The Korean Journal of Applied Statistics
2022, Vol. 35, No. 3, 335–346

DOI: https://doi.org/10.5351/KJAS.2022.35.3.335
Print ISSN1225-066X / Online ISSN 2383-5818

Linear profile monitoring with random covariate

Daeun Kim[a], Sungim Lee[1,a], Johan Lim[b]

[a]Department of Statistics, Dankook University; [b]Department of Statistics, Seoul National University

Abstract

Profile control chart aims to detect a change in the functional relationship of multivariate characteristics in the statistical process control. In monitoring two variables, a linear profile is of interest composed of the intercept and slope of one variable (response variable) against the other (explanatory variable). The previous studies on monitoring of the linear profile mostly assume that the explanatory variables are the same for all profiles. However, there are also cases where they vary depending on profiles. This paper intends to extend the monitoring method to where explanatory variables are different for each profile. We compare the new method's performance through simulation and apply it to monitoring a network intrusion using NSL-KDD data.

Keywords: profile control chart, simple linear profile, intrusion detection, NSL-KDD dataset

1. 서론

대부분의 통계적 공정관리에서 공정이나 제품의 품질은 하나 또는 다수의 품질 특성치의 분포변화를 모니터링함으로써 관리상태인지 이상상태인지 판정할 수 있다. 만약 품질 특성치 간 함수관계가 존재한다면, 시간에 따라 이러한 함수관계가 일정한지를 모니터링함으로써 관리상태 여부를 확인할 수 있는데 이것을 프로파일 모니터링(profile monitoring)이라고 한다. 품질 특성치 간 선형 또는 비선형으로 나타낼 수 있는 함수 관계를 프로파일(profile)이라고 부른다. 이상원인이 발생하면 프로파일이 변화하게 되며 공정의 이상유무는 관리도를 사용하여 모수의 변화를 탐지함으로써 파악될 수 있다. 관리도는 중앙선과 관리한계선으로 구성되어 있으며, 관리 통계량이 관리한계선을 벗어나면 공정에 이상이 발생했다는 것을 의미한다. 일반적으로 관리도를 통한 모니터링은 일단계(Phase I)와 이단계(Phase II)로 나누어진다. 일단계에서는 과거 관측치를 통해 관리상태하의 관리모수를 추정하고 관리한계선을 정하게 된다. 이단계에서는 일단계에서 추정된 관리한계선으로부터 미래 관측치의 관리상태를 모니터링하게 된다.

본 논문에서는 단순 선형 프로파일 관리도를 다루고자 한다. 단순 선형 프로파일 관리도를 통한 모니터링은 관심 있는 두 특성치 간에 단순 선형관계가 존재하고, 공정이 관리상태일 때 선형관계가 일정하게 나타난다는 사실에 기초한다. 단순 선형 프로파일 모니터링과 관련한 대표적인 연구로는 Kang과 Albin (2000), Kim 등 (2003)과 Mahmoud와 Woodall (2004) 등이 있다. 기존의 방법들은 각 프로파일에서 고정된 설명변수를 사용한다. 즉, 설명변수의 관측값이 프로파일마다 일정하다는 가정이 존재한다. 하지만 실제 문제에서 설명변수는 고정이 아니라 랜덤하게 관측되는 경우가 많고, 따라서 프로파일 모니터링을 좀 더 다양한 응용 분야에서 활용하기 위해서는 프로파일마다 랜덤하게 관측되는 설명변수로 관리통계량을 확장할 필요가 있다.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C1003257).

[1] Corresponding author: Department of Statistics, Dankook University, 152, Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do 16890, Korea, Repulic of. E-mail: silee@dankook.ac.kr

Published 30 June 2022 / journal homepage: http://kjas.or.kr

이에 본 연구에서는 랜덤한 설명변수를 갖는 프로파일 관리도를 제안하고 모의실험을 통해 그 성능을 고찰하고자 한다. 또한 실제 자료 분석으로 NSL-KDD 자료(Sklavounos 등, 2008)에 적용하여 Root to Local(R2L) 사이버 공격 탐지 문제에 프로파일 관리도를 적용해 보고자 한다.

본 논문의 요약은 다음과 같다. 2절에서는 설명변수가 고정인 경우에 선형 프로파일 모니터링을 위한 관리도를 소개한다. 3절에서는 2절에서 소개한 방법을 랜덤한 설명변수로 확장한 관리통계량을 제안한다. 4절에서는 모의실험을 통해 각 관리도의 성능을 비교한다. 5절에서는 실제 데이터에 선형 프로파일 관리도를 적용하고, 마지막으로 6절에서는 연구 결과를 요약하고 앞으로의 연구 방향에 대해 고찰해보기로 한다.

2. 기존의 프로파일 관리도

먼저 n개의 데이터로 구성된 j번째 프로파일에 대해 다음과 같은 선형 프로파일을 가정한다.

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right), \ (i = 1, 2, \ldots, n; \ j = 1, 2, \ldots), \tag{2.1}$$

여기서 j번째 프로파일이 관리상태하에 있다면 $\beta_{0j} = \beta_0$이고, $\beta_{1j} = \beta_1$을 만족한다고 가정한다. β_0와 β_1은 관리상태하에서의 절편과 기울기를 나타낸다. 식 (2.1)에서 각 프로파일의 절편과 기울기는 최소제곱법을 통해

$$\hat{\beta}_{1j} = \frac{S_{xy(j)}}{S_{xx(j)}}, \qquad \hat{\beta}_{0j} = \bar{y}_{\cdot j} - \hat{\beta}_{1j}\bar{x}_{\cdot j}$$

로 추정한다. 단, $S_{xx(j)} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_{\cdot j})^2$이고 $S_{xy(j)} = \sum_{i=1}^{n}(y_{ij} - \bar{y}_{\cdot j})(x_{ij} - \bar{x}_{\cdot j})$이다. 본 연구에서는 과거 m개의 프로파일 데이터로부터 관리상태하의 $\beta_0, \beta_1, \sigma^2$값을 추정한 후 새로운 선형 프로파일이 관리상태라고 할 수 있는지 모니터링 하고자 한다.

