EFFECTS OF FEATURE TRANSFORMATION AND SELECTION ON CLASSIFICATION OF NETWORK TRAFFIC ACTIVITIES

LIM WEN YING

FACULTY OF COMPUTING AND INFORMATICS

UNIVERSITI MALAYSIA SABAH

2015

EFFECTS OF FEATURE TRANSFORMATION AND SELECTION ON CLASSIFICATION OF NETWORK TRAFFIC ACTIVITIES

LIM WEN YING

THESIS SUBMITTED IN PARTIAL FULFILMENT FOR THE BACHELOR OF COMPUTER SCIENCE (NETWORK ENGINEERING)

FACULTY OF COMPUTING AND INFORMATICS

UNIVERSITI MALAYSIA SABAH

2015

DECLARATION

I hereby declare that this thesis, submitted to Universiti Malaysia Sabah as partial fulfilment of the requirements for the degree of Bachelor of Computer Science (Network Engineering), has not been submitted to any other university for any degree. I also certify that the work described herein is entirely my own, except for quotations and summaries sources of which have been duly acknowledged.

This thesis may be made available within the university library and may be photocopied or loaned to other libraries for the purposes of consultation.

22 JUNE 2015

.....

LIM WEN YING

BK 1111 0156

CERTIFIED BY

_____

Dr. Mohd Hanafi Ahmad Hijazi

SUPERVISOR

ACKNOWLEDGEMENT

First and foremost, I am grateful to God for the good health and well-being that were necessary to complete this research paper. I must thank my parents and family for their understanding and support. They have always given their kindness, patience and tolerance when I had rough times.

I wish to express my utmost appreciation and deepest gratitude to my supervisor, Dr Mohd Hanafi Ahmad Hijazi. He constantly provided me with constructive comments for improvement to this project. Weeks after weeks and consultations after consultations, he continuously enlightened me when I was in doubt and when there were areas that I lacked knowledge. In addition, he always gave words of wisdom that encouraged me to continuous work on this project. Without his continuous dedication in guiding me, I would never have completed this research paper.

Last but not least, Associate Professor Dr Rayner Alfred who provided me advice on improving the quality of this research paper for which I am thankful to. I also place on record, my gratitude to one and all, who directly or indirectly, have lent their hand in this venture.

LIM WEN YING
22 JUNE 2015

ABSTRACT

As new technologies are emerging day by day, network, regardless of the Internet or Intranet within a corporation often plays a crucial role in connecting people from all around the world. From military use to achieving business goals and household need, data security often get attention from computer scientists. Traditional security measures that include the installation of firewall and antivirus software are commonly utilised to prevent intrusion. However, such types of defence are merely sufficient to secure a network and data travelling across it. Thus, second lines of defence like Intrusion Detection System (IDS) and Intrusion Prevention System (IPS) are introduced to overcome the inadequacy of traditional security measures. Generally, IDS uses two approaches, the Anomaly Detection (A-IDS) and the Misuse Detection in order to identify patterns of intrusion. A-IDS often perform comparison of the model of normal and anomalous model. Depending on the ability to measure similarity or distance between a target and a known type, comparison is made to determine whether to establish a new target anomalous or not. This research aims to investigate the effects of feature transformation on the classification of network activities; the focus is to represent the data into point series form to permit the application of Time Series Classification (TSC). The TSC technique used is *k*-Nearest Neighbour (KNN) coupled with Dynamic Time Warping. Effects of using different similarity measures, Euclidean Distance (ED) and Cosine similarity algorithm are also investigated. Experiments conducted involve conversion of the categorical data by three different conversion techniques to generate point series data – simple, probability and entropy conversion. Comparison between different classifiers is also conducted. The performance of the classifier is best using 1NN with Euclidean distance and entropy conversion for categorical data, where the recorded accuracy is 99.19%.

