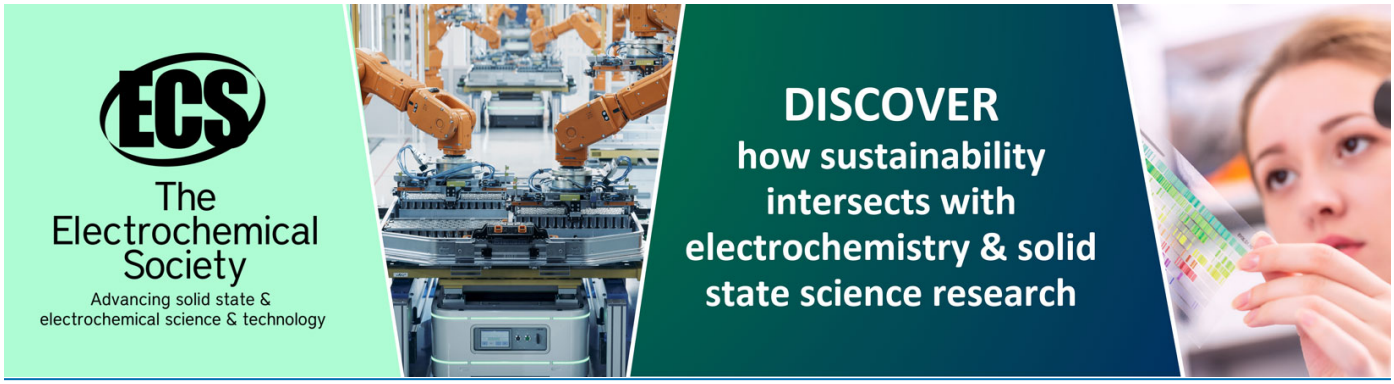
T^2 Control Chart based on Successive Difference Covariance Matrix for Intrusion Detection System



IOP Conf. Series: Journal of Physics: Conf. Series **1028** (2018) 012220 doi:10.1088/1742-6596/1028/1/012220

T^2 Control Chart based on Successive Difference Covariance Matrix for Intrusion Detection System

Muhammad Ahsan[1], Muhammad Mashuri[1*], Heri Kuswanto[1], Dedy Dwi Prastyo[1], and Hidayatul Khusna[1]

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

[*]m_mashuri@statistika.its.ac.id

Abstract: The Intrusion detection is a process to monitor the events taking place in a computer system or network and analyze the monitoring results to find signs of intrusion. One of alternative solutions for intrusion detection is the usage of statistical methods that Statistical Process Control especially the control charts.. In this research, the Hotelling's T^2 chart performance for intrusion detection is improved using the Successive Difference Covariance Matrix where the control limits will be calculated using Kernel Density Estimation. The proposed method using T^2 based on Kernel Density Estimation control limit outperforms other approaches both in training and testing dataset.

1. Introduction

Intrusion detection can also be performed using a statistical approach. One 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. SPC has an advantage where it does not require knowledge of an unprecedented attack. SPC based Intrusion Detection System (IDS) can also guarantee the real time attack detection process [1]. The most commonly used multivariate control chart for intrusion detection is Hotelling's T^2. Several researchers have developed T^2 control chart for individual observations[2,3]. The comparison of T^2 control chart power value based on different kind of covariance matrix estimator had been investigated by Chou, Mason, and Young[4]. Cambanis, Huang, and Simons[5] probe the necessary and sufficient requirement under those underlying multivariate normal distribution.

