An Efficient Network Traffic Classification Based on Vital Random Forest for High Dimensional Dataset

by

ALHAMZA MUNTHER WARDI ALALOUSI

(1340211014)

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

School of Computer and Communication Engineering
UNIVERSITI MALAYSIA PERLIS

2017

UNIVERSITI MALAYSIA PERLIS

DECLARATION OF THESIS

Author's full name : Alhamza Munther Wardi Alalousi ……………………….

Date of birth : 03/12/1980……………………………………..…………………...

Title : An Efficient Network Traffic Classification Based on Vital……………...

Random Forest for High Dimensional Dataset ……..……………..

……………………………………………..………………………………..…

. Academic Session : 2016/2017………………….……………………………………..............

I hereby declare that this thesis becomes the property of University Malaysia Perlis (UniMAP) and to be placed at the library of UniMAP. This thesis is classified as:

☐ **CONFIDENTIAL** (Contains confidential information under the Official Secret Act 1972)

☐ **RESTICTED** (Contains where research was done) restricted Information as specified by the organization

☑ **OPEN ACCESS** I agree that MY thesis is to be made immediately available as hard copy or online open access (full text)

I, the author, give permission to the UniMAP to reproduce this thesis in whole or in part for the purpose of research or academic exchange only (except during a period of —— years, if so requested above).

Certified by:

SIGNATURE **SIGNATURE OF SUPERVISOR**

(g1340811014/ A10833650)

(NEW IC NO. / PASSPORT NO.) **NAME OF SUPERVISOR**

Date: Date:

NOTES : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentially or restriction.

ACKNOWLEDGEMENTS

بسم الله الرحمن الرحيم

نَرْفَعُ دَرَجَٰتٖ مَّن نَّشَآءُۗ وَفَوْقَ كُلِّ ذِي عِلْمٍ عَلِيمٞ ٧٦ (سورة يوسف)

((الْحَــــمْدُ لِلَّهِ الَّذِي بِنِعْمَتِــهِ تَتِمُّ الصَّــالِحَاتُ))

The favor, above all, before all, and after all, is entirely Allah‟s (SWT), to whom my never-ending thanks and praise are humbly due.

I would like to take this opportunity to convey my sincere thanks and deepest gratitude to my supervisor, Dr. Rozime Razif Bin Othman for all the help and valuable guidance provided to me during throughout my period of research and especially for his confidence in me. Also, for his insight and the precious scientific guidance, this greatly helped me in the research's progress and to accomplish of this thesis through his academic advice. I would also like to thank my co-supervisor Dr. Mohammed Anbar for his support, and invaluable suggestions during my research.

I would like to express my appreciation to Prof. Ir. Dr. Badlishah Ahmed the Dean of Computer and Communication Engineering School for his support and help towards my postgraduate affairs. My deep appreciation and special thanks go to the University Malaysia Perlis (UniMAP) and to the staff of the UniMAP for their cooperation, support and good treatment for foreign students; especially school members my colleagues. My Acknowledgement also goes to the Center of Graduate Studies, and the university library for their help and support.

Also, my deep appreciation and special thanks go to Prof. Dr. Syed Alwee Aljunid Bin Syed Junid the Dean of Research, Management and Innovation Centre for his unlimited helps and support during my study.

Moreover, I would like to thank those who are always in my heart; my father for his endless and continuous encouragement and constant support, my mother for her continuous prayers and inspiration. My sincere gratitude goes to my dearest brothers (Numan, Alabass) for their continuous supporting and encouragement. In addition to, my sisters for always keeping a smile on my face and motivating me all the time. My special acknowledgement goes to my cousin (Ammar) for his support and encouragement through my study journey. My thankful and grateful goes to Dr. Salah Noori for his sincere brotherly advices and continuous support during my study.

