Journal of Theoretical and Applied Information Technology
30th April 2018. Vol.96. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

INTRUSION DETECTION SYSTEM USING BOOTSTRAP RESAMPLING APPROACH OF T^2 CONTROL CHART BASED ON SUCCESSIVE DIFFERENCE COVARIANCE MATRIX

[1]MUHAMMAD AHSAN, [2]MUHAMMAD MASHURI, [3]HIDAYATUL KHUSNA

[1,2,3] Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia

E-mail: [1]ahsan4th@gmail.com, [2]m_mashuri@statistika.its.ac.id, [3]khusna16@mhs.statistika.its.ac.id

ABSTRACT

The multivariate control chart is one of SPC method that is often used in intrusion detection. The Hotelling's T^2 control chart with Successive Difference Covariance Matrix (SDCM) is the robust method that can detect outliers in the process data for individual observation. This method will effective to be applied in Intrusion Detection System (IDS) because it can detect the anomaly or outliers in the network. The problem arise when the exact distribution of this method has not determined. Bootstrap is one of the nonparametric method that widely used to estimate the parameter without any distribution assumption applied to overcome the problem. In this research, the Hotelling's T^2 control chart is improved using the SDCM while its control limits is calculated using bootstrap resampling method. The proposed method is applied in IDS and its performance is compared to the other control chart approaches. The performance evaluation result shows that the proposed IDS with bootstrap control limit performs better than the other control chart approaches for testing dataset. Moreover, the proposed IDS outperforms the other classification methods.

Keywords: *T^2 control chart, successive difference covariance matrix, kernel density estimation, bootstrap, intrusion detection system*

1. INTRODUCTION

The flow of information is getting crowded due to the fast development of internet technology and its implementation in a computer network. Since the important value of information, the ability to access and provide information quickly and accurately is required. However, the rapid flow of such information allows the potential for a security hole that can increase crime in cyber network. The security system in question is a precautionary measure from a computer attack or an irresponsible network user.

An early warning system or usually called intrusion detection is one of the methods that can be used in the security mechanism. Intrusion detection is a process to monitor the events taking place in a computer system or network and analyze the monitoring results to find the signs of intrusion [1]. An endeavor that can compromise the confidentiality, integrity, availability of a computer system or network is called intrusion [2]. Furthermore, the intrusion detection system (IDS) has thrived into a very important component in computer network architecture. IDS is usually analogous to a burglar alarm since the IDS can include the hardware, software, or combination of both that can monitor system and network.

Intrusion detection can be carried out using a statistical approach. One of the statistical approach that can be used in intrusion detection is Statistical Process Control (SPC) that has been widely used in various fields, especially in industry and services. Some control charts had been developed to monitor univariate data such as $\bar{X}$ chart [3], Exponentially Weighted Moving Average (EWMA) chart [4] and Cumulative Sum (CUSUM) chart [5]. In addition, some control charts such as p chart, np chart [6], and Laney p' chart [7,8] are developed to monitor the attribute process. The advantage of SPC method is not require the knowledge of an unprecedented attack. Furthermore, SPC based SDI can ensure the online detection process [9]. SPC can also be used as a powerful method to guarantee the system security and stability in network monitoring and intrusion detection [10]. There are many researches on SPC that has been implemented in IDS in both univariate and multivariate processes [11].

In multivariate case, reference [12] using Markov Chain, Hotelling's T^2 and chi-square multivariate test for intrusion detection. To detect both counter-relations and mean-shift anomalies,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

reference [13] proposed a technique based on the Hotelling's T^2 test. Reference [14] employed Hotelling's T^2 to detect intrusion on a network. An online system which called Multivariate Analysis for Network Attack (MANA) detection algorithm has the control limits that will be updated at certain interval. The Chi Square Distance Monitoring (CSDM) method had been developed by reference [15] to monitor uncorrelated, high correlated, autocorrelated, normally distributed, and non-normally distributed of data. Reference [16] had acquainted a method to detect the anomalies on computer networks named Support Vector Clustering (SVC) based control chart. Covariance Matrix Sign (CMS) had been implemented by reference [17] to detect Denial of Service (DoS) attacks. The high accuracy of Hotelling's T^2 for all types of attack classes is found after comparing the performance of its control chart with Support Vector Machine (SVM) and Triangle Area based Nearest Neighbours (TANN) methods [18].

The Hotelling's T^2 is the most commonly used control chart for intrusion detection. The hotelling's T^2 which uses the conventional mean and covariance matrix is sensitive to outlier, however the method is not effective to use for multiple outliers case due to masking effect [19]. The masking effect in monitoring process happen due to the outlier cannot be detected by the control chart. To overcome the problem arise some methods has been proposed to decrease the effect of multiple outlier by change the estimators with the robust estimators especially for the estimator for covariance matrix.

The determination of sample covariance matrix in phase I monitoring process become an important aspect in the establishment of Hotelling's T^2 control chart. If the data composes of rational subgroups then sample covariance matrix and the average over all the subgroups are used as common procedures. The Hotelling's T^2 control chart for individual observations had been developed by several researchers such as Tracy et al. [19] and Lowry and Montgomery [20]. Reference [22] had investigated the power comparison of T^2 control chart based on different kinds of covariance matrix estimator in phase I monitoring process and under multivariate normal distribution. The necessary and sufficient requirement under those underlying multivariate normal distribution was investigated by Cambanis et al. [23].

The problem is arise while taking the sample covariance matrix from the data consist of individual observation. The control chart developed using this covariance matrix will lead to a poor performance in detecting shift in the mean vector [24]. Moreover, the performance of Hotelling's T^2 control chart in detecting shift of mean vector will increase if robust covariance matrix estimator was utilized [25]. Successive Difference Covariance Matrix (SDCM) is one of the robust covariance matrix estimator. Hotelling's T^2 control chart based on SDCM is effective in detecting shift of mean vector [24,26]. Moreover, SDCM can also be applied to autocorrelated data such as T^2 control chart based on SDCM for multivariate process using residuals of Vector Autoregressive (VAR) model [27].

