DISTRIBUTED ONLINE AVERAGED ONE DEPENDENCE ESTIMATOR ALGORITHM FOR NETWORK ANOMALY DETECTION SYSTEMS

by

MUKRIMAH BINTI NAWIR
(1530211891)

A thesis submitted in fulfillment of the requirements for the degree of Master of Science in Computer Engineering

School of Computer and Communication Engineering
UNIVERSITI MALAYSIA PERLIS

2019

ACKNOWLEDGMENT

This thesis becomes a reality with the kind support and help of many individuals. I would like to extend my sincere thanks to all of them. This journey is the most precious things that I had treasure in my life with great life lessons, experiences of immense knowledge, friendship, and kindness beyond boundaries. Foremost, I want to offer this endeavor to Allah the Almighty for the wisdom HE bestowed upon me, the strength, peace of my mind, and good health in order to complete my postgraduate studies.

I am very grateful to my great and supportive supervisor, Dr Amiza Amir, who expertise, generous guidance, and support made it possible for me to work on a topic that was a great interest to me. It was a pleasure to working with her to explore and opportunity to learn and gain new knowledge. Her unflinching wisdom and patience with me in completing my studies really appreciated. Also, not to be forgotten my Co-supervisor, Dr Ong Bi Lynn, for imparting her knowledge in this study.

For the ancestors who paved the path before me upon whose shoulders I stand. This also dedicated to my beloved parents, Nawir bin Salim and Hasnah binti Ladadi, for their continued prayers and doa for my successful till the end of my journey in this research study. Both always keep remind and give motivations on me to stay strong and never give up on what I am doing. With all my siblings and my big family behind me and they are peoples that I share my feelings and stories to express out my stress.

The research reported sponsored by Research Acculturation Grant Scheme (RAGS) and I take these opportunities to thanks them, Malaysian Ministry of Higher Education (MOHE) – MyBRAIN15 and University of Malaysia Perlis (UniMAP) for the financial support and facilities provided. My big thanks to my friend Ng Hui Qun, the one who always ready to help me in coding, technical problem, and her sincerely sharing her knowledge. Of course, really appreciated the help from kak Nuramina Ramli who sit beside me spend her time to me writing coding in JAVA language, Lee Yee Ann who I refer to study the network course.

Madam Ismahayati, Madam Nur Baya, and Kak Fatimah Noni who always teach, guide, and advise me during my studies. There are so many names that I really want to be mentioned but I know all of you know who you are that so meaningful in my journey. All my Embedded Network Advanced Computing (ENAC) cluster colleagues – Nur Waheeda Basharudin, Wan Aida Nadia Wan Abdullah, Siti Asilah Yah, Farah Wahida Zulkefli, and Ng Yen Phing. All I can say is Thank You so much! You all are my awesome besties.

©This item is protected by original copyright

TABLE OF CONTENTS

©This item is protected by original copyright

©This item is protected by original copyright

©This item is protected by original copyright

LIST OF TABLES

©This item is protected by original copyright

©This item is protected by original copyright

LIST OF FIGURES

©This item is protected by original copyright

©This item is protected by original copyright

LIST OF ABBREVIATIONS

ARFF	Attribute-Relation File Format
AODE	Averaged One Dependence Estimator
BN	Bayesian Network
CBAODE	Centralized Batch AODE
CBNB	Centralized Batch NB
COAODE	Centralized Online AODE
CONB	Centralized Online NB
.csv	comma separator vector
DBAODE	Distributed Batch AODE
DBNB	Distributed Batch NB
DOAODE	Distributed Online AODE
DONB	Distributed Online NB
FPR	False Positive Rate
FNR	False Negative Rate
GUI	Graphical User Interface
IDS	Intrusion Detection System
IoT	Internet of Things
ML	Machine Learning
MLP	Multi-Layer Perceptron
NADS	Network Anomaly Detection System
NB	Naïve Bayes
NN	Neural Network
P2P	Peer-to-Peer
RBFN	Radial Basis Function Network
SVDD	Support Vector Data Description
SVM	Support Vector Machine
TPR	True Positive Rate
TNR	True Negative Rate
UNSW-NB15	University of New South Wales-Network Benchmark 2015
WEKA	Waikato Environment for Knowledge Analysis

