Kernel principal component analysis (PCA) control chart for monitoring mixed non-linear variable and attribute quality characteristics

Muhammad Ahsan *, Muhammad Mashuri, Hidayatul Khusna, Wibawati

Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

A R T I C L E   I N F O

Keywords:
Kernel PCA
Hotelling’s chart
Mixed quality characteristics
Kernel Density Estimation
Nonlinearity

A B S T R A C T

The products are commonly measured by two types of quality characteristics. The variable characteristics measure the numerical scale. Meanwhile, the attribute characteristics measure the categorical data. Furthermore, in monitoring processes, the multivariate quality characteristics may have a nonlinear relationship. In this paper, the Kernel PCA control chart is applied to monitor the mixed (attribute and variable) characteristics with the nonlinear relationship. First, the Average Run Length (ARL) is utilized to evaluate the performance of the proposed chart. The simulation studies show that the proposed chart can detect the shift in process. For this case, the Radial Basis Function (RBF) kernel demonstrates the consistent performance for several cases studied. Second, the performance comparison between the proposed chart and the conventional PCA Mix chart is performed. Based on the results, it is known that the proposed chart performs better in detecting the small shift in process. Finally, the proposed chart is applied to monitor the well-known NSL KDD dataset. The proposed chart shows good accuracy in detecting intrusion in the network. However, it still produces more False Negatives (FN).

1. Introduction

Two types of control charts have been developed based on the monitored quality characteristics. These charts are named as the attribute and variable charts. The variable control chart is developed to monitor the variable quality characteristics (in variable or ratio scale) such as length, temperature, or height (Montgomery, 2009). Meanwhile, to monitor the attribute quality characteristics (in categorical scale) the attribute chart was applied (Ahsan et al., 2018). When the characteristics quality is correlated or cannot be monitored separately, the multivariate control chart has been developed. There are three main types of multivariate variable control charts namely Shewhart, multivariate exponentially weighted moving average (MEWMA), and multivariate cumulative sum (MCUSUM).

The product quality characteristics are not only gauged individually by the attribute or variable characteristics but also can be monitored using a mixed scheme. In order to facilitate a mixed procedure of the monitoring process, several works have studied the development of the mixed characteristics charts. The mixed scheme by employing the combination between $\bar{X}$ and $\text{np}$ charts has been proposed and has a good performance in monitoring mixed characteristics (Aslam et al., 2015).

The mixed chart proposed by Aslam et al. (2015) is compared with Hybrid Exponential Weighted Moving Average (HEWMA) (Aslam et al., 2016). The spatial-sign covariance matrix-based control chart has been proposed by integrating the standardized ranks and spatial signs in calculating the mixed statistics (Wang et al., 2018). Furthermore, the principal component analysis for mixed data is applied in inspecting the process (Ahsan et al., 2018) and in detecting outliers (Ahsan et al., 2019). To overcome the PCA Mix chart drawbacks, Ahsan et al. (2020) proposed the Kernel PCA (KPCA) Mix chart for monitoring the mixed variable and attribute quality characteristics.

The problem arises when the PCA Mix chart (Ahsan et al., 2018) is applied to inspect the nonlinear multivariate processes. In monitoring processes, the multivariate quality characteristics may have a nonlinear relationship. Some studies about the utilization of control charts in detecting a shift in nonlinear data have been conducted. A multivariate chart based on KPCA and Exponentially Weighted Moving Average (EWMA) is proposed to monitor nonlinear biological processes (Yoo and Lee, 2006). Khediri et al. (2010) suggested Support Vector Regression (SVR) control charts for multivariate nonlinear processes with dependency on its samples. Fan et al. (2014) proposed a control chart based on filtering kernel independent component analysis–principal compo-
Table 1. The recent development of multivariate variable control charts.