Many studies had proved the effectiveness of Hotelling's T^2 control chart based on SDCM. However, the distribution of this control chart has not been exactly determined. Sullivan and Woodall [23] and Williams et al. [24] proposed approximate distribution for Hotelling's T^2 control chart based on SDCM. Some studies had improved Hotelling's T^2 based on SDCM control limit by using nonparametric approaches in order to overcome the limitation knowledge of Hotelling's T^2 based on SDCM distribution. Several researches had been conducted to improve the control limit of Hotelling's T^2 based on SDCM using Kernel Density Estimation (KDE) approach [22,28,29]. However, KDE approach is not effective while applied to a very skewed distribution of data. Reference [29] proved that bootstrap approach is more effective for calculating the control limits when the monitoring statistics are skewed. Bootstrap is one of the nonparametric method that widely used to estimate the parameter without any distribution assumption [30,31]. Therefore, bootstrap can be used to establish the control limits of the control chart which statistic does not follow any distribution pattern. Reference [28] developed Hotelling's T^2 control limit based on bootstrap technique.

The aim of this study is to propose Hotelling's T^2 control chart based on SDCM using bootstrap approach. By using bootstrap method is expected to yield the more accurate control limit of Hotelling's T^2 based on SDCM. The application of the proposed control chart to IDS is expected to produce more accurate monitoring system. The performance of proposed method would be compared with the other control chart approaches. In addition, the performance of T^2 based on SDCM using bootstrap approach is compared with the other classification methods as in [32].

This paper is organized as follows. Section 2 describes T^2 control chart based on SDCM, while the control limit of Hotelling's T^2 control chart

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

using bootstrap and KDE approach is explained at Section 3. Section 4 presents the dataset and methodology that used in this research. Next, the evaluation performance of IDS is displayed at Section 5. The performance comparison of the proposed IDS and other classifier methods is presented in section 6. Finally, section 7 summarizes the obtained results and presents a future research.

2. HOTELLING'S T2 CONTROL CHART BASED ON SDCM

Hotelling's T^2 is one of multivariate the control charts that could be used to monitor the mean of a process [33].

Let $\mathbf{x}_i$, where $i = 1,2,\text{K},n$ number of observations, are identic and independently random vectors which follow multivariate normal distribution with common mean vector and covariance matrix, i.e. $\mathbf{x}_i \sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Otherwise, those $n \times p$ dataset could be defined as: $\mathbf{X} = \left[\mathbf{x}_1^{'},\mathbf{x}_2^{'},\text{K},\mathbf{x}_n^{'}\right]^{'}$. Using $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ and $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^{'}$, T^2 statistic [34] can be calculated as follows:

$$T_i^2 = \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^{'}\mathbf{S}^{-1}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right). \quad (1)$$

If the data follow multivariate normal distribution then the control limit of Hotelling's T^2 can be obtained with following equation:

$$CL = \frac{p(n+1)(n-1)}{n^2 - np}F_{(\alpha,p,n-p)}, \quad (2)$$

where n is number of observations, p is number of variables and α is false alarm rate. The process is said to be in-control if T^2 statistic in equation (1) is not greater than the control limit CL.

SDCM is another alternative method to estimate the covariance matrix that firstly introduced by reference [35] and [36]. The T^2 based on SDCM can be calculated as follows:

$$T_{D,i}^2 = \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^{'}\mathbf{S}_D^{-1}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right), \quad (3)$$

where

$$\mathbf{S}_D = \frac{1}{1(n-1)}\sum_{i=2}^{n}\left(\mathbf{x}_i - \mathbf{x}_{i-1}\right)\left(\mathbf{x}_i - \mathbf{x}_{i-1}\right)^{'}. \quad (4)$$

Under in-control condition, the SDCM covariance matrix $\mathbf{S}_D$ is an unbiased estimator for $\boldsymbol{\Sigma}$ [24]. There are some approaches to construct the control limit of the data that follow multivariate normal distribution such as control limit based on Sullivan and Woodall (CLSW) [24], control limit based on Mason and Young (CLMY) [37], and control limit based on chi-square distribution (CL_{χ^2}). Those control limits can be calculated using equation (5) until equation (7) as follows:

$$CL_{SW} = \frac{(n-1)^2}{n}BETA_{(1-\alpha),\frac{p}{2},\frac{(g-p-1)}{2}}, \quad (5)$$

$$CL_{MY} = \frac{(f-1)^2}{f}BETA_{(1-\alpha),\frac{p}{2},\frac{(g-p-1)}{2}}, \quad (6)$$

$$CL_{\chi^2} = \chi^2_{(1-\alpha),v}, \quad (7)$$

where $BETA_{(1-\alpha),p,g}$ is $[1-\alpha]$-th quantile of beta distribution with shape parameter p and g while $\chi^2_{(1-\alpha),v}$ is $[1-\alpha]$-th quantile of chi-square distribution with v degree of freedom and let $g = \frac{2(n-1)^2}{3n-4}$.

3. T^2 CONTROL LIMIT BASED ON KERNEL DENSITY ESTIMATION AND BOOTSTRAP

In this section, two computational approaches to calculate the unknown distribution control limit is presented. These two approaches are Kernel Density Estimation (KDE) and Bootstrap method.

3.1 T^2 Control Limit Based On KDE

Kernel density estimation (KDE) method is non parametric method to estimate the probability density function of a random variable. This method was first introduced by Rosenblatt [38] and Parzen [39] so that its name is called the Rosenblatt-Parzen kernel density estimator which is the development of the histogram estimator.