©This item is protected by original copyright

LIST OF SYMBOLS

$\forall$	For all observed values
$\in$	Elements of
$\boldsymbol{n}$	Total number of features
$\boldsymbol{x_j}$	Feature's value
$\wedge$	Logical and
$\boldsymbol{\hat{p}}$	Hat p

©This item is protected by original copyright

Algoritma Edaran Penganggar Pergantungan Purata dalam Talian untuk Sistem Pengesanan Anomali Rangkaian

ABSTRAK

Sistem pengesanan anomali rangkaian (NADS) digunakan secara meluas untuk memantau aplikasi yang melibatkan data aliran seperti Internet benda (IoT) dengan menentukan normal dan anomali dalam rangkaian. Menguruskan aliran data yang bersifat rangkaian data yang sering mengemas kini data disebabkan trafik rangkaian masuk dengan cepat dalam sistem memerlukan algoritma klasifikasi pembelajaran yang pantas untuk mengesan corak. Oleh itu, penyelidikan ini menggunakan algoritma penaksir pergantungan secara purata dalam talian (AODE) bagi data penstriman yang besar untuk klasifikasi binari dan pelbagai kelas untuk pembelajaran pantas data untuk memastikan pengeluar sentiasa dikemas kini. Selain itu, untuk menangani sejumlah besar data dan isu pemusat, pengelas adalah perlu untuk dibangunkan dalam algoritma pengkelasan yang diedarkan untuk mengesan corak trafik rangkaian dengan menggunakan dataset rangkaian. Oleh itu, tesis ini membangunkan sistem pengesanan anomali rangkaian dengan menggunakan algoritma AODE (DOAODE) dalam talian yang diedarkan dengan menggunakan dataset UNSW-NB15 yang berkaitan dengan beberapa isu seperti pengkelasan skala besar, data kerap dikemas kini, dan pemusatan. Algoritma DOAODE akan mengesan serangan rangkaian dengan menjalin beberapa stesen. Kemudian, pengelas tempatan pada setiap nod akan bergabung dengan menggunakan pengundian majoriti untuk mempunyai pengelas global dan membuat ramalan terakhir trafik rangkaian. Penemuan pertama dari eksperimen yang dijalankan menunjukkan bahawa algoritma AODE dalam talian mempunyai ketepatan yang tinggi dengan peratusan yang sama dengan 97.26% untuk klasifikasi binari dan 83.32% untuk klasifikasi pelbagai kelas. Juga, pengeluar dalam talian belajar lebih cepat daripada pengelas terkumpul. Di mana AODE dalam talian mengambil kurang daripada 10 saat bagi kumpulan binari serta klasifikasi pelbagai kelas kumpulan data rangkaian. Penemuan kedua menunjukkan bahawa kerja yang dicadangkan (algoritma DOAODE) memperolehi ketepatan yang tinggi dan tidak banyak berbeza dengan ketepatan algoritma terpusat di mana hasilnya direkodkan dalam julat 95% hingga 97% untuk pengkelasan binari dan kira-kira 83% untuk kelas berbilang klasifikasi. Walaupun, ketepatan DOAODE terdegradasi tetapi hasil yang diperoleh masih boleh dibandingkan dengan pengelas berpusat dan ia boleh mengelakkan titik kegagalan tunggal berlaku dalam sistem rangkaian kerana seni bina berkongsi pengetahuan di antara mereka dan semua nod mempunyai tahap yang sama di mana mereka semua boleh membuat ramalan pada trafik rangkaian.