ABSTRACK

Pembaharuan teknologi berlaku setiap hari, rangkaian, tidak kira daripada Internet mahupun Intranet yang terdapat dalam sebuah korperasi sering memainkan peranan penting dalam menghubungkan orang ramai dari seluruh dunia. Daripada penggunaan oleh pihak tentera atau dalam bidang perniagaan untuk mencapai matlamat harian dan keperluan isi rumah, keselamatan untuk data yang mengalir di seluruh rangkaian sering mendapat perhatian daripada ahli-ahli sains komputer. Langkah keselamatan tradisional termasuk pemasangan "firewall" dan perisian antivirus biasanya menggunakan untuk mencegah pencerobohan. Walau bagaimanapun, jenis pertahanan tersebut semata-mata adalah tidak cukup untuk memastikan keselamatan rangkaian dan data yang merentasinya. Oleh itu, pertahanan peringkat kedua seperti "Intrusion Detection System (IDS)" dan "Intrusion Prevention System (IPS)" diperkenalkan untuk mengatasi kekurangan langkah-langkah keselamatan tradisional. Secara umumnya, IDS menggunakan dua pendekatan, Pengesanan Anomali (A-IDS) dan Pengesanan Penyalahgunaan untuk mengenal pasti corak pencerobohan. Secara umumnya, A-IDS mengenal pasti pencerobohan dengan membuat perbandingan sasaran bersama modal biasa. Bergantung kepada keupayaan untuk mengukur persamaan atau jarak antara sasaran dan jenis yang dikenali, perbandingan dibuat untuk menentukan sama ada untuk memastikan sasaran baru anomali atau tidak. Kajian ini bertujuan untuk menyiasat kesan perubahan ciri klasifikasi aktiviti rangkaian; tumpuan adalah untuk mewakili data sebagai siri titik bagi membenarkan "Time Series Classification" (TSC) aplikasi. TSC teknik yang digunakan adalah "k-Nearest Neighbour" (KNN) berserta dengan "Dynamic Time Warping" (DTW). Kesan menggunakan pengukuran persamaan yang berbeza, "Euclidean Distance" (ED) dan "Cosine similarity" algoritma juga disiasat. Eksperimen yang dijalankan melibatkan penukaran data berkategori dengan menggunakan tiga teknik penukaran yang berbeza untuk menghasilkan data siri titik - mudah, kebarangkalian dan entropy. Perbandingan antara klasifikasi berbeza juga dijalankan. Prestasi klasifikasi itu adalah yang terbaik apabila menggunakan 1NN dengan pengukuran jarak Euclidean dan penukaran entropy untuk data berkategori, di mana ketepatan yang direkodkan adalah 99.16%.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Chapter Overview

This chapter serves to present a brief background and introduction so as to aid readers into better understanding of this research paper. Section 1.2 presents the problem background. Section 1.3 and 1.4 describe research statement and objectives. Section 1.5 presents the research scope. The methodology used in this research paper is briefly lined out in Section 1.6 whereas the organization of this report is described in Section 1.7.

1.2 Problem Background

In this 21st century that is dominated by social networking, the Internet has surged to reveal itself as one of the most promising technologies that affect human in numerous ways; it has become increasingly critical to human. Private and confidential data that are propagated through the network are exposed and made vulnerable to attacks. Recent attacks such as the Cyber-attack on U.S. Public utility and its control system network and also the leakage of celebrity private photos ("Apple confirms accounts compromised but denies security breach," 2014) that are believed to be obtained from Apple iCloud backup services again give prominence to the importance of network security.

Traditional network traffic monitor detects regular network performance, recognizing application's identity by assuming that most applications constantly use 'well known' or common TCP/UDP port numbers (visible in the TCP or UDP headers). While this convention has been active in the early days of the Internet, this however, are merely sufficient in our modern days. Port-based estimates are currently significantly not reliable; as unpredictable (or at least obscure) port numbers are

increasingly being used for various applications, and also with the continuous emergence of new protocols, it has become increasingly difficult to get the details of the network traffic component. For this reason, researchers propose a new method to identify current sophisticated traffic data generated from various newly emerging network-based applications.

