Intrusion Detection method Based on Nonlinear Correlation Measure

Mohammed Ambusaidi, Zhiyuan Tan, Xiangjian He*, Priyadarsi Nanda, Liang Fu Lu, Aruna Jamdagni

Center for Innovation in IT Services and Applications (iNEXT),
School of Computing and Communication,
Faculty of Engineering and Information Technology,
University of Technology, Sydney, Australia
E-mail: Mohammed.A.AmbuSaidi@student.uts.edu.au
E-mail: Zhiyuan.Tan@uts.edu.au
E-mail: Xiangjian.He@uts.edu.au
E-mail: Priyadarsi.Nanda@uts.edu.au
E-mail: liangfulv@gmail.com
E-mail: a.jamdagni@uws.edu.au
*Corresponding author

Biographical:
Mohammed Ambusaidi is a PhD student at the Faculty of Engineering and Information Technology (FEIT) of the University of Technology, Sydney (UTS), also a research member of Research Centre for Innovation in IT Services and Applications (iNEXT). His primary research interests include Computer and Network Security and on Pattern Recognition techniques for efficient Network Intrusion Detection and anomalous behavior detection.

Zhiyuan Tan is a Research Associate in the School of Computing and Communications and a research member of the Research Centre for Innovation in IT Services and Applications (iNEXT), University of Technology, Sydney (UTS), Australia. Zhiyuan received his PhD degree from UTS in 2014. His research interests are network security, pattern recognition, machine learning and distributed systems.

Xiangjian He received the Bachelor of Science degree in Mathematics from Xiamen University in 1982, the Master of Science degree in Applied Mathematics from Fuzhou University in 1986, the Master of Science degree in Information Technology from the Flinders University of South Australia in 1995, and the PhD degree in Computing Sciences from the University of Technology, Sydney, Australia in 1999. From 1982 to 1985, he was with Fuzhou University. From 1991 to 1996, he was with the University of New England. Since 1999, he has been with the University of Technology, Sydney, Australia. He is the Director of Computer Vision and Recognition Laboratory and a Deputy Director of the Research Centre for Innovation in IT Services and Applications at the University of Technology, Sydney.

Priyadarsi Nanda joined UTS in 2001. He is Senior Lecturer, in the School of Computing and Communications, and Core Research Member, Centre for Innovation in IT Services Applications (iNEXT). His research interests are Network QoS, Network securities, Assisted health care using sensor networks, and Wireless networks. He has published 38 referred publications including two journals, four book chapters, one conference tutorial and 31 referred conference papers.

Liangfu Lu received the Bachelor of Science, in Mathematics and Applied Mathematics Dept. of Mathematics Yantai Normal University, Yantai, China, the Master of Science, in Computational Mathematics Dept. of Mathematics, School of Science Nanjing University of Aeronautics and Astronautics, Nanjing, China and the Ph.D in Visual Analytics, Visual Data Mining Faculty of Engineering and IT University of Technology, Sydney, Australia.

Aruna Jamdagni received her PhD degree from University of Technology Sydney, Australia in 2012. She is a lecturer in the School of Computing and Mathematics, University of Western Sydney (UWS), Australia, and a research member of Research Centre for Innovation in IT Services and Applications (iNEXT) at University of Technology Sydney (UTS), Australia. Her research interests include Computer and Network Security and on Pattern Recognition techniques and fuzzy set theory.

Intrusion Detection method Based on Nonlinear Correlation Measure

Abstract- Cyber crimes and malicious network activities have posed serious threats to the entire internet and its users. This issue is becoming more critical, as network-based services, are more widespread and closely related to our daily life. Thus, it has raised a serious concern in individual internet users, industry and research community. A significant amount of work has been conducted to develop intelligent anomaly-based Intrusion Detection Systems (IDSs) to address this issue. However, one technical challenge, namely reducing false alarm, has been along with the development of anomaly-based IDSs since 1990s. In this paper, we provide a solution to this challenge. A Nonlinear Correlation Coefficient (NCC) based similarity measure is proposed to help extract both linear and nonlinear correlations between network traffic records. This extracted correlative information is used in our proposed IDS to detect malicious network behaviours. The effectiveness of the proposed NCC-based measure and the proposed IDS are evaluated using NSL-KDD data set. The evaluation results demonstrate that the proposed NCC-based measure not only helps reduce false alarm rate, but also helps discriminate normal and abnormal behaviours efficiently.

Keywords-Intrusion Detection; Nonlinear Correlation Coefficient (NCC); Mutual Information (MI);

I. Introduction

Network technologies have made significant progress in development, while the security issues alongside with these technologies have not been well addressed. Current research on network security mainly focuses on developing preventative measures, such as security policies and secure communication protocols. Meanwhile, attempts have been made to protect computer systems and networks against malicious behaviours by using Intrusion Detection Systems (IDSs). Clearly, the collaboration of IDSs and preventative measures can provide a safe and secure communication environment.

Intrusion detection has been a hot topic since 1990s. Intrusion detection techniques can be generally classified into two main categories: signature-based and anomaly-based detection. Signature-based or misuse-based detection systems detect on-going anomalies by looking for a match with any pre-defined attack signature (Vigna and Kemmerer, 1998). Signature-based detection systems can identify types of malicious attacks based on the pre-defined signatures and even activate responses to some particular intrusions. These systems are widely used because they are simple and efficient. More importantly, they have a small number of false positive alarms. Bro is a well-known example, and it is a stand-alone system for monitoring real-time traffic and detecting incoming intrusions (Paxson, 1999). However, one of the disadvantages of these systems is that the detection accuracy and efficiency heavily depend on the quality of attack signatures. Furthermore, to extract such high quality signatures, it requires the involvement of experts in extensive study of malicious behaviours, which is costly and time consuming. Moreover, the signature of an intrusion is required before the system can detect the respective. Consequently, this type of IDS cannot detect any previously unknown attacks due to the lack of attack signatures.

