Comparing the performance of $T^2$ chart based on PCA Mix, Kernel PCA Mix, and Mixed Kernel PCA for Network Anomaly Detection

M Mashuri*, M Ahsan¹, H Kuswanto¹, D D Prastyo¹, H Khusna¹, and Wibawati¹
Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
*e-mail: m_mashuri@statistika.its.ac.id

Abstract. Statistical Process Control (SPC) is not only used to monitor the quality of manufacturing processes and services but also is applied to detect intrusions in the network. Hotelling's $T^2$ chart is the SPC method that has been widely developed for intrusion detection. However, in its application, the conventional Hotelling's $T^2$ chart has several drawbacks such as less effective when used to monitor large observations and quality characteristics. Conventional Hotelling's $T^2$ chart is not perform-well for non-Gaussian distributed data. Also, the current conventional control chart has not been able to monitor the processes which have mixed quality characteristics. To overcome these weaknesses, two types of the control chart is proposed in this study, namely, the multivariate control chart based on Principal Component Analysis (PCA) Mix and Kernel PCA. For Kernel PCA chart, two schemes are developed, that is Kernel PCA Mix and Mixed Kernel PCA control charts. Kernel Density Estimation (KDE) is employed to estimate the control limits of the developed charts. In monitoring the network intrusion, the proposed control charts are applied to well-known NSL-KDD dataset. The evaluation performance shows that the PCA Mix chart can detect attacks occurred on the network more accurate and faster compared to the Kernel PCA Mix and Mixed Kernel PCA charts.

Keywords: Mixed Quality Characteristics, PCA Mix, Kernel PCA, Hotelling’s $T^2$ Chart, Kernel Density Estimation

1. Introduction

The metric data can be handled by the variable chart [1,2]. On the other hand, the attribute chart [3–5] can be applied for the non-metric data. In the manufacturing process, it is sometimes necessary to monitor metric and non-metric data together [6]. However, so far, there are still a few multivariate control charts that can be used to monitor the quality of the mixed data simultaneously. Some previous research about mixed data can be found in [7–9].

In its application to monitor the network traffic, multivariate control charts have several drawbacks. First, the performance of a multivariate control chart will decrease when it is used to monitor a process with large-quality characteristics. In fact, most of the benchmark datasets for Intrusion Detection System (IDS) method have large quality characteristics or variables. Second, the number of false alarms that occur due to the network traffic data is not multivariate normally distributed [10] with a nonlinear relationship. Third, the network traffic data consists of metric and non-metric variables. Meanwhile, the conventional control charts are developed using certain distribution assumptions (usually multivariate normal distribution) and cannot accommodate differences in data types.
To overcome these drawbacks, several methods can be applied. First, to overcome the large quality characteristics in a process, the PCA method can be used [11]. By using this method, the large quality characteristics involved in monitoring the process can be reduced. Second, to overcome the high false alarms caused by the non-multivariate normal distributed data, the Kernel Density Estimation (KDE) method can be used to estimate control limits [12–14]. Finally, to overcome the nonlinearity problem between the quality characteristics, the KPCA method [15] can be used. Finally, to overcome the different types of metric and non-metric data, the PCA Mix method is used [16,17].

Based on the description above, this research is focused on developing the multivariate control charts that can deal with large-quality characteristics with mixed data, non-multivariate normal distributed data, and nonlinear relationship between the quality characteristics. Therefore, this research will develop two types of control charts based on PCA Mix and Kernel PCA. For the Kernel PCA chart, two schemes are developed, namely Kernel PCA Mix and Mixed Kernel PCA. The control limits of the proposed charts are estimated using the KDE method. Furthermore, the developed control charts are then applied to monitor NSL-KDD 20% dataset. Also, the performance of each proposed chart is compared in monitoring the network intrusion detection. The rest of this paper is arranged as follows. Section 2 presents the proposed charts. Section 3 contains the methodology and algorithm of the proposed IDS. Section 4 shows the results and discussions about the performance comparison of the proposed charts. Finally, Section 5 is devoted to the conclusion and future works.

2. Proposed Methods

2.1. PCA Mix Control Chart

Figure 1 presents the algorithm of PCA Mix control chart. The main idea of this chart is utilizing the PCA Mix method [18] to create the Principal Component Scores (PCs). The next step is calculating the statistic \( T^2 \). Finally, the control limit is estimated using the KDE method with Gaussian kernel [19], where the Gaussian kernel is defined as follows:

\[
K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right), -\infty < u < \infty.
\]

---

![Figure 1 Algorithm of PCA Mix chart](image-url)
2.2. Kernel PCA Mix Control Chart
The algorithm of Kernel PCA Mix control chart is depicted in Figure 2. The main idea of this chart is applying the Kernel PCA procedure [15] to create the PCs by using the kernel function. In this research, Kernel PCA is calculated using the Radial Basis Function (RBF) as follows:

\[ K(x_i, x_j) = \exp(-\sigma \| x_i - x_j \|^2), \]

where \( \sigma \) is the hyperparameter of RBF.