기존 연구는 모형 (2.1)에서 서로 다른 프로파일 j와 j'에서 설명변수가 모든 $i = 1, \ldots, n$에서 $x_{ij} = x_{ij'}$을 만족한다. 이때 새로운 l번째 프로파일에서 관계의 변화유무는 절편과 기울기에 관한 두 개의 가설 $H_{0,1} : \beta_{0l} = \beta_0$, $H_{0,2} : \beta_{1l} = \beta_1$을 동시에 검정함으로써 알 수 있다. Kim 등 (2003)은 절편과 기울기에 대한 두 개의 Shewhart 관리도를 통해 개별적으로 모니터링하는 방법을 제안하였고, Stover와 Brill (1998), Kang과 Albin (2000)은 절편과 기울기에 대한 이변량 T^2 관리도를 제안하였다. l번째 새로운 프로파일에 대해서도 모형 (2.1)에서 $x_{il} = x_i$ $(i = 1, 2, \ldots, n)$을 가정하면 $\bar{x}_{\cdot j} = \bar{x}$, $S_{xx(j)} = S_{xx}$임이 성립한다.

2.1. 절편과 기울기에 대한 Shewhart 관리도

이 방법은 Kim 등 (2003)이 제안한 것으로 선형 프로파일의 절편과 기울기에 대한 개별 Shewhart 관리도를 작성한다. Shewhart 관리도란 관리통계량 W에 대하여 관리상한선(UCL)과 관리하한선(LCL)을 다음과 같이 결정한다.

$$\text{UCL} = \text{E}(W) + L\sqrt{\text{Var}(W)}, \qquad \text{LCL} = \text{E}(W) - L\sqrt{\text{Var}(W)}.$$

이때 L은 관리한계선의 폭을 결정하며, 관리도의 성능을 결정하는 중요한 관리모수이다. 만약 $W \in (\text{LCL}, \text{UCL})$이면 관리상태하에 있다고 판정한다. 절편과 기울기에 대한 Shewhart 관리도를 작성하는 경우, 두 관리 통계량 중 한 경우라도 관리한계선을 벗어나면 해당 프로파일은 이상상태라고 판정한다. 식 (2.1)에서 $\beta_0, \beta_1, \sigma^2$은 미지의 모수로 일단계 모니터링(phase I monitoring)을 통하여 m개의 관리상태하의 프로파일로부터

$$\bar{\hat{\beta}}_0 = \frac{1}{m}\sum_{j=1}^{m}\hat{\beta}_{0j}, \quad \bar{\hat{\beta}}_1 = \frac{1}{m}\sum_{j=1}^{m}\hat{\beta}_{1j}, \quad \hat{\sigma}^2 = \frac{1}{m}\sum_{j=1}^{m}\text{MSE}_j = \frac{1}{m}\sum_{j=1}^{m}\left\{\frac{\sum_{i=1}^{n}(e_{ij} - \bar{e}_{\cdot j})^2}{n-2}\right\} \tag{2.2}$$

로 추정한다. m개의 프로파일 데이터와, 이와는 서로 독립으로 새롭게 관측된 $l(= m+1, m+2, \ldots)$번째 프로파일에서 추정한 절편 $\hat{\beta}_{0l} = \bar{y}_{\cdot l} - \hat{\beta}_{1l}\bar{x}$과 기울기 $\hat{\beta}_{1l} = S_{xy(l)}/S_{xx}$의 관리상태 유무를 결정하기 위해서 Shewhart 관리도의 관리한계선을 알아 보기로 한다. 두 관리도에 대하여 제1종의 오류를 α로 제어하기 위해 Bonferroni 보정을 사용하면 각 관리한계선은 다음과 같이 추정된다.

$$\left(\mathrm{UCL}_{\beta_0}, \mathrm{LCL}_{\beta_0}\right) = \bar{\hat{\beta}}_0 \pm t_{\frac{\alpha}{4}}(m(n-2)) \sqrt{\frac{m+1}{m}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\hat{\sigma}^2}, \tag{2.3}$$

$$\left(\mathrm{UCL}_{\beta_1}, \mathrm{LCL}_{\beta_1}\right) = \bar{\hat{\beta}}_1 \pm t_{\frac{\alpha}{4}}(m(n-2)) \sqrt{\frac{m+1}{m}\frac{\hat{\sigma}^2}{S_{xx}}}, \tag{2.4}$$

여기서 $t_{\alpha/4}(m(n-2))$은 자유도 $m(n-2)$인 t 분포의 $100(1-\alpha/4)\%$의 분위수 값이다.

2.2. 프로파일 호텔링 T^2 (PT) 관리도

이 방법은 절편과 기울기의 추정량에 대하여 이변량 관리도를 적용한 것이다. Kang과 Albin (2000)은 l ($l = m+1, m+2, \ldots$)번째 프로파일로 부터 추정한 회귀계수 벡터 $\hat{\boldsymbol{\beta}}_l = (\hat{\beta}_{0l}, \hat{\beta}_{1l})^\top$에 대한 이변량 호텔링 T^2 (PT_l) 통계량에 대하여 다음과 같이 정의했다.

$$\mathrm{PT}_l = \left(\hat{\boldsymbol{\beta}}_l - \bar{\hat{\boldsymbol{\beta}}}\right)^\top \hat{\Sigma}_{\hat{\boldsymbol{\beta}}}^{-1} \left(\hat{\boldsymbol{\beta}}_l - \bar{\hat{\boldsymbol{\beta}}}\right), \quad l = m+1, m+2, \ldots \tag{2.5}$$