| Sources                      | Proposed scheme                                      | Findings                                                                 |
|------------------------------|------------------------------------------------------|--------------------------------------------------------------------------|
| Chiang et al. (2021)         | New scheme of multivariate auxiliary-information-based (AI) chart | The performance of the proposed chart is evaluated using Monte-Carlo simulation and applied to cement data |
| Ahmad and Ahmed (2021)       | $T^2$ control chart to inspect the high dimensional data | The proposed method is unable without preprocessing or dimension reduction with high accuracy detection |
| Haddad (2021)                | $T^2$ control charts using modified Mahalanobis distance  | The proposed method has better performance in detecting more outliers compared to the traditional chart |
| Cabana and Lillo (2021)      | Robust multivariate chart for individual observations using reweighted shrinkage estimators | The proposed chart has a better performance for high dimensional and high contaminated data |
| Maleki et al. (2020)         | Median estimators of the $T^2$ control chart          | The proposed method outperforms performance compared to the conventional chart |
| Haddad et al. (2019)         | Bivariate Hotelling's $T^2$ charts with bootstrap data | The proposed method shows a better performance compared to the conventional method |
| Tiengket et al. (2020)       | Bivariate Copulas on the Hotelling's $T^2$ Control Chart | The bivariate copulas method can be used in the Hotelling's $T^2$ chart |
| Mashuri et al. (2019)        | $\hat{\gamma}$ ($\hat{R}$) control charts with Kernel Density Estimation (KDE) control limit  | The proposed control chart method presents better performance to detect the shift for the large characteristics and sample size |
| Mehmood et al. (2019)        | Hotelling $T^2$ control chart based on bivariate ranked set schemes | Proposed control chart schemes demonstrate an outstanding performance compared to the classical Hotelling $T^2$ |
| Haq and Khoo (2019)          | Adaptive MEWMA chart                                  | The proposed chart surpasses the performances of the existing adaptive multivariate charts |
| Flury and Quaglino (2018)    | MEWMA chart for asymmetric gamma distributions       | The proposed MEWMA chart outperforms the performance of the conventional $T^2$ chart in all the cases |
| Haq et al. (2020)            | Dual MCUSUM charts with auxiliary information for the process mean | The proposed chart has a better performance compared to the DMCUSUM and MDMCUSUM charts when detecting different sizes of a shift in the process mean vector |

Table 2. The recent development of attribute control charts.

| Sources                      | Proposed scheme                                      | Findings                                                                 |
|------------------------------|------------------------------------------------------|--------------------------------------------------------------------------|
| Yeganeh et al. (2021)        | Combined novel run rules and MEWMA control chart     | The proposed method has better performance for small and moderate shifts in monitoring linear profiles |
| Xie et al. (2021)            | MCUSUM control chart for monitoring Gumbel's bivariate exponential data | The proposed chart outperforms the other charts for most shift domains |
| Mashuri et al. (2020)        | Fuzzy bivariate chart                                 | The proposed chart is more sensitive than the conventional bivariate Poisson chart |
| Zhou et al. (2020)           | Synthetic control chart for attribute inspection     | The proposed chart demonstrates a higher detection performance for small and large mean shifts |
| Quinino et al. (2020)        | Attribute chart for the joint monitoring of mean and variance | The proposed method is easier to be implemented compared to the conventional approach |
| Aldosari et al. (2019)       | Attribute control chart for multivariate Poisson distribution using multiple dependent state repetitive sampling (MDSRS) | The proposed method has a better performance than the conventional one based on repetitive sampling |
| Aslam et al. (2019)          | Shewhart attribute chart with the neutrosophic statistical interval | The proposed attribute control chart has a good ability to detect a shift in the process |
| Chong et al. (2019)          | Multi-attribute CUSUM-np chart                       | The proposed procedure has a better or equal performance compared to the conventional chart |
| Aslam (2019)                 | Attribute control chart using the repetitive sampling under the fuzzy neutrosophic system | The proposed chart with repetitive sampling under the fuzzy neutrosophic system is more sensitive in detecting a shift in the process as compared with the existing chart |
| Lee et al. (2017)            | Multinomial generalized likelihood ratio (MGLR) chart | The proposed chart has better performance than the set of 2-sided Bernoulli CUSUM charts |

ment analysis (FKICA–PCA) to monitor multivariate industrial processes. The nonparametric Revised Spatial Rank Exponential Weighted Moving Average (RSREWMA) control chart is developed to assess the multivariate nonlinear profile data (Pan et al., 2019). Kernel PCA can be applied in monitoring such cases mentioned above by using the control chart approach.

