Network Traffic Classification Techniques and Comparative Analysis Using Machine Learning Algorithms

Muhammad Shafiq, Xiangzhan Yu, Asif Ali Laghari, Lu Yao, Nabin Kumar Karn, Foudil Abdessamia
School of Computer Science and Technology
Harbin Institute of Technology
Harbin 15001 PR, China
e-mail:{muhammadshafiq, yxz, asiflaghari, luyao, karnnabin, foudil.abdessamia}@hit.edu.cn

***Abstract*—Network Traffic Classification is a central topic nowadays in the field of computer science. It is a very essential task for internet service providers (ISPs) to know which types of network applications flow in a network. Network Traffic Classification is the first step to analyze and identify different types of applications flowing in a network. Through this technique, internet service providers or network operators can manage the overall performance of a network. There are many methods traditional technique to classify internet traffic like Port Based, Pay Load Based and Machine Learning Based technique. The most common technique used these days is Machine Learning (ML) technique. Which is used by many researchers and got very effective accuracy results. In this paper, we discuss network traffic classification techniques step by step and real time internet data set is develop using network traffic capture tool, after that feature extraction tool is use to extract features from the capture traffic and then four machine learning classifiers Support Vector Machine, C4.5 decision tree, Naïve Bays and Bayes Net classifiers are applied. Experimental analysis shows that C4.5 classifiers gives very good accuracy result as compare to other classifies.**

Keywords-traffic classification; machine learning; methods.

I. INTRODUCTION

Network Traffic Classification is an important topic nowadays in the field of Computer Science. It is very essential for Internet Service Providers (ISPs) to manage the overall performance of a network. Traffic classification is the first step to identify 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 (IANA) [2]. But 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 (IANA). Then second one is Payload Based technique. This technique gives accurate results in network traffic classification. This technique is also called Deep Packet Inspection (DPI) technique. But 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. So, 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. And then we discuss comparative analysis of four machine learning classifiers. We first capture network traffic using packet capturing application Wireshark [15]. After that using NetMate 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 II introduces the basic introductory information about developed techniques. Section III demonstrates Internet Traffic Classification Model. Finally, we draw conclusion in section IV.

II. NETWORK TRAFFIC CLASSIFICATION TECHNIQUES

Network Traffic Classification is the process to identify the network applications or protocol that exists in a network [1]. Network traffic classification has got great significance in the last two decades. Researchers have proposed many methods to classify network applications. In this section, we discuss Port-based Technique, Payload Based Technique and Machine Learning (ML) techniques.

A. Port-Based Technique

As we mentioned in section I that traditionally, in this technique, a classification of network applications is executed using the well-known ports number. Moreover, we also discussed that first network applications are registered in

978-1-4673-9026-2/16/$31.00 ©2016 IEEE

their ports in the Internet Assigned Number Authority (AINA). In this way, the traffic is identified corresponding to the registered ports number registered in AINA. Table 1 shows different types of applications and its ports numbers assigned by AINA. For instance, E-mail applications use 25 (SMTP) port number to send emails and to receive email 110 (POP3) port is used. In this way, web applications use 80 ports number.

TABLE I. IANA ASSIGNED PORT NUMBER FORMAT FOR SOME WELL-KNOWN APPLICATIONS

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

Thus, this technique does not provide good classification accuracy results [6], [7]. Moreover, this technique fails due to using dynamic port number of new applications to evade being detected.

B. *Payload-Based Technique*

This method is also called Deep Packet Inspection technique (DPI). In this technique, the contents of the packets are examined looking characteristics signatures of the network applications in the traffic. This is the first alternative to ports-based method. This technique is specially proposed for Peer to Peer (P2P) applications. It means applications which use dynamic port number to identify traffic in a network. Below is the table, which illustrates examples, used by Karagiannis et al. in [5].

TABLE II. KARAGIANNIS DESCRIBE STRINGS AT THE BEGINNING P2P PROTOCOL PAYLOAD

P2P Protocol	String	Trans. Protocol
Edonkey 2000	0xe319010000	TCP/UDP
	0xe53f010000	
Fasttrack	"Get /.hash"	TCP
	0x2700000002980	UDP
BitTorrent	"0x13Bit"	TCP
Gnutella	"GNUT" "GIV"	TCP
Aress	"GET hash"	UDP
	"Get Shal"	

But in this technique, we stumble upon some problems. The first problem in this technique is that it needs a very expensive hardware for pattern searching in a payload. The second problem in this technique is that it does not work in encrypted network application traffic. Finally, this approach needs continuous update of signature pattern of new applications.