First, define \( Z = [Z_1, Z_2] \) sized \( n \times (p + m) \), where \( Z_1 \) sized \( n \times p \) is the centered version of \( X_1 \) and \( Z_2 \) sized \( n \times m \) is the centered version of binary coding of \( X_2 \). The metric kernel of \( Z \) is obtained as

\[ K = K(z_i, z_j) = \langle \Phi(z_i), \Phi(z_j) \rangle. \]

By conducting the KPCA on \( Z \) the principal component \( t \) is formed. Furthermore, the statistic \( \tilde{T}^2_i \) is determined by the \( l \) first principal component. The final step is calculating the KDE control limit.

![Figure 2 Algorithm of Kernel PCA Mix chart](image-url)
2.3. Mixed Kernel PCA Control Chart

The procedures of Mixed Kernel PCA control chart is illustrated in Figure 3. Similar to the previous method, this scheme uses Kernel PCA to create PCs. However, in this scheme, the non-metric data is treated differently from the Kernel PCA Mix control chart. If in the previous method non-metric data is formed into the dummy variable, in this approach non-metric data is formed into the new non-metric kernel matrix $K_n$ [20,21]. The non-metric kernel $K_n$ is calculated as

$$K_n = \frac{\sum_{d=1}^{m} k_{0,d}}{m},$$

where $k_0$ is defined as follows:

$$k_0 = \{\begin{array}{ll}
1, & x_{2,j} = x_{2,j} \\
0, & x_{2,j} \neq x_{2,j}.
\end{array}$$

After that, the mixed kernel $K_{\text{mixed}}$ can be obtained as follow:

$$K_{\text{mixed}} = 0.5(K_m + K_n),$$

with $K_m$ is metric kernel calculated using RBF. After conducting Kernel PCA on $K_{\text{mixed}}$, the statistic $T_i^2$ is calculated from the $l$ first principal component. The final step is calculating the KDE control limit.

**Figure 3** Algorithm of Mixed Kernel PCA chart
3. Methodology

3.1. NSL-KDD 20% Dataset
In this research, the NSL-KDD 20% is used to evaluate the performance of the proposed chart. This dataset can be found on https://www.unb.ca/cic/datasets/nsl.html. The characteristics of this dataset are described in Table 1.

| Label   | Number of observations | Percentage |
|---------|------------------------|------------|
| Normal  | 13,449                 | 53.39      |
| DOS     | 9,234                  | 36.65      |
| Probe   | 2,289                  | 9.09       |
| U2R     | 11                     | 0.04       |
| R2L     | 209                    | 0.83       |
| Total   | 25,192                 | 100.00     |

3.2. Intrusion Detection System Algorithm

Phase I: Normal Profile
In this phase, the normal profiles (mean vector, covariance matrix, and the KDE control limit) are calculated from the normal labelled data. The estimated normal profile from this phase will then be used in phase II to monitor the new data. The procedures from phase I are given as follows:

Step 1 Form matrix $X_{1,normal}$, which contains the metric data that is labelled as the normal connection.

Step 2 Form matrix $X_{2,normal}$, which contains the non-metric data that is labelled as the normal connection.

Step 3 Conduct PCA Mix, Kernel PCA Mix, or Mixed Kernel PCA to $X_{1,normal}$ and $X_{2,normal}$.

Step 4 Calculate mean vector $\mu_{normal}$, eigenvalue matrix $\Lambda_{normal}$, and eigenvector matrix $\alpha_{normal}$ from the principal component.

Step 5 Calculate statistic $T_i^2, T_k^2, or \hat{T}_{mk}$.

Step 6 Determine $\alpha$ and calculate the KDE control limit $CL$.

Phase II: Detection
In phase II, $\mu_{normal}, S_{normal}$, and $CL$ estimated from phase I are used to monitor the new connection. The steps of phase II are defined as follows:

Step 1 Form matrix $X_{1,test}$, which contains the metric data of the new connection.

Step 2 Form matrix $X_{2,test}$, which contains the non-metric data of the new connection.

