IDENTIFICATION THE BEST ALGORITHM AND FEATURES FOR SKYPE TRAFFIC CLASSIFICATION

KHALED MOHAMMED OBAID BAWAKED

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Engineering (Electrical - Electronics & Telecommunications)

Faculty of Electrical Engineering
Universiti Teknologi Malaysia

January 2013

Alhamdullah that Allah give the power to finish this work. This project is dedicated to my parents, Mr.Mohammed Bawaked & Ms. Fawzia Baazim ,,, to my brothers & sisters ,,, to my family ,,, and all my friends ,,, Thank you

ACKNOWLEDGEMENT

In the name of God, the Most Gracious, the Most Merciful. This thesis could not have been accomplished without the assistance of many people whose contributions I gratefully acknowledge. I am heartily thankful to my supervisor, Dr. Izzeldin Ibrahim, whose encouragement, patience, guidance and support from the initial to the final level enabled me to develop an understanding of the subject. His immense support and encouragement has kept me going during the times when I was encountering problems at every turn. Also to Mr. Hamza who had given me valuable help and advice throughout this project. Lastly, I offer my regards and blessings to all who have supported me.

ABSTRACT

Skype uses strong encryption to secure communications inside the whole Skype network. Clients choose communication ports randomly. Therefore traditional port based or payload based identification of Skype traffic is not feasible. In this project we used a Machine Learning identification method to discover Skype host and voice calls as well. In this method, we test the whole algorithms in Weka application with five groups of features to show the most effective features and algorithm for Skype classification. Results indicate the Random forest and REPtree based approach perform much better than other algorithms on the identification of Skype traffic with accuracy 96.90% and 95.40% respectively.

ABSTRAK

Skype menggunakan penyulitan yang kuat untuk mendapatkan komunikasi di dalam rangkaian Skype secare keseluruhan. Pelanggan memilih pengkalan komunikasi secara rawak. Oleh itu sistem pengenalan secara tradisional atau payload trafik Skype tidak dapat dilaksanakan. Dalam projek ini kami menggunakan pencarian dangon kgedan Machine Learning untuk menemui pengkalan Skype serta panggilan suara. Dalam kaedah ini, kita menguji keseluruhan algoritma dalam aplikasi Weka dengan lima kumpulan tertentu untuk menunjukkan ciri-ciri yang paling berkesan dan algoritma bagi klasifikasi Skype. Keputusan menunjukkan pendekatan berasaskan Random Forest dan REPtree adalah jauh lebih baik berbanding algoritma lain dalam pengenalan trafic Skype dengan ketepatan 96,90% dan 95,40% masing-masing.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Network and service providers can only deliver a service to their customers with a certain level of quality if they know how their platforms are being used. Currently ISPs try to achieve a seamless experience for their users by ensuring that enough bandwidth is available in all parts of the network and to adjacent peers at all times. In practice this means that links are typically over-provisioned according to the peak traffic expected and are upgraded as soon as a certain threshold is exceeded. This is not very efficient and does not take into account the different kinds of IP traffic that fill up the links.

The success of the internet is mainly based on its versatility and flexibility, allowing for the development of network applications ranging from simple text based utilities to complex systems for e-commerce and multi-media content. The on-going expansion of the internet is the cause of continuous unitization and traffic behaviour changes. Due to this diversity and the fast changing properties the internet is a moving target. At present, the internet is far from being well understood in its entirety. However, constantly changing internet characteristics associated with both time and location make it imperative for the internet community to understand the nature and behaviour of current internet traffic. Measuring and understanding data traffic is essential for ensuring the reliable operation and smooth growth of computer

networks. Through the measurement and analysis of traffic the internet can be better understood because of the over-all impact of these traffic classes on internet traffic behaviour.

To study the data and the applications in the traffic we need to make a classification operation. By making the classification, we will be able to study the behaviour of each application on the traffic. There are several methods of internet traffic classification such as Machine Learning, Payload and Port base among others. Machine Learning (ML) is one of the most popular methods used for classifications [4].

One of the applications which are commonly used among people and consider as the most popular voice over IP applications is Skype. [9]. Skype is encrypted application and tune itself throw different ports, therefore identifying Skype traffic becomes even more challenging.

1.2 problem statement

This paper started with the importance of making the classification for all applications in the traffic. Since we study Skype traffic, we faced some problems in classify this application. The first problem is Skype can not be classified efficiently by using payload method because it is encrypted. Moreover, we can not use port base method because it establishes the connection by using dynamic ports. The second problem is there are many algorithms it can be used for the classification but the most accurate algorithm for Skype classification is still debatable. The third problem is there are many features it can be used to identify the applications, but the features that can be selected for Skype classification are still under study.

1.3 Project Objectives

The objectives and goals of this paper can be briefly summarized in the following points :

- Specify the best algorithm that can be used for Skype classification, since there are many algorithms it can be used for classification. All the algorithms are available under machine learning method.
- Specify the best features that can be used with Skype classifications since there are more than 240 features for the traffic.