이때, 관리상태 하의 회귀계수 벡터에 대한 평균 벡터의 추정량 $\bar{\hat{\boldsymbol{\beta}}}$과 공분산행렬의 추정값 $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}$은

$$\bar{\hat{\boldsymbol{\beta}}} = \left(\bar{\hat{\beta}}_0, \bar{\hat{\beta}}_1\right)^\top, \quad \hat{\Sigma}_{\hat{\boldsymbol{\beta}}} = \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} & -\frac{\bar{x}}{S_{xx}} \\ -\frac{\bar{x}}{S_{xx}} & \frac{1}{S_{xx}} \end{pmatrix} \hat{\sigma}^2 \tag{2.6}$$

와 같고, 이때 $\bar{\hat{\beta}}_0, \bar{\hat{\beta}}_1, \hat{\sigma}^2$ 추정량은 식 (2.2)와 같다. 식 (2.5)의 PT 통계량은 Kang과 Albin (2000)에서와 마찬가지로 자유도 2인 카이제곱 분포를 따르며 관리한계선은 다음과 같이 계산된다.

$$\mathrm{UCL}_{\mathrm{PT^{KA}}} = \frac{m+1}{m}\chi_\alpha^2(2). \tag{2.7}$$

이에 반하여 Stover과 Brill (1998)은 회귀계수 벡터의 공분산행렬을 다음과 같이 표본 공분산행렬로 정의하였다.

$$\hat{\Sigma}_{\hat{\boldsymbol{\beta}}} = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{\boldsymbol{\beta}}_j - \bar{\hat{\boldsymbol{\beta}}}\right)\left(\hat{\boldsymbol{\beta}}_j - \bar{\hat{\boldsymbol{\beta}}}\right)^T \tag{2.8}$$

식 (2.5)에서 분산-공분산을 식 (2.8)로 추정한 T^2 ($\mathrm{PT^{SB}}$) 통계량은 관리상태하에서 분모의 자유도 2, 분자의 자유도 $m-2$인 F-분포를 따르며 관리한계선은 다음과 같다.

$$\mathrm{UCL}_{\mathrm{PT^{SB}}} = \frac{2(m+1)(m-1)}{m(m-2)}F_\alpha(2, m-2) \tag{2.9}$$

설명변수가 프로파일마다 고정인 경우에는 식 (2.6)의 분산-공분산 추정량이 식 (2.8)의 추정량보다 이상신호에 대한 탐지성능이 우수하다고 알려져 있다.

3. 랜덤한 설명변수를 갖는 프로파일 관리도

이 절에서는 모형 (2.1)에서 설명변수가 프로파일마다 서로 다른 $x_{ij} \neq x_{ij'}$ $(i = 1, 2, \ldots, n\ j \neq j')$인 경우에 대해 다루기로 한다. 이 경우 프로파일별로 설명변수가 랜덤하게 관측되므로, 2절에서 가정한 $\bar{x}_{\cdot j} = \bar{x}$, $S_{xx(j)} = S_{xx}$가 성립하지 않는다. 또, 이 경우에는 기존 연구에서 다룬 통계량의 분포를 그대로 확장 적용할 수 없기 때문에, 이 절에서는 $x_{ij} \neq x_{ij'} (i = 1, 2, \ldots, n\ j \neq j')$인 경우에 프로파일 모니터링을 위한 관리통계량에 대해 살펴보기로 한다.

3.1. 절편과 기울기에 관한 수정된 Shewhart 관리도

설명변수가 랜덤하게 관측되는 경우에는 관리상태하에서의 새로운 프로파일에 대한 절편 $\hat{\beta}_{0l}$ $(l = m+1, m+2, \ldots)$의 기댓값과 분산에 대하여 m개의 프로파일에서 구한 $\hat{\beta}_{0j}$ $(j = 1, 2, \ldots, m)$으로부터 다음과 같이 추정한다.

$$\hat{\mathrm{E}}\left(\hat{\beta}_{0l}\right) = \frac{1}{m}\sum_{j=1}^{m}\hat{\beta}_{0j} \equiv \bar{\hat{\beta}}_{0\cdot}, \tag{3.1}$$

$$\widehat{\mathrm{Var}}\left(\hat{\beta}_{0l}\right) = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{\beta}_{0j} - \bar{\hat{\beta}}_{0\cdot}\right)^2. \tag{3.2}$$

이를 바탕으로 $\hat{\beta}_{0l}$을 표준화하여 이 값이 근사적으로 표준정규분포를 따른다고 가정한다. 식 (3.1)과 (3.2)에서와 마찬가지로 기울기 $\hat{\beta}_{1l}$에 대한 기대값과 분산에 대한 추정량을 구한 후 표준화한다. 그러므로 절편과 기울기에 대한 관리한계선은

$$\left(\mathrm{UCL}_{\beta_0}, \mathrm{LCL}_{\beta_0}\right) = \bar{\hat{\beta}}_0 \pm z_{\frac{\alpha}{4}}\sqrt{\frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{\beta}_{0j} - \bar{\hat{\beta}}_{0\cdot}\right)^2} \tag{3.3}$$

$$\left(\mathrm{UCL}_{\beta_1}, \mathrm{LCL}_{\beta_1}\right) = \bar{\hat{\beta}}_1 \pm z_{\frac{\alpha}{4}}\sqrt{\frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{\beta}_{1j} - \bar{\hat{\beta}}_{1\cdot}\right)^2} \tag{3.4}$$

이 된다. l번째 프로파일에서 추정한 절편과 기울기 추정량 $(\hat{\beta}_{0l},\ \hat{\beta}_{1l})$에 대하여, $\hat{\beta}_{0l} \notin (\mathrm{UCL}_{\beta_0}, \mathrm{LCL}_{\beta_0})$ 또는 $\hat{\beta}_{1l} \notin (\mathrm{UCL}_{\beta_1}, \mathrm{LCL}_{\beta_1})$이면 l번째 프로파일은 이상상태라고 판정한다.