Based on the previous study, the KPCA Mix chart (Ahsan et al., 2020) can be extended to monitor the multivariate nonlinear data. Therefore, this research suggests a mixed multivariate control chart based on the KPCA algorithm that can accommodate the mixed type of quality characteristics with the nonlinear relationship. The estimated PCs Mix from KPCA are then transformed into Hotelling’s $T^2$ statistics. The control limit of $T^2$ statistics is calculated using the kernel density estimation (KDE), the same method used in Ahsan et al. (2020). Moreover, to show the benefits and drawbacks of the proposed chart, its performance is compared with the conventional PCA Mix chart. The rest of this article is arranged as follows: Some related studies are shown in section 2. Section 3 describes the Kernel PCA method. The charting procedures of the proposed KPCA Mix control chart are displayed in section 4. Section 5 presents the performance assessment of the proposed chart in detecting a shift in the process along with the comparison with the PCA Mix chart. The utilization of the proposed chart in simulated and real data is shown in Section 6. Some conclusions and possible future research are presented in Section 7.

2. Related research

The recent studies of the control charts are presented in this section. There are three main categories of control charts discussed in this section such as a multivariate variable chart, attribute chart, and mixed chart. The recent developments in multivariate variable charts are displayed in Table 1. Table 2 shows the recent developments of multivariate attribute charts. Meanwhile, the recent developments in mixed characteristics are presented in Table 3.

Based on the recent development of the mixed control chart, it can be seen that there are a few works that studied the mixed monitoring variable and attribute characteristics. Therefore, more development in this area is needed especially for nonlinear data. This work proposes the mixed control chart based on the Kernel PCA Mix algorithm. The control limit of the $T^2$ statistics from PCs Mix is estimated using the KDE method which has better performance in estimating the non-normal data. The proposed chart is expected to have better performance to monitor the nonlinear mixed data. To show this, the performance of the proposed chart is compared with the conventional PCA Mix chart. Also, the application to the real data is conducted.
3. Kernel PCA

PCA is the basis of transformation to diagonalize the estimated covariance matrix $C$ from input data. PCA was originally proposed for linear data. Therefore, this method is not powerful for nonlinear data. To overcome this nonlinearity problem, Schölkopf et al. (1997) proposed the Kernel PCA scheme.

The basic idea of Kernel PCA is calculating the Principal Component Scores in higher dimensional space by conducting a nonlinear mapping $\Phi : \mathbb{R}^p \rightarrow \mathbb{R}^q$, $y \rightarrow \Psi$ as displayed in Fig. 1. This mapping can be executed by utilizing the kernel functions known from the Support Vector Method (SVM) (Boser et al., 1992).

Assume that the centered data are mapped to feature space $F$, $\Phi(y_1), ..., \Phi(y_n)$. The feature space covariance matrix with a size of $n \times n$ can be written as in Equation (3.1).

$$C^F = \frac{1}{n} \sum_{j=1}^{n} \Phi(y_j) \Phi(y_j)^T. \quad (3.1)$$

The next step is estimating the eigenvalues $\lambda \geq 0$ and eigenvector that satisfies Equation (3.2).

$$\lambda \mathbf{V} = C^F \mathbf{V}. \quad (3.2)$$

In general, the mapping $\Phi(.)$ is not always can be calculated. To solve the problem, the dot product calculation from to vector in feature space is performed. Let $K$ with a size of $n \times n$ defined as $K_{ij} = \langle \Phi(y_i), \Phi(y_j) \rangle$. The Principal component score (PCs) $\mathbf{t}$ is computed using projection of $\Phi(y_i)$ to eigenvector $V_i$, where $v = 1, 2, ..., l$, as expressed in Equation (3.3).

$$\mathbf{t}_v = \langle \mathbf{v}_v, \Phi(y) \rangle = \sum_{i=1}^{n} \mathbf{a}_{i_v} \langle \Phi(y_i), \Phi(y) \rangle. \quad (3.3)$$

To solve the eigenvalue problem and principal component calculation, nonlinear mapping is not needed to be conducted. To replace this, the kernel function can be constructed $K(y_j, y) = \langle \Phi(y_j), \Phi(y) \rangle$.