An Intrusion Detection System (IDS) is a network-monitoring system that is passive in nature. It is configured mainly to monitor, identify and initiate alerts for attacks or compromise on the network. Unlike Intrusion Prevention System (IPS), it does not do any direct action or measure to the potential breach. Signature-based and Anomaly detection are two general approaches to computer IDS. An IDS that is signature-based (also known as knowledge-based) uses pre-defined set of rule to identify intrusion. By comparing the current traffic pattern of known and documented attacks, signature-based IDS determines attack when there is a match to the signature in the attack database. Signature-based is the most widely use type of IDS currently (Chowdharyet *al.*, 2014). However systems employing Signature-based detection method has a limitation of being unable in detecting intrusion when the signature of an attack is not recorded in the database. Furthermore, these systems are incapable of integrating information that comes from heterogeneous sources where the latter can provide informative details on the on-going network activities of the system (More et *al.*, 2012). In anomaly detection, the IDS capture the network traffic activity and based on that create a profile representing its stochastic behaviour. During the anomaly detection process, two data sets of network activity are involved, with one as the real-time profile recorded over time and another would be the previously trained profile. IDS function by attempt estimating the behaviour of the network traffic activity, normal or abnormal and trigger anomaly alarms whenever a predefined threshold (pre-defined abnormalities) is exceeded (García-Teodoro et *al.*, 2009). In general, two phases - the learning phase and the detection phase - made up the algorithm performed within an Anomaly-based Intrusion Detection system (A-IDS). The detector learns the normal behaviour of a network system by recording the data representing normal or "non-malicious" system activity in the training phase. Meanwhile, in the detection phase, the detector compares the input data to its learnt model of nominal behaviour to report any deviations as anomalies or attacks. García-

Teodoro et al., 2009 in their research paper highlighted some of the most significant challenges and issues in Anomaly-based Intrusion Detection:

(i) Low detection efficiency

This aspect is generally explained as arising from the lack of good studies on the nature of the intrusion events. The problem calls for the exploration and development of new, accurate processing schemes, as well as better structured approaches to modelling network systems.

(ii) Low throughput and high cost

Mainly due to higher data rates (Gbps) that characterize current wideband transmission technologies. Some proposals intended to optimize intrusion detection are concerned with grid techniques and distributed detection paradigms.

As mentioned earlier, A-IDS often performs a comparison of the model of normal and anomalous model. Depending on the ability to measure similarity or distance between a target and a known type, comparison is made to determine whether to establish a new target anomalous or not. Thus the distance or similarity employed will greatly affect the effectiveness of an A-IDS.

Data pre-processing is required in all knowledge discovery tasks, including network-based intrusion detection, which attempts to classify network traffic as normal or anomalous. Pre-processing converts network traffic into a series of observations, where each observation is represented as a feature vector. Observations are optionally labelled with its class, such as "normal" or "anomalous". These feature vectors are then suitable as input to data mining or machine learning algorithms (Davis and Clark, 2011). Feature construction aims to create additional features with a better discriminative ability than the initial feature set. This can bring significant improvement to machine learning algorithms. A well-defined feature extraction algorithm makes the classification process more effective and efficient (Datti and Verma, 2010). To decrease the time needed for an IDS to detect an intrusion, data dimension for a particular network traffic need to be reduced,

insignificant features should be removed or omitted, subsequently improving the performance of the IDS. The goal of features extraction lies in shrinking a relative huge data dimension to a smaller size and increasing the accuracy of classifier by preserving the features that have the most significance on the class label and omitting features that contribute less.

1.3 Problem Statement

From the previous section, the main question of this research paper would be "How feature transformation and selection affects the performance of the classifier" This question gives rise to two sub questions:

i. How to represent network traffic data that contains numerical and categorical features into point series form?

ii. How does different similarity measures affect the performance of classifier

1.4 Objective

Four objectives have been identified to answer the questions identified in the foregoing sub-section, which are:

a) To investigate and identify feature transformation technique that can generate point series data for network activities classification.

b) To investigate the feasibility of Time Series Classification techniques by using k-NN coupled with DTW to classify network traffic activities.

c) To investigate the effects of using different similarities measurement, Euclidean Distance (ED) and Cosine similarity algorithm.

d) To compare the performance of network traffic classifier produced in (b) and (c) with other machine learning techniques, Self-Organization Map (SOM) Artificial Neural Network (ANN) by (Ibrahim, Basheer, and Mahmod, 2013)

and Discriminative Multinomial Naive Bayes (NB) proposed by (Panda, Abraham, and Patra, 2010).

1.5 Research Scope

The scope of this research consists of examining the feasibility of representing network traffic data into point series form so as to be classified using Time Series Classification (TSC). Conversion of categorical data using three different approach which is simple conversion, probability conversion and lastly entropy conversion technique are also explored in this research paper. Two feature selection approaches - Information Gain (IG) Feature Selection and Correlation Feature Selection (CFS) are also being used to reduce the dimension of the dataset.