Reference [40] proposed KDE to estimate the distribution of T^2 statistic. Let T^2 is a Hotelling's statistic which obtained under in-control condition. The distribution of T^2 statistic could be calculated with following kernel function:

$$\hat{f}_h(t) = \frac{1}{n}\sum_{i=1}^{n}K\left[\frac{\left(t - T_{D,i}^2\right)}{h}\right], \quad (8)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

where K and h define kernel function and smoothing parameter respectively.

There are some kernel functions displayed in [41], such as:

i. Uniform Kernel:

$$K(x)=\frac{1}{2}\mathrm{I}(|x|\le 1)$$

ii. Triangle Kernel:

$$K(x)=(1-|x|)\mathrm{I}(|x|\le 1)$$

iii. Epanechnikov Kernel:

$$K(x)=\frac{3}{4}(1-x^2)\mathrm{I}(|x|\le 1)$$

iv. *Quartik* Kernel:

$$K(x)=\frac{15}{16}(1-x^2)^2\mathrm{I}(|x|\le 1)$$

v. Triweight Kernel:

$$K(x)=\frac{35}{32}(1-x^2)^3\mathrm{I}(|x|\le 1)$$

vi. Cosinus Kernel:

$$k(x)=\frac{\pi}{4}\cos\left(\frac{\pi}{2}x\right)\mathrm{I}(|x|\le 1)$$

vii. Gaussian Kernel:

$$k(x)=\frac{1}{\sqrt{2\pi}}\exp\left(\frac{1}{2}(-x^2)\right),-\infty<x<\infty,$$

where I is indicator. The most used kernel is Gaussian Kernel so in this paper it is used in analysis.

The control limit in equation (9) can be calculated using tables of integrals, in closed form distribution,. However, the control limit might be not efficient to be calculated if the distribution is not closed form. Thus, the kernel control limit is solved using trapezoidal rule [42], one of the numerical integration method to approximate definite value of integral equation.

Furthermore, the control limit of T^2 based on KDE could be estimated by taking the percentile of kernel distribution. Hence, the control limit of T^2 based on KDE equal to $[100(1-\alpha)]$-th percentile of T^2 distribution which could be calculated using as follows:

$$CL_{\text{kernel}}=\hat{f}_h(t)^{-1}(1-\alpha). \qquad (9)$$

3.2 T^2 Control Limit Based On Bootstrap

Bootstrap is one of the resampling method that most widely used to estimate the parameter of unknown distribution. The bootstrap method was first introduced by [30]. This method is easy to perform because it requires neither specification of the parameters nor a procedure for numerical integration as in KDE method [28].

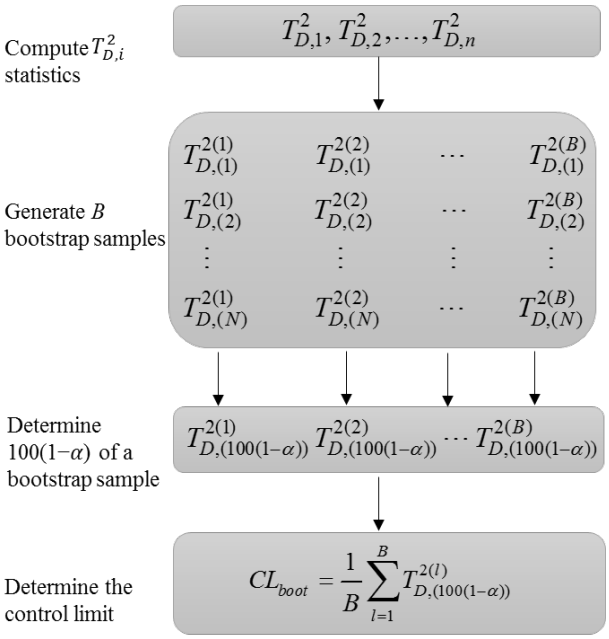

Figure 1: Bootstrap Procedure for calculating control limit of T^2_D statistics

Figure 1 illustrates an overview of the bootstrap procedure for calculating the control limit of T^2_D statistic as in [28,29] to calculate the control limit. The T^2_D control limit is calculated by resampling $T^2_{D,i}$ statistic, where $i=1,2,\mathrm{K},n$, for B times, where B is the large number that usually larger than 1000. Let $T^{2(l)}_{D,1},T^{2(l)}_{D,2},\mathrm{L}\ T^{2(l)}_{D,N}$, $l=1,2,\mathrm{K}\ B$, $N\ge n$ is a set of B bootstrap samples that randomly drawn from $T^2_{D,1},T^2_{D,2},\mathrm{K},T^2_{D,n}$ statistics with replacement for l-th replication. For each replication of B bootstrap samples, determine the $[100(1-\alpha)]$-th percentile of T^2 distribution, where α is significance level. Then, the control limit can be calculated by taking mean from the value of $[100(1-\alpha)]$-th percentile as follows:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$CL_{boot} = \frac{1}{B}\sum_{l=1}^{B} T^{2(l)}_{D,(100(1-\alpha))} \qquad (10)$$

4. METHODOLOGY

In this section, methodology of the research will be illustrated. First, the methodology of using control chart as IDS is presented. After that, the procedures of proposed T^2 SDCM control chart for IDS is described.

4.1 Intrusion Detection System (IDS) using Control Chart

The intrusion detection system (IDS) using control chart is presented in this section. Control chart has advantage to use in IDS, especially if there are no historical data of network traffic. Using this method will allow the user to create the online monitoring system. To construct this system there are two main procedures. These two procedures are data preparation and construction of the control chart.