4. Kernel PCA Mix chart

4.1. Statistics calculation

The main concept of the Kernel PCA Mix chart is to form the $Z$ as a representation of the mixed variable. There are two main steps in the KPCA Mix chart procedure. First, the $T^2$ statistics are computed from matrix $Z$. Second, the control limit calculation is performed by applying the KDE. These procedures are illustrated by the flowchart in Fig. 1. Furthermore, detailed procedures are given as follows:

Statistics $T^2$ calculation

1. Create matrix $Z = [Z_1, Z_2] \text{ sized } n \times (p + m)$ where:
   a. $Z_1$ is the centered version of a matrix $Y_1$ which is contained the variable characteristics (numeric data).
   b. $Z_2$ is the centered version of a matrix $B$ which is contained the dummy from each category in attribute characteristics (categorical data) $Y_2$.
2. Define $N = \frac{1}{2} \mathbf{I}_n$, where $\mathbf{I}_n$ is the identity matrix with the size of $n \times n$.
3. Define $\mathbf{M} = \text{diag}(1, ..., 1, \frac{1}{s_1}, ..., \frac{1}{s_m})$, where the first $p$ columns are specified as by 1 and the last $m$ columns are weighted by $\frac{1}{s}$, for $s = 1, 2, ..., m$.
4. Compute $\mathbf{Z} = N^2 \mathbf{Z} \mathbf{M}^2$.
5. Calculate the matrix kernel $K = K(z_j, z_j) = \langle \Phi(z_j), \Phi(z_j) \rangle$.
6. Calculate Principal Component Scores (PCs) $\mathbf{t}$ using the formula as shown in Equation (4.4).

$$\mathbf{t}_v = \sum_{i=1}^{n} \mathbf{a}_{i_v} \langle \Phi(z_i), \Phi(z_v) \rangle = \sum_{i=1}^{n} \mathbf{a}_{i_v} \bar{K}(z_i, z_v). \quad (4.4)$$

7. From the first $l$ principal component $\mathbf{t}$, calculate the $T^2$ statistics using Equation (4.5).

$$T^2_{\chi} = \sum_{i=1}^{l} \mathbf{t}_v \mathbf{a}_v^{-1} \mathbf{t}_v^T. \quad (4.5)$$

where $v = 1, 2, ..., l$, and $\lambda_v$ eigenvalues that correspond to $v$-th PCs.
4.2. Control limit calculation

The control limit is estimated using the KDE approach due to its ability to follow the unknown distribution of data input. The procedures of control limit calculation are presented as:

1. Estimate the empirical density of \( T^2_k \) statistics using Equation (4.6).

\[
\hat{f}_h(T^2_k) = \frac{1}{nh} \sum_{i=1}^{n} \left( \frac{T^2_k - \bar{T}^2}{\hat{h}} \right).
\]  

(4.6)

2. Calculate \( \tilde{F}_h(T^2_k) = \int_{\hat{T}^2_h}^{T^2_k} f_h(T^2_k) dT^2_k \) using the numerical integration trapezoid rule as in Equation (4.7).

\[
\int_{\hat{T}^2_h}^{T^2_k} f_h(T^2_k) dT^2_k \approx \frac{n}{\hat{T}^2_h} \sum_{i=1}^{n} \left( \hat{T}^2_{h,i} + \frac{\hat{T}^2_h}{2} \right) \hat{f}_h(T^2_k).
\]  

(4.7)

where \( \hat{T}^2_{h,i} \) and \( \hat{T}^2_h \) are the maximum and minimum values of \( T^2_k \).

3. Calculate the control limit using the expression as shown in Equation (4.8).

\[
\bar{CL} = \hat{F}_h^{-1}(I_k)(1 - \alpha).
\]  

(4.8)

5. Performance evaluation

5.1. Simulation set-up

The performance of the proposed control chart is assessed for the variable characteristics (numeric data) which have a nonlinear relationship. The nonlinear data is generated using the following procedures:

1. Generate vector \( y_0 \sim N(0, 1) \) and \( a_0 \sim U(0, 0.1) \).

2. Define five nonlinear variable characteristics as:

\[
y_1^{nl} = a_0 + y_0 \\
y_2^{nl} = 2x_1^{nl} - (y_1^{nl})^2 + 4(y_1^{nl})^3 + y_0 \\
y_3^{nl} = \exp(y_1^{nl}) + y_0 \\
y_4^{nl} = \sin(y_1^{nl}) - 3\sin((y_1^{nl})^2) + y_0 \\
y_5^{nl} = 2(y_1^{nl})^2 - 2\cos((y_1^{nl})^2) + y_0.
\]

The visualizations of those five generated characteristics are presented in Fig. 2.