1.4 Scope of study

There are five points we considered them in this project:

1- Obtain stored data Skype and non-Skype data offline.
2- Identify the features for classifying Skype and non-Skype data.
3- Select the whole algorithms in WEKA application to run and test the selected features.
4- Determine the five best results based on the highest accuracy and the time that consumed to build each model.
5- The general traffic data has been taken from Università degli Studi di Brescia in Italy.

1.5 Organization of thesis

At the beginning, chapter one shows an introduction on the necessary to understand the behaviour of the applications in the traffic and how that can affect on the bandwidth. Also, it will include some characteristics about Skype application and

why it is difficult to classify it. Moreover, this chapter will contain the problem statement, the objective and the scope of this project.

While in chapter two, the literature review of the thesis will be stated. First, after the introduction, it will talk about the Skype and what are the services which are provided by Skype. Also, it will include the Skype components. Moreover, this chapter will contain several methods it can be used for classification and what is the best method for Skype classification. Finally, at the end of this chapter we will show some related studies for Skype classification.

Chapter three will talk about the methodology of the research and how the thesis had been organized and how the data had collected from Università degli Studi di Brescia in Italy and how do we select the group of featurs and test all the featurs and the algorithms. Also, it will cover how to compare the results and select the best features and algorithms.

Chapter four will cover the results and the discussion of the results which had obtained by using WEKA software Such as, a comparison between the accuracy with several algorithms, and the time taken to build each model.

Chapter five will talk about the conclusion of the thesis according to the result which had been obtained. Also, it will list the point that could be covered in future work for the other research.

REFRANCES

1. *Abuagla Babiker Mohd & Dr. Sulaiman bin Mohd Nor. Towards a Flow-based Internet Traffic Classification for Bandwidth Optimization. International Journal of Computer Science and Security (IJCSS), Volume (3) : Issue (2)*
2. *Skype Network Administrator's Guide. Skype 3.0 Beta. 2. 2006-10-31 Document version 2.0 Beta*
3. *Li Jun; Zhang Shunyi; Xuan Ye; Sun Yanfei; , "Identifying Skype Traffic by Random Forest," Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on , vol., no., pp.2841-2844, 21-25 Sept. 2007 doi: 10.1109/WICOM.2007.705*
4. *Ziemniak, T.; , "Use of Machine Learning Classification Techniques to Detect Atypical Behavior in Medical Applications," IT Security Incident Management and IT Forensics (IMF), 2011 Sixth International Conference on , vol., no., pp.149-162, 10-12 May 2011 doi: 10.1109/IMF.2011.20*
5. *Perenyi, M.; Gefferth, A.; Trang Dinh Dang; Molnar, S.; , "Skype Traffic Identification," Global Telecommunications Conference, 2007. GLOBECOM '07. IEEE , vol., no., pp.399-404, 26-30 Nov. 2007 doi: 10.1109/GLOCOM.2007.81*
6. *Cisco (2008). WAN and Application Optimization Solution Guide. USA . Americas Headquarters*
7. *http://tstat.tlc.polito.it/traces-skype.shtml*
8. *Lu, H. and H. Liu (2000). Decision Tables: Scalable Classification Exploring RDBMS Capabilities. Proceedings of the 26th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc.: 373-384.*

9. *Angevine, D.; Zincir-Heywood, A.N.; , "A Preliminary Investigation of Skype Traffic Classification Using a Minimalist Feature Set," Availability, Reliability and Security, 2008. ARES 08. Third International Conference on , vol., no., pp.1075-1079, 4-7 March 2008 doi: 10.1109/ARES.2008.158*
10. *Ian H. Witten, Eibe Frank(2005). Data mining (second edition). Morgan Kaufmann. San Francisco, CA*
11. *Neukirchner, M.; Stein, S.; Ernst, R.; , "A Lazy Algorithm for Distributed Priority Assignment in Real-Time Systems," Object/Component/Service-Oriented Real-Time Distributed Computing Workshops (ISORCW), 2011 14th IEEE International Symposium on , vol., no., pp.126-132, 28-31 March 2011*
    doi: 10.1109/ISORCW.2011.22
12. *Williams, N., S. Zander, et al. (2006). "A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification." SIGCOMM Comput. Commun. Rev. 36(5): 5-16.*
13. *Alshammari, R.; Zincir-Heywood, A.N.; , "Machine learning based encrypted traffic classification: Identifying SSH and Skype," Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on , vol., no., pp.1-8, 8-10 July 2009 doi: 10.1109/CISDA.2009.5356534*
14. *http://www.ing.unibs.it/ntw/tools/traces/*
15. *Alshammari, R.; Zincir-Heywood, A.N.; , "Machine learning based encrypted traffic classification: Identifying SSH and Skype," Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on , vol., no., pp.1-8, 8-10 July 2009*
    doi: 10.1109/CISDA.2009.5356534