3.2. 수정된 프로파일 호텔링 T^2 관리도

설명변수가 고정인 경우에는 각 회귀계수의 공분산 행렬에 대한 추정을 식 (2.5)와 같이 할 수 있지만, 설명변수가 랜덤인 경우에는 통계량의 표본분포를 구하는 것이 간단하지 않다. 따라서, 식 (2.8)에서와 같이 회귀계수 벡터에 대한 분산-공분산 행렬을 추정하여 식 (3.5)와 같은 호텔링 $T^2(\mathrm{PT}^{\mathrm{SB}})$ 통계량을 구하도록 한다. 먼저 회귀계수 벡터 $\hat{\boldsymbol{\beta}}_l = (\hat{\beta}_{0l}, \hat{\beta}_{1l})^\top$ 에 대한 분산-공분산 행렬 추정량은

$$S = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{\boldsymbol{\beta}}_j - \bar{\hat{\boldsymbol{\beta}}}\right)\left(\hat{\boldsymbol{\beta}}_j - \bar{\hat{\boldsymbol{\beta}}}\right)^\top$$

로 할 수 있고, 프로파일이 관리상태하에 있다면 다음의 통계량은

$$\mathrm{PT}_l^{\mathrm{SB}} = \left(\hat{\boldsymbol{\beta}}_l - \bar{\hat{\boldsymbol{\beta}}}\right)^\top S^{-1}\left(\hat{\boldsymbol{\beta}}_l - \bar{\hat{\boldsymbol{\beta}}}\right) \sim \frac{2(m+1)(m-1)}{m(m-2)}F(2, m-2) \tag{3.5}$$

을 따르게 되어, 관리한계선은 다음과 같이 추정할 수 있다.

$$\mathrm{UCL}_{\mathrm{PT}^{\mathrm{SB}}} = \frac{2(m+1)(m-1)}{m(m-2)}F_\alpha(2, m-2) \tag{3.6}$$

3.3. 이변량 호텔링 T^2 (BT) 관리도

절편과 기울기에 대한 이변량 관측치 대신에 프로파일 관계를 고려하지 않는 이변량 관측치 $\mathbf{Z}_l = (\bar{x}_{\cdot l}, \bar{y}_{\cdot l})^\top$에 대한 T^2 통계량을 BT라 하면 다음과 같이 나타낼 수 있다.

$$\mathrm{BT}_l = n(\mathbf{Z}_l - \hat{\boldsymbol{\mu}}_{\mathbf{Z}})^\top S_{\mathbf{Z}}^{-1}(\mathbf{Z}_l - \hat{\boldsymbol{\mu}}_{\mathbf{Z}}), \quad l = m+1, m+2, \ldots. \tag{3.7}$$

즉, 새로운 l번째 프로파일에서 랜덤한 설명변수와 반응변수에 대한 관측치의 평균 $\mathbf{Z}_l = (\bar{x}_{\cdot l}, \bar{y}_{\cdot l})^\top$에 대한 평균벡터 $\boldsymbol{\mu}_Z$와 분산-공분산 행렬 $\boldsymbol{\Sigma}_Z$은 m개의 프로파일로부터 다음과 같이 추정한다.

$$\hat{\boldsymbol{\mu}}_Z = (\bar{\bar{x}}, \bar{\bar{y}}), \quad S_{\mathbf{Z}} = \begin{pmatrix} \bar{s}_x^2 & \bar{s}_{xy} \\ \bar{s}_{xy} & \bar{s}_y^2 \end{pmatrix} \tag{3.8}$$

이때

$$\begin{aligned}
\bar{\bar{x}} &= \frac{1}{m}\sum_{j=1}^{m} \bar{x}_{\cdot j}\ ,\ \bar{\bar{y}} = \frac{1}{m}\sum_{j=1}^{m} \bar{y}_{\cdot j}, \\
\bar{s}_x^2 &= \frac{1}{m}\sum_{j=1}^{m}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_{\cdot j}\right)^2\right], \\
\bar{s}_{xy} &= \frac{1}{m}\sum_{j=1}^{m}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_{\cdot j}\right)\left(y_{ij} - \bar{y}_{\cdot j}\right)\right], \\
\bar{s}_y^2 &= \frac{1}{m}\sum_{j=1}^{m}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(y_{ij} - \bar{y}_{\cdot j}\right)^2\right]
\end{aligned}$$

가 되고, 관리한계선은 다음과 같이 추정할 수 있다.

$$\mathrm{UCL}_{\mathrm{BT}} = \frac{2(m+1)(n-1)}{nm - m - 1} F_\alpha(2, nm - m - 1). \tag{3.9}$$

4. 모의실험

이 절에서는 모의실험을 실시하여 3절에서 소개한 랜덤한 설명변수를 갖는 선형 프로파일 관리도의 성능을 비교해 보고자 한다. 모의실험을 위해 식 (4.1)과 같은 관계를 가정한다. 이 관계는 5절의 실제 자료분석에서 소개할 NSL-KDD 데이터로부터 추정한 것으로 정상 네트워크로만 이루어진 두 변수의 선형 프로파일 관계를 나타낸다.

$$y_{ij} = 153.46 + 2.71 x_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{iid}}{\sim} N\left(0, 80^2\right),\ (i = 1, 2, \ldots, n;\ j = 1, 2, \ldots) \tag{4.1}$$

