Implementation of Network Traffic Classification by Using MLA

Miss. Ghodake shubhangi, Miss. Raut Sarika, Miss. Ghuge Shital

Department of Computer Engineering, SVPM's College of Engineering, Malegaon, Tal-Baramati, Dist-Pune, M.S., India

Abstract- Network traffic classification is challenging task in high speed network. Network monitoring is required for quality of service and analysis, therefore it generate network traffic. Existing system has some drawback, to overcome that drawback we have develop our system i.e classification of network traffic using machine learning algorithm. According to generated traffic information by client we have constructed boosted classifier with high accuracy. This system is used to classify application like FTP, Skype, TCP ,etc. For constructing c5.0 classifier we have to provide unique dataset and training set to algorithm. This paper shows how we implement machine learning algorithm and how we use that algorithm for classification of network traffic.

Keywords- FTP, Skype, TCP, Network Monitoring, HTTP, Traffic, Browser Traffic, c5.0.

I. INTRODUCTION

Network traffic classification create challenge for whole network, network traffic increase the load of network, because of that one cant do work in proper or in fast way. Each network carries data for different applications, which have different usage. Therefore providing information about the quality level requires knowledge of what kind of data is flowing in the network at the present time. Generally, methods for traffic classification use a concept of flow defined as a group of packets having the same end IP addresses, using the same transport protocol, and its port numbers. Flows are considered bidirectional-packets going from the local machine to the remote server and from the remote server to the local machine are part of the same flow. Using application ports for traffic classification is a very simple idea widely used by network administrators to limit traffic generated by worms and unwanted services. This method is fast, and can be applied to almost all the routers existing on the market. This method is good to classify some protocols operating on fixed port numbers , but this method is not applicable for dynamic port no e.g skpe,gaming etc. because its operate on dynamic port no and one cant detect that application. Deep Packet Inspection (DPI) solutions are quite slow Furthermore they relay on inspecting the user data and therefore privacy and confidentiality issues can appear because user data is always in private manner. exiting methods like C4.5, J48, Random Forests have much wider coverage. They can be used in any point of the network, providing very fast statistical detection of the application, to which the traffic belongs. Achievable detection rate correctness is over 95The goal of machine learning is to design and develop algorithms that allow systems to use empirical data, experience, and training to evolve and adapt to changes that occur in their environment. A major focus of machine learning research is to automatically induce models, such as rules and patterns, from the training data it analyses. Keeping that goal in mind, this paper describes previous related work and then focuses on implementation and evaluation of new machine learning algorithm i.e c5.0 algorithm which has more accuracy than previous.

II. RELATED W ORK

There are many methods to classification of network traffic, they all tries for classify network traffic accurately but classification accuracy is less. In existing method like c4.5 algorithm, some packets are lost during capturing . that's why accuracy of correct classification is less. Classification based on well-known TCP or UDP ports is becoming increasingly less effective growing numbers of networked applications are port-agile (allocating dynamic ports as needed), end users are deliberately using non-standard ports to hide their traffic, and use of network address port translation (NAPT) is widespread (for example a large amount of peer-ton-peer file sharing traffic is using non-default ports).Payload-based classification relies on some knowledge about the payload formats for every application of interest: protocol decoding requires knowing and decoding the payload format while signature matching relies on knowledge of at least some characteristic patterns in the payload. This approach is limited by the fact that classification rules must be updated whenever an application implements even a trivial protocol change, and privacy laws and encryption can effectively make the payload inaccessible.

In exiting tool like Wireshark, it is a cross-platform analyser that does deep inspection of hundreds of protocols. It does live capture and capture save (for offline browsing), which can be viewed in GUI. when we use wireshark then we have to stop all running application. wireshark have to install on each machine this is the main drawback of wireshark. Another tool is Angry Ip scanner, it is one of the easiest to use of all the IP scanners. It has a user-friendly GUI that can scan IP addresses. Angry IP Scanner is cross platform and doesnt require installation, so you can use it as a portable scanner. In our proposed system we are using accurate data sets for both training and testing the boosted classifier. The classifier is able to distinguish traffic which appears to be similar, like web browser traffic and radio streamed via a web page. our system is used to extract the features of packet like application port no, direction of flow of packet,ip address etc.

III. SYSTEM OVERVIEW

1) Packet Capture

For capturing packets and extract features from that packet there are number of tools are present, but we are using jNetPcap and WinPcap. Because jNetPcap is latest version and it is used for supporting jdk environment. Before jNetPcap there was capturing tool which is known as jPcap. jPcap is not supported for capturing packets from high speed links, thats why it is not used in future. WinPcap is used for supporting to windows environment.

2) Training for live packets

After capturing live packets important task is to train that captured packets. In training phase, packet features are extracted in the form of information, after that packet data is converted into byte format and add dynamical into text file, which is used as training data.

3) Detection

In detection task ,it match features of packet in present dataset, if it is not match then it detect as unknown packet.

4) Import Dataset

Import dataset is import the newly created training dataset into system. It load the training text file as input for classification.