However, taking the sample covariance matrix from the data consist of individual observation leads to poor performance in detecting shift in mean vector[6]. Moreover, the utilization of robust covariance matrix estimator would improve T^2 control chart performance in detecting shift of mean vector[7]. Successive Difference Covariance Matrix (SDCM) is one of the robust covariance matrix estimators. The T^2 control chart based on SDCM proved effective in detecting shift of mean vector[6,8]. Moreover, VAR based residual of T^2 control chart using SDCM for multivariate autocorrelated data is powerful[9]. Although effectively used, the distribution of T^2 control chart based on SDCM has not been exactly determined. Some literatures propose approximate distribution

for T^2 control chart based on SDCM [6,7]. In order to overcome this limitation, some studies improved T^2 based on SDCM control limit by using nonparametric approaches. The T^2 based on SDCM control limit could be improved significantly by Kernel Density Estimation (KDE)[10-12]. Hence, this study is aimed to propose T^2 control chart based on SDCM using KDE approach. The utilization of KDE method is expected to yield more accurate control limit of T2 based on SDCM. The performance of proposed method would be compared with the other approaches.

2. Hotelling's T^2 Control Chart Based on SDCM

The Hotelling's T^2 is one of multivariate the control charts that could be used to monitor the mean of production process[13]. Let $\mathbf{x}_i$, where $i = 1, 2, \ldots, n$ define number of observation, are random vectors follow multivariate normal i.i.d with common mean vector and covariance matrix, i.e. $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. On the other hand, those *n*x*p* dataset could be defined as: $\mathbf{X} = \left[\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n'\right]'$. T^2 statistics [14] can be calculated according to the following equation:

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

where $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$ and $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)'$. Under the assumption that the data follow multivariate distribution, the control limit of can be obtained as follows:

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

Another alternative method to estimate the covariance matrix is SDCM that firstly introduced by Hawkins and Merriam[15] and Holmes and Mergen[16]. 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 $\mathbf{S}_D = \frac{1}{1(n-1)}\sum_{i=2}^{n}\left(\mathbf{x}_i - \mathbf{x}_{i-1}\right)\left(\mathbf{x}_i - \mathbf{x}_{i-1}\right)'$. In Phase I, the $\mathbf{S}_D$ is an unbiased estimator for $\boldsymbol{\Sigma}$ [6]. Under the assumption that the data follow multivariate distribution, there are some approaches to construct control limit, i.e. CL_{SW}[6], CL_{MY}[17], and CL_{χ^2}, that could be obtained as follows:

$$CL_{SW} = \frac{(n-1)^2}{n} BETA_{(1-\alpha), \frac{p}{2}, \frac{(g-p-1)}{2}}, \quad (4)$$

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

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

where $BETA_{(1-\alpha), p, g}$ is $1-\alpha$ quantile of beta distribution with shape parameter p and g, $\chi^2_{(1-\alpha), v}$ is $(1-\alpha)$ 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

Chou, Mason, and Young[12] introduced KDE to estimate the distribution of T^2 statistics. Given *n* value of T^2 statistics obtained from in-control conditions, then T^2 distribution could be calculated using following kernel function:

$$\hat{f}_h(t) = \frac{1}{n}\sum_{i=1}^{n} K\left[\frac{\left(t - T_i^2\right)}{h}\right], \quad (3)$$

where K and h define kernel function and smoothing parameter respectively.
Furthermore, the control limit of T^2 based on KDE could be determined by percentile of kernel distribution. Thus, control limit T^2 based on KDE equal to $100(1-\alpha)^{\text{th}}$ percentile which could be calculated using following equation:

$$CL_{\text{kernel}} = \hat{f}_h(t)^{-1}(1-\alpha). \quad (4)$$

4. Methodology

Dataset that used in this study is NSL-KDD. This dataset proposed by Tavallaee *et al.*[18] as a solution for obsolete KDD-99 dataset[19]. 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.

In this study, NSL-KDD data is analyzed using conventional T2 and T2 based on SDCM control chart. Furthermore, the control limit of T2 based on SDCM is estimated using several approaches, i.e. F distribution control limit Sullivan and Woodall approach (SDCMSW) based on (4), Mason and Young approach (SDCMMY) according to (5), chi-square control limit based on (6), and proposed KDE control limit according to (8). Moreover, the performance of IDS is evaluated by confusion matrix as shown in Table 1.

