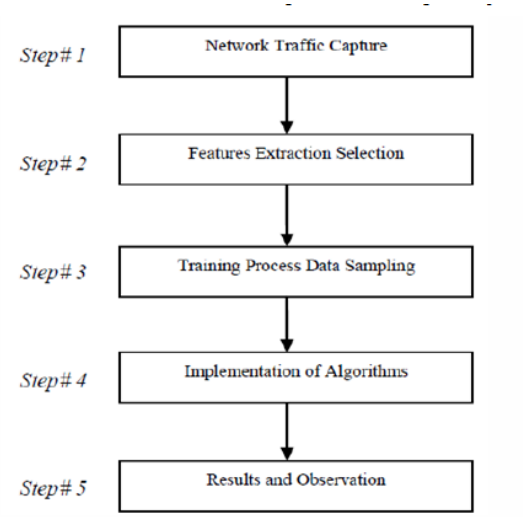

Figure 3. Network traffic classification model.

3.1 Network Traffic Capture

This is the first and most important step, which includes data collection. In this step, the real time network traffic captured. Known as data collection step. There are many tools for network traffic capturing, but Tcpdump tool, which is used to capture the real time network traffic. To capture network traffic, we use Wire Shark tool [IS] for packet capturing and analyzing. We captured the traffic the duration of one minute of WWW, DNS, FTP, P2P and Telnet application.

3.2 Feature Extraction Selection

After capture network traffic data, the feature selection and extraction step follow. In this step, the features are extracted from the captured data such as packet duration, packet length: inter arrival packet time protocol etc. Then extracted features are used to train the machine learning classifier. For feature extraction, Perl script can be used to extract the feature from captured data set. However, we use Netmate tool [20 for feature extraction and we extract 23 features. We use MS Excel for

saving the dataset for Weka tool as a Comma Separated Values (CSV) file format.

3.3 Training Process Sampling

In this stage, data sets are sampled for supervised learning technique. In supervised learning, data are first labeled to classify unknown network applications.

3.4 Implementation of Machine Learning Algorithms

This is the implementation step, which includes applying machine learning algorithm or classifiers on the instances.

For example, applying supervised, unsupervised and semi supervised learning algorithm. For implementation of machine learning algorithm, there are many tools available on internet, but most commonly nowadays are used MatLab [14] and Weka classification simulation tools [13]. In this paper, we use Weka tool and apply four machine-learning algorithm C4.S, Support Vector Machine, Bayes Net and Nave Bayes to build classification model using 10 Folder Cross-validation

4 Result And Observation

After the implementation of machine learning algorithms, the simulation tool gives detailed results about the applied algorithms such as accuracy detailed information, training time and recall etc. In this work, we use four classifiers C4.S, Support Vector Machine, Bayes Net and Naive Bayes. However, C4.S algorithm gives very high result accuracy as compare to other algorithm. In table 3 shown the accuracy, training time results and figure 4 shown, the comparison of accuracy results of using 4 machine-learning algorithms.

TABLE 3. Accuracy Result and Comparison of MILA

Classifiers	Accuracy	Time(Second)
C43	78.9189	0
SVSI	740541	0.03
ByesNet	68.1081	0.01
NaiveBayes	71 8919	0.01

In the Table 3 it is clear that C4.S machine learning classifiers gives accuracy result better than other applied machine learning classifies. In addition, Fig. 4 shows the accuracy result to show which classifier gives very accurate accuracy result as well as Fig. 5. Shows the recall and precision comparison result in which C4.S classifiers result is very good as compare to other ML classifiers.

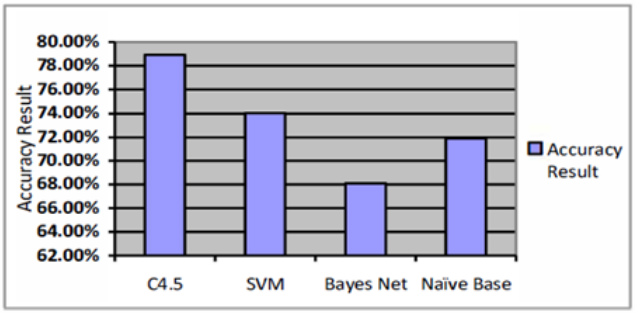

Figure 4. Accuracy result and Comparison.

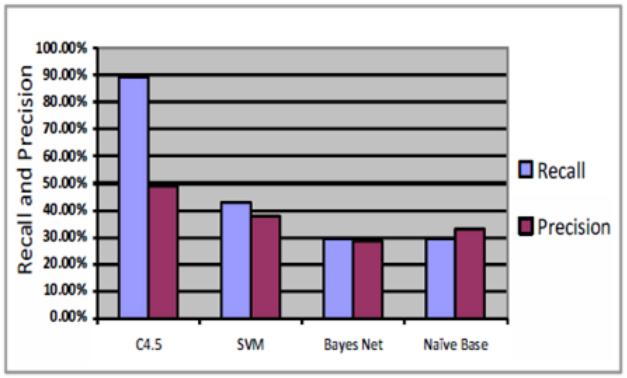

Figure 5. Recall and Precision of four Machine Learning Classifiers.

Figure 6 and 7 shows the recall and precision result of captured WWW, DNS, FTP, P3P and TELNET applications. From which it is clear that which application recall and precision results are good which are not. From these figure it is clear that DNS and WWW application recall and precision result is very poor as compare to other applications.