The second category is anomaly-based detection systems, which has been in favour of research community. Anomaly-based detection makes an assumption that intruders' behaviours are different from those of normal network traffic (Hassanzadeh and Sadeghian, 2008). Therefore, intrusions can be defined as network traffic patterns that do not confirm the expected pattern of normal traffic behaviour (Barford et al., 2002). In comparison with signature-based detection systems, anomaly-based detection systems enjoy the advantage of detecting unknown attacks and variants of known attacks. That is because they make use of statistical analysis to evaluate the deviations of the behaviours of observed traffic flows from those of the normal traffic. They study normal traffic behaviours on a network and then create models for normal flows. After that any deviations from the normal flows are considered as suspicious behaviours.

Recent research works have widely used machine learning techniques to build anomaly-based IDSs. Machine learning techniques are capable of improving the performance of detection algorithms, namely higher detection rates and lower false positive rates. Numerous machine learning techniques, including Bayesian network (Kruegel et al., 2003), Support Vector Machine (SVM) (Ying et al., 2012), Mutual Information (Amiri et al., 2011) and Markov models (Ye et al., 2004), have been used in anomaly-based detection systems. The main advantages of these approaches are the ability of recognising known and unknown attacks, and no need of continuous update of the attack knowledge base. However, the major weaknesses of these techniques include that they are prone to high false positive rates with newly occurring normal network traffic, and low detection rates with attacks that mimic normal network traffic behaviours. These limitations encourage us to focus on developing anomaly-based detection systems that can overcome such weaknesses.

In this paper, we intend to provide a solution to the problem of false positive alarm. Although various techniques have been proposed to address this issue in the recent decade (Yu, 2012), there is still no perfect solution. The main contribution of this paper is to propose a scheme, named Nonlinear Correlation Coefficient (NCC) based Mutual Information (MI) extraction. It can accurately extract the correlation between network traffic records. Moreover, our proposed method is sensitive to both linear

correlation and nonlinear correlation. Theoretically, NCC-based measure is more rational than Pearson's Correlation Coefficient (PCC) based measures. NCC-based measure can help improve the performance of intrusion detection (Shen et al., 2011). In addition, to demonstrate the effectiveness of our proposed NCC-based measure in extracting the correlation between network traffic records, we compare our results with the results using the PCC-based scheme.

The rest of the paper is organized as follows. Section II conducts a review on the works which are closely related to our research. The linear-based correlation (i.e., PCC) measure and the nonlinear-based correlation (i.e., NCC) measure are introduced in Section III. Section IV proposes an intrusion detection algorithm based on correlation coefficient measures. Experimental results, analysis and performance comparison are given in Section V. Finally, the conclusion and future work are presented in Section VI.

II. RELATED WORKS

Despite many intrusion detection methods are used to detect different types of attacks within networks, reducing false positive rate is still a major issue. Over time, large number of intrusion detection techniques have been proposed to overcome this problem and maintain the reliability of networks (Liao et al., 2013). Recent literatures on intrusion detection techniques have shown that correlation analysis is one of the effective ways to improve the detection ability and reduce false positive rate. Next-generation Intrusion Detection Expert System (NIDES) was one of the earliest statistical intrusion detection algorithms that could operate in real-time for continuous monitoring of user activities (Anderson et al., 1995). It uses statistical measures to build the normal profile and then use this profile to detect anomalies. Beauquier and Hu proposed a model named Pearson's Correlation Coefficients-Rank (PCC-R), which applied PCC to evaluate distances between different known methods including Uniqueness, Nave Bayes, Bayes one-step Markov and Probabilistic Finite State Automata (PFSA) (Beauquier and Hu, 2007). They built their intrusion detection model based on the combination of those methods. While this model showed improvement in the detection rate, the false alarm rates were still high. Jin et al. in (Jin et al., 2007) utilized covariance matrix of sequential samples to detect multiple network attacks. In order to investigate the performance of their model, they applied two different statistical pattern recognition approaches, namely threshold based detection approach and traditional decision tree approach, to detect anomalies. Experimental results show that both approaches can distinguish multiple known attacks in the covariance feature space effectively. However, one of the disadvantages of this model is that such a scheme is susceptible to any attacks which linearly change the monitored features.

Lately, Anuar et al. proposed a hybrid statistical approach using combination of data mining and decision tree classification in identifying the false alarms (Anuar et al., 2008). Hu et al. in 2008 proposed a detection method based on AdaBoost algorithm, where decision stumps are used as weak classifiers and decision rules are provided for categorical and continuous features (Hu et al., 2008). By combining the weak classifiers for continuous and categorical features into a strong classifier, the relations between these two different types of features are handled naturally. Their experimental results have shown that this method has low false rates.

Some new ideas were proposed lately to deal with the problem of linear changes to the monitored features and to reduce false positive rate. Tsai and Lin (Tsai and Lin, 2010) proposed a method based on Triangle Area based Nearest Neighbours (TANN). TANN combined clustering and classification techniques to detect attacks. Compared with the previously proposed methods, TANN shows significant enhancement in detection rate and false positive rates. Jamdagni et al. in (Jamdagni et al., 2010) has proposed a Geometrical Structure Anomaly Detection (GSAD) model. GSAD is a pattern recognition method using Mahalanobis Distance Map (MDM) to extract correlations between packet payload features. To reduce the processing overhead of GSAD model, Tan et al. (2010) proposed a two-tier system based on linear discriminant method (Tan et al., 2010). More recently, Tan et al. in (Tan et al., 2012) proposed an effective Multivariate Correlation Analysis (MCA) technique that investigates geometrical correlations (triangle areas) between features in a single network traffic record.