©This item is protected by original copyright

Distributed Online Averaged One Dependence Estimator Algorithm for Network Anomaly Detection Systems

ABSTRACT

Network anomaly detection systems (NADS) are widely used for monitoring the applications involving the streaming data such as the Internet of Things (IoT) by determining the normal and anomalies in the network. Streaming data deal with the nature of network data that is frequent updating data due to the fast-incoming network traffic in the system requires a fast learning classification algorithm to detect the patterns. Hence, this research used an online averaged one dependence estimator (AODE) algorithm of large streaming data for binary and multi-class classification for fast learning the data to ensure the classifier always updated. Furthermore, to deal with a large amount of data and centralization issues the classifier is necessary to develop in a distributed classification algorithm to detect the pattern of network traffic by using network dataset. Therefore, this thesis developed a network anomaly detection system by using a distributed online AODE (DOAODE) algorithm by using UNSW-NB15 dataset that concerning several issues such as the large-scale classification, frequently updated data, and centralization. DOAODE algorithm will detect the network attack by collaborating several stations. Then, the local classifier at each node is combined by using majority voting to have a global classifier and make a final prediction of network traffic. First finding from the conducted experiment showed that the online AODE algorithm is high in accuracy with the percentage equal to 97.26% for binary classification and 83.32% for multi-class classification. Also, online classifier learns faster than a batch classifier. Where online AODE took about less than 10 seconds for binary as well as multi-class classification the network dataset. Second, the finding shows that the proposed work (DOAODE algorithm) obtained high accuracy and not much diverged from the accuracy of centralized algorithm where the results recorded in the range 95% to 97% for binary classification and approximately 83% for multi-class classification. Although, the accuracy of DOAODE algorithm degraded the obtained result still comparable to the centralized classifier and it can avoid the single failure point to occur in a network system due to the architecture share their knowledge among them and all the nodes have the same level where all of them can make a prediction on the network traffic.

©This item is protected by original copyright

CHAPTER 1 : INTRODUCTION

The thesis is about a distributed online classification algorithm for network anomaly detection system (NADS). Chapter 1 is to establish the context, the background of work, indicate the issues or problems regard NADS by using machine learning (ML) approaches, and the proposed work that had been used in this research.

©This item is protected by original copyright

1.1 Research Background

The growth of technologies in this present time become worrisome when the huge amount of data information nowadays can be accessed easily on the Internet without a limitation and any authorization. As recorded by Malaysian Communications and Multimedia Commission (MCMC), in the year 2017 there are 24.5 million Internet users (76.9%) out of 32 million number of Malaysians (M. F. Ahmad, 2017). This condition brings up the economical, scalability, and security issues.

For example, there is a lot of social media that connects people around the world to share their information as well as a software distribution tool such as Instagram, Facebook, E-mails, and Yahoo Messenger. All these technologies paradigm are open, publicly, and accessible that endangers the users to be susceptible to various network threats (Nawir, Amir, Yaakob, & Lynn, 2016) including spam, phishing, or inappropriate contents if no protection systems (i.e firewall, Intrusion Prevention System, and NADS) that concerning security matters.

In such cases, this thesis interested to develop a network anomaly detection system (NADS) for large-scale data that can distinguish the pattern of normal or anomalous data in a network system. For instance, the NADS can be built in the desktop through simulation to determine the network access log for Intrusion Detection System (IDS). Unfortunately, they still lack focus on the security issues to protect their applications such as video streaming, online gaming, and online social network. In the context of a computer network, anomalies are known as network attacks where the patterns behave differently from the normal traffic.

Figure 1.1 is an example of anomalies (red points) in a network. They are categorized as anomalies because there is a far distance from most of all data points. As illustrated the points *O*1 and *O*2 are far away from the data points in Class 2. It is possible to build NADS within different domains such as medical diagnostic (Pachauri & Sharma, 2015), telecommunication (Niemelä, 2010), and water distribution system (Kühnert, Baruthio, Bernard, Steinmetz, & Weber, 2015). Thus, the ML approach is necessary to be employed for NADS.