DEDICATION

((وَقُل رَّبِّ ٱرۡحَمۡهُمَا كَمَا رَبَّيَانِي صَغِيرٗا ٢٤))

سورة القصص

I would like to dedicate this thesis to the two persons (My father and my mother) who you always feel their presence beside me despite the long distance by introducing the continuous Doa'a, encourages and unlimited support during my study.

((قَالَ سَنَشُدُّ عَضُدَكَ بِأَخِيكَ وَنَجۡعَلُ لَكُمَا سُلۡطَٰنٗا ٣٥))

سورة الإسراء

I would also like to dedicate this thesis to my dearest brother Alabass for his exceptional sacrificing, continues help, prayers and support during my study.

((وصلى اللهم على سيدنا محمد وعلى اله وصحبه وسلم))

TABLE OF CONTENTS

CHAPTER 2 LITERATURE REVIEW

CHAPTER 3 METHODOLOGY AND DESIGN

CHAPTER 5 RESULTS AND DISCUSSION

CHAPTER 6 CONCLUSION AND FUTURE WORK

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABRF	Active Build-model Random Forest
BFW	Broadband Fixed Wireless
CBMG	Customer Behavior Model Graph
Chi2	Chi-Squared
CPU	Central Processing Unit
DCCP	Datagram Congestion Control Protocol
DNS	Domain Name System
DoS	Denial of Service
DPI	Deep-Packet Inspection
DR	Data Reduction
DSL	Digital Subscriber Line
EM	Expectation Maximization
FCBF	Fast Correlation-Based Filter
FSCGRV	Fast correlation based filter, Significance feature evaluator, Chi-squared and Gain ratio: and Ranking and Borda Count Voting
FTP	File Transfer Protocol
GR	Gain Ratio
HMM	Hidden Markov Model
HTTP	Hypertext Transfer Protocol
IANA	Internet Assigned Numbers Authority
IBL	Instance-Based Learning
ID3	Iterative Dichotomiser 3
IDS	Intrusion Detection System

IETF	Internet Engineering Task Force
IG	Information Gain
IP	Internet Protocol
IPTs	Inter-Arrival Time Packets
IRC	Internet Relay Chat
ISP	Internet service providers
KNN	K-Nearest Neighbors
ML	Machine Learning
MySQL	My Structured Query Language
NAT	Network Address Translation
NB	Naïve Bayes
NBKE	Naïve Bayes Kernel Estimation
OAA	One-Against-All
OAO	One-Against-One
P2P	Peer-to-Peer
PDF	Probability Density Function
PNN	Probabilistic Neural Network
POP3	Post Office Protocol 3
PS	Packet Size
QoS	Quality of Service
RF	Random Forest
SCTP	Stream Control Transmission Protocol
SFE	Significance Feature Evaluator
SMTP	Simple Mail Transfer Protocol
SSH	Secure Shell

SU	Symmetrical Uncertainty
SVM	Support Vector Machine
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
VoIP	Voice over Internet Protocol
VPN	Virtual Private Network
VRF	Vital Random Forest
WWW	World Wide Web

Pengelasan Trafik Rangkaian Berkesan berdasarkan Vital Random Forest bagi Set Data Dimensi Tinggi