Table 1. Intrusion detection confusion matrix

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

The *FP* causes a false alarm while *FN* allows an attack on the system. The level of accuracy used is the hit rate that can be calculated as follows:

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

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 displays the performance of IDS for NSL-KDD dataset using conventional T^2 and T^2 based on SDCM control chart.

Table 2. Performance of various IDS for training data

IDS	Hit Rate	FALSE Negative	FALSE Positive	FN Rate	FP Rate
T^2	0.9133	5428	5494	0.081	0.094
$SDCM_F$	0.9134	5417	5495	0.080	0.094
$SDCM_{SW}$	0.9170	4280	6170	0.064	0.105
$SDCM_{MY}$	0.9133	5429	5492	0.081	0.094

IDS	Hit Rate	FALSE Negative	FALSE Positive	FN Rate	FP Rate
$SDCM_{CH}$	0.9133	5427	5492	0.081	0.094
$SDCM_{KDE}$	0.9171	4124	6319	0.061	0.108

Table 2 displays the performance of T^2 and T^2 based on SDCM control chart with various control limit approaches for training data. While, the performance of T^2 and T^2 based on SDCM control chart with various control limit approaches for testing data is shown at Table 3.

Table 3. Performance of various IDS for testing data

IDS	Hit Rate	FALSE Negative	FALSE Positive	FN Rate	FP Rate
T^2	0.8049	814	3584	0.084	0.279
$SDCM_F$	0.8049	814	3585	0.084	0.279
$SDCM_{SW}$	0.7911	731	3978	0.075	0.310
$SDCM_{MY}$	0.8049	814	3584	0.084	0.279
$SDCM_{CH}$	0.8049	814	3584	0.084	0.279
$SDCM_{KDE}$	0.8558	1236	2014	0.127	0.157

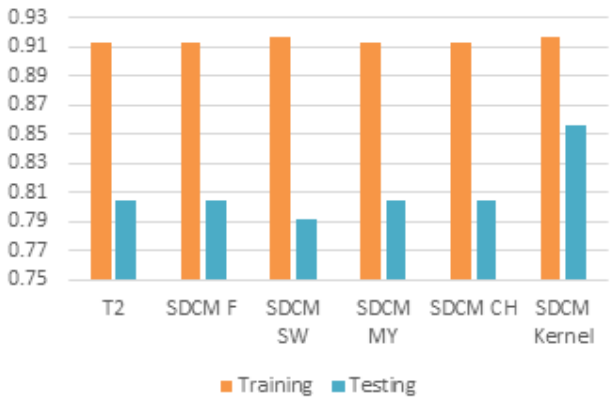


Figure 1. Hit rate comparison for various IDS type

The comparison of hit rate values from various control limit approaches need to be visualized in single graphic so the performance of each control chart could be compared easily. Figure 1 exhibits the hit rate comparison of various control limit approaches for both training and testing dataset. It can be known that for training dataset, the highest hit rate is possessed by T^2 based on $SDCM_{KDE}$ and T^2 based on $SDCM_{SW}$ respectively. T^2 based on $SDCM_{KDE}$ has highest hit rate for testing data. The *FN* rate and *FP* rate comparison for various control limit approaches in training dataset is performed at Figure 2(a). The two lowest *FN* rate is owned by T^2 based on $SDCM_{KDE}$ and T^2 based on $SDCM_{SW}$ respectively.