IV. C5.0 ALGORITHM

The general algorithm for building decision trees is:

(a) Check for base cases

(b) For each attribute a:Find the normalized information gain from splitting on a

(c) Let a best be the attribute with the highest normalized information gain

(d) Create a decision node that splits on a best

(e) Recurse on the sublists obtained by splitting on a best, and add those nodes as children of node

V. C5.0 CLASSIFIER

For constructing c5.0 classifier, we use concept of machine learning algorithm. Machine learning means machine learn itself by using different resultset. The goal of machine learning is to design and develop algorithms that allow systems to use empirical data, experience, and training to evolve and adapt to changes that occur in their environment. A major focus of machine learning research is to automatically induce models, such as rules and patterns, from the training data it analyzes.

For Training c5.0 classifier, we use text file that is already imported from import dataset. According to that text file machine learn itself about newly added packet features and check that features matched with kdd dataset. Kdd means knowledge discoverable dataset which is unique dataset. kdd discovery of previously unknown knowledge that evaluated with respect to know knowledge. While in typical kdd task, supervised method can not be used due to unavailability training data.

That training dataset used to classification of network traffic into two group class A and class B. That is class of allowed and blocked packets

VI. APPLICATION

Network planning and dimensioning
Performance evaluation
Charging and billing
QoS policies
Research purposes

VII. ADVANTAGES

1. Verifies success or failure of an attack: Since uses system logs containing events that have actually occurred, they can determine whether an attack occurred or not with greater accuracy and fewer false positives than a network based system.

2. Monitors System Activities: A sensor monitors user and file access activity including file accesses, changes to file permissions, attempts to install new executables etc.

3. Lower entry cost: C5.0- are far more cheaper than the network based IDS sensors.

VIII. EXPERIMENTAL R ESULT

In our system, rules are generated by using packet capturing from client side. According to that dataset system classify the network traffic on the basis of c5.0 algorithm. classified result is as follows.

IX. CONCLUSION

The C5.0 MLA for distinguishing different kinds of traffic in computer networks. It was demonstrated that our method is feasible to classify traffic belonging to 7 different applications with an average accuracy of above 98%, when using accurate data sets for both training and testing the boosted classifier. Our results proved that the classifier is able to distinguish traffic which appears to be similar, like web browser traffic and radio streamed via a web page.

ACKNOWLEDGEMENT

The completion of our project brings with it a sense of satisfaction, but it is never complete without them those people who made it possible and whose constant support has crowned our efforts with success. One cannot even imagine the power of the force that guides us all and neither can we succeed without acknowledging it. It is the great pleasure that we acknowledge the enormous assistance and excellent co-operation to us by the following respected personalities.

REFERENCES

[1] Bujlow tombaz, Tahir, Jenns Peddersen, A method for classification of network traffic based on C5.0 Machine Learning Algorithm, to appear in International Conference on Networking and Communications (ICNC 2012).

[2] Tomasz Bujloow, Taahir Riaz, Jens Myrup Pedersen, A method forAssessing Quality of Service in Broadband Networks, submitted to ICACCT 2010 (International Conference on Advanced Communication.

[3] Tomasz Bujlow, Karrtheeapan Balacchandran, Tahirr Riaz, Jens Myrup Pedersen, Volunteer-Based System for classification of traffic in local area network, submitted to 19th Telecommunications Forum TELFOR 011.

[4] Sven Ubikk, Peetr Zejdl, Evaluating Application-Layer Classification Using a Machine Learning Technique Over Different High Speed Networks, 2011 Fifth International Conference on Systems and Network Communications, IEEE 2010, pp.

[5] Li, Shoounyi Zhang, Lu, Junorong san, Real-time P2P,GLOaBECOM 2007 PROCEEDINGS, pp. 14.

[6] Jiiong Cai, Zhibin Zhang, Xinbo Song, An Analysis of UDP Traffic Classification.

[7] Riyad Alshamma, A. Nur Zincir-Heywood, Machine Learning Based Encrypted Traffic Classification: Identifying SSH and Skype, Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defreense Applications (CISDA 2009).

[8] Sven Ubik, Petr Zejdl, Evaluating Application-Layer Classification Using a Machine Learning Technique Over Different High Speed Networks, 2010 Fifth International Conference on Systems and Network Communications, IEEE 2010, pp. 387391.

[9] Riyad Allshammari, A. Nur Zincir-Heywood, Unveiling Skype En-crypted Tunnels using GP, IEEE Congress on Evolutionary Computation(CEC),IEEE 2010, pp. 16.

AUTHORS

First Author – Miss. Ghodake shubhangi, BE Comp, shubhangi.ghodake27@gmail.com.
Second Author – Miss. Raut Sarika, BE Comp, sararaut7750@gmail.com.
Third Author – Miss. Ghuge Shital, BE Comp, ghuge.shital@gmail.com.

Correspondence Authors – Prof. Mhaske V.D. Prof. Khalate Y.R, yogeshkhalate@gmail.com.