이때 랜덤한 설명변수를 가정하기 위해 x는 평균이 800이고 표준편차가 400인 정규분포를 따른다고 가정한다. 관리도의 성능은 모의실험을 통한 평균 런길이(Average Run Length; ARL)를 이용하여 평가하였으며, 여기서 런길이(run length)란 선형 프로파일 관계에 변화가 생겨 처음으로 관리한계선을 벗어날 때까지 추출한 프로파일의 개수를 의미하고, 반복을 통해 구한 런길이의 평균값을 ARL이라고 정의한다. 즉, Shewhart 관리도의 경우 j번째 프로파일의 기울기(β_{0j})와 절편(β_{1j})이 관리상태라면 $H_0 : \beta_{0j} = \beta_0, \beta_{1j} = \beta_1$라고 나타낼 수 있다. 관리상태하에서 j번째 기울기와 절편의 추정치가 관리한계선 안과 밖에 존재할 확률은 다음과 같다.

$$\begin{aligned}
P\left(\hat{\beta}_{0j} \in \left[\mathrm{LCL}_{\beta_0}, \mathrm{UCL}_{\beta_0}\right], \hat{\beta}_{1j} \in \left[\mathrm{LCL}_{\beta_1}, \mathrm{UCL}_{\beta_1}\right] \,\middle|\, H_0\right) &= 1 - \alpha, \\
P\left(\hat{\beta}_{0j} \notin \left[\mathrm{LCL}_{\beta_0}, \mathrm{UCL}_{\beta_0}\right], \hat{\beta}_{1j} \notin \left[\mathrm{LCL}_{\beta_1}, \mathrm{UCL}_{\beta_1}\right] \,\middle|\, H_0\right) &= \alpha.
\end{aligned}$$

따라서, 관리상태일때 런길이(RL)의 분포는

$$P(\mathrm{RL} = k \mid H_0) = \alpha(1-\alpha)^{k-1}, \quad k = 1, 2, \ldots$$

이 된다. 또, 이상상태하에서 j번째 기울기와 절편의 추정치가 관리한계선 안과 밖에 존재할 확률은

$$P\left(\hat{\beta}_{0j} \in \left[\mathrm{LCL}_{\beta_0}, \mathrm{UCL}_{\beta_0}\right], \hat{\beta}_{1j} \in \left[\mathrm{LCL}_{\beta_1}, \mathrm{UCL}_{\beta_1}\right] \,\middle|\, \text{not } H_0\right) = \beta,$$
$$P\left(\hat{\beta}_{0j} \notin \left[\mathrm{LCL}_{\beta_0}, \mathrm{UCL}_{\beta_0}\right], \hat{\beta}_{1j} \notin \left[LCL_{\beta_1}, \mathrm{UCL}_{\beta_1}\right] \,\middle|\, \text{not } H_0\right) = 1-\beta$$

이고, 런길이(RL)의 분포는

$$P(\mathrm{RL} = k \mid \text{not } H_0) = \beta^{k-1}(1-\beta)$$

이 된다. 따라서, 절편과 기울기에 변화가 없을 때의 평균 런길이를 ARL_0라 하고, 절편 또는 기울기에 변화가 생겼을 때의 평균 런길이를 ARL_1이라 하면 각각은

$$\mathrm{ARL}_0 = E(\mathrm{RL} \mid H_0) = \sum_{k=1}^{\infty} kP(\mathrm{RL} = k) = \frac{1}{\alpha},$$
$$\mathrm{ARL}_1 = E(\mathrm{RL} \mid \text{ not } H_0) = \sum_{k=1}^{\infty} kP(\mathrm{RL} = k) = \frac{1}{1-\beta}$$

로 정의된다. 다른 관리도에서도 평균 런길이는 비슷하게 정의된다. 여기서는 $\alpha = 0.5\%$로 $\mathrm{ARL}_0 = 200$에 관한 관리도를 설정하였다. 구체적인 모의실험 절차는 다음과 같다.

(1) n개의 설명변수 x와 오차항을 랜덤발생한 후, 식 (4.1)의 관계로부터 y를 생성하고 이를 m번 반복한다.

(2) (1)에서 발생한 m개의 프로파일 데이터로부터 이단계 모니터링을 위해 $\mathrm{ARL}_0 = 200$에 관한 관리한계선 (3.3)와 (3.4), (3.6), 그리고 (3.9)을 설정한다.

(3) 새로운 $l(l = m+1, m+2, \ldots)$번째 프로파일에 대하여 $\hat{\boldsymbol{\beta}}_l^{\top} = (\hat{\beta}_{0l}, \hat{\beta}_{1l})$ 통계량을 계산한다.

(4) (3)의 통계량이 (2)에서 설정한 관리한계선을 넘어갈 때까지의 런길이를 기록한다.

(5) (3)−(4)를 10,000번 반복하여 ARL을 계산한다.

먼저 관리상태하에서의 회귀계수 벡터 값이 알려지지 않은 경우를 가정하고 있으므로, 프로파일마다 n개의 데이터로 이루어진 m개의 프로파일 데이터로부터 회귀계수 벡터를 추정하여 사용할 때, ARL_0의 크기가 맞는지 살펴보고자 한다. Table 1은 3절에서 소개한 3가지 관리도에 대하여 n, m의 크기에 따라 관리상태 하의 런길이에 대한 평균런길이(ARL_0)가 미리 설정한 수준으로 관리되는지 살펴 보았다. 또한 런길이의 표준편차 SDRL을 나타낸다. 각 프로파일에 대한 표본크기로 $n = 5, 10, 20, 50, 100$을 설정하고 과거 프로파일의 수는 $m = 20, 50, 100, 200$을 가정하였다.