However, these approaches either ignore the correlations between traffic records or do not take nonlinear correlation into account. Considering that in real world communication the correlations can also be nonlinear, we propose IDS that can measure both linear and nonlinear correlations of multiple records which are presented in Section IV.

III. LINEAR AND NONLINEAR CORRELATION ANALYSIS

In this section, two correlation analysis methods are introduced. They are Pearson's Correlation Coefficient (PCC) and Nonlinear Correlation Coefficient (NCC). Correlation coefficient is a type of statistical measure that indicates the magnitude of relationship between the two variables. It also shows how the two variables interact with each other.

A. Pearson's Correlation Coefficient

Pearson's Correlation Coefficient (PCC) (Rodgers and Nicewander, 1988) is one of the basic linear correlation methods used to measure dependence between two variables. Given two random variables *X* and *Y*, shown in (1) and (2) respectively,

$$X = \{x_1, x_2, \ldots, x_n\}, \quad (1)$$

$$Y = \{y_1, y_2, \ldots, y_n\}, \quad (2)$$

where; n is the total number of samples in X and Y. The PCC of the variables X and Y is defined in (3).

$$PCC(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})^2 \sum_{i=1}^{n}(y_i - \bar{Y})^2}}, \quad (3)$$

where $\bar{X} = 1/n \sum_{i=1}^{n} x_i$ and $\bar{Y} = 1/n \sum_{i=1}^{n} y_i$ indicate the means of X and Y respectively.

The value of a PCC is ranged from -1 to 1. It indicates the degree of the linear correlation between two random variables. As it has been claimed in (Rodgers and Nicewander, 1988, Shen et al., 2009), when the PCC value is close to 1 or -1, it denotes a strong relationship. If the value is close to 0, it means a weak relationship between the two variables. A positive correlation coefficient denotes that the two variables are in the same direction, and a negative one indicates that the two variables are in the opposite direction.

B. Nonlinear Correlation Coefficient

Although PCC can reveal the linear correlation between two dependent random variables, in real world the correlation between two variables can be nonlinear. Thus, we need an approach to analyze the nonlinear correlation between variables.

Nonlinear Correlation Coefficient (NCC) is a method based on Mutual Information (MI), which is a quantity measuring the relationship between two discrete random variables. MI provides a generalized correlation analogous to linear correlation coefficient, but it is sensitive to both linear and nonlinear correlations (Roulston, 1999). Given the same random variables X and Y, the MI is denoted by (4).

$$I(X;Y) = H(X) + H(Y) - H(X,Y), \quad (4)$$

where $H(X)$ and $H(Y)$ are the information entropies of the variables X and Y, which are defined in (5) and (6) respectively,

$$H(X) = -\sum_{i=1}^{n} P(x_i) \ln P(x_i), \quad (5)$$

$$H(Y) = -\sum_{j=1}^{n} P(y_j) \ln P(y_j), \quad (6)$$

and $H(X,Y)$ is the joint entropy of X and Y shown in (7),

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{n} P(x_i, y_j) \ln P(x_i, y_j). \quad (7)$$

In (5) and (6), $P(x_i)$ and $P(y_j)$ denote the probabilities that the random variable X is in state x_i and the random variable Y is in state y_i respectively. In (7), $P(x_i, y_j)$ denotes the joint probability that X is in state x_i and Y is in state y_i.

The disadvantage of MI is that it does not range in a definite closed interval [-1, 1] as the correlation coefficient does (Wang et al., 2005). Therefore, Wang et al. (Wang et al., 2005) developed a revised version of the MI, named nonlinear correlation coefficient. The revised joint entropy of the two variables X and Y is given by (8).

$$H^r(X,Y) = -\sum_{i=1}^{b}\sum_{j=1}^{b} \frac{n_{ij}}{N} \log_b \frac{n_{ij}}{N}, \quad (8)$$

where $b \times b$ rank grids are used to place the sample pairs $\{(x_i, y_j)\}$, for $1 \le i \le n$ and $1 \le j \le n$, to the rank sequences of X and Y, n_{ij} is the number of samples located in the $(i,j)-th$ rank grid, and N is the total number of samples. The NCC is defined in (9).

$$NCC(X;Y) = H^r(X) + H^r(Y) - H^r(X,Y), \quad (9)$$

where $H^r(X)$ and $H^r(Y)$ are the revised entropies of the variables X and Y, which are given in (10) and (11).

$$H^r(X) = -\sum_{i=1}^{b} \frac{n_{ij}}{N} \log_b \frac{n_{ij}}{N}, \quad (10)$$

$$H^r(Y) = -\sum_{j=1}^{b} \frac{n_{ij}}{N} \log_b \frac{n_{ij}}{N} \quad (11)$$

Therefore, the $NCC(X;Y)$ in (9) can be rewritten into (12).

$$NCC(X;Y) = 2 + \sum_{i=1}^{b}\sum_{j=1}^{b} \frac{n_{ij}}{N} \log_b \frac{n_{ij}}{N}, \quad (12)$$

and describes the correlation between two discrete random variables, which is within a definite closed interval [-1, 1] where -1 and 1 indicate the weakest and the strongest relationships respectively.

For a multi-record scenario, the correlation matrix S of n observed records can be written as

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{pmatrix}, \quad (13)$$

The elements of the matrix S are the correlation coefficients between distinct pairs of records. The values of elements can be obtained using (3) and (12) for linear correlation and

nonlinear correlation respectively. It is noticed that S is a symmetric matrix and the elements' value along its diagonal are equal to one this is because $s_{ij} = s_{ji}$, where $i \neq j$, $1 \leq i \leq n$ and $1 \leq j \leq n$

The aforementioned Pearson's Correlation Coefficient, Nonlinear Correlation Coefficient and correlation matrix S are applied in our intrusion detection algorithm proposed in Section IV.