5.2. Performance evaluation

The number of variable quality characteristics \( Y_i \) (generated from the Multivariate Normal distribution) involved is five. Meanwhile, the number of principal components / evaluated is 2, 3, and 4. The performance is evaluated for three cases, namely the case of attribute characteristics \( Y_2 \) (generated from the Multinomial distribution) with extreme imbalanced, imbalanced, and balanced proportions as defined below:

a. Balanced case with parameter \( \theta_1, \theta_2 = 0.3 \) and \( \theta_3 = 0.4 \)

b. Imbalanced case with parameter \( \theta_1, \theta_2 = 0.1 \) and \( \theta_3 = 0.8 \)

c. Extreme Imbalanced case with parameter \( \theta_1, \theta_2 = 0.05 \) and \( \theta_3 = 0.9 \)

Furthermore, three categories of kernel functions utilized in this research are defined as follows:

a. Linear: \( K(x, x') = \langle x, x' \rangle \).
Table 4. ARLs of an extreme imbalanced case for l = 2.

| Shift | Kernel functions | $\delta_2$ | $\delta_1$ | $\delta_0$ |
|-------|------------------|-------------|-------------|-------------|
| 0     | RBF              | 376.820     | 374.850     | 379.000     |
| 0.1   | 0.0025           | 367.375     | 377.570     | 375.855     |
| 0.2   | 0.0050           | 357.063     | 354.560     | 368.283     |
| 0.3   | 0.0075           | 313.063     | 345.998     | 365.330     |
| 0.4   | 0.0100           | 284.322     | 330.686     | 346.508     |
| 0.5   | 0.0125           | 264.272     | 317.742     | 327.998     |
| 0.6   | 0.0150           | 250.244     | 302.643     | 310.600     |
| 0.7   | 0.0175           | 236.421     | 286.088     | 293.735     |
| 0.8   | 0.0200           | 226.051     | 268.144     | 274.916     |
| 0.9   | 0.0225           | 220.402     | 252.661     | 261.438     |
| 1.0   | 0.0250           | 219.707     | 238.942     | 246.952     |
| 1.1   | 0.0275           | 224.183     | 225.516     | 233.486     |
| 1.2   | 0.0300           | 239.949     | 213.429     | 221.341     |
| 1.3   | 0.0325           | 272.919     | 202.299     | 209.421     |
| 1.4   | 0.0350           | 310.705     | 191.916     | 199.267     |
| 1.5   | 0.0375           | 352.232     | 182.158     | 189.546     |

c. Radial Basis Function (RBF): $K(x_i, x_j) = \exp(-\sigma^2 \|x_i - x_j\|^2)$.

5.2.1. Extreme imbalanced case

The performance of the Kernel PCA Mix chart in handling nonlinear data with an extreme imbalanced proportion of attribute characteristics is tabulated in Tables 4-6. For the small number of the principal component score used, it is seen that the RBF kernel performs poorer compared to the other kernels. Meanwhile, for the larger number of the principal component score used, the RBF kernel displays better results compared to the other functions. Also, for this case, the KDE control limit produces stable ARL$_0$ at about 370.

5.2.2. Imbalanced case

Tables 7-9 show the Kernel PCA Mix chart performance in inspecting the nonlinear for an extreme imbalanced proportion of attribute char-
Similar to the previous results, the control limit produces stable ARL\textsubscript{0} at about 370. For all number of principal component scores used, the RBF kernel has a preferable performance compared to the other functions. It is also known that the linear kernel displays poorer results in this case.

### 5.2.3. Balanced case

Kernel PCA Mix chart performance in assessing the nonlinear data with a balanced proportion of attribute characteristics is displayed in Tables 10-12. Similar to the previous results, the control limit produces consistent ARL\textsubscript{0} at about 370. The RBF kernel performs better compared to the others for all number of principal component scores used. Also, the RBF kernel reaches its peak performance when inspecting the balanced proportion of attribute characteristics. For this case, the Polynomial and Linear kernel functions have similar performance.

### 5.2.4. Comparison with PCA Mix chart

The Kernel PCA Mix performance chart is compared with the performance of the PCA Mix chart in inspecting the nonlinear data. The performance comparisons for extreme imbalanced, imbalanced, and balanced cases are tabulated in Tables 13-15, respectively. Meanwhile, the visualizations of these comparisons are displayed in Figs. 3-5.