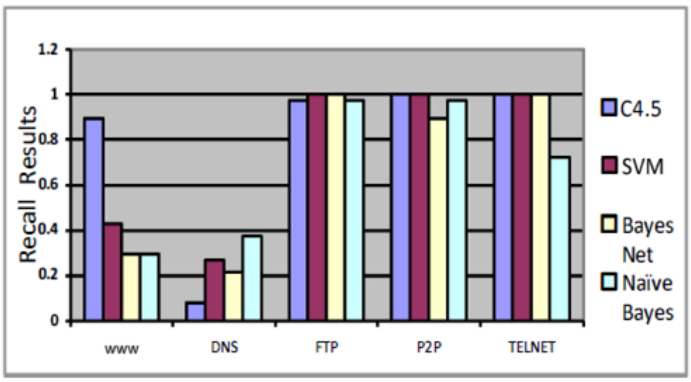

Figure 6. Recall of four machine learning classifiers on five applications

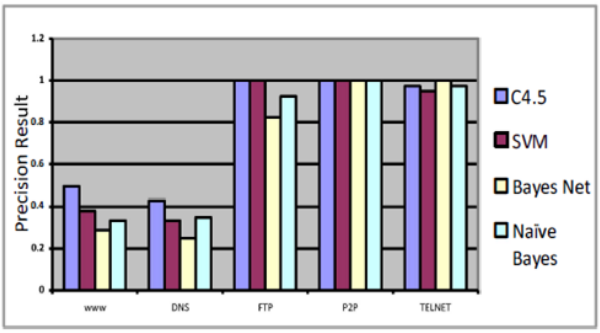

Figure 7. Precision of four machine-learning classifiers of five applications.

5 CONCLUSION

In this paper, we discuss Network traffic classification techniques and discuss How new researchers or new network operators will apply the network traffic classification technique using machine learning algorithm to classify unknown applications and manage performance of network. Then we perform comparative analysis of four machine-learning classifiers. Firstly, we demonstrate Network Traffic Classification Techniques (Port Based, Pay Load Based and Machine Learning Based technique) and their limitation. Then we structure model of network traffic classification from traffic capture to end result.

For comparative analysis of four algorithms, we capture five WWW, DNS, FTP, P3P and TELNET applications traffic duration of I minute using Wire Shark tool and extract 23 features using Netmate tool. After that, traffic is classified using four machine-learning algorithms.

Experimental result show that C4.5 decision algorithm gives high accuracy result as compare to other Support Vector Machine, ByesNet and Naive Baes machine learning classifiers.

References

[1] A. Moore and K. Papagiannaki (2005): "Toward the accurate identification of network applications," in Porco Of PAM Conf., March, 2005.

[2] Arthur Callado, Carlos Kamienski, Geza Szabo, Balazs Peter GerYo, Judith Kelner,Stenio Fernandes, and Djamel Sadok. (2009): "A Survey on Internet Traffic Identification," IEEE Communications Survey tutorials, Vol. II, No. 3, pp. 37-52, Third Quarter 2009.

[3] Cao, Jie, et al. (2015): "Network Traffic Classification Using Feature Selection and Parameter Optimization.Journal of Communications 10.10.

[4] http://www.mathworks.com/help/stats/supervised- learning-machinelearning-workflow-and-algorithms.html

[5] Ian H. Witten and Eibe Frank (2005): Data Mining: Practical Machine Learning Tools and Techniques, 2th edition, Morgan Kaufmann Publishers, San Francisco, CA.

[6] Internet Assigned Numbers (2008): http://www.iana.orglassignments/port . Authority (lANA), numbers, as of August 12, 2008.

[7] Introduction to NetMate tool, download information https://dan.arndt.calnims/calculating-flow-statistics-using-netmatelcomment-page-II

[8] Jie Cheng, Russell Greiner (2015): "Learning Bayesian Belief Network Classifiers: Algorithms and System, "Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada. 2455

[9] Knowledge Analysis http://www.mathworks.com/downloads/Matlab.

[10] Kuldeep Sing, Sunil Agrawal (2011): "Comparative Analyssis of Five Machine Learning Algorithms for IP Traffic Classification" IEEE International Conference on 2011 Emerging Trends in Network and Computer Communication.

[11] Pawel Foremski (2013): On different ways to classify Internet, traffic: a short review of selected publications Theoretical and Applied Informatics, 2013.

[12] T. Karagiannies, A. Broido, and M. Faloutsos (2004): 'Transport layer identification of P2P traffic," Proc. of ACM SIGCOMM IMC, August 2004.

[13] T. Karagiannies, A. Broido, N. Brownlee, K. Claffy and M. Faloutsos (2004): "File-sharing in the internet: a characterization of p2p traffic in the backbone," Proc. of ACM SIGCOMM IMC, August 2004.

[14] T. Nguyen, and G. Armitage (2008): A Survey of Techniques for Internet Traffic Classification using Machine Learning, IEEE Surveys and Tutorials, 10(4), pp. 56-76.

[15] T. Auld, A. Moore, and S. Gull (2007): Bayesian neural networks for Internet track classification," IEEE Transactions on Neural Networks, vol. 18, no. 1, 2007.

[16] Thuy (2008): Introduction to Network Traffic Classification, http://www.cisco.com/c/eu/us/td/docs/nsite/.../chap05.pdf

[17] Thuy T.T. Nguyen and Grenville Armitage (2008): "A Survey of Techniques for Internet Traffic Classification using Machine Learning," IEEE Communications Survey tutorials, Vol. 10, No. 4, pp. 56-76, Fourth Quarter 2008.

[18] To capture online traffic, Wire shark tool, Application: http://www.wireshark.org.

[19] Thales Sehn Korting (2015): "C4.5 algorithm and Multivariate Decision Trees," Image Processing Division, National Institute for Space Research - INPE, SP, Brazil. 2454

[20] Waikato Environment for Knowledge Analysis (WEKA) 3.4.4, http://www.cs.waikato.ac.nz/ml/weka