C. *Machine Learning (ML) Technique*

Machine learning (ML) technique [8],[9],[10] is based on data set (Labeled Data Set) . In this technique, a machine learning classifier is trained as input and then using the trained sample prediction, unknown classes are classified. There are two main areas in machine learning technique: the supervised and unsupervised learning technique.

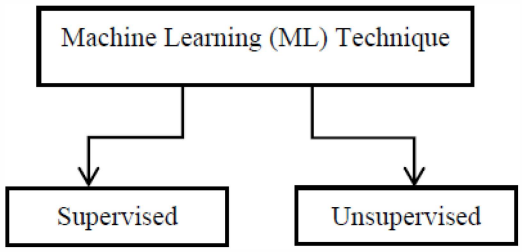

Figure 1. Kinds of machine learning

A. *Supervised Learning Technique*

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.

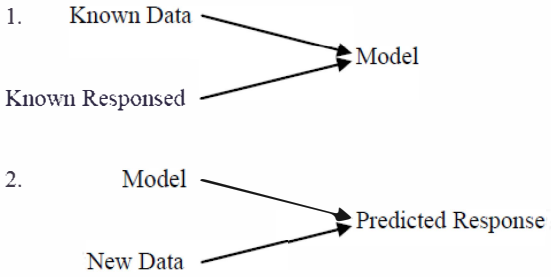

Figure 2. Method description by mathworks

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

$$TS = <x_1,y_1>, <x_2,y_2>, \ldots\ldots\ldots <x_N, y_M>,$$

Such that x_i is the feature vector which belongs to i^{th} and y_i is its output predicted value.

B. *Unsupervised Technique*

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 identify or classify instances in predefined classes.

III. Network Traffic Classification Model

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

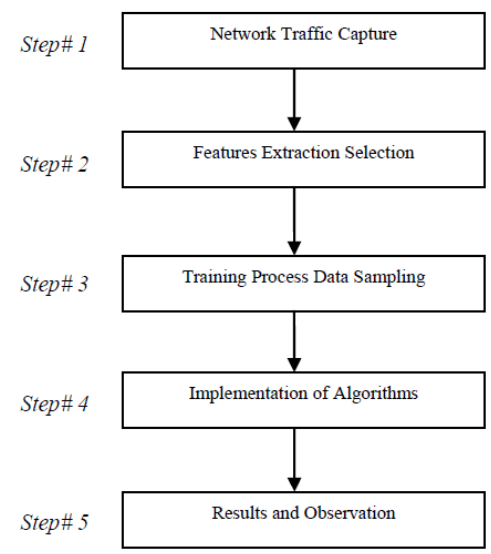

Figure 3. Network traffic classification model.

A. Network Traffic Capture

This is the first and most important step, which includes data collection. In this step, the real time network traffic is captured. It is also known as data collection step. There are many tools for network traffic capturing, but Tcpdump tool can be used to capture the real time network traffic. To capture network traffic, we use Wire Shark tool [15] for packet capturing and analyzing. We captured the traffic the duration of one minute of WWW, DNS, FTP, P2P and Telnet application.

B. 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. But 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.

C. 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.

D. 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 C4.5, Support Vector Machine, BayesNet and NaïveBayes to build classification model using 10 Folder Cross-validation.

E. 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 C4.5, Support Vector Machine, BayesNet and NaïveBayes. But C4.5 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 III. Accuracy Result and Comparison of MLA.

CLASSIFIERS	ACCURACY (%)	T TIME (Second)
C4.5	78.9189	0
SVM	74.0541	0.03
ByesNet	68.1081	0.01
NaiveBayes	71.8919	0.01

In the Table III it is clear that C4.5 machine learning classifiers gives accuracy result better than other applied machine learning classifies and also 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 C4.5 classifiers result is very good as compare to other ML classifiers.

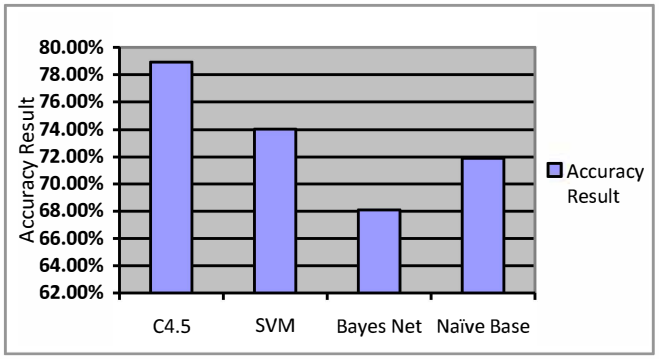

Figure 4. Accuracy result and comparison.