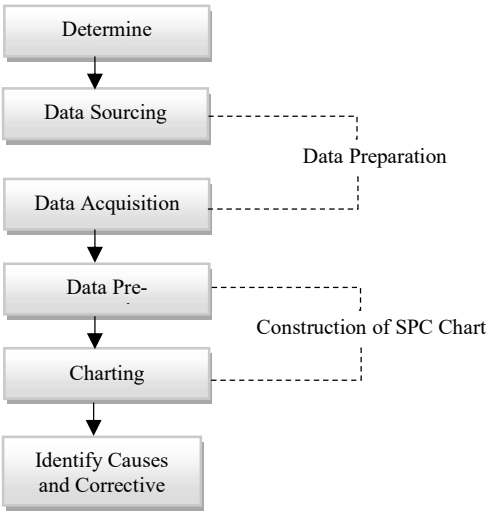

Figure 2: Intrusion Detection System using Control Chart Method [11]

Figure 2 illustrates the intrusion detection process using the control chart [11]. The first step is to determine the purpose of constructing IDS. The main purpose of an IDS is to detect the intrusion as fast as possible with low rate of false alarms. The next step is the preparation of data. The most difficult part in IDS process is preparation of data. The preparation of data is also consuming more time. In preparation of data there are two step that must be done such as, data sourcing and data acquisition. Data sourcing step is process to identify the sources and select the target of the data. While, the data acquisition refers to transform the target data into the input data which can be used in control chart method.

After preparing the network data, the next step is construction of control chart. In control chart construction, the procedure is characterized into two steps such as, data pre-processing and create control chart. In this step, the control limits previously constructed are applied to the monitoring network process. The final procedures in this method are identifying causes and taking corrective actions.

4.2 *IDS based T^2 SDCM Control Chart*

In this section, algorithm of proposed IDS based on T^2 *SDCM Control Chart* is described. The dataset and confusion matrix that is used to evaluate the performance proposed IDS also explained in this section.

NSL-KDD is the dataset that used in this research. This dataset is first proposed by reference [43] as a solution for obsolete KDD-99 dataset [44] that had been available for more than 15 years. NSL-KDD dataset consist of 41 variables with 34 quantitative variables and 7 qualitative variables. Nevertheless, this study only uses 32 quantitative variables because the value of the rest quantitative variables is equal to zero. Reference [45] reviewed that NSL-KDD dataset has some advantages as follow:

1. The redundant records is not found in the training dataset, that makes the IDS would not yield any biased result.
2. The duplicate record in not found for the testing dataset which have better reduction rates.
3. The number of selected records from each attack label is inversely proportional to the percentage of records in the original KDD data set.

In this study, NSL-KDD data is monitored using conventional Hotelling's T^2 and Hotelling's T^2 based on SDCM control chart. Moreover, for the Hotelling's T^2 based on SDCM, control limit is estimated using several methods, such as F distribution control limit according to (2), Sullivan and Woodall control limit approach based on (4), Mason and Young control limit approach according to (5), chi-square control limit based on (6), KDE control limit according to (8), and proposed bootstrap control limit as in (10).

The proposed IDS is not only compared with the other approaches of control chart and control limit but also with the other classifiers that

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

presented in [32]. These classifiers are Random Forest Classification (RFC), Logistic Regression (LR), Gaussian Naive Bayes (GNB), and Support Vector Machine (SVM) method

The algorithm for IDS with T^2 based SDCM using bootstrap control limit can be divided into two phase as follows:

Phase I: Building Normal Profile

1. Form matrix $\mathbf{X}_{normal}$ which is the normal connection data.
2. Calculate vector $\bar{x}_{normal,l}, l = 1, 2, ..., p$ which is the average of each column of normal connection data $\mathbf{X}_{normal}$.
3. Calculate the matrix of $\mathbf{S}_{DN}$ as in equation (4) which is the matrix variance covariance of normal connection data $\mathbf{X}_{normal}$.
4. Calculate statistics $T^2_{DN,i}$ as in equation (3) using normal connection data $\mathbf{X}_{normal}$.
5. Determine α and calculate the bootstrap control limit using CL_{boot} as in equation(10).

Phase II: Detection

1. Form matrix $\mathbf{X}_{test}$ which is the new connection data.
2. Calculate statistics $T^2_{DT,i}$ from new connection data $\mathbf{X}_{test}$ as follows:

$$T^2_{DT,i} = \left(\mathbf{x}_i - \bar{\mathbf{x}}_{normal}\right)' \mathbf{S}^{-1}_{DN} \left(\mathbf{x}_i - \bar{\mathbf{x}}_{normal}\right),$$

where $\bar{\mathbf{x}}_{normal}$ and $\mathbf{S}_{DN}$ are taken from normal connection data in phase I.

3. If $T^2_{DT,i} > CL_{boot}$ then the connection is intrusion and $T^2_{DT,i} < CL_{boot}$ the connection is normal.

Moreover, to show the effectiveness of each approach, the performance of each IDS approach is evaluated by confusion matrix as shown in Table 1. The performance of a IDS method could be measured by the degree of accuracy and degree of error. The accuracy in detecting intrusion can be divided into two types:

1. True Positives (TP) is number of successful attack that concluded as an attack.
2. True Negatives (TN) is number of normal activities that successfully detected as normal activity.

On the other hand, the errors in intrusion detection can also be divided into two types:

1. False Positives (FP) is number of normal activities that detected as an attack.
2. False Negatives (FN) is number of successful attacks that detected as normal activity.

The FP that occur in network causes a false alarm that can disturb the system. While, FN that occur in network will allow an attack on the system.