1.5.1 Dataset

This research paper will use a set of secondary data which was acquired from the Internet. The chosen dataset is the NSL-KDD dataset, the improved version of KDD'99 data set. Figure 1.1 illustrates the snapshot of the NSL-KDD original data set. Features with different types and values are also shown in the figure below. Note that the data shown in Figure 1.1 is the original dataset which have not been pre-processed for the experiment. Data pre-processing of selected dataset will be further discussed in Chapter 4 which focused on experimental settings.

NSL-KDD is a data set suggested to solve some of the inherent problems of the KDD'99 data set. The NSL-KDD data set has the following advantages over the original KDD data set (Tavallaee, Bagheri, Lu, and Ghorbani, 2009):

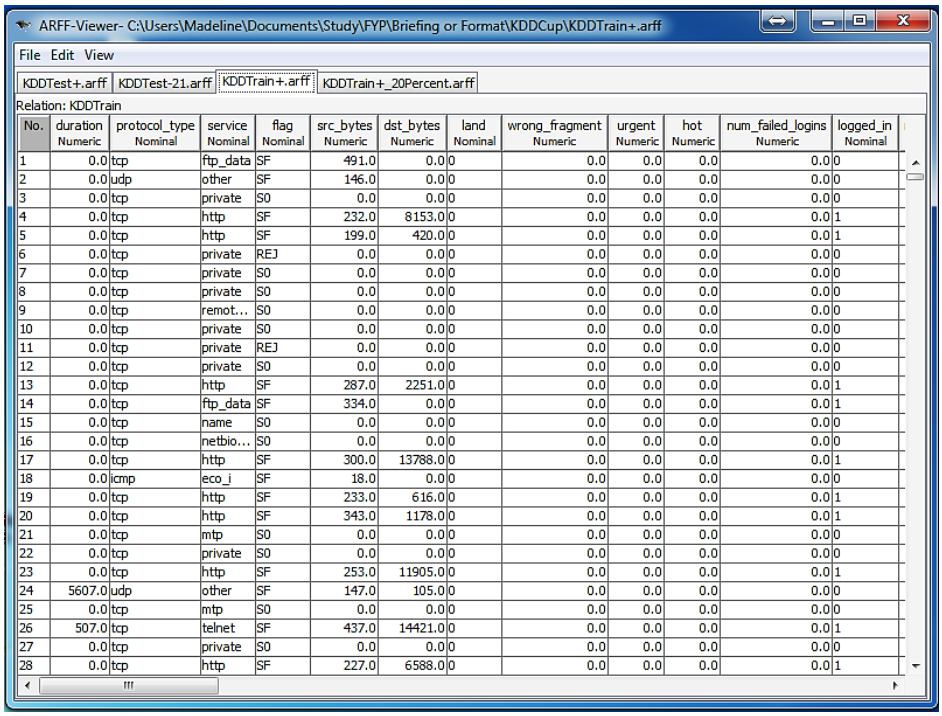
i. Redundant records are not included in the dataset, making the classifier unbiased to frequently appear records.

ii. It does not include redundant records in the train set, so the classifiers will not be biased towards more frequent records.

iii. There is no duplicate records in the proposed test sets; therefore, the performance of the learners is not biased by the methods which have better detection rates on the frequent records.

iv. The number of selected records from each difficulty level group is inversely proportional to the percentage of records in the original KDD data set. As a result, the classification rates of distinct machine learning methods vary in a wider range, which makes it more efficient to have an accurate evaluation of different learning techniques.

v. The number of records in the train and test sets are reasonable, which makes it affordable to run the experiments on the complete set without the need to randomly select a small portion. Consequently, evaluation results of different research works will be consistent and comparable.