IV. INTRUSION DETECTION BASED ON CORRELATION COEFFICIENT

In this section, we propose an intrusion detection algorithm, which can apply Pearson's Correlation Coefficient and Nonlinear Correlation Coefficient to extract linear correlations and nonlinear correlations of network traffic records and form correlation matrices using (13). It is to be noticed that our detection algorithm includes two main components: normal profile and a pre-defined threshold σ . The process of generating the normal profile and choosing the threshold value σ are described in the following subsections.

A. Normal Profile Generation Using Nonlinear Correlation Coefficient

To determine the similarity between the normal record and new incoming record, we separate the detection method into two different stages. Firstly, normal profile is built for normal records and obtain the mean value of correlation coefficient among the normal ones, which can extract the linear and nonlinear correlation of the traffic records. Secondly, we set threshold value to determine whether the new incoming record is normal or not.

Given a set of m normal training traffic samples $X^{normal} = \{x_1^{normal}, x_2^{normal}, ..., x_m^{normal}\}$, we first calculate the NCC using (12), between the n normal records and then, generate the correlation matrix S using (13) for the normal records. After that, the mean $\overline{S_c^n}$ of each column c in S is defined as

$$\overline{S_c^n} = \frac{1}{h} \sum_{i,j=1}^{h,q} NCC_{c,j}^{n,i} , \tag{14}$$

where, h indicates the NCC values in each column in the correlation matrix S and q indicates the row number in matrix S .

The mean $\overline{NCC}^n$ for the $\overline{S_c^n}$ is finally calculated using:

$$\overline{NCC}^n = \frac{1}{g} \sum_{k=1,h=1}^{g} \overline{S_{c,h}^{n,k}} , \tag{16}$$

where, g indicates the number of means in the correlation matrix $\overline{S_c^n}$.

B. Threshold Selection

The selection of the threshold value σ is a delicate task when designing IDS. It directly influences the false positive and detection rates. In other words, larger value of the threshold generates less false positive alarms and small value of the threshold leads to higher detection rates and vice versa. In this paper, the value of the threshold σ is ranged from 0 to 1.

In fact, the key point of the method between the PCC measure and NCC measure, as explained in section III-A, is the correlation between the two variables. To the best of our knowledge, there is no exact mathematical solution to determine the threshold value as the degree of strong or weak correlations. Therefore, without loss of generality, we can utilize the mean of the two extreme values 1 and 0, i.e., 0.5, to differentiate the strong and weak correlation of the two variables. This value is rational because it is similar to the Hurst parameter which is a value to reveal the network self-similarity (Rose, 1996). During the training phase we have tested various values for the threshold range from 0 to 1. The experimental result shows that the large threshold value leads to less false positive alarm but less detection rate as well. Therefore, empirically we found the threshold values between 0.1 and 0.5 give higher detection rates and low false positive rate. More explanation about the threshold selection is given in Section V-B.

C. Detection Algorithm

Similar to the normal profile development process, for any new incoming record $n+1$, the $NCC^{n,n+1}$ between the new incoming record and the records in the normal profile is calculated using (12). Then, the mean $\overline{NCC}^{n,n+1}$ of the $NCC^{n,n+1}$ values is defined as

$$\overline{NCC}^{n,n+1} = \frac{1}{r} \sum_{l=1}^{r} NCC_l^{n,n+1} , \tag{17}$$

where r is the number of NCC values between the new incoming record and the normal profile records.

After that, the difference between the mean of the normal profile given in (15) and the mean in (17) is defined as

$$\left| \overline{NCC}^n - \overline{NCC}^{n,n+1} \right| , \tag{18}$$

Finally, the incoming record is considered as an attack or abnormal if the difference between $\overline{NCC}^n$ and $\overline{NCC}^{n,n+1}$ is greater than a pre-defined threshold σ or not.

$$\left| \overline{NCC}^n - \overline{NCC}^{n,n+1} \right| > \sigma , \tag{19}$$

The flow chart given in Fig. 1 illustrates the aforementioned processes of the detection algorithm. This detection algorithm has been applied to PCC as well. The

comparison results between NCC detection algorithm and PCC detection algorithm are explained in the next section.

Figure1. The flow chart of the proposed algorithm

V. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we describe the results obtained by applying the proposed algorithm in section IV-C and the selected threshold as described in section IV-B and Table IV, to detect the normal records and six different types of DOS attacks. We set detection, false positive and accuracy rates to evaluate the performance of our proposed algorithm.

A. Data Set Selection

In this experimentation, NSL-KDD data set (http://iscx.ca/NSL-KDD), as an enhanced version of KDD Cup 99 data set, is utilized to demonstrate our approach. Even though KDD Cup 99 data set is a well-known data set and widely used for network-based intrusion detection techniques, it contains some problems such as including huge number of redundant records, which affect the effectiveness of evaluated systems greatly and consequently lead to poor detection results. To overcome these issues, Tavallaee et al. in (Tavallaee et al., 2009) presented a new revised version of KDD Cup 99 termed as NSL-KDD. The training and testing data sets of NSL-KDD data set consist of about 125,973 and 22,544 connection records, respectively. Each record is labeled as either normal or attack, and has 41 features. There are 22 types of attacks available in the data set. They can be categorized into four classes, namely Probe, Denial of Service (DOS), User to Root (U2R) and Remote to User (R2U). The attacks distribution is listed in Table I.

TABLE I. ATTACK DISTRIBUTION IN NSL-KDD DATASET (OLUSOLA ET AL., 2010)

B. Experimental Results

In our experiments, detection and false positive rates are used as standard metrics to evaluate the performance and effectiveness of our algorithm. During the training and testing phases, six types of attacks including Smurf, Neptune, Land, Teardrop, Back and Pod attacks are randomly selected for training and testing. The distribution of records of various types in training and testing phases are listed in Table II and Table III respectively.