©This item is protected by original copyright

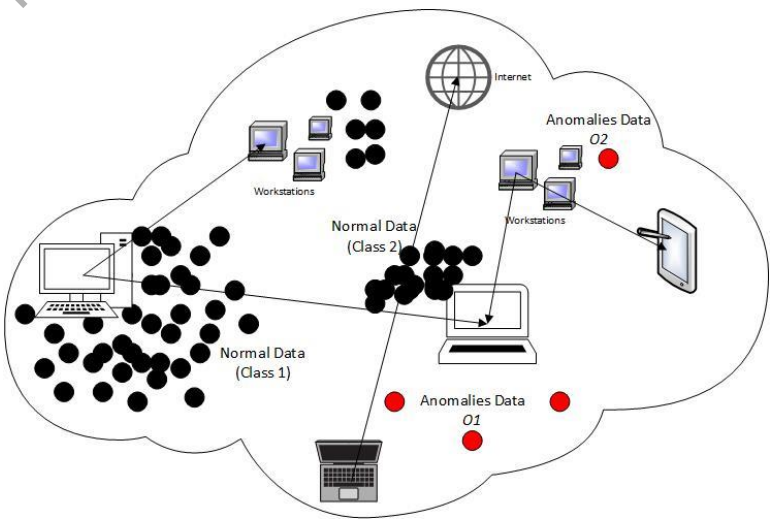

Figure 1.1 Anomalies in Computer Network

The ML approaches have two types of learning, supervised or unsupervised learning. Supervised learning involves labelled data where the data can be binary or multi-class data. Whereas, unsupervised does not require labelled data. In the context of network data, binary classification classifies the network pattern is normal or attack. However, multi-class classification is applied when the data represented in a complex pattern that consists of ten classes that include normal traffic and subcategories of attacks.

Network Anomaly Detection System (NADS) is a defense system for monitoring network-based security by employing the training and testing toward the behavior in a computer network using a powerful tool that can determine the normal and anomalous data according to the learned pattern. From the learned process (training stage) the knowledge can classify the data according to the same behavior. Anomalous data (network attacks) are the deviation of normal patterns in a network. NADS is used to determine the normal as well as anomalous data in a network by using a powerful tool such as machine learning approaches (Bhattacharyya & Kalita, 2013) as shown in Figure 1.2 where Distributed Denial of Service (DDoS) is an example of a network attack that compromise in the system causes the failure of a network. A taxonomy of security attacks in various domains was discussed in the paper (Nawir et al., 2016).

©This item is protected by original copyright

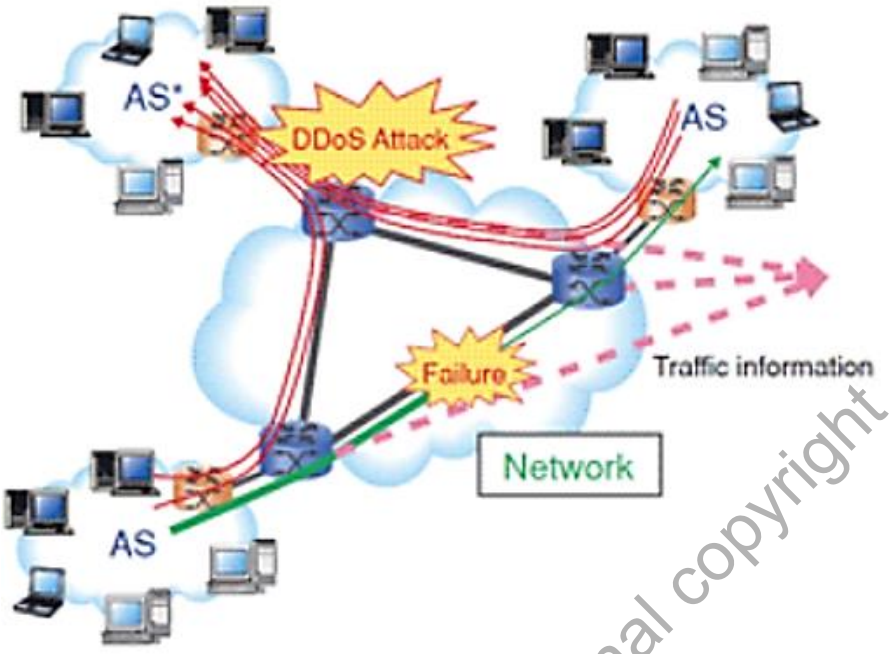

Figure 1.2 Network Anomaly Detection System

©This item is protected by original copyright

Generally, anomaly detection widely used in many applications either using statistical or machine learning (ML) technique. For NADS, it is suitable to use machine learning technique because they can automatically learn the large-scale data within networks meanwhile statistical learning emphasize on "mathematical modelling" (probability) that requires more time to detect abnormal data. The machine learning techniques, the user does not intervene in telling a model how to build the classifier as statistical learning does. Yet, the roughly speaking ML is a technique to automatically recognize the worthwhile patterns in data by learning and adapting to the knowledge that gained.