No.	duration Numeric	protocol_type Nominal	service Nominal	flag Nominal	src_bytes Numeric	dst_bytes Numeric	land Nominal	wrong_fragment Numeric	urgent Numeric	hot Numeric	num_failed_logins Numeric	logged_in Nominal
1	0.0	tcp	ftp_data	SF	491.0	0.0	0	0.0	0.0	0.0	0.0	0
2	0.0	udp	other	SF	146.0	0.0	0	0.0	0.0	0.0	0.0	0
3	0.0	tcp	private	S0	0.0	0.0	0	0.0	0.0	0.0	0.0	0
4	0.0	tcp	http	SF	232.0	8153.0	0	0.0	0.0	0.0	0.0	1
5	0.0	tcp	http	SF	199.0	420.0	0	0.0	0.0	0.0	0.0	1
6	0.0	tcp	private	REJ	0.0	0.0	0	0.0	0.0	0.0	0.0	0
7	0.0	tcp	private	S0	0.0	0.0	0	0.0	0.0	0.0	0.0	0
8	0.0	tcp	private	S0	0.0	0.0	0	0.0	0.0	0.0	0.0	0
9	0.0	tcp	remot...	S0	0.0	0.0	0	0.0	0.0	0.0	0.0	0
10	0.0	tcp	private	S0	0.0	0.0	0	0.0	0.0	0.0	0.0	0
11	0.0	tcp	private	REJ	0.0	0.0	0	0.0	0.0	0.0	0.0	0
12	0.0	tcp	private	S0	0.0	0.0	0	0.0	0.0	0.0	0.0	0
13	0.0	tcp	http	SF	287.0	2251.0	0	0.0	0.0	0.0	0.0	1
14	0.0	tcp	ftp_data	SF	334.0	0.0	0	0.0	0.0	0.0	0.0	1
15	0.0	tcp	name	S0	0.0	0.0	0	0.0	0.0	0.0	0.0	0
16	0.0	tcp	netbio...	S0	0.0	0.0	0	0.0	0.0	0.0	0.0	0
17	0.0	tcp	http	SF	300.0	13788.0	0	0.0	0.0	0.0	0.0	1
18	0.0	icmp	eco_i	SF	18.0	0.0	0	0.0	0.0	0.0	0.0	0
19	0.0	tcp	http	SF	233.0	616.0	0	0.0	0.0	0.0	0.0	1
20	0.0	tcp	http	SF	343.0	1178.0	0	0.0	0.0	0.0	0.0	1
21	0.0	tcp	mtp	S0	0.0	0.0	0	0.0	0.0	0.0	0.0	0
22	0.0	tcp	private	S0	0.0	0.0	0	0.0	0.0	0.0	0.0	0
23	0.0	tcp	http	SF	253.0	11905.0	0	0.0	0.0	0.0	0.0	1
24	5607.0	udp	other	SF	147.0	105.0	0	0.0	0.0	0.0	0.0	0
25	0.0	tcp	mtp	S0	0.0	0.0	0	0.0	0.0	0.0	0.0	0
26	507.0	tcp	telnet	SF	437.0	14421.0	0	0.0	0.0	0.0	0.0	1
27	0.0	tcp	private	S0	0.0	0.0	0	0.0	0.0	0.0	0.0	0
28	0.0	tcp	http	SF	227.0	6588.0	0	0.0	0.0	0.0	0.0	1

Figure 1.1 Snapshot of NSL-KDD Original Dataset

Table 1.1 Name of Features for NSL-KDD Data Set

1	duration	22	is_guest_login
2	protocol_type	23	count
3	service	24	srv_count
4	flag	25	serror_rate
5	src_bytes	26	srv_serror_rate
6	dst_bytes	27	rerror_rate
7	land	28	srv_rerror_rate
8	wrong_fragment	29	same_srv_rate
9	urgent	30	diff_srv_rate
10	hot	31	srv_diff_host_rate
11	num_failed_logins	32	dst_host_count
12	logged_in	33	dst_host_srv_count
13	num_compromised	34	dst_host_same_srv_rate
14	root_shell	35	dst_host_diff_srv_rate
15	su_attempted	36	dst_host_same_src_port_rate
16	num_root	37	dst_host_srv_diff_host_rate
17	num_file_creations	38	dst_host_serror_rate
18	num_shells	39	dst_host_srv_serror_rate
19	num_access_files	40	dst_host_rerror_rate
20	num_outbound_cmds	41	dst_host_srv_rerror_rate
21	is_host_login		

Table 1.1 contains a more detailed list of the features for the NSL-KDD data. There are a total 41 features for each data entry.

1.5.2 Time Series Classification (TSC) using K-Nearest Neighbour Algorithm with Dynamic Time Warping (DTW) as similarity measure

The Time Series Classification technique that will be used in this research paper is the Dynamic Time Warping (DTW) technique incorporated in the K-Nearest Neighbour Algorithm (*k*-NN).