During the training phase, we have applied both NCC measure and PCC measure. By following the proposed detection algorithm shown in Fig. 1, we calculate the correlation coefficients (s_{ij}) between the selected records to generate the normal profile using both PCC measure and NCC measure respectively. Fig. 2 illustrates the two different correlation matrices S^{PCC} and S^{NCC} of normal profiles for the same samples. Each element s_{ij} in the matrices describes the linear correlation coefficient or the nonlinear correlation coefficient between the $i-th$ and $j-th$ records.

(a) Pearson's Correlation Coefficient-based Correlation Matrix

(b) Nonlinear Correlation Coefficient-based Correlation Matrix

Figure 2. Matrices Expressions of Two Different Measures for Normal Profiles

However, to differentiate normal and abnormal records exactly, it is necessary to define the pre-defined sensitive threshold σ firstly. To our best knowledge, there is no good way to solve this value theoretically. Hence, we adopt the conventional method to determine this value by setting different values for the threshold. The value varies from 0.1 to 0.5 with the step length 0.1, and these are discussed in Table IV and Fig. 3. Here, we mainly focus on detection and false positive rates during the training stage, and it is noticed that good detection results depend on σ greatly when it is neither too small nor too large, such as $\sigma = 0.2$ or 0.3. From the comparison between the various threshold values and results illustrated in Table IV and Fig. 3, when $\sigma = 0.3$, the detection rate for normal records decreases slightly from 100% to when $\sigma = 0.5$ with a detection rate of 99.85%. However, the average false positive rate between different types of attacks has a significant decrease from approximately 3.08 to 0.28 with the decrease of σ from 0.5 to 0.3. In addition, even though there is a slight difference in the detection rate when $\sigma = 0.3$ and $\sigma = 0.1$, the false positive rate between normal records when $\sigma = 0.1$ is obviously higher than when $\sigma = 0.3$. To sum up, it is the best way that we choose $\sigma = 0.3$ as a fixed threshold value for our detection algorithm.

Here, we describe the detection method in detail. During the test process, the mean correlation coefficient $\overline{NCC}_{m+1,i}$ among each new record and the corresponding normal profile which is built based on the normal traffic records is calculated. If the distance between mean coefficient of normal profile and $\overline{NCC}_{m+1,i}$ exceeds the pre-defined threshold 0.3 it would be taken into account as abnormal record. Otherwise it would be considered as legitimate traffic.

Considering the selected threshold value 0.3 , the confusion matrix presented in Table V shows that our intrusion detection algorithm using NCC measure achieves high accuracy in detecting both normal records (99.84%) and attack records (99.55%).

TABLEII. SAMPLE DISTRIBUTION ON THE TRAINING DATASET

TABLE III. SAMPLE DISTRIBUTION ON THE TESTING DATASET

TABLE IV. DETECTION RATE FOR VARIOUS THRESHOLD VALUES ON THE TRAINING DATASET

Figure 3. False positive rate for various threshold values on the training dataset

TABLE V. CONFUSION MATRIX FOR NCC-TRAINING SET

The performance of intrusion detection technique is defined by its ability to make correct predictions. Comparing an event with the predications from the IDS, there are four possible outcomes shown in Table VI. These outcomes are known as the confusion matrix. The performance comparison between NCC-based IDS and PCC-based IDS is given in Section V-C.

TABLE VI. CONFUSION MATRIX

C. Performance Comparison and Analysis

The effectiveness of a detection system is defined by Detection Rate (DR) and False Positive Rate (FPR). While the DR represents the capability of IDS in detecting attacks, the FPR refers to the probability of IDS triggering an alarm when there is no attack occurring. The DR and FPR can be obtained by

$$DR = \frac{TP}{TP + FN} \times 100\% , \quad (20)$$

$$FPR = \frac{FP}{FP + TN} \times 100\% , \quad (21)$$

where

- True Positives (TP): the number of actual attacks classified as attack.
- True Negatives (TN): the number of actual normal classified as normal.
- False Positives (FP): the number of actual normal classified as attack.
- False Negatives (FN): the number of actual attack classified as normal.

Considering the selected threshold value $\sigma = 0.3$, the results presented in Table VII show that during the testing phase our detection algorithm achieves higher detection rate and lower false positive rate than PCC measure.

As shown in Table VII the DR of NCC-based IDS (98.754%) outperforms the DR of PCC-based and other proposed methods. More importantly, for the FPR of NCC-based IDS (1.246%) also performs better than other proposed methods.

TABLE VII. COMPARISON OF DETECTION AND FALSE ALARM BETWEEN DEFFIRENT IDS USING NSL-KDD DATASET

Additionally, given the columns in PCC and NCC matrices are S_j^{PCC} and S_j^{NCC} respectively. We use the covariance of these two columns to illustrate the significant difference as shown in (22).

$$Cov(S_j^{PCC}, S_j^{NCC}) = E[(S_j^{PCC} - ES_j^{PCC})(S_j^{NCC} - ES_j^{NCC})] \quad (22)$$

It should be noticed that the correlation coefficient matrices are symmetric. Therefore, the dimension of columns that we need to calculate the covariance decreases gradually.

VI. Conclusion and Future Work

This paper has introduced a Nonlinear Correlation Coefficient (NCC) measure for discrete variables which is designed based on Mutual Information (MI). We have proposed an intrusion detection algorithm based on the assumption that intrusion behaves differently between normal network traffic. To equip our intrusion detection algorithm with high accuracy in recognizing deviation of an attack from the normal traffic flow, we have adopted the NCC into detection algorithm to extract the correlation between network traffic records. This makes our algorithm not only feasible in linear correlation extraction but also nonlinear correlation extraction.

We have also verified our findings by experimentation and comparison with Pearson's Correlation Coefficient (PCC) measure. The experimental results have shown that NCC-based intrusion detection algorithm achieves not only lower false positive rate but also higher detection rate than those of PCC-based intrusion detection algorithm.