There are two learning strategies in the context of ML: online and batch learning (Hussein, 2014). An online anomaly detection becomes prominent especially when it involves the processing of streaming data in many intelligent systems from simple applications to more complex applications. Online processing in ML can directly inform

the system and process the data immediately (Gumus, Sakar, Erdem, & Kursun, 2014). An online classification by using machine learning framework is an effort to control the systems that comprises large-scale data or data streams (continuous large data scale) which keep changing the nature of data (S. Ahmad & Purdy, 2016; Pham, Venkatesh, Lazarescu, & Budhaditya, 2014; Rettig, Khayati, Cudré-Mauroux, & Piórkowski, 2015). Unlike online learning, in batch processing, the data is group first before being process periodically.

Combining geographically-distributed data from multiple monitoring sites results in a large amount of collective knowledge within the network (Kühnert et al., 2015). Hence, manipulating this knowledge may lead to better results. The most common method, all the raw data at the single central site where the data mining process can be performed in a conservative way as shown in Figure 1.3. In a centralized monitoring system, a single server is used to keep track of and analyze the large-scale, global data from many different monitoring sites.

As depicted in Figure 1.3, an ensemble of sensing elements will communicate with their connected routers or servers (PC1, PC2, PC3, and PC4) that are called local monitoring sites. These sensing elements are not only limited to sensor devices but also mobile devices such as smartphones, tablets, and laptops. The local monitoring sites will send data to a centralized monitoring site (PC5) where the data from all local sites are analyzed and a single classifier model is developed. This requires a powerful computing facility to store and perform computations on a large combined dataset.