To perform classification, the *k*-NN algorithm takes an unlabelled data and compares to a population observations to obtain class label. The unlabelled data, x is classified by a majority vote of its neighbours, with x being labelled to the class most common amongst its *k*-NN measured by a similarity or distance measure. In this research paper, DTW algorithm is used to compute the similarity between two sequences and further classify and label the test data using the k-NN algorithm.

Based on the related work reviewed, DTW is believed to have a better accuracy as compared to other distance metric like Euclidean Distance. However, to the best of my knowledge, no one has implemented KNN-DTW in the context of network traffic so as in IDS. In this research paper, one of the challenges highlighted by García-Teodoro et al., 2009 in Anomaly-based Intrusion Detection System, which is the low detection efficiency is hope to be tackled by implementing the KNN-DTW in the context of network traffic activities.

1.6 Research Methodology

The following section will discuss briefly on the research methodology used in this research paper. A more detailed explanation will be provided in Chapter 4 Implementation of Proposed Approach. Four stages of experiments are divided in order to achieve the objectives stated in Section 1.4.

The first stage of experiments is the extraction of numerically represented features into point series format. In the first experiment, the categorical data are left out. Data pre-processing of normalization using min-max normalization method is performed. The dataset is then prepared in ten sets for ten-fold cross validation using Time Series Classification (TSC) K-Nearest Neighbour classifier with three different similarity measures which are the Euclidean Distance, Cosine Similarity and also the Dynamic Time Warping (DTW).

Second experiment involved the conversion of categorical data using simple conversion technique which is establishing a correspondence between each category and a sequence of integer.

Third stage of experiment is performing TSC on dataset which have undergone two different approach of categorical data conversion, namely Probability Conversion and Entropy Conversion.

Feature selection technique, Information Gain (IG) and Correlation Feature Selection (CFS) are implemented in the last stage of the experiments to reduce the dimensionality of the dataset.

After all the stages of experiment are carried out, the results produced are compiled and will be further discussed in Chapter 5 Result and Analysis. Comparison of performance in terms of accuracy, sensitivity and specificity (if applicable) will be made between different similarity measures and also with other machine learning approach that are stated in Chapter 2 Literature Review.

1.7 Organization of Report

The remainder of this paper is organized as below. For Literature Review in Chapter 2, Intrusion Detection System (IDS), Time Series Classification (TSC) and Classification of Network Traffic Data will be discussed.

In Chapter 3 Methodology, discussion is on the methodology used in the research in order to achieve research objectives. Procedure of carrying out this research is listed out with the aid of flow charts.

Chapter 4 Experimental setting covers in detail the steps involved to run experiments in stages for this research paper. Data pre-processing including categorical data conversion, and the experimental setup are discussed here.

All the result of the experiments carried out in this research paper is stated in Chapter 5 Result and Analysis. Followed by the detailed explanation and analysis of the result.

In the final chapter, Chapter 6 Conclusion summarizes all the works in this research paper. Future works will also be discussed here. All the references that aided in this paper are stated in the appendix in the last section of this research paper.

CHAPTER 2

LITERATURE REVIEW

2.1 Chapter Overview

This chapter reviews past similar work done on the classification of network traffic and application of time series analysis of different data set and are not confined to network traffic only. The reviewed findings and works will serve as the framework which is used as main reference in this paper. Beside, discussion in this chapter also focuses on the extraction of features that affect the performance and accuracy of Intrusion Detection System (IDS). Section 2.2 presents a fundamental understanding towards IDS whereas Section 2.3 will be discussing Time Series Analysis (TSA), Time Series Classification (TSC) and more specifically Dynamic Time Warping (DTW). Section 2.4 covers the classification techniques – the k-NN algorithm that will be used in this research paper.

2.2 Intrusion Detection System (IDS)

"An Intrusion Detection System (IDS) is a device or a software application that monitors network or system activities for malicious activities or policy violations and produces reports to a management station." (Chowdhary, Suri, and Bhutani, 2014).

2.2.1 Introduction of IDS

IDS concept was first introduced by (Anderson, 1980) in the effort of improving the computer security auditing and surveillance capability. He proposed the user, data

set and program profiles can provide security personnel with information regarding abnormal usage of a system.

According to (Robbins, 2002), ntrusion detection is the process of identifying computing or network activity that is malicious or unauthorized. Generally, IDS has comprise of common structure and components. He mentioned that an IDS comprise of an agent (sensor) that observe one or more network traffic activities and apply various types of detection algorithm. Thus, zero or more reaction will be activated.