Table 1: Intrusion Detection Confusion Matrix

	Prediction	
	Intrusion	Normal
Intrusion	True Positives (TP)	False Negatives (FN)
Normal	False Positives (FP)	True Negatives (TN)

The level of accuracy used is the hit rate that can be calculated as follows:

$$\text{Hit Rate} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Based on the type of error, the level of error in intrusion detection can be divided into two types, namely FP rate and FN rate. The FP and FN rate formula is calculated as follows:

$$FP \text{ Rate} = \frac{FP}{TN + FP}.$$

$$FN \text{ Rate} = \frac{FN}{TP + FN}.$$

5. RESULT AND DISCUSSION

This section is aimed to present the performance evaluation of proposed IDS compared to the other control limit approaches. In addition, this section also be equipped with the performance evaluation comparison of proposed IDS with the other classification methods.

5.1 Result

This section displays the performance evaluation of IDS for NSL-KDD dataset using conventional Hotelling's T^2 and Hotelling's T^2 based on SDCM control chart. The control limit of Hotelling's T^2 based on SDCM is estimated using several approaches such as F distribution control limit ($SDCM_F$), Sullivan and Woodall approach ($SDCM_{SW}$), Mason and Young approach ($SDCM_{MY}$), chi-square control limit ($SDCM_{CH}$),

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

KDE control limit ($SDCM_{KDE}$) and the proposed bootstrap control limit ($SDCM_{boot}$)

The performance of Hotelling's T^2 and Hotelling's T^2 based on SDCM control chart with various control limit approaches for training data is presented at Table 2. Hotelling's T^2 based on $SDCM_{KDE}$ control chart has hit rate 0.9171, FN rate 0.0612, and FP rate 0.1078 by alpha equal to 0.061. The conventional T^2, the T^2 based on $SDCM_{MY}$, and the T^2 based $SDCM_{CH}$ control chart have similar performance with hit rate 0.9133, FN rate 0.0806, and FP rate 0.0937. The hit rate, FN rate, and FP rate for T^2 based on $SDCM_{SW}$ is 0.9170, 0.0636, and 0.1052, respectively. Moreover, the proposed $SDCM_{boot}$ control chart has has hit rate 0.91706, FN rate 0.0629, and FP rate 0.1059 by alpha equal to 0.063.

Table 2 : Performance Of Various IDS For Training Data

IDS	Hit Rate	FN	FP	FN Rate	FP Rate
T^2	0.91330	5428	5494	0.0806	0.0937
$SDCM_F$	0.91338	5417	5495	0.0804	0.0937
$SDCM_{SW}$	0.91705	4280	6170	0.0636	0.1052
$SDCM_{MY}$	0.91331	5429	5492	0.0806	0.0937
$SDCM_{CH}$	0.91332	5427	5492	0.0806	0.0937
$SDCM_{KDE}$	0.91710	4124	6319	0.0612	0.1078
$SDCM_{Boot}$	0.91706	4238	6210	0.0629	0.1059

Table 3 explains the performance of Hotelling's T^2 and Hotelling's T^2 based on SDCM control chart with various control limit approaches for testing dataset. The Conventional Hotelling's T^2, Hotelling's T^2 based on $SDCM_F$, $SDCM_{MY}$, and $SDCM_{CH}$ control chart have similar performance with hit rate 0.8049, FN rate 0.0838, and FP rate 0.279. In addition, the hit rate, FN rate, and FP rate for T^2 based on $SDCM_{KDE}$ is 0.8558, 0.1273, and 0.1569 respectively. Furthermore, the performance of proposed $SDCM_{boot}$ control chart for testing data set is outperform the other methods with hit rate 0.8562, FN rate 0.1257, and FP rate 0.1574.

The hit rate of various control limit approaches is need to be visualized in single graphic so the performance of each control chart can be easily compared. The hit rate comparison of various control limit approaches for both training and testing dataset are depicted at Figure 3. It can be shown that for training dataset, T^2 based on $SDCM_{KDE}$ has the highest hit rate. However, T^2 based on $SDCM_{boot}$ has the highest hit rate for testing dataset. Moreover, the hit rate of both T^2 based on $SDCM_{KDE}$ and T^2 based on $SDCM_{boot}$ are significantly higher than that of the other methods.

Table 3 : Performance of Various IDS For Testing Data

IDS	Hit Rate	FN	FP	FN Rate	FP Rate
T^2	0.8049	814	3584	0.0838	0.2793
$SDCM_F$	0.8049	814	3585	0.0838	0.2794
$SDCM_{SW}$	0.7911	731	3978	0.0753	0.3100
$SDCM_{MY}$	0.8049	814	3584	0.0838	0.2793
$SDCM_{CH}$	0.8049	814	3584	0.0838	0.2793
$SDCM_{KDE}$	0.8558	1236	2014	0.1273	0.1569
$SDCM_{Boot}$	0.8562	1221	2020	0.1257	0.1574

Figure 4 displays the FN rate and FP rate comparison for various control limit approaches in training dataset. The two lowest FN rate is owned by T^2 based on $SDCM_{KDE}$ and T^2 based on $SDCM_{boot}$ respectively. On the contrary, these two methods have highest FP rate. Therefore, T^2 based on $SDCM_{KDE}$ and T^2 based on $SDCM_{boot}$ have good performance based on FN rate criteria of training dataset, but they do not have good performance according to FP rate criteria of training dataset.