©This item is protected by original copyright

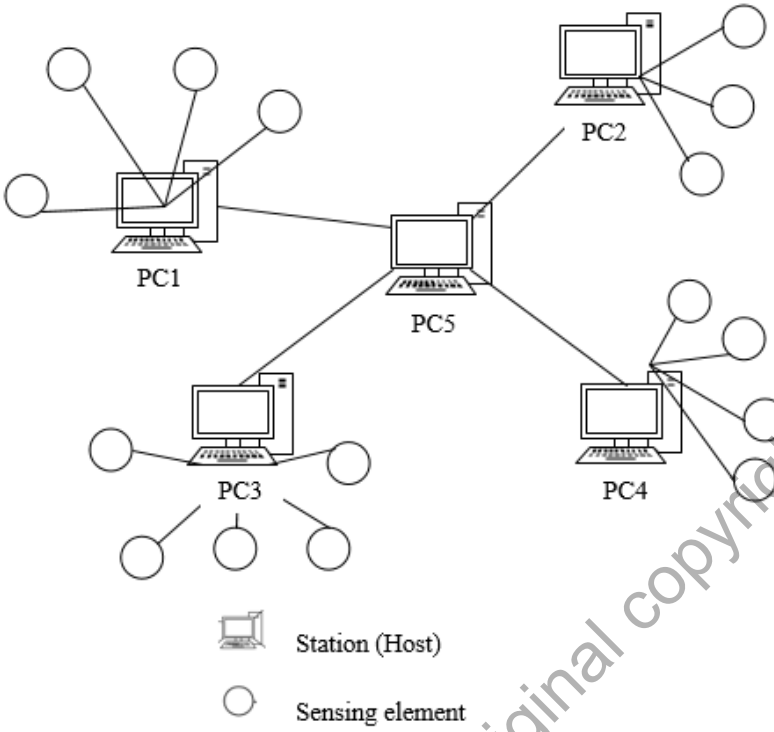

Figure 1.3 Global Monitoring Systems using Centralized Classifier

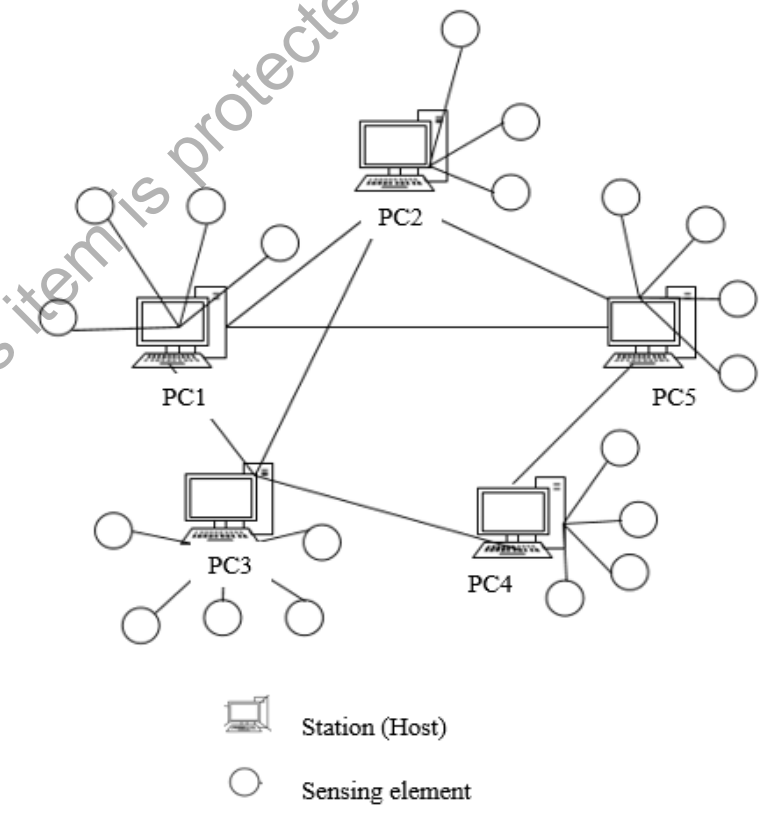

Figure 1.4 Global Monitoring Systems using Distributed Classifier

©This item is protected by original copyright

On the other hand, this thesis proposes a fully-distributed architecture of monitoring system as shown in Figure 1.4 where each local station develops its own classifier by using its local data. This approach provides an economical way to handle large data streams by storing and analyzing parts of the data streams separately at multiple distributed sites, and the local data analyzes (which are performed at every site) are combined through aggregation. Figure 1.4 is an example of a global monitoring system by using a distributed classifier where the knowledge from all local monitoring sites is combined by using distributed aggregation to produce a global prediction.

This research contributes a distributed online classification algorithm by applying an Averaged One Dependence Estimator (AODE) algorithm over a new network dataset, UNSW-NB15, for NADS. The framework lies to update local learners that will interpret the streaming data that associated with a common event required to be classified by using online learning. Each learner habits a local classifier to form a local prediction. Then that local prediction is amassing from each learner to be mingled to get an output of final prediction. For updating, the aggregation rule as in ensemble learning algorithms to ensure the classifier keep updating their status in the network.

The design considers the issues such as large-scale data, frequent update classifier, and centralization (that might cause a destroy whole system when there is an error occur). In this work, the simulation of several supervised ML algorithms in batch as well as an online manner for classifications (binary and multi-class data) and distributed learning that high detection rate and fast processing of ML algorithm. Therefore, the distributed online classifier has been proposed as an alternative scalable ML for NADS.

©This item is protected by original copyright

1.2 Problem Statement

With billions of mobile devices and sensors in the world today, computer networks are flooded with an enormous amount of data streams (network traffic). Nonetheless, the intelligent classification of large volumes of data streams with multi-dimensional data is non-trivial. Intelligent classification is often considered a computationally demanding task (Fatih Ertam and Engin Avci, 2016; Ganapathy, Kulothungan, Muthurajkumar, & Vijayalakshmi, 2013; Sohaib, Ahmad, & Khan, 2013). Solving this problem becomes more challenging considering the limited computational resources, memory, and power of these systems (Baig, Shaheen, & Abdelaal, 2011; Pachauri & Sharma, 2015). Hence, the first problem is the difficulty to classify network data as either normal or anomalous considering the large and frequently updated network data.