In a research by (Deepa and Kavitha, 2012), the authors defines intrusion detection as the field of trying to detect intrusions like computer break-ins, misuse and unauthorized access to system resources and data. Activities of a given network are monitored by an IDS and determine the behaviour of these activities as malicious (intrusive) or legitimate (normal) based on system integrity, confidentiality and the availability of the information resources. An IDS is mainly categorized by their processing method which is detecting intrusion by misuse detection and anomaly detection. Deepa and Kavitha (2012) in their research state that, in misuse detection, the IDS search for specifying patterns or sequence of programs and user behaviour that match well known intrusion scenarios. Whereas models of normal network patterns is developed and by evaluating significant deviations from normal behaviour, the new intrusions are detected is the method used in anomaly detection IDS.

Sabahi and Movaghar (2008) in their research further elaborate the misuse detection into three sub-categories, which are the signature based, rule based and state transition. In signature based misuse detection, intrusions are detected by matching observed data from network activities to available signatures in its database. For rule-base method, characterisation of intrusions is based on a set of "if-then" implication rules. In state transition approach, from the network, a finite state machine is deduced and the intrusions are identified using above states. The finite state machine will contain various states of the network and an event will mark a transit. Stateful protocol analysis is also defined as an additional method used in IDS (Sabahi and Movaghar, 2008). Commonly recognizes definitions of good or

REFERENCES

Amr, T. (2012). Survey on Time-Series Data Classification, 1–10.

Anderson, J. P. (1980). *Computer Security Threat Monitoring and Surveillance*.

Apple confirms accounts compromised but denies security breach. (2014, September 2). *BBC*. Retrieved from http://www.bbc.com/news/technology-29011850

Bouzida, Y., & Cuppens, F. (2004). Efficient intrusion detection using principal component analysis. *Proceedings of the*. Retrieved from http://yacine.bouzida.free.fr/Articles/2004SAR.pdf

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to Time Series and Forecasting , Second Edition Springer Texts in Statistics*.

Chaovalitwongse, W. A., Fan, Y., & Sachdeo, R. C. (2007). On the Time Series <formula formulatype="inline"><tex>K</tex></formula>-Nearest Neighbor Classification of Abnormal Brain Activity. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *37*(6), 1005–1016. doi:10.1109/TSMCA.2007.897589

Chowdhary, M., Suri, S., & Bhutani, M. (2014). Comparative Study of Intrusion Detection System. *International Journal of Computer Sciences and Engineering*, *2*(4), 197–200.

Datti, R., & Verma, B. (2010). Feature Reduction for Intrusion Detection Using Linear Discriminant Analysis. *International Journal on Computer Science and Engineering (IJCSE)*, *02*(04), 1072–1078.

Davis, J. J., & Clark, A. J. (2011). Data preprocessing for anomaly based network intrusion detection: A review. *Computers & Security*, *30*(6-7), 353–375. doi:10.1016/j.cose.2011.05.008

Deepa, a. J., & Kavitha, V. (2012). A Comprehensive Survey on Approaches to Intrusion Detection System. *Procedia Engineering*, *38*, 2063–2069. doi:10.1016/j.proeng.2012.06.248

Elsayed, A., Hijazi, M. H. A., Coenen, F., Garc´ıa-Fi˜Nana˜, M., Sluming, V., & Zheng, Y. (2011). Time Series Case Based Reasoning for Image Categorisation. In *Case-Based Peasoning Research and Development* (pp. 423–436). doi:10.1007/978-3-642-23291-6_31

García-Teodoro, P., Díaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers & Security*, *28*(1-2), 18–28. doi:10.1016/j.cose.2008.08.003

Gillian, N., Knapp, R. B., & Modhrain, S. O. (2011). Recognition Of Multivariate Temporal Musical Gestures Using N-Dimensional Dynamic Time Warping, (June), 337–342.

He, W., Hu, G., Yao, X., Gangyuan, K., Wang, H., & Hongmei, X. (2008). Applying multiple time series data mining to large-scale network traffic analysis. *2008 IEEE Conference on Cybernetics and Intelligent Systems*, 394–399. doi:10.1109/ICCIS.2008.4670844

Hernández-Pereira, E., Suárez-Romero, J. a., Fontenla-Romero, O., & Alonso-Betanzos, a. (2009). Conversion methods for symbolic features: A comparison applied to an intrusion detection problem. *Expert Systems with Applications*, *36*(7), 10612–10617. doi:10.1016/j.eswa.2009.02.054

Ibrahim, L. M., Basheer, D. T., & Mahmod, M. S. (2013). A Comparison Study for Intrusion Database (KDD99, NSL-KDD) Based on Self Organization Map(SOM) Artificial Neural Network. *Journal of Engineering Sience and Technology*, *8*(1), 107–119.