Step 3 Conduct PCA Mix, Kernel PCA Mix, or Mixed Kernel PCA to $X_{1,test}$ and $X_{2,test}$.

Step 4 Calculate statistic $T_i^2, T_k^2, or \hat{T}_{mk}$ from the principal component score of the new connection.

Step 5 If $T_i^2 > CL$ then the observation is an intrusion, otherwise, if $T_i^2 \leq CL$ then the observation is a normal connection.
3.3. Performance metric
The performance of the proposed chart is assessed by the confusion matrix as shown in Table 2. There are three performance metric used in this study, namely Hit Rate, FP Rate, and FN Rate. The Hit Rate is calculated as follows:

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

Meanwhile, the FP rate and FN Rate formula can be expressed as follows:

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

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

### Table 2 Intrusion detection confusion matrix

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

4. Results and discussions
In this section, the three types of the proposed control charts are applied to monitor NSL-KDD 20% dataset. Using three performance metric stated before, the performance of the proposed charts is compared to each other.

4.1. PCA Mix Control Chart
Table 3 shows the performance of PCA Mix chart in monitoring the NSL-KDD 20% dataset. It can be concluded that the best results are obtained by the KDE control limit of 1.6328 for the number of principal components $l = 2$. The KDE control limits are computed for the range values of $\alpha$ from 0.001 to 1. The control limits that produced the highest hit rate are tabulated in Table 2. From these results, it can be seen that this chart is still unable to detect some attacks which can be seen in the high FN rate.

### Table 3 Performance of PCA Mix Control Chart in monitoring the NSL-KDD 20% dataset

| $l$ | KDE control limit | hit rate | FP rate | FN rate |
|-----|-------------------|----------|---------|---------|
| 2   | 1.63280           | **0.91136** | 0.05688 | 0.12501 |
| 3   | 2.65236           | 0.89151  | 0.09785 | 0.12067 |
| 4   | 2.05552           | 0.80406  | 0.31712 | 0.05714 |
| 5   | 2.06507           | 0.78692  | 0.35215 | 0.05382 |
| 7   | 2.13866           | 0.79636  | 0.35668 | 0.02836 |
| 10  | 2.15982           | 0.70137  | 0.54651 | 0.01473 |

4.2. Kernel PCA Mix Control Chart
Table 4 reports the performance of Kernel PCA Mix Control Chart in monitoring the NSL-KDD 20% dataset with RBF kernel with hyperparameter $\sigma = 0.001$. It can be concluded that the best results are obtained for $l = 4$ and the estimated KDE control limit of 3698.07 with a hit rate value of 0.858.
Furthermore, after finding out that \( l = 4 \) yields the highest hit rate, the most optimum hyperparameter \( \sigma \) selection is tabulated in Table 5. Based on the table, it can be concluded that the most optimum hyperparameter value for \( l = 4 \) is 0.001.

### Table 4 Performance of Kernel PCA Mix Control Chart in monitoring the NSL-KDD 20% dataset with RBF kernel \( (\sigma = 0.001) \)

| \( l \) | KDE control limit | hit rate | FP rate | FN rate |
|-----|-------------------|----------|---------|---------|
| 2   | 2563.06           | 0.82744  | 0.0675  | 0.29285 |
| 3   | 2590.45           | 0.84741  | 0.06714 | 0.25044 |
| 4   | 3698.07           | **0.85769** | 0.08305 | 0.21016 |
| 5   | 13545.83          | 0.84653  | 0.07361 | 0.24491 |
| 7   | 20975.08          | 0.82347  | 0.13183 | 0.22771 |
| 10  | 2590.45           | 0.84741  | 0.06714 | 0.25044 |
| 20  | 43718.26          | 0.68986  | 0.42724 | 0.17601 |

### Table 5 Performance of Kernel PCA Mix Control Chart in monitoring the NSL-KDD 20% dataset for \( l = 4 \) and several values of \( \sigma \)

| \( \sigma \) | KDE control limit | hit rate | FP rate | FN rate |
|-------------|-------------------|----------|---------|---------|
| 0.10000     | 383.64            | 0.58772  | 0.02632 | 0.85429 |
| 0.01000     | 821.21            | 0.84522  | 0.06825 | 0.25385 |
| 0.00100     | 3698.07           | **0.85769** | 0.08305 | 0.21016 |
| 0.00500     | 1445.94           | 0.84590  | 0.06022 | 0.26160 |
| 0.00010     | 7883.36           | 0.63492  | 0.52643 | 0.18027 |
| 0.00001     | 25895.20          | 0.53385  | 0.00000 | 1.00000 |