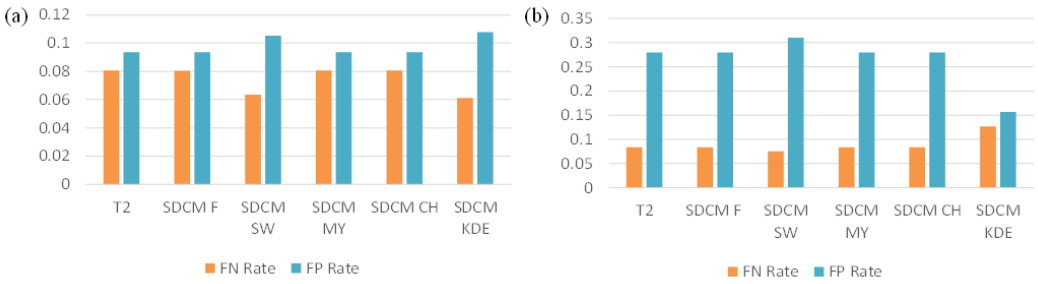


Figure 2. FN and FP rate comparison for (a) training dataset, (b) testing dataset

Figure 2(b) explains the *FN* rate and *FP* rate comparison for various control limit approaches in testing dataset. It could be understood that for testing dataset, T^2 based on $SDCM_{KDE}$ has lowest *FP* rate but the *FN* rate is highest.

Therefore, T^2 based on $SDCM_{KDE}$ has the highest hit rate both for training and testing dataset. The 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 of $SDCM_{KDE}$ also have low value. Thus, IDS constructed by KDE control limit yields low false alarm and superior to detect the attacks in network.

6. Conclusion and Future Research

The evaluation performance of IDS for NSL-KDD dataset had been conducted using conventional T^2 and T^2 based on SDCM control chart using some approaches to estimate the control limit of T^2 based on SDCM. Those control limit are *F* distribution control limit, Sullivan and Woodall approach, Mason and Young approach, chi-square control limit, and KDE control limit. The performance evaluation result shows that the proposed IDS using T^2 based on SDCM control chart using KDE control limit outperforms the other approaches both in training and testing dataset. Furthermore, IDS using T^2 based on SDCM with computational approaches such as bootstrap might be useful for future researches.

References

[1] C A Catania, C G Garino, 2012 *Computers & Electrical Engineering* 38 1062
[2] N D Tracy, J C Young and R L Mason 1992 *Journal of Quality Technology* **24**, 88.
[3] C A Lowry, D C Montgomery 1995 *IIE Transactions* 27, 800
[4] Y Chou, R L Mason and J C Young, 1999 Communications in Statistics - Simulation and Computation 28 1031.
[5] S Cambanis, S Huang and G Simons 1981 *Journal of Multivariate Analysis* **11** 368
[6] J H Sullivan and W H Woodall 1996 *Journal of Quality Technology* **28** 398
[7] J D Williams, W H Woodall, J B Birch and J O E H Sullivan 2006 *Journal of Quality Technology* **38** 217
[8] N J Vargas 2003 *Journal of Quality Technology* 35, 367
[9] J K Wororomi, M Mashuri, Irhamah, A Z Arifin 2014 *Applied Mathematical Sciences* **8** 3491
[10] P Phaladiganon, S B Kim, V C P Chen, J G Baek and S K Park, 2011 *Communications in Statistics - Simulation and Computation* **40** 645
[11] P Phaladiganon, S B Kim, V C.P Chen and W Jiang, 2013 *Expert Systems with Applications* **40** 3044
[12] Y M Chou, R Mason and J Young 2001 *Communications in Statistics: Theory & Methods* **30** 1937
[13] D Montgomery 2009 *Introduction to Statistical Quality Control*
[14] H Hotelling 1974 *Techniques of Statistical Analysis* (New York: McGraw-Hill)
[15] D M Hawkins and D F Merriam, 1974 Journal of the International Association for

Mathematical Geology 6, 263
[16] D S Holmes and A E Mergen, 1993 *Quality Engineering* **5** 619
[17] R L Mason and J C Young 2002 *Multivariate Statistical Process Control with Industrial Applications* Society for Industrial and Applied Mathematics
[18] M Tavallaee, E Bagheri, W Lu and A A Ghorbani 2009 *in; IEEE Symposium on Computational* Intelligence for Security and Defense Applications, CISDA
[19] S J Stolfo 1999 *UCI KDD Repository* http://kdd.ics.uci.edu.