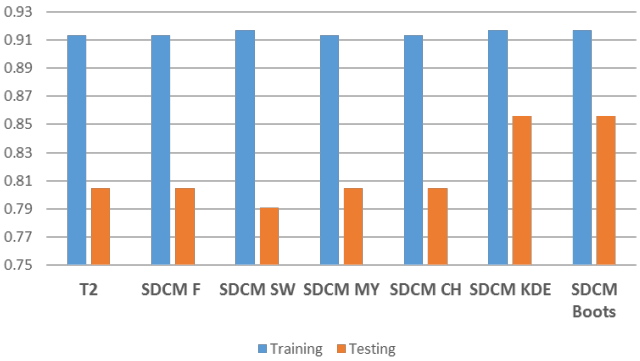

Figure 3: Hit Rate Comparison for Various IDS Type

The FN rate and FP rate comparison for various control limit approaches in testing dataset is

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

exhibited at Figure 5. It could be understood that for testing dataset, T^2 based on $SDCM_{boot}$ and T^2 based on $SDCM_{KDE}$ have the two lowest FN rate respectively. For testing dataset, both T^2 based on $SDCM_{boot}$ and T^2 based on $SDCM_{KDE}$ have the smallest difference between FP rate and FN rate. Thus, it could be concluded that T^2 based on $SDCM_{boot}$ and T^2 based on $SDCM_{KDE}$ have the better performance compared to others.

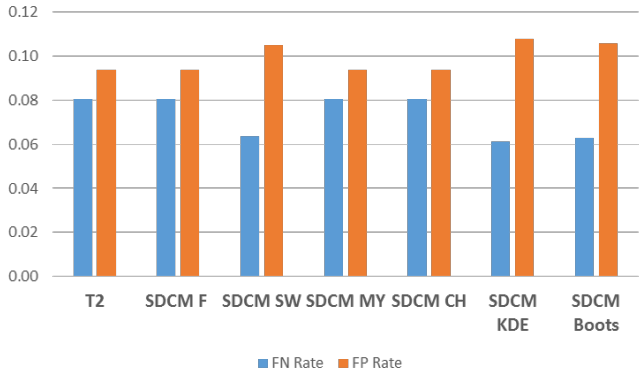

Figure 4: FN and FP Rate Comparison of Training Data

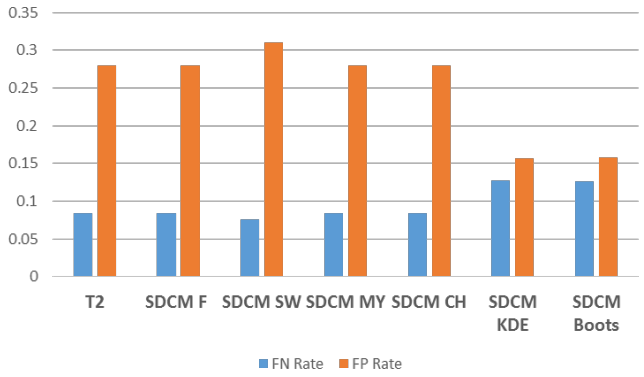

Figure 5: FN and FP Rate Comparison of Testing Data

5.2 Discussion

Based on the performance evaluation of conventional T^2 and T^2 based on SDCM with various control limit approaches, it could be known that conventional T^2, the T^2 based on $SDCM_F$, $SDCM_{SW}$, $SDCM_{MY}$ and $SDCM_{CH}$ have similar hit rate value. In addition, hit rates of training dataset from those approaches are higher than the hit rate of testing dataset. The low value of hit rate from testing dataset might be caused by the high value of FP rate from testing dataset. The high value of FP rate from testing dataset happens due to oversensitivity of IDS to detect an attack while attack is not actually happened in network. On the other hand, those approaches are superior due to the low value of FN rate. Consequently, IDS by those approaches would detect an attack while attack is actually happened in network but would produce high false alarm.

T^2 based on $SDCM_{KDE}$ and $SDCM_{boot}$ has the highest hit rate for training dataset while T^2 based on $SDCM_{boot}$ has the highest hit rate for testing dataset. Different from the other approaches, $SDCM_{boot}$ has high value of testing hit rate might be caused by low value of testing FP rate. The low value of FP rate from testing dataset happens due to superiority of control limit to detect an attack while real attacks happen in network. Similarly, FN rate also have low value. Thus, IDS constructed by bootstrap control limit yields low false alarm and superior to detect the attacks in network.

6. COMPARISON PERFORMANCE WITH EXISTING CLASSIFIRERS

In the previous section, the performance of proposed IDS based on T^2 SDCM using bootstrap approach is compared with the other control limits. Several researchers has proposed IDS using some classification methods for NSL KDD dataset. In this section, the performance of proposed chart is compared with the other classifiers such as Random Forest Classification (RFC), Logistic Regression (LR), Gaussian Naive Bayes (GNB), and Support Vector Machine (SVM) method. The proposed T^2 based on $SDCM_{boot}$ is compared with the other classification methods such as in Belavagi and Muniyal [32]. Belavagi and Muniyal work is using to compare the proposed method not only because it is using the same NSL-KDD dataset with various method but this work is a recent publication in this field.

The hit rate of proposed IDS using T^2 based on $SDCM_{boot}$ is compared with the hit rate of the other classification methods as displayed at Figure 6. From the figure the hit rate of proposed IDS based on $SDCM_{boot}$ is 0.92. The hit rate for Logistic Regression (LR) method is 0.84 while the Gaussian Naïve Bayes (GNB) method has 0.79 hit rate. The Support Vector Machine (SVM) and Random Forest Classification (RFC) have hit rate of 0.75 and 0.99 respectively.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

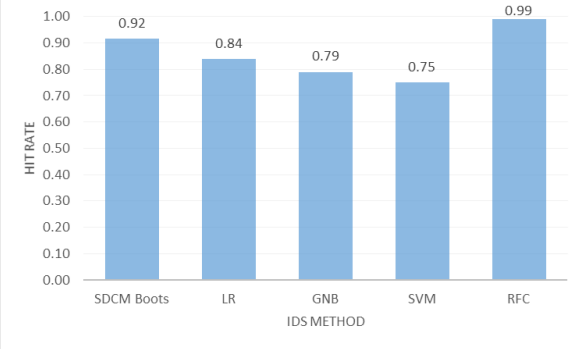

Figure 6: Hit Rate Comparison of Proposed IDS with the other classification methods in Belavagi and Muniyal [32]

The hit rate of proposed T^2 based on $SDCM_{boot}$ is higher than that of the other classification methods, except for Random Forest Classification (RFC). Otherwise, the hit rates of Logistic Regression (LR), Gaussian Naive Bayes (GNB), and Support Vector Machine (SVM) method are smaller than that of T^2 based on $SDCM_{boot}$.