However, the proposed intrusion detection scheme still needs to be further studied in some aspects. For instance, we will consider when an attack occurs and what type of an attack is. These are the research objectives in our future works. In addition, we are going to apply our approach on a large sized enterprise network.

References

AMIRI, F., REZAEI YOUSEFI, M., LUCAS, C., SHAKERY, A. & YAZDANI, N. 2011. Mutual information-based feature selection for intrusion detection systems. *Journal of Network and Computer Applications,* 34, 1184-1199.

ANDERSON, D., FRIVOLD, T. & VALDES, A. 1995. *Next-generation intrusion detection expert system (NIDES): A summary*, SRI International, Computer Science Laboratory.

ANUAR, N. B., SALLEHUDIN, H., GHANI, A. & ZAKARIA, O. 2008. Identifying false alarm for network intrusion detection system using hybrid data mining and decision tree. *Malaysian journal of computer science,* 21, 110-115.

BARFORD, P., KLINE, J., PLONKA, D. & RON, A. A signal analysis of network traffic anomalies. Internet Measurement Conference: Proceedings of the 2 nd ACM SIGCOMM Workshop on Internet measurment, 2002. 71-82.

BEAUQUIER, J. & HU, Y. 2007. Intrusion detection based on distance combination. *CESSE07, Venice, Italy, World Acacemy of Sciences, WAS*.

BHAT, A. H., PATRA, S. & JENA, D. 2013. Machine Learning Approach for Intrusion Detection on Cloud Virtual Machines.

DE LA HOZ, E., ORTIZ, A., ORTEGA, J. & DE LA HOZ, E. 2013. Network Anomaly Classification by Support Vector Classifiers Ensemble and Non-linear Projection Techniques. *Hybrid Artificial Intelligent Systems.* Springer.

HASSANZADEH, A. & SADEGHIAN, B. Intrusion detection with data correlation relation graph. Availability, Reliability and Security, 2008. ARES 08. Third International Conference on, 2008. IEEE, 982-989.

HU, W., HU, W. & MAYBANK, S. 2008. Adaboost-based algorithm for network intrusion detection. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on,* 38, 577-583.

JAMDAGNI, A., TAN, Z., NANDA, P., HE, X. & LIU, R. P. Intrusion detection using GSAD model for HTTP traffic on web services. Proceedings of the 6th International Wireless Communications and Mobile Computing Conference, 2010. ACM, 1193-1197.

JIN, S., YEUNG, D. S. & WANG, X. 2007. Network intrusion detection in covariance feature space. *Pattern Recognition,* 40, 2185-2197.

KRUEGEL, C., MUTZ, D., ROBERTSON, W. & VALEUR, F. Bayesian event classification for intrusion detection. Computer Security Applications Conference, 2003. Proceedings. 19th Annual, 2003. IEEE, 14-23.

LIAO, H.-J., LIN, C.-H. R., LIN, Y.-C. & TUNG, K.-Y. 2013. Review: Intrusion detection system: A comprehensive review. *J. Netw. Comput. Appl.*, 36, 16-24.

OLUSOLA, A. A., OLADELE, A. S. & ABOSEDE, D. O. Analysis of KDD'99 Intrusion Detection Dataset for Selection of Relevance Features. Proceedings of the World Congress on Engineering and Computer Science, 2010. 20-22.

PANDA, M., ABRAHAM, A. & PATRA, M. R. Discriminative multinomial Naïve Bayes for network intrusion detection. Information Assurance and Security (IAS), 2010 Sixth International Conference on, 2010. IEEE, 5-10.

PAXSON, V. 1999. Bro: a system for detecting network intruders in real-time. *Computer networks,* 31, 2435-2463.

RODGERS, J. L. & NICEWANDER, W. A. 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician,* 42, 59-66.

ROSE, O. 1996. Estimation of the hurst parameter of long-range dependent time series. *University of Wurzburg, Institute of Computer Science Research Report Series.—February*.

ROULSTON, M. S. 1999. Estimating the errors on measured entropy and mutual information. *Physica D: Nonlinear Phenomena,* 125, 285-294.

SHEN, Z., WANG, Q. & SHEN, Y. A new non-liner correlation measure. Information, Computing and Telecommunication, 2009. YC-ICT '09. IEEE Youth Conference on, 20-21 Sept. 2009 2009. 11-14.

SHEN, Z., WANG, Q. & SHEN, Y. Effects of statistical distribution on nonlinear correlation coefficient. Instrumentation and Measurement Technology Conference (I2MTC), 2011 IEEE, 2011. IEEE, 1-4.

TAN, Z., JAMDAGNI, A., HE, X., NANDA, P., LIU, R., JIA, W. & YEH, W.-C. 2010. A two-tier system for web attack detection using linear discriminant method. *Information and Communications Security*, 459-471.

TAN, Z., JAMDAGNI, A., HE, X., NANDA, P. & LIU, R. P. Triangle-Area-Based Multivariate Correlation Analysis for Effective Denial-of-Service Attack Detection. Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on, 2012. IEEE, 33-40.

TAVALLAEE, M., BAGHERI, E., LU, W. & GHORBANI, A.-A. A detailed analysis of the KDD CUP 99 data set. Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009, 2009.

TSAI, C.-F. & LIN, C.-Y. 2010. A triangle area based nearest neighbors approach to intrusion detection. *Pattern Recognition,* 43, 222-229.

VIGNA, G. & KEMMERER, R. A. NetSTAT: A network-based intrusion detection approach. Computer Security Applications Conference, 1998. Proceedings. 14th Annual, 1998. IEEE, 25-34.

WANG, Q., SHEN, Y. & ZHANG, J. Q. 2005. A nonlinear correlation measure for multivariable data set. *Physica D: Nonlinear Phenomena,* 200, 287-295.