Karagiannis, T., Papagiannaki, K., & Faloutsos, M. (2005). BLINC : Multilevel Traffic Classification in the Dark. In *ACM SIGCOMM Conference 2005* (pp. 229–240). ACM.

Kia, A., SamanHaratizadeh, & HadiZare. (2013). Prediction of USD / JPY Exchange Rate Time Series Directional Status by KNN with Dynamic Time. *Bonfring International Journal of Data Mining*, *3*(2), 12–16. doi:10.9756/BIJDM.4658

Kumar, S. (2007). Survey of Current Network Intrusion Detection Techniques. *Citeseer*, 1–18. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.129.7105&rep=rep1&type=pdf\npapers2://publication/uuid/42BBD57C-EACC-4349-AC44-F69CDF10E018

Li, H., Chen, C. L. P., & Huang, H.-P. (2000). *Fuzzy Neural Intelligent Systems: Mathematical Foundation and the Applications in Engineering*. Taylor & Francis. Retrieved from https://books.google.com/books?id=IzvqngEACAAJ&pgis=1

More, S., Matthews, M., Joshi, A., & Finin, T. (2012). A knowledge-based approach to intrusion detection modeling. *Proceedings - IEEE CS Security and Privacy Workshops, SPW 2012*, 75–81. doi:10.1109/SPW.2012.26

Muscillo, R., Schmid, M., Conforto, S., & D'Alessio, T. (2011). Early recognition of upper limb motor tasks through accelerometers: real-time implementation of a DTW-based algorithm. *Computers in Biology and Medicine*, *41*(3), 164–72. doi:10.1016/j.compbiomed.2011.01.007

Panda, M., Abraham, A., & Patra, M. (2010). Discriminative multinomial naive bayes for network intrusion detection. In *2010 Sixth International Conference on*

Information Assurance and Security (IAS) (pp. 5–10). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5604193

Robbins, R. (2002). *Distributed Intrusion Detection Systems: An Introduction and Review*.

Roobaert, D., Karakoulas, G., & Chawla, N. V. (2006). Information gain, correlation and support vector machines. Retrieved June 21, 2015, from http://www.springerlink.com/index/KJ45153333192803.pdf

Sabahi, F., & Movaghar, a. (2008). Intrusion Detection: A Survey. *2008 Third International Conference on Systems and Networks Communications*, 23–26. doi:10.1109/ICSNC.2008.44

Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., & Chang, L. (2003). A Novel Anomaly Detection Scheme Based on Principal Component Classifier. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA465712\nhttp://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA465712

Stolfo, S. J., Fan, W., Lee, W., Prodromidis, A., Street, W., York, N., & Chan, P. K. (1999). *Cost-based Modeling and Evaluation for Data Mining With Application to Fraud and Intrusion Detection : Results from the JAM Project* ∗.

Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. a. (2009). A detailed analysis of the KDD CUP 99 data set. *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, (Cisda), 1–6. doi:10.1109/CISDA.2009.5356528

Wang, W., & Battiti, R. (2006). Identifying intrusions in computer networks with principal component analysis. *Proceedings - First International Conference on Availability, Reliability and Security, ARES 2006*, *2006*, 270–277. doi:10.1109/ARES.2006.73

Weller-Fahy, D., Borghetti, B., & Sodemann, A. (2014). A Survey of Distance and Similarity Measures used within Network Intrusion Anomaly Detection. *IEEE Communications Surveys & Tutorials*, *PP*(99), 1–1. doi:10.1109/COMST.2014.2336610

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., … Dan, J. H. (2008). *Top 10 algorithms in data mining*. doi:10.1007/s10115-007-0114-2

Xu, X. (2006). Adaptive intrusion detection based on machine learning: Feature extraction, classifier construction and sequential pattern prediction. *International Journal of Web Services Practices*, *2*(1), 49–58. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.107.9575&rep=rep1&type=pdf