Based on the hit rate comparison, the proposed T^2 based on $SDCM_{boot}$ is outperforms the performance of IDS based of LR, GNB IDS method, and SVM. This fact reveals that the proposed T^2 based on $SDCM_{boot}$ is effective to be used as IDS better than the other classification methods in Belavagi and Muniyal [31] except the Random Forest Classification (RFC).

7. CONCLUSION

In this paper, multivariate Hotelling's T^2 control chart is improved by SDCM based on bootstrap control limit and applied into IDS. The evaluation performance of IDS for NSL-KDD dataset is conducted using conventional T^2 and T^2 based on SDCM control chart with various control limit approaches. Furthermore, the performance of proposed IDS is compared with some existing classifiers.

The performance evaluation by confusion matrix shows that the proposed IDS using T^2 based on SDCM with bootstrap control limit outperforms the other control chart approaches in testing dataset. For training dataset, T^2 based on $SDCM_{KDE}$ and $SDCM_{boot}$ has the highest hit rate.

Moreover, the proposed T^2 based on $SDCM_{boot}$ outperforms the other existing classification methods except random forest classification. Thus, the proposed method is effective to be applied in IDS based on its ability to detect anomaly in the network which is confirmed by the value of hit rate. The multiclass detection for each type of attack with incremental algorithm can be considered as a future research.

DISCLOSURE STATEMENT

The authors declare that there are no potential conflict of interest

ACKNOWLEDGEMENT

This work was supported by Research, Technology, and Higher Education Ministry, Republic of Indonesia through PMDSU scheme year 2017

REFRENCES

[1] R. Bace, P. Mell, NIST special publication on intrusion detection systems, 2001. doi:10.1016/S1361-3723(01)00614-5.

[2] R.G. Bace, Intrusion detection, Macmillan Technical Publishing, Indianapolis, IN, 2000.

[3] W.A. Shewhart, Some Applications of Statistical Methods to the Analysis of Physical and Engineering Data, Bell Labs Tech. J. 3 (1924) 43–87.

[4] S.W. Roberts, Control Chart Tests Based on Geometric Moving Averages, Technometrics. 1 (1959) 239–250. doi:10.1080/00401706.1959.10489860.

[5] E.S. Page, Cumulative Sum Charts, Technometrics. 3 (1961) 1–9. doi:10.1080/00401706.1961.10489922.

[6] W.H. Woodall, Control charts based on attribute data: Bibliography and review, J. Qual. Technol. 29 (1997) 172. http://proquest.umi.com/pqdweb?did=11613494&Fmt=7&clientId=43036&RQT=309&VName=PQD.

[7] D.B. Laney, Improved control charts for attributes, Qual. Eng. 14 (2002) 531–537.

[8] M. Ahsan, M. Mashuri, H. Khusna, Evaluation of Laney p' Chart Performance, Int. J. Appl. Eng. Res. 12 (2017) 14208–14217.

[9] C.A. Catania, C.G. Garino, Automatic network intrusion detection: Current techniques and open issues, Comput. & Electr. Eng. 38 (2012) 1062–1072. doi:10.1016/j.compeleceng.2012.05.013.

[10] S. Bersimis, A. Sgora, S. Psarakis, The application of multivariate statistical process monitoring in non-industrial processes, Qual. Technol. Quant. Manag. 3703 (2016) 1–24. doi:10.1080/16843703.2016.1226711.

[11] Y. Park, A Statistical Process Control Approach for Network Intrusion Detection, Georgia Insitute of Technology, 2005.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[12] N. Ye, X. Li, Q. Chen, S.M. Emran, M. Xu, Probabilistic techniques for intrusion detection based on computer audit data, IEEE Trans. Syst. Man, Cybern. Part ASystems Humans. 31 (2001) 266–274. doi:10.1109/3468.935043.

[13] N. Ye, S.M. Emran, Q. Chen, S. Vilbert, Multivariate statistical analysis of audit trails for host-based intrusion detection, IEEE Trans. Comput. 51 (2002) 810–820. doi:10.1109/TC.2002.1017701.

[14] G. Qu, S. Hariri, M. Yousif, Multivariate Statistical Analysis for Network Attacks Detection, 3rd ACS/IEEE Interenational Conf. Comput. Syst. Appl. (2005) 9–14. doi:10.1109/AICCSA.2005.1387011.

[15] N. Ye, D. Parmar, C.M. Borror, A Hybrid SPC Method with the Chi-Square Distance Monitoring Procedure for Large-scale, Complex Process Data, Qual. Reliab. Eng. Int. 22 (2006) 393–402. doi:10.1002/qre.717.

[16] Z. Zhang, X. Zhu, J. Jin, SVC-Based Multivariate Control Charts for Automatic Anomaly Detection in Computer Networks, in: IEEE, 2007: p. 56. doi:10.1109/CONIELECOMP.2007.99.

[17] M. Tavallaee, W. Lu, S.A. Iqbal, A. Ghorbani, A Novel Covariance Matrix based Approach for Detecting Network Anomalies, in: Sixth Annu. Conf. Commun. Networks Serv. Res., 2008.