YE, N., ZHANG, Y. & BORROR, C. M. 2004. Robustness of the Markov-chain model for cyber-attack detection. *Reliability, IEEE Transactions on,* 53, 116-123.

YING, K.-C., LIN, S.-W., LEE, C.-Y. & LEE, Z.-J. 2012. An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection. *Applied Soft Computing*.

YU, Y. 2012. A survey of anomaly intrusion detection techniques. *J. Comput. Sci. Coll.*, 28, 9-17.

Appendix

A. Figures

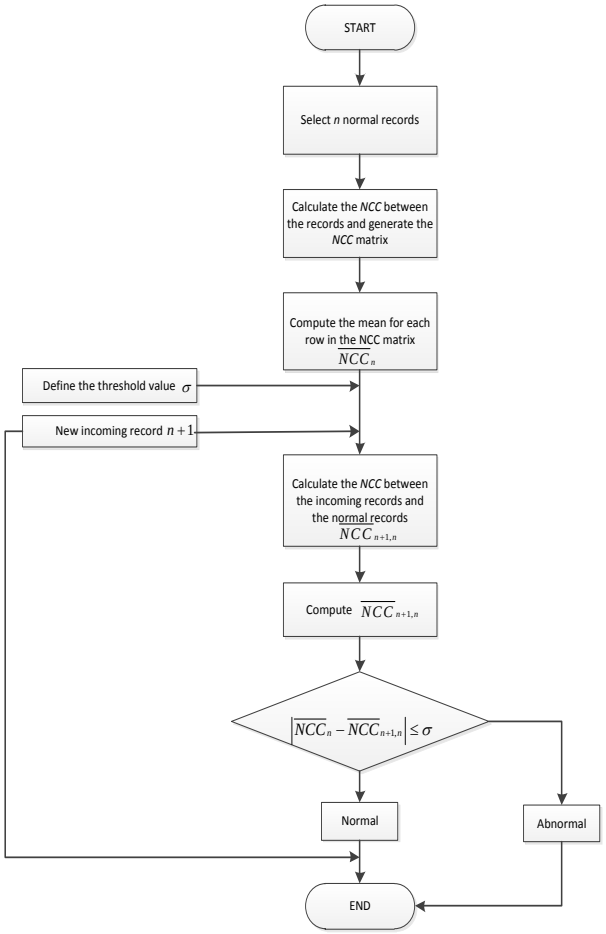

Figure 1. The flow chart of the proposed algorithm

$$S^{PCC} = \begin{pmatrix} 1.0000 & 0.8972 & 0.8892 & 0.9548 & 0.9841 & 0.4312 & 0.3832 & 0.4599 & 0.8789 & 0.7032 & 0.6321 & 0.6721 \\ 0.8972 & 1.0000 & 0.9997 & 0.7694 & 0.8131 & 0.0320 & 0.0536 & 0.0434 & 0.6265 & 0.4233 & 0.3632 & 0.3771 \\ 0.8892 & 0.9997 & 1.0000 & 0.7623 & 0.8023 & 0.0215 & 0.0337 & 0.0278 & 0.6147 & 0.4130 & 0.3445 & 0.3560 \\ 0.9548 & 0.7694 & 0.7623 & 1.0000 & 0.9636 & 0.6152 & 0.4005 & 0.6289 & 0.9340 & 0.8142 & 0.6016 & 0.6388 \\ 0.9841 & 0.8131 & 0.8023 & 0.9636 & 1.0000 & 0.5677 & 0.4622 & 0.5998 & 0.9388 & 0.7915 & 0.6777 & 0.7401 \\ 0.4312 & 0.0320 & 0.0215 & 0.6152 & 0.5677 & 1.0000 & 0.4942 & 0.9517 & 0.7553 & 0.8314 & 0.4528 & 0.5442 \\ 0.3832 & 0.0536 & 0.0337 & 0.4005 & 0.4622 & 0.4942 & 1.0000 & 0.5443 & 0.3967 & 0.2477 & 0.9494 & 0.9208 \\ 0.4599 & 0.0434 & 0.0278 & 0.6289 & 0.5998 & 0.9517 & 0.5443 & 1.0000 & 0.8017 & 0.8756 & 0.5039 & 0.6145 \\ 0.8789 & 0.6265 & 0.6147 & 0.9340 & 0.9388 & 0.7553 & 0.3967 & 0.8017 & 1.0000 & 0.9515 & 0.5528 & 0.6586 \\ 0.7032 & 0.4233 & 0.4130 & 0.8142 & 0.7915 & 0.8314 & 0.2477 & 0.8756 & 0.9515 & 1.0000 & 0.3461 & 0.4842 \\ 0.6321 & 0.3632 & 0.3445 & 0.6016 & 0.6777 & 0.4528 & 0.9494 & 0.5039 & 0.5528 & 0.3461 & 1.0000 & 0.9763 \\ 0.6721 & 0.3771 & 0.3560 & 0.6388 & 0.7401 & 0.5442 & 0.9208 & 0.6145 & 0.6586 & 0.4842 & 0.9763 & 1.0000 \end{pmatrix}$$