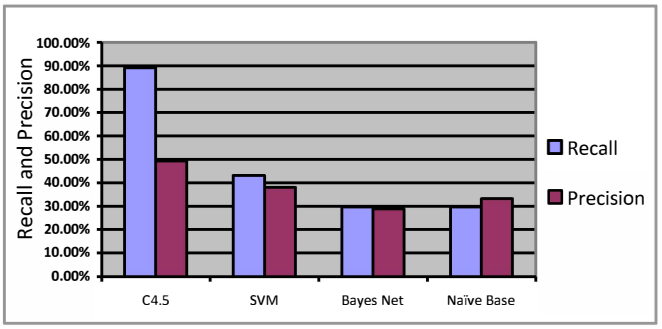

FIGURE 5. RECALL AND PRECISION OF FOUR MACHINE LEARNING CLASSIFIERS

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.

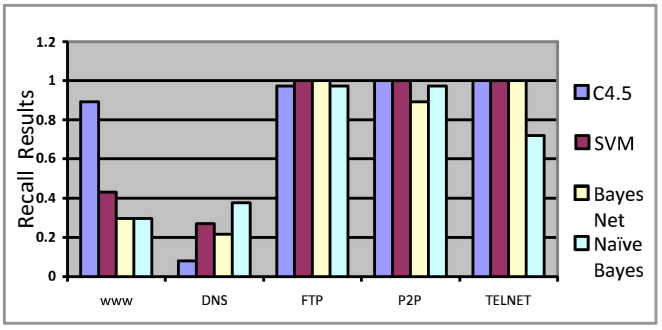

Figure 6. Recall of four machine learning classifiers on five applications.

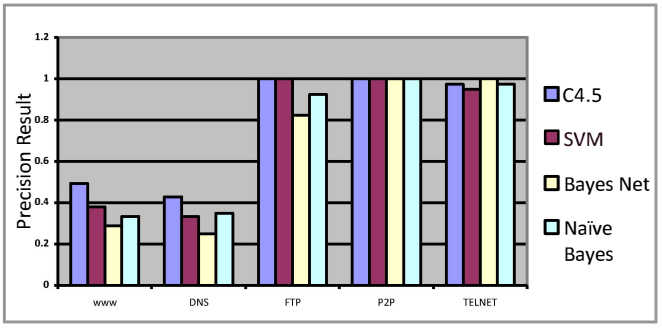

Figure 7. Precision of four machine learning classifiers of five applications.

IV. 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. And 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 1 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 NaiveBaes machine learning classifiers.

REFERENCES

[1] Thuy Introduction to Network Traffic Classification, http://www.cisco.com/c/en/us/td/docs/nsite/.../chap05.pdf

[2] Internet Assigned Numbers Authority (IANA), http://www.iana.org/assignments/port- numbers, as of August 12, 2008.

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

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

[5] T. Karagiannies, A. Broido, and M. Faloutsos, "Transport layer identification of P2P traffic," Proc. of ACM SIGCOMM IMC, August, 2004.

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

[7] A. Moore and K. Papagiannaki, "Toward the accurate identification of network applications," in Porc. Of PAM Conf., March, 2005.

[8] Thuy T.T. Nguyen and Grenville Armitage. "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.

[9] Arthur Callado, Carlos Kamienski, Géza Szabó, Balázs Péter GerÝo, Judith Kelner,Stênio Fernandes ,and Djamel Sadok. "A Survey on Internet Traffic Identification," IEEE Communications Survey & tutorials, Vol. 11, No. 3, pp. 37-52, Third Quarter 2009.

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

[11] T. Auld, A. Moore, and S. Gull, \Bayesian neural networks for Internet tracclassication," IEEE Transactions on Neural Networks, vol. 18, no. 1,2007.

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

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

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

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

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

[17] Thales Sehn Korting, "C4.5 algorithm and Multivariate Decision Trees," Image Processing Division, National Institute for Space Research – INPE, SP, Brazil.

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

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

[20] Introduction to NetMate toll, download information https://dan.arndt.ca/nims/calculating-flow-statistics-using-netmate/comment-page-1/

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