Machine learning algorithms are effectively used for classification tasks. However, most of the popular machine learning algorithms such as Support Vector Machine (SVM) (Yong Xie, 2012) and neural networks are computationally expensive and require the data to be trained in batch processing that is unsuitable for streaming data (Al-Janabi & Saeed, 2011; Baig et al., 2011; Chowdhury, Ferens, & Ferens, 2010). Therefore, online training that performs fast training without the need to buffer the data is important in analyzing large data streams of network traffic. The NADS is commonly conducted in binary classification which is insufficient to determine the behavior of network data in a system as conducted in the papers (Belouch, Mustapha, Hadaj, Salah, El, Idhammad, 2017; Nawir, Amir, Lynn, & Yaakob, 2018). Therefore, the design of an online learning ML algorithm for multi-class classification of NADS is required.

©This item is protected by original copyright

The second problem is the decentralization issues in distributed systems results in a difficulty to develop an ML classifier to detect network attacks (Butun, Kantarci, & Erol-Kantarci, 2015; Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2012). Often, in distributed systems, multiple computers (e.g. base stations, servers, or software defined networks controller) are geographically located all over the world and connected through the Internet. Monitoring resource in a network system is a difficult task where the development of this paradigm endangers the user's computer into the safety and security where the data that exposed into the Internet make the data easier to be accessed, non-authorization, and open (Hussein, 2014). In addition, the problem of inefficient distributed monitoring due to local monitoring in distributed system. Hence, local monitoring by a local classifier at the local site is preferred than sending data to a centralized monitoring site. However, this leads to the question of how the knowledge from many different local classifiers can be shared so that the knowledge can be manipulated for better results. Therefore, this thesis proposed that the network attack detection in these distributed systems can be achieved through a distributed monitoring system.

©This item is protected by original copyright

Finally, the last problem is the issue in the performance evaluation of the proposed network anomaly detection method. The challenges and issues of network anomaly detection datasets are presented in (Zuech, Khoshgoftaar, Seliya, Najafabadi, & Kemp, 2015). The CAIDAs datasets not effective to be used for network anomaly detection since their data resources might be removed during the simulation and it is not adaptable to any environment. The DARPA datasets (KDDCup1999) is impractical because there are a lot of redundant data instances in a set, perceived to skewness, and biases classification toward the training data (Gumus et al., 2014; Moustafa & Slay, 2015c; Osanaiye et al.,

2016). The existing dataset consists of irrelevant and redundant feature (Idhammad, Afdel, & Belouch, 2018; Janarthanan & Zargari, 2017). Hence, this thesis suggests the performance evaluation of the proposed network anomaly detection by using a benchmark dataset that represents the real network traffic dataset and consists of large number of data instances.

1.3 Research Questions

Whilst to complete this research, several questions can be understood and considered. They are including:

1. Which ML algorithm is suited to be used that well performed with high accuracy and fast training time?
2. How to build an online classification algorithm for NADS that is high in accuracy and time efficient when dealing with the large and rapid updated data stream?
3. How to build an online multi-class classification algorithm for NADS that classifies network traffic data into different types of attacks (more than two types)?
4. How to develop a distributed monitoring where knowledge from many different local classifiers can be shared so that the knowledge can be manipulated for better results?
5. How to evaluate the proposed network anomaly detection algorithm?

©This item is protected by original copyright

1.4 Objectives of the Research

The primary objective of this thesis is to develop an accurate and fast distributed online classification algorithm for distributed network anomaly detection system. The sub-objectives of the research are:

1. To propose an online network anomaly detection system for multi-class classification of the network dataset by using machine learning algorithms.
2. To propose a distributed monitoring technique by using distributed online averaged one dependence estimator (DOAODE) algorithm in the network anomaly detection system.
3. To evaluate the performances of the distributed online network anomaly detection system by using the UNSW-NB15 benchmark dataset.

1.5 Contributions

The contribution is a novel method of solving an old problem. The contribution of this research is obvious as the resulting outcomes can be capitalized as guidelines for network security monitoring, machine learning algorithms, and anomaly detection. Cyber-security is one of the examples that bring up the opportunities to design a better system that enable to handle the security issues. The distributed online classification algorithm for large-scale network anomaly detection system using the ML approach is implemented. As a summary of these contributions as follows:

©This item is protected by original copyright