#### 4.3. Mixed Kernel PCA Control Chart

Table 6 tabulates the performance of Mixed Kernel PCA Control Chart in monitoring the NSL-KDD 20% dataset using RBF kernel with hyperparameter \( \sigma = 0.001 \). From the monitoring results, it can be concluded that the best results are obtained for \( l = 25 \) and the KDE control limit of 103980.4. This combination produces a hit rate of 0.689. The optimum hyperparameter \( \sigma \) selection for \( l = 25 \) is carried out as tabulated in Table 7. According to the table, it can be concluded that the most optimum hyperparameter \( \sigma \) value for \( l = 4 \) is 0.005 with a hit rate of 0.783. This chart gives a fairly balanced result between the FN rate and FP rate. However, the value of hit rate is still lower than the other charts.

#### 4.4. Discussions

After searching for the best level of hit rate or accuracy for each chart in the previous section, this section provides some discussions about the performance of each chart. Figure 4a shows the hit rate comparison between the proposed charts. From these results, it can be said that the best accuracy level is owned by the PCA Mix chart with a hit rate above 0.9 then followed by the Kernel PCA Mix and Mixed Kernel PCA charts. From Figure 4b and c, it can also be seen that the PCA Mix chart yields the smallest FP and FN rate. According to this figure, it can also be concluded that the Kernel PCA chart obtains a smaller level of accuracy due to the inability of this chart to detect the intrusions which can be seen by the high FN rate produced.
Table 6 Performance of Mixed Kernel PCA Control Chart in monitoring the NSL-KDD 20% dataset with RBF kernel ($\sigma = 0.001$)

| $l$ | KDE control limit | hit rate | FP rate | FN rate |
|-----|-------------------|----------|---------|---------|
| 2   | 957.74            | 0.46241  | 0.9829  | 0.02759 |
| 3   | 1305.78           | 0.46943  | 0.9916  | 0.00255 |
| 4   | 1760.29           | 0.45681  | 0.97256 | 0.05143 |
| 5   | 2762.90           | 0.48126  | 0.27839 | 0.79400 |
| 7   | 89732.07          | 0.53581  | 0.00335 | 0.99200 |
| 10  | 103373.70         | 0.54779  | 0.02127 | 0.94575 |
| 15  | 85891.42          | 0.59844  | 0.16975 | 0.66704 |
| 20  | 91070.26          | 0.65422  | 0.21637 | 0.49400 |
| 25  | 103980.40         | 0.68970  | 0.23608 | 0.39530 |

Table 7 Performance of Mixed Kernel PCA Control Chart in monitoring the NSL-KDD 20% dataset for $l=25$ and several values of $\sigma$

| $\sigma$ | KDE control limit | hit rate | FP rate | FN rate |
|---------|-------------------|----------|---------|---------|
| 0.0100  | 43412.89          | 0.74865  | 0.39415 | 0.08779 |
| 0.0500  | 8348.824          | 0.73464  | 0.48435 | 0.01456 |
| 0.0075  | 55489.44          | 0.78172  | 0.29541 | 0.12995 |
| 0.0050  | 66549.23          | 0.78354  | 0.21154 | 0.22209 |
| 0.0010  | 103980.40         | 0.68970  | 0.23608 | 0.39532 |
| 0.0001  | 100525.70         | 0.59527  | 0.27043 | 0.55855 |

Figure 4 Comparison of PCA Mix, Kernel PCA Mix, and Mixed Kernel PCA for: a) Hit rate, b) FP Rate, c) FN Rate

From these facts, it can be concluded that the PCA Mix chart has a better performance than the other charts based on two criteria. The first criterion is the ease of use. When using the PCA Mix chart, the user is not necessary to find the optimum hyperparameter. The second criterion is the speed of computation time which is a very important aspect in monitoring the network intrusion. The PCA mix chart has a much faster computational time compared to the other charts. This happens because the PCA mix chart only needs to calculate the covariance matrix variance sized $p \times p$, while the other charts have to calculate the kernel matrix sized $n \times n$. Calculating the kernel matrix takes a quite high computational cost. However, PCA Mix Kernel and Mixed Kernel PCA charts are still possible to obtain better accuracy if the optimum combination of $l$, $\alpha$, and hyperparameter is found.
5. Conclusion and future research
In this research, three types of multivariate control charts for mixed quality characteristics are presented to monitor the intrusion in the network. The control limits of the proposed control charts are estimated using KDE procedure. These charts are then applied to monitor the NSL-KDD 20% dataset. The comparative performance of these charts shows that the PCA Mix chart can detect attacks more accurately than Kernel PCA and Mixed Kernel PCA charts. Not only in accuracy, the PCA Mix chart also shows its superiority in terms of ease of use and the low computational cost. The future work will find the optimum hyperparameter of the kernel function. To speed up the computing process, the Fast KPCA method [22] can be used. Also, the bootstrap confidence interval can be another alternative to estimate the empirical control limit [23,24].