Table 1에 따르면 프로파일 모니터링 방법의 경우 절편과 기울기에 대한 Shewhart 관리도는 n, m이 작을 때 ARL_0이 잘 맞지 않았으나 n, m이 커질수록 원래 관리도를 설계한 수준인 $\mathrm{ARL}_0 = 200$에 수렴함을 알 수 있다. 이에 반하여 회귀계수 추정량 기반이 아닌 평균값 $(\bar{x}_{\cdot j}, \bar{y}_{\cdot j})$을 사용하는 BT 관리도는 n, m에 상관없이 ARL_0가 200에 가까운 것을 알 수 있다. PT 관리도는 선형 프로파일의 표본 크기가 $n = 10$이어도 m이 충분히 크면 ARL_0가 잘 맞는 모습을 보인다. 따라서 이상상태에서의 ARL_1을 비교하기 위해 $n = 10$인 프로파일 $m = 100$개를 가정하고 관리도의 성능을 비교해 보았다. 이상상태로는 식 (4.1)에서 절편 β_0가 $\beta_0 + \delta\sigma$로 변화하는 경우와 기울기 β_1이 $\beta_1 + \delta\sigma$로 변화하는 경우를 가정했고, 이때의 ARL_1은 Table 2와 Table 3에 각각 주어졌다. 단,

Table 1: ARL_0 (SDRL) for various sizes of n and m

	m	n				
		5	10	20	50	150
Shewhart charts	20	100.7(100.2)	129.1(130.0)	167.9(166.0)	211.1(210.9)	292.0(292.2)
	50	84.8(83.7)	108.2(108.0)	144.9(140.6)	203.2(200.8)	224.8(227.0)
	100	60.8(59.5)	85.7(85.8)	101.3(100.8)	134.1(133.1)	213.7(215.8)
	200	57.5(57.2)	67.8(66.9)	87.6(87.6)	88.6(87.6)	202.6(201.5)
PT^{SB} chart	20	30.1(28.8)	95.3(93.9)	104.9(106.2)	169.9(167.3)	187.3(185.6)
	50	37.5(37.4)	155.5(156.8)	174.1(159.7)	184.7(190.2)	193.0(178.0)
	100	43.9(42.8)	177.0(175.4)	184.4(178.9)	191.9(189.4)	202.1(199.4)
	200	54.4(54.0)	182.0(186.2)	192.1(190.9)	206.0(211.2)	208.7(204.0)
BT chart	20	248.3(243.5)	243.3(240.6)	230.5(229.9)	218.8(219.0)	214.3(212.2)
	50	233.2(233.0)	230.9(224.1)	221.7(211.7)	208.9(206.7)	201.8(200.2)
	100	214.3(209.7)	211.3(209.7)	210.3(203.1)	205.3(199.1)	193.6(192.1)
	200	201.9(191.8)	193.2(194.2)	185.0(184.3)	180.7(176.0)	177.2(176.8)

Table 2: ARL_1 when β_0 changes to $\beta_0 + \delta\sigma$ in case $n = 10$ and $m = 100$

	$\delta\sigma$										
	0	8	16	24	32	40	48	56	64	72	80
Shewhart charts	199.3	171.2	150.4	132.0	113.8	95.9	76.2	51.3	31.9	23.4	12.7
PT^{SB}	199.4	163.9	118.2	73.0	45.6	25.2	14.5	8.3	5.8	2.8	2.3
BT	199.5	149.5	96.3	49.8	26.7	15.5	9.4	5.7	3.8	2.6	2.0

Table 3: ARL_1 when β_1 changes to $\beta_1 + \delta\sigma$ in case $n = 10$ and $m = 100$

	$\delta\sigma \times 10^{-2}$										
	0	0.8	1.6	2.4	3.2	4.0	4.8	5.6	6.4	7.2	8.0
Shewhart charts	199.3	189.5	174.6	169.4	146.5	132.8	105.1	91.4	60.1	42.9	23.6
PT^{SB}	199.4	171.7	128.4	85.5	53.8	34.5	19.9	12.7	8.6	6.0	4.2
BT	199.5	165.3	113.3	72.4	42.4	27.4	16.6	9.6	6.9	4.7	3.7

$n = 10$, $m = 100$일 때 BT 관리도의 ARL_0는 200에 가까운 반면 Shewhart 관리도와 PT^{SB} 관리도의 ARL_0가 작아 이상상태하에서의 평균 런길이를 공정하게 비교하기 위해, $\delta\sigma = 0$일때 이들 두 관리도의 관리한계선을 수정하여 Table 2와 Table 3에서 보는 것처럼 관리상태일 때 세 관리도의 ARL_0을 비슷한 수준으로 정하였다. 모의실험 결과 이상상태에서 Shewhart 관리도, PT^{SB} 관리도, 그리고 BT 관리도의 순으로 ARL_1이 더 짧아져, 이변량 호텔링 T^2 (BT) 관리도가 프로파일 호텔링 T^2 (PT^{SB}) 관리도보다 변화의 탐지가 빠른 것으로 보인다. 하지만 위 모의실험의 설계에서 단순선형 프로파일은 회귀계수의 추정치가 변화하면 평균값 $(\bar{x}_{\cdot l}, \bar{y}_{\cdot l})$ 또한 같이 변화할 가능성이 높은 설정으로 Table 2와 3은 PT^{SB} 관리도의 효과가 뚜렷하지 않음을 알 수 있었다. 따라서 식 (4.1)과 같이 이변량 정규분포를 따르는 $\{(x_{il}, y_{il}), i = 1, \ldots, n\}$을 추출하여 $(\bar{x}_{\cdot l}, \bar{y}_{\cdot l})$값은 일정하면서 회귀계수 벡터의 관계만 변화하는 모의실험을 실시하였다.