(a) Pearson's Correlation Coefficient-based Correlation Matrix

$$S^{NCC} = \begin{pmatrix} 1.0000 & 0.9696 & 0.8087 & 0.8087 & 0.8441 & 0.7341 & 0.6166 & 0.6493 & 0.5764 & 0.5426 & 0.7341 & 0.7341 \\ 0.9696 & 1.0000 & 0.9696 & 0.9696 & 0.9339 & 0.8424 & 0.7744 & 0.7249 & 0.6633 & 0.6166 & 0.8424 & 0.6633 \\ 0.8087 & 0.9696 & 1.0000 & 0.8087 & 0.8441 & 0.7341 & 0.8326 & 0.6493 & 0.5764 & 0.5426 & 0.7341 & 0.5764 \\ 0.8087 & 0.9696 & 0.8087 & 1.0000 & 0.8441 & 0.7341 & 0.6166 & 0.6493 & 0.5764 & 0.5426 & 0.7341 & 0.5764 \\ 0.8441 & 0.9339 & 0.8441 & 0.8441 & 1.0000 & 0.7249 & 0.7041 & 0.7530 & 0.7249 & 0.8148 & 0.9339 & 0.9339 \\ 0.7341 & 0.8424 & 0.7341 & 0.7341 & 0.7249 & 1.0000 & 0.7744 & 0.9339 & 0.6633 & 0.6166 & 0.6633 & 0.6633 \\ 0.6166 & 0.7744 & 0.8326 & 0.6166 & 0.7041 & 0.7744 & 1.0000 & 0.7041 & 0.7744 & 0.5463 & 0.7744 & 0.5747 \\ 0.6493 & 0.7249 & 0.6493 & 0.6493 & 0.7530 & 0.9339 & 0.7041 & 1.0000 & 0.9339 & 0.8148 & 0.7249 & 0.7249 \\ 0.5764 & 0.6633 & 0.5764 & 0.5764 & 0.7249 & 0.6633 & 0.7744 & 0.9339 & 1.0000 & 0.7716 & 0.8424 & 0.6633 \\ 0.5426 & 0.6166 & 0.5426 & 0.5426 & 0.8148 & 0.6166 & 0.5463 & 0.8148 & 0.7716 & 1.0000 & 0.6166 & 0.7716 \\ 0.7341 & 0.8424 & 0.7341 & 0.7341 & 0.9339 & 0.6633 & 0.7744 & 0.7249 & 0.8424 & 0.6166 & 1.0000 & 0.6633 \\ 0.7341 & 0.6633 & 0.5764 & 0.5764 & 0.9339 & 0.6633 & 0.5747 & 0.7249 & 0.6633 & 0.7716 & 0.6633 & 1.0000 \end{pmatrix}$$

(b) Nonlinear Correlation Coefficient-based Correlation Matrix

Figure 2. Matrices Expressions of Two Different Measures for Normal Profiles

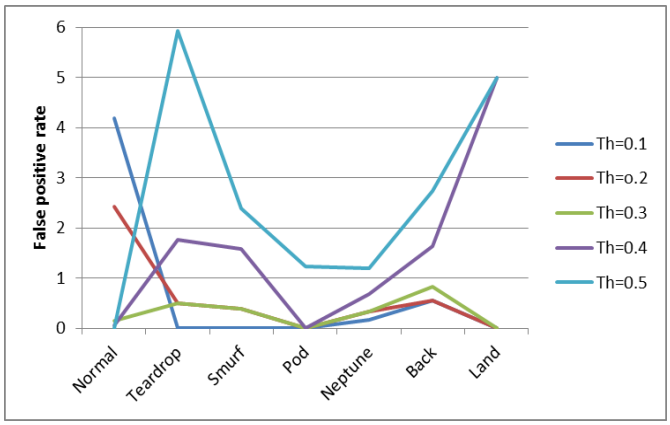

Figure 3. False positive rate for various threshold values on the training dataset

B. Tables

TABLE I. ATTACK DISTRIBUTION IN NSL-KDD DATASET (OLUSOLA ET AL., 2010)

Attack type	attack name
Probing	nmap
	ipsweep
	portsweep
	satan
DoS	back
	land
	neptune
	pod
	smurf
	teardrop
U2R	rootkit

	perl
	loadmodule
	buffer-overflow
R2L	ftp-write
	spy
	phf
	guess-passwd
	imap
	warezclient
	wrezmaster
	multihop

TABLE II. SAMPLE DISTRIBUTION ON THE TRAINING DATASET

Normal	Attack						Total
	Neptune	Land	Smurf	Teardrop	Back	Pod	
	590	19	251	169	365	162	
1980	1556						3536

TABLE III. SAMPLE DISTRIBUTION ON THE TESTING DATASET

Normal	Attack						Total
	Neptune	Land	Smurf	Teardrop	Back	Pod	
	1840	19	1566	1313	988	1761	
14590	7487						22077

TABLE IV. DETECTION RATE FOR VARIOUS THRESHOLD VALUES ON THE TRAINING DATASET

Type of records	Threshold σ				
	0.1	0.2	0.3	0.4	0.5
Normal	95.81%	97.57%	99.85%	99.94%	100%
Teardrop	100%	99.41%	99.41%	98.22%	94.08%
Smurf	100%	99.60%	99.60%	98.41%	97.61%
Pod	100%	100%	100%	100%	98.76%
Neptune	99.83%	99.66%	99.66%	99.32%	98.81%
Back	99.45%	99.45%	99.18%	98.35%	97.26%
Land	100%	100%	100%	95%	95%

TABLE V. CONFUSION MATRIX FOR NCC-TRAINING SET

Predicted Actual	Normal	Attack	Correct (%)
Normal	1977	3	99.84
Attack	7	1549	99.55

TABLE VI. CONFUSION MATRIX

Actual	Prediction	
	Normal	Attack
Normal	TP	FN
Attack	FP	TN

TABLE VII. COMPARISON OF DETECTION AND FALSE ALARM BETWEEN DEFFIRENT IDS USING NSL-KDD DATASET

Methods	False Positive Rates (%)	Detection Rates (%)
NCC-MI (Proposed method)	1.246	98.754
PCC	2.367	97.632
Naïve Bayes Tree (Bhat et al., 2013)	2.0	**
SVM (de la Hoz et al., 2013)	93.4	14
DM- Naïve Bayes (Panda et al., 2010)	3.0	96.5
Random Forest (Tavallaee et al., 2009)	**	80.67

** indicates data not provided by the author in their paper.