### 5.3. Discussion

In this subsection, some discussion about the performance of the proposed chart is provided. First, the best kernel used is the RBF kernel. This happened because the other kernel is developed based on a linear kernel. As we know that the process is generated to follow the nonlinear relationship. The RBF kernel is renowned to have a better performance in inspecting the nonlinear process and under general smoothness assumptions (Zhicheng et al., 2012). Therefore, it makes sense that the RBF kernel performs better in this study.

Table 16 tabulates the summary of the performance comparison between the Kernel PCA Mix chart and PCA Mix chart. In general, both charts yield good performance in detecting the process shift. However, for the specific case, the Kernel PCA Mix chart demonstrates better performance for the small process shift. Meanwhile, the PCA Mix chart has a better performance for a large shift in process. This result indicates that the proposed method is better to be used for nonlinear data with a small shift. This happened because the PCA Mix chart is only developed for the linear process. In contrast, the proposed Kernel PCA Mix chart is developed to overcome the nonlinearity problem so that it has good performance.

### 6. Application to the real data

In this section, the Kernel PCA Mix chart is applied to monitor intrusion in the real dataset. The dataset used is the famous NSL KDD. This research only analyzes 20% of the NSL KDD dataset which can be found at [https://www.unb.ca/cic/datasets/nsl.html](https://www.unb.ca/cic/datasets/nsl.html). The summary of this dataset is displayed in Table 17. From Figs. 6, it is known that the normal connection of the NSL KDD dataset is not normally distributed. The RBF kernel is used in this analysis due to its performance consistency in simulation studies.

Table 18 shows the accuracy rate of the Kernel PCA Mix chart in detecting intrusion in the NSL KDD dataset for several principal component scores. From the results, it is seen that the optimal number of principal components is 4. After finding the optimal number of principal components, this analysis is continued by searching for the optimal value of \( \sigma \). Based on the result in Table 19, it can be known that the optimal value of \( \sigma \) is 0.001. From the detection results, it can be seen that the proposed method has a detection accuracy of about 0.85769. The misdetection happens due to the large value of the FN rate which indicates that more attacks cannot be accurately detected as an attack.

The performance comparison with the other methods is shown in Table 20. The proposed method is compared with several machine learning algorithms (Decision Tree, Naïve Bayes, Logistic Regression, and Support Vector Machine) and control chart method (Hotelling’s \( T^2 \) and PCA Mix chart). According to the table, it is clear that the proposed method has higher accuracy compared to the other machine learning methods and control chart method for the same number of quality characteristics monitored. Also, we can see that the proposed method yields...
Table 13. Performance comparison between KPCA Mix and PCA Mix charts for extreme imbalanced case.

| Shift | $\delta_1$ | $\delta_2$ | PCA Mix | KPCAMix | PCA Mix | KPCAMix | PCA Mix | KPCAMix | PCA Mix |
|-------|------------|------------|---------|---------|---------|---------|---------|---------|---------|
| $p = 5$, $l = 2$ | 376.820 | 383.490 | 370.920 | 376.110 | 362.860 | 385.690 |
| $p = 5$, $l = 3$ | 376.820 | 383.490 | 370.920 | 376.110 | 362.860 | 385.690 |
| $p = 5$, $l = 4$ | 376.820 | 383.490 | 370.920 | 376.110 | 362.860 | 385.690 |

Table 14. Performance comparison between KPCA Mix and PCA Mix charts for imbalanced case.

| Shift | $\delta_1$ | $\delta_2$ | PCA Mix | KPCAMix | PCA Mix | KPCAMix | PCA Mix | KPCAMix |
|-------|------------|------------|---------|---------|---------|---------|---------|---------|
| $p = 5$, $l = 2$ | 386.060 | 360.580 | 374.770 | 372.990 | 371.100 | 381.750 |
| $p = 5$, $l = 3$ | 386.060 | 360.580 | 374.770 | 372.990 | 371.100 | 381.750 |
| $p = 5$, $l = 4$ | 386.060 | 360.580 | 374.770 | 372.990 | 371.100 | 381.750 |

Table 15. Performance comparison between KPCA Mix and PCA Mix charts for balanced case.

| Shift | $\delta_1$ | $\delta_2$ | PCA Mix | KPCAMix | PCA Mix | KPCAMix | PCA Mix | KPCAMix |
|-------|------------|------------|---------|---------|---------|---------|---------|---------|
| $p = 5$, $l = 2$ | 380.770 | 378.110 | 374.770 | 370.220 | 373.580 | 383.910 |
| $p = 5$, $l = 3$ | 380.770 | 378.110 | 374.770 | 370.220 | 373.580 | 383.910 |
| $p = 5$, $l = 4$ | 380.770 | 378.110 | 374.770 | 370.220 | 373.580 | 383.910 |
Fig. 3. ARLs comparison for extreme imbalanced case for: a) $p = 5, l = 2$, b) $p = 5, l = 3$, and c) $p = 5, l = 4$.