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim BN\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right). \tag{4.2}$$

식 (4.2)는 식 (4.1)의 설정과 동일하게 $\mu_x = 800$, $\mu_y = 2321.46$, $\sigma_x = 400$ 그리고 $\sigma_y = 1086.948$으로 설정하였

Table 4: ARL_1 in the case of changing the relationship of the regression coefficient vector by changing only the correlation coefficient with $n = 10$ and $m = 100$

	ρ								
	0.990	0.989	0.988	0.987	0.986	0.985	0.984	0.983	0.982
PT^{SB}	229.1	165.1	116.0	97.1	77.4	62.8	52.2	45.2	34.8
BT	228.6	175.3	133.2	107.3	80.1	67.7	56.2	49.4	41.0

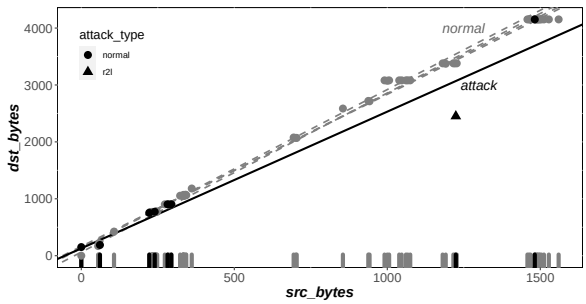
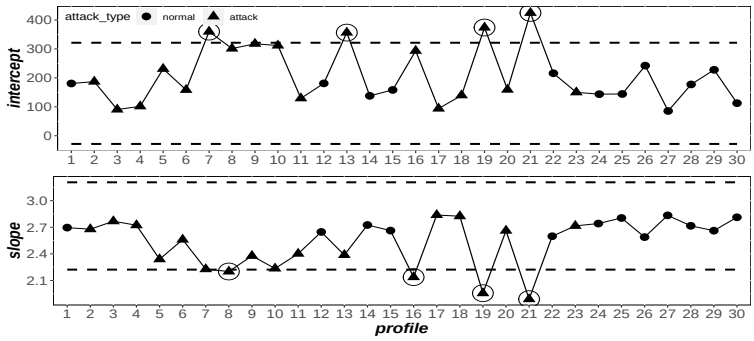


Figure 1: *Simple linear profiles with normal and attack networks in the NSL-KDD dataset.*

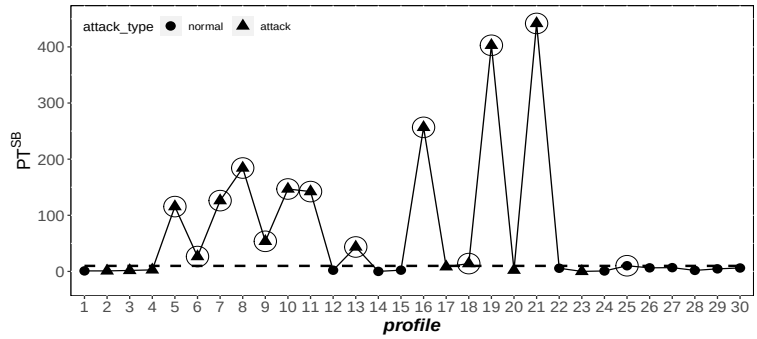
다. 이제 평균벡터는 고정하고 ρ값을 변화하면서 구한 ARL_1은 Table 4와 같다.
프로파일 호텔링 T^2의 ARL_1이 이변량 호텔링 T^2보다 짧다. 이는 관계 변화의 탐지가 프로파일 호텔링 T^2가 더 빠른 것을 의미하므로 평균값은 변화가 없고 관계만 변화할 때 프로파일 모니터링이 더 효과적임을 알 수 있다.

5. 실제 자료 분석

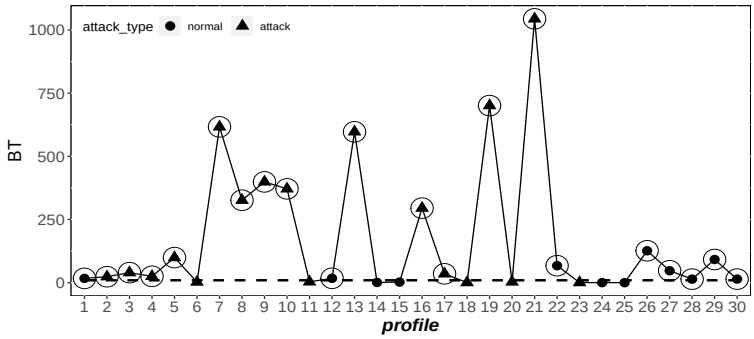
이 절에는 네트워크 침입 탐지와 관련하여 프로파일 모니터링 방법을 적용해 보기로 한다. 본 논문에서 사용한 데이터는 NSL-KDD 데이터로 침입 탐지시스템을 위해 공개된 대표적인 데이터로 학습 데이터와 평가 데이터로 구성되어 있으며, 41개의 특징과 정상 또는 4가지 공격 유형(DoS, Probe, R2L, U2R)을 나타내는 속성을 포함하고 있다. 본 절에서는 R2L 공격에 한정하여 프로파일 관리도를 적용하고자 한다. 이 데이터의 41개 속성 중 `source bytes`와 `destination bytes`는 주어진 시간 내 컴퓨터 네트워크의 송수신 포트 간 이동하는 데이터의 양을 의미하는데, FTP 서버를 이용할 때 이들 사이에는 선형 관계가 존재하기 때문이다. Figure 1은 $n = 8$인 선형 프로파일을 구성한 것으로 `source bytes`를 설명변수로, `destination bytes`를 반응변수로 하여 나타낸 것이다. 한 프로파일내에서 $n = 8$개가 모두 정상 네트워크인 경우(회색)와 프로파일 내 하나의 공격데이터가 포함된 경우(검정색)의 선형 프로파일은 기울기와 절편이 서로 다르게 나타남을 알 수 있다. 특히 이 데이터의 경우 설명변수의 관측값이 프로파일마다 일정하지 않음을 알 수 있다. 먼저 관리상태하에서의 선형 프로파일 관계를 추정하기 위해 모두 정상인 네트워크로 이루어진 $n = 8$의 $m = 100$개의 프로파일을 구성하였고, 새로운 프로파일에 대한 네트워크 공격 탐지 성능을 알아보기 위해 정상 네트워크 데이터로 구성된 12개의 프로파일과, 공격 네트워크를 포함하고 있는 18개의 프로파일을 구성하였다. 이제 3절에서 소개한 Shewhart 관리도, PT^{SB} 관리도, 그리고 BT 관리도를 적용한 결과를 살펴보면 Figure 2와 같다. 앞서 소개한 것처럼 NSL-KDD 데이터에는 네트워크 공격 탐지 알고리즘 개발을 위해 모의로 생성한 데이터로 실제 공격유무가 알려져 있기 때문에, Figure 2에서 각 프로파일마다 정상이면 '●', 공격 네트워크를 포함한 프로파일은 '▲'으로 표시하였다. 정오분류표를 이용해 탐지 성능을 요약하면 Table 5와 같다. PT관리도가 특