ABSTRAK

Tesis ini mencadangkan serta melaksanakan satu kaedah pengelasan trafik rangkaian yang berkesan berdasarkan *vital random forest* (VRF) baru bagi pemeriksaan data berdimensi tinggi. Kejuruteraan trafik rangkaian merupakan satu daripada teknologi penting yang memaparkan pertumbuhan yang pantas dalam revolusi teknologi seantero dunia. Pengelasan trafik rangkaian memberikan faedah yang boleh dipertimbangkan sebagai suatu wadah kejuruteraan rangkaian yang penting bagi sekuriti rangkaian, reka bentuk rangkaian dan juga bagi pemantauan dan pengurusan rangkaian. Ia memberikan perkhidmatan yang berbeza seperti mengenal pasti aplikasi yang paling banyak menggunakan sumber rangkaian, ia mewakili bahagian teras daripada sistem pengesanan instrusi secara automatik, ia membantu mengesan aplikasi anomali, dan ia juga membantu mengenali aplikasi yang digunakan di seantero dunia bagi tujuan penawaran produk baru. Dalam kata lain, pelbagai cabaran yang dihadapi oleh para jurutera rangkaian dalam usaha mereka mengelaskan trafik. Yang paling biasa adalah pertambahan jenis aplikasi dan saiz data trafik yang besar. Oleh itu, berdasarkan kajian lepas, ramai penyelidik berlumba-lumba memperkenalkan kaedah kaedah yang berkesan bagi pengelasan trafik. Keberkesanan pengelasan trafik bergantung pada beberapa faktor penting seperti ketepatan pengelasan, penggunaan memori dan masa pemprosesan Tesis ini mencadangkan vital random forest VRF sebagai pengelasan trafik rangkaian yang berkesan, yang merupakan satu pakej yang memperkenalkan teknik pemilihan penapis baru, pengurangan input data dan model binaan baru bagi kaedah *random forest* baru untuk mengelaskan trafik rangkaian bagi set data yang besar. VRF juga bermatlamat mengurangkan masa pemprosesan, meningkatkan ketepatan pengelasan, serta mengurangkan penggunaan memori. Rangka kerja VRF yang dicadangkan memberikan tiga sumbangan, pertama, reka bentuk *teknik pemilihan penapis baru* berdasarkan penggunaan teknik empat dan dua penapis bagi memilih set penapis yang paling signifikan. Kedua, pengurangan input data yang bertujuan menyingkirkan semua input rekod yang berlebihan dengan pengkatogerian kelas. Ketiga, mereka bentuk model binaan baru bagi *random forest* standard, yang dikenali sebagai ABRF (Active Build model Random Forest). ABRF dibina berdasarkan pokok aktif (pengelas), sebaliknya, pokok pasif didiagnosis, dikeluarkan dan diganti dengan pokok aktif. Keputusan eksperimen adalah berdasarkan beberapa daripada set data tanda aras (data set global). Data ini dikumpul daripada beberapa rangkaian bagi mencapai semua paket yang berkaitan dengan TCP, UDP dan IP dalam kedua-dua arah. Keputusan menunjukkan penambahbaikan yang ketara dari segi faktor yang signifikan, iaitu ketepatan pengelasan, masa pemprosesan dan penggunaan memori. Secara purata, keputusan VRF berhubung dengan faktor ini dijalankan berdasarkan. 16 set data tanda aras, ketepatan keseluruhan VRF meningkat sebanyak 6% berbanding dengan hutan rawak asal untuk mencapai 99.58%, masa pemprosesan telah menurun dengan perbezaan 62% manakala VRF menggunakan hanya 38% daripada jumlah purata masa dan purata penggunaan memori dikurangkan sebanyak 40%. Selanjutnya, VRF telah disahkan melalui perbandingan keputusan daripada faktor yang dinyatakan di atas dengan penyelidikan terdahulu.

An Efficient Network Traffic Classification Based on Vital Random Forest for High Dimensional Dataset