Fig. 4. ARLs comparison for imbalanced case: a) $p = 5, l = 2$, b) $p = 5, l = 3$, and c) $p = 5, l = 4$. 
Fig. 5. ARLs comparison for balanced case: a) \( p = 5, l = 2 \), b) \( p = 5, l = 3 \), and c) \( p = 5, l = 4 \).  

Table 16. Summary of performance comparison.  

| Parameter data non-metric | \( l \) | Kernel PCA Mix | PCA Mix |
|---------------------------|-------|----------------|---------|
| \( \theta_1, \theta_2 = 0.3 \) and \( \theta_3 = 0.4 \) | 2     | \*            | \*       |
|                           | 3     | \*            | \*       |
|                           | 4     | \*            | \*       |
| \( \theta_1, \theta_2 = 0.1 \) and \( \theta_3 = 0.8 \) | 2     | \*            | \*       |
|                           | 3     | \*            | \*       |
|                           | 4     | \*            | \*       |
| \( \theta_1, \theta_2 = 0.05 \) and \( \theta_3 = 0.9 \) | 2     | \*            | \*       |
|                           | 3     | \*            | \*       |
|                           | 4     | \*            | \*       |

\* represents better performance for a small shift.  
** represents better performance for a large shift.

Table 17. Summary of NSLKDD 20% dataset.  

| Attack types | 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         |

Fig. 6. NSL-KDD 20% QQ Plot for normal connection.

Table 18. Performance of Kernel PCA Mix Control Chart in monitoring the NSL-KDD dataset for different numbers of principal components.  

| \( l \) | Accuracy  | FP rate  | FN rate |
|--------|-----------|----------|---------|
| 2      | 0.82744   | 0.06751  | 0.29285 |
| 3      | 0.84741   | 0.06714  | 0.25044 |
| 4      | 0.85769   | 0.08305  | 0.21016 |
| 5      | 0.84653   | 0.07361  | 0.24491 |
| 7      | 0.82347   | 0.13183  | 0.22771 |
| 10     | 0.84741   | 0.06714  | 0.25044 |
| 20     | 0.68986   | 0.42724  | 0.17601 |

7. Conclusion and future research  

In this research, the control chart which has the ability in monitoring the mixed variable and attribute characteristics with nonlinear relation-
ships is proposed. The performance of the proposed chart is evaluated for several types of attribute characteristics and several kernel functions. Through simulation studies, it can be seen that the Kernel PCA Mix chart can detect the shift in process. It also can be known that the better kernel function is PolyBF due to its consistency in detecting a shift in process. The comparison with the PCA Mix chart shows that the proposed chart has better performance for a small shift in the process. On the other hand, the PCA Mix chart has better performance for a large shift. This method can be applied in monitoring the process with a nonlinear relationship such as in manufacture and industry, chemical process, biological process, and network anomaly detection. Furthermore, the proposed chart is also applied to monitor the real dataset. The well-known NSL KDD dataset is used as the benchmark for the proposed chart. The monitoring results show that the proposed chart has a good accuracy detection at about 0.85769. Compared to the other methods the proposed demonstrates a better performance by producing higher accuracy and lower false alarms. For future research, the Generative Principal Component Analysis (K. Liu et al., 2020, 2021) can be used in order to improve the performance of the proposed method. Also, the Bayesian-based PCA method (Y. Liu et al., 2018) can be applied for imbalanced cases.

Declarations

Author contribution statement

Muhammad Ahsan: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.
Muhammad Mashuri: Conceived and designed the experiments; Wrote the paper.
Hidayatul Khusna: Analyzed and interpreted the data; Wrote the paper.
Wibawati: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Funding statement

This work was supported by Direktorat Jenderal Pendidikan Tinggi (3/81/KP.PTNBH/2021).

Data availability statement

Data associated with this study is available at https://www.unb.ca/cic/datasets/NSL.html.