References
[1] Shewhart W A 1926 Quality Control Charts Bell System Technical Journal 5 593–603
[2] Page E S 1961 Cumulative Sum Charts Technometrics 3 1–9
[3] Ahsan M, Mashuri M and Khusna H 2017 Evaluation of Laney p’ Chart Performance International Journal of Applied Engineering Research 12 14208–17
[4] Wibawati, Mashuri M, Purhadi, Irhamah and Ahsan M 2018 Performance Fuzzy Multinomial Control Chart Journal of Physics: Conference Series 1028 12120
[5] Woodall W H 1997 Control charts based on attribute data: Bibliography and review Journal of Quality Technology 29 172
[6] Pu X, Li Y and Xiang D 2011 Mixed variables-attributes test plans for single and double acceptance sampling under exponential distribution Mathematical Problems in Engineering 2011
[7] Aslam M, Khan N, Aldosari M S and Jun C H 2016 Mixed Control Charts Using EWMA Statistics IEEE Access 4 8286–93
[8] Aslam M, Azam M, Khan N and Jun C H 2015 A mixed control chart to monitor the process International Journal of Production Research 53 4684–93
[9] Wang J, Su Q, Fang Y and Zhang P 2018 A multivariate sign chart for monitoring dependence among mixed-type data Computers & Industrial Engineering 126 625–36
[10] Zhu X 2006 Anomaly Detection Through Statistics-Based Machine Learning For Computer Networks (The University of Arizona)
[11] Jackson J E 1959 Quality Control Methods for Several Related Variables Technometrics 1 359–77
[12] Phaladiganon P, Kim S B, Chen V C P, Baek J-G and Park S-K 2011 Bootstrap-Based T 2 Multivariate Control Charts Communications in Statistics - Simulation and Computation 40 645–62
[13] Phaladiganon P, Kim S B, Chen V C P and Jiang W 2013 Principal component analysis-based control charts for multivariate nonnormal distributions Expert Systems with Applications 40 3044–54
[14] Ahsan M, Mashuri M, Lee M H, Kuswanto H and Prastyo D D 2020 Robust adaptive multivariate Hotelling’s T2 control chart based on kernel density estimation for intrusion detection system Expert Systems with Applications
[15] Schölkopf B, Smola A and Müller K-R 1997 Kernel principal component analysis Artificial Neural Networks—ICANN’97 583–8
[16] Ahsan M, Mashuri M, Kuswanto H, Prastyo D D and Khusna H 2018 Multivariate Control Chart based on PCA Mix for Variable and Attribute Quality Characteristics Production & Manufacturing Research 6 364–84
[17] Ahsan M, Mashuri M, Kuswanto H, Prastyo D D and Khusna H 2019 Outlier detection using PCA mix based T2 control chart for continuous and categorical data Communications in Statistics - Simulation and Computation 1–28
[18] Chavent M, Kuentz-Simonet V, Labenon A and Saracco J 2014 Multivariate analysis of mixed
data: The PCAmixdata R package

[19] Botev Z I, Grotowski J F and Kroese D P 2010 Kernel density estimation via diffusion *The Annals of Statistics* **38** 2916–57

[20] Belanche Muñoz L A and Villegas M 2013 Kernel functions for categorical variables with application to problems in the life sciences *Artificial intelligence research and development: proceedings of the 16 International Conference of the Catalan Association of Artificial Intelligence* pp 171–80

[21] Villegas García M A 2013 An investigation into new kernels for categorical variables

[22] Ozawa S, Takeuchi Y and Abe S 2010 A fast incremental kernel principal component analysis for online feature extraction *Pacific Rim International Conference on Artificial Intelligence* (Springer) pp 487–97

[23] Khusna H, Mashuri M, Ahsan M, Suhartono S and Prastyo D D 2018 Bootstrap Based Maximum Multivariate CUSUM Control Chart *Quality Technology & Quantitative Management*

[24] Ahsan M, Mashuri M and Khusna H 2018 Intrusion Detection System Using Bootstrap Resampling Approach Of T2 Control Chart Based On Successive Difference Covariance Matrix *Journal of Theoretical and Applied Information Technology* **96** 2128–38