ABSTRACT

This thesis proposes and implements an efficient network traffic classification method based on a new vital random forest for high dimensional data. Network traffic engineering is one of the most important technologies that have witnessed a rapid growth in the revolution of worldwide technologies. Network traffic classification has added considerable interest as an important network engineering tool for network security, network design, as well as network monitoring and management. It can introduce different services such as identifying the applications which are most consuming for network resources, it represents the core part of automated intrusion detection systems, it helps to detect anomaly applications and it helps to know the widely-used applications for the intention of offering new products. On the other hand, several challenges faced by network engineers on their course to classify traffic. The most common of which are increasing application types and the huge size of data traffics. Therefore, many researchers have been competing in literature to introduce an efficient method for traffic classification. The efficiency is dependent on important factors such as classification accuracy, memory consumption and processing time. This thesis presents a *Vital Random Forest* (*VRF)* as efficient network traffic classification which is a one package that introduces a new features-selection technique, data inputs reduction and a new build model for original random forest method to classify network traffic for huge datasets. VRF aims to reduce processing time, increase classification accuracy and decrease memory consumption. The proposed framework of VRF consists of three contributions; first one is design of a *new features-selection technique* based on adopting four techniques and two filters for selecting most significant features set. Second is a data input reduction aiming to remove all redundant record inputs with class categorization. Third is to design a new build model for standard random forest called *Active Build model Random Forest* (*ABRF)*. ABRF is built based on active trees (classifiers) while the passive trees are diagnosed, omitted and replaced with active trees. The results from the experiments conducted are based on several benchmark datasets (global dataset). These data were collected from the edge of a network to access all packets associated with a TCP, UDP and IP connections in both directions. The results exhibit noticeable improvement in terms of three significant factors, namely; classification accuracy, processing time and memory consumption. The averaging results of VRF with regard to these factors were conducted based on 16 benchmark datasets where the classification accuracy of VRF is increased by 6% compared with original random forest to reach 99.58%, the processing time was decreased down with difference 62% while VRF consumed only 38% from the total average time and the averaging memory consumption is reduced by 40%. Furthermore, VRF has been validated by comparing the results for the above-mentioned factors with previous works.

CHAPTER 1

INTRODUCTION

1.1 Introduction

Network traffic engineering is one of the most important technologies that have witnessed a rapid growth in the revolution of global technologies. Network traffic identification and classification have recently gained considerable interest as an important network engineering tool for network security, network design, as well as network monitoring and management (Dainotti, Pescape, & Claffy, 2012).

Network traffic classification is a process that categorizes network traffic according to various parameters (e.g. port number, arrival time, type of protocol, packet length, etc) into a number of application classes such as (P2P, WWW, Mail, etc.). This method introduces multi-beneficial solutions in different avenues, such as network security, network management, network measurement and quality-of-service (QoS) (Callado et al., 2009). Many of network engineers have had started to inspect and analyze network traffic but they faced several novel challenges the most significant of which are the huge amount of classified data traffic and the variety of applications. As a result, numerous studies have been proposed to face the above mentioned challenges and they started competing in terms of different factors such as accuracy of classification, memory consumption, CPU consumption and time consumption of processing (Dainotti, Pescapé, & Sansone, 2011; Nguyen & Armitage, 2008). The studies introduced various solutions that involved two major types: First, representing traditional solutions that include well-known port number, deep packet inspection and behavior-based approach. Second is Machine Learning (ML): there was a considerable interest in several disciplines such as philosophical, logical and conceptual issues, and then research interest shifted to computational and algorithmic aspects of ML that is driven mainly by practical application. That's why many studies were rerouted towards ML. This thesis focuses on supervised machine learning method and its contribution in the area of network traffic classification.

1.2 Background

This section provides a general background to the work presented in this thesis. It briefly introduces the principal technologies referenced throughout this thesis as network traffic engineering, Machine Learning and Random Forest.

1.2.1 Network Traffic Engineering

First and foremost, to understand the network traffic engineering we need to define the network engineering, it's a method of manipulating your network to suit your traffic. Network traffic engineering is defined as that aspect of network engineering of dealing with the issue of performance evaluation and performance optimization of operational IP networks (Callado et al., 2009). Traffic Engineering encompasses the application of technology and scientific principles to the measurement, characterization, modeling, and control of network traffic. The enhancement of the network is achieved at both levels: traffic and resources, with regard to different performance requirements and utilizing network resources economically and reliably. The performance is measured in terms of delay, jitter, packet loss and throughput. The main purpose of network traffic engineering is to facilitate reliable network operations. This is accomplished using policies to keep network survivability and minimizing the vulnerability of the network to service outages