Table 19. Performance of Kernel PCA Mix Control Chart in monitoring the NSL-KDD dataset for $l = 4$ and several values of $\sigma$.

| $\sigma$ | Accuracy | FP rate | FN rate |
|----------|----------|---------|---------|
| 0.10000 | 0.58772  | 0.02632 | 0.85429 |
| 0.01000 | 0.84522  | 0.06925 | 0.25385 |
| 0.00100 | **0.85769** | 0.08365 | 0.21016 |
| 0.00050 | 0.84590  | 0.06022 | 0.26160 |
| 0.00010 | 0.63492  | 0.52643 | 0.18027 |
| 0.00001 | 0.53385  | 0.00000 | 1.00000 |

Table 20. Performance comparison with the other methods.

| Method                        | Accuracy | FP rate |
|-------------------------------|----------|---------|
| Hybrid Decision Tree (Farid et al., 2014) | 0.8192 | 0.1740 |
| Hybrid Naive Bayes (Farid et al., 2014) | 0.8239 | 0.1640 |
| Logistic Regression (Belavagi and Muniyal, 2016) | 0.8400 | 0.1700 |
| Support Vector Machine (Belavagi and Muniyal, 2016) | 0.7500 | 0.2400 |
| Hotelling’s $T^2$ chart       | 0.7023  | 0.1433  |
| PCA Mix                       | 0.8041  | 0.3171  |
| **Proposed method**           | **0.8577** | **0.0831** |

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

There is no additional information.

Appendix A. Source code

```r
rm(list=ls(all=TRUE))
alpha=0.00273
uc1t=15942.24

n=1000
n1=70
n2=30
p=5
k=2
q=3
rho=0.3
p0=0.3
sigmaz=0.001
degreez=1
UCl0=NULL
t2pc=NULL
tpcr=NULL
kdhsl1=NULL
uc1l=NULL
shift=0.0025
t2pcr=NULL
RL=NULL
daq=NULL
eig=NULL
mup=NULL

for (ju in 1:2){
t2pc=NULL
tpcr=NULL
Sigma <- matrix(rho,p,p)-diag(rho,p)+diag(1,p)
mu<-rep(0, p)
a1=mvrnorm(n,mu,Sigma)
prob=c(rep(p0,q-1),1-sum(rep(p0,q-1)))
rcat=t(multinom(n, 1,prob))
a=data.frame(cbind(a1,rcat))
kpcac=kpcac(~.,data=a,kernel="rbfdot",kpar=list(sigma=sigmaz),features=k)
#kpcac=kpcac(~.,data=a,kernel="polydot",kpar=list(degree=degreez, scale=1, offset=0),features=k)
#kpcac=-kpcac(~., data=data.frame(a), kernel="tanhdot", kpar=list(scale=1, offset=1),
#features=k)
#pcs=pcv(kpcac)
pcs=pcv(kpcac)
eig=as.vector(kpcac@eig)
eig=rbind(eig,eig)
siglam=diag(eig,k,k)
mup=colMeans(pcs)
mup=rbind(mup,mup)

for (ka in 1:n) {
 t2pcs=pcs[ka,]%*%solve(sigmalam)%*%transpose(pcs[ka,]))
tpcr=rbind(tpcr,t2pcs)
}
```
Sigma <- matrix(rho,p,p)-diag(rho)+diag(1,p)
mue<-rep(0, p)
mu1<-rep(2,p)
a1=mvrnorm(n1,mue,Sigma)
a2=mvrnorm(n2,mue,Sigma)
prob=c(rep(p0,q-1),1-sum(rep(p0,q-1)))
rcat1=t(rmultinom(n1, 1,prob))
rcat2=t(rmultinom(n2, 1,prob))
aa=data.frame(cbind(a1, rcat1))
ab=data.frame(cbind(a2, rcat2))
a0=rbind(aa,ab)

pcs=predict(kpca,a0)
eig1=as.vector(kpca$eig)

mulp=colMeans(rotated(kpca))
mulp=rbind(mulp,mulp)
eig=rbind(eig,eig)
siglam=diag(eig1,k,k)
t2pcs=NULL
t2pctr=rep(0,1)
for (ka in 1:n1+n2) {
   t2pcs[ka]=solve(sigmam)$%*%transpose(pcs[ka,])
}
plot(t2pctr)
lines(t2pctr)
lines(rep(uclt,n1+n2), col = "blue")