[18] A. Avalappampatty Sivasamy, B. Sundan, A Dynamic Intrusion Detection System Based on Multivariate Hotelling's T ² Statistics Approach for Network Environments, Sci. World J. 2015 (2015) 1–9. doi:10.1155/2015/850153.

[19] J.L. Alfaro, J.F. Ortega, A comparison of robust alternatives to Hotelling's T2 control chart, J. Appl. Stat. 36 (2009) 1385–1396. doi:10.1080/02664760902810813.

[20] N.D. Tracy, J.C. Young, R.L. Mason, Multivariate Control Charts for Individual Observations, J. Qual. Technol. 24 (1992) 88. https://ezproxy.bibl.ulaval.ca/login?url=http://search.proquest.com/docview/214483701?accountid=12008%5Cnhttp://sfx.bibl.ulaval.ca:9003/sfx_local??url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&genre=article&sid=ProQ:ProQ%3Aabiglobal&atitle=.

[21] C.A. LOWRY, D.C. MONTGOMERY, A review of multivariate control charts, IIE Trans. 27 (1995) 800–810. doi:10.1080/07408179508936797.

[22] Y. Chou, R.L. Mason, J.C. Young, Power comparisons for a hotelling's t 2 STATISTIC, Commun. Stat. - Simul. Comput. 28 (1999) 1031–1050. doi:10.1080/03610919908813591.

[23] S. Cambanis, S. Huang, G. Simons, On the theory of elliptically contoured distributions, J. Multivar. Anal. 11 (1981) 368–385. doi:10.1016/0047-259X(81)90082-8.

[24] J.H. Sullivan, W.H. Woodall, A Comparison of Multivariate Control Charts for Individual Observations, J. Qual. Technol. 28 (1996) 398–408.

[25] J.D. Williams, W.H. Woodall, J.B. Birch, J.O.E.H. Sullivan, On the Distribution of Hotelling ' s T 2 Statistic Based on the Successive Differences Covariance Matrix Estimator, J. Qual. Technol. 38 (2006) 217–229.

[26] N.J. Vargas, Robust estimation in multivariate control charts for individual observations, J. Qual. Technol. 35 (2003) 367–376.

[27] J.K. Wororomi, M. Mashuri, Irhamah, A.Z. Arifin, On monitoring shift in the mean processes with vector autoregressive residual control charts of individual observation, Appl. Math. Sci. 8 (2014) 3491–3499. doi:10.12988/ams.2014.44298.

[28] P. Phaladiganon, S.B. Kim, V.C.P. Chen, J.-G. Baek, S.-K. Park, Bootstrap-Based T 2 Multivariate Control Charts, Commun. Stat. - Simul. Comput. 40 (2011) 645–662. doi:10.1080/03610918.2010.549989.

[29] P. Phaladiganon, S.B. Kim, V.C.P. Chen, W. Jiang, Principal component analysis-based control charts for multivariate nonnormal distributions, Expert Syst. Appl. 40 (2013) 3044–3054. doi:10.1016/j.eswa.2012.12.020.

[30] B. Efron, Bootstrap Methods: Another Look at the Jacknife, Ann. Stat. 7 (1979) 1–26. doi:10.1214/aoms/1177692541.

[31] B. Efron, R.J. Tibshirani, An introduction to the bootstrap, CRC press, 1994.

[32] M.C. Belavagi, B. Muniyal, Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection, in: Procedia Comput. Sci., 2016: pp. 117–123. doi:10.1016/j.procs.2016.06.016.

[33] D. Montgomery, Introduction to statistical quality control, 2009. doi:10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C.

[34] H. Hotelling, Multivariate quality control, in: Tech. Stat. Anal., McGraw-Hill., New York, 1974.

[35] D.M. Hawkins, D.F. Merriam, Zonation of multivariate sequences of digitized geologic data, J. Int. Assoc. Math. Geol. 6 (1974) 263–269. doi:10.1007/BF02082892.

[36] D.S. Holmes, A.E. Mergen, Improving the performance of the T2 control chart, Qual. Eng. 5 (1993) 619–625. doi:10.1080/08982119308919004.

[37] R.L. Mason, J.C. Young, Multivariate Statistical Process Control with Industrial Applications, Society for Industrial and Applied Mathematics,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2002. http://epubs.siam.org/doi/book/10.1137/1.9780898718461.

[38] Murray Rosenblatt, Remarks on Some Nonparametric Estimates of a Density Function, Ann. Math. Stat. Volume 27 (1956) 832–837. doi:10.1214/aoms/1177728190.

[39] E. Parzen, On Estimation of a Probability Density Function and Mode, Ann. Math. Stat. 33 (1962) 1065–1076. doi:10.1214/aoms/1177704472.

[40] Y.-M. Chou, R. Mason, J. Young, the Control Chart for Individual Observations From a Multivariate Non-Normal Distribution, Commun. Stat. Theory Methods. 30 (2001) 1937. doi:10.1081/STA-100105706.

[41] W. Härdle, O. Linton, Applied nonparametric methods, Handb. Econom. 4 (1994) 2295–2339. doi:10.1016/S1573-4412(05)80007-8.

[42] R.L. Burden, J.D. Faires, Numerical Analysis, 2011. doi:10.1017/CBO9781107415324.004.

[43] M. Tavallaee, E. Bagheri, W. Lu, A.A. Ghorbani, A detailed analysis of the KDD CUP 99 data set, in: IEEE Symp. Comput. Intell. Secur. Def. Appl. CISDA 2009, 2009. doi:10.1109/CISDA.2009.5356528.

[44] S.J. Stolfo, KDD cup 1999 dataset, UCI KDD Repos. Http//kdd.ics.uci.edu. (1999) 0.

[45] S. Revathi, A. Malathi, A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection, (n.d.).