(a) Shewhart charts



(b) PT^{SB} chart



(c) BT chart

Figure 2: *Phase II monitoring results. Points outside the control limits are indicated by circles.*

이도가 91.7%, 민감도는 66.7%로 나타났으며 정확도는 76.7%로 다른 방법들에 비해 네트워크 침입탐지에 유용한 결과를 보였다.

한편, NSL-KDD 데이터는 기존의 21번의 적합모형으로부터 각 네트워크 별 공격여부를 올바르게 탐지한 횟수를 함께 제공한다. 실제 자료 분석에서 사용한 18개의 공격 프로파일에는 1개 또는 2개의 공격 데이터가 존재하는데, Table 6에서는 공격별로 21개의 모형 중 공격을 탐지한 횟수를 요약해 나타내었다. 예를 들어, 4번 프로파일의 경우 1번의 공격데이터가 존재하는데, 이 공격은 21개의 모형 중 오직 2개의 모형만이 이 공격을

Table 5: Performance of various control charts

Chart	Accuracy(%)	Sensitivity(%)	Specificity(%)
Shewhart	60.0	33.3	100.0
PT^{SB}	76.7	66.7	91.7
BT	56.7	72.2	33.3

Table 6: Performance of NSL-KDD dataset. N indicates that the data in that profile are all normal networks

Profile no.	1	2	3	4	5	6	7	8	9	10
	N	2/21 14/21	7/21	2/21	6/21 10/21	2/21	11/21 3/21	14/21 7/21	15/21	7/21 11/21
Profile no.	11	12	13	14	15	16	17	18	19	20
	8/21	N	5/21	N	N	10/21 4/21	11/21 7/21	13/21 4/21	6/21	13/21
Profile no.	21	22	23	24	25	26	27	28	29	30
	4/21	N	13/21	N	N	N	N	N	N	N

탐지를 했다는 것을 나타낸다. 이와 같이 실제 자료분석에서 사용된 R2L 공격은 평균 38.3%의 탐지능력을 보인 것으로, 비교적 공격 탐지가 쉽지 않음을 알 수 있다. 그런데, 선형 프로파일을 구성하여 공격을 탐지한 PT 관리도의 경우 민감도가 66.7%로 높게 나타나 프로파일을 활용한 모니터링이 유용함을 알 수 있다. BT 관리도의 경우 민감도는 72.2%로 나타났지만, 특이도가 낮아 상대적으로 관리성능이 떨어지는 것을 확인할 수 있다.

6. 결론 및 토의

본 논문에서는 선형 프로파일의 설명변수가 고정이 아닌 랜덤한 경우로 확장하여 적용할 수 있는 프로파일 관리도에 대해 소개하고, 모의실험을 통해 그 타당성을 살펴보았다. 또한 실제 데이터 분석에 적용해 보았다. 설명변수가 고정인 경우 선형 프로파일의 절편과 기울기에 대한 이변량 T^2 관리도에 대하여, Kang과 Albin (2000)의 통계량이 더 관리성능이 좋음이 알려져 있지만, 랜덤한 설명변수에 대해서는 Stover와 Brill (1998)의 방법을 사용하는 것이 유용함을 알 수 있었다.

Sklavounos 등 (2008)은 `source bytes`변수만 이용하여 관리도를 적용하였는데, 이를 위해 값에 대한 제약을 가정하였다. 그러나 프로파일 관리도에서는 이러한 제약조건을 사용하지 않아 공격탐지를 위해 좀 더 유용하게 사용될 수 있다. 다만, 프로파일 관리도 작성을 위해서는 n개의 관측치를 하나의 프로파일로 구성해 사용하므로 공격 탐지에 대한 시차가 늦어 질 수 있다는 단점이 있다. 이 연구는 네트워크 침입 데이터를 기반으로 프로파일 관리도 방법의 적용 가능성을 확인한 것으로, 추후 설명변수의 분포에 대한 가정 및 특성치 간 프로파일 관계에 대한 추가 연구가 필요할 것이다.

References

Kang L and Albin SL (2000). On-line monitoring when the process yields a linear profile, *Journal of Quality Technology*, **32**, 418–426.

Kim K, Mahmoud MA, and Woodall WH(2003). On the monitoring of linear profiles, *Journal of Quality Technology*, **35**, 317–328.

Mahmoud MA and Woodall WH (2004). Phase I analysis of linear profiles with calibration applications, *Technometrics*, **46**(4), 380–391.

Sklavounos D, Edoh A, and Plytas M (2008). Statistical process control method for cyber intrusion detection(DDoS, U2R, R2L, Probe), *International Journal of Cyber-Security and Digital Forencisc*, **8**, 82–88.

Stover FS and Brill RV (1998). Statistical quality control applied to ion chromatography calibrations, *Journal of Chromatography A*, **804**, 37–43.

Received January 3, 2022; Revised February 14, 2022; Accepted February 19, 2022

설명변수가 랜덤인 선형 프로파일 연구

김다은[a], 이성임[1,a], 임요한[b]

[a]단국대학교 응용통계학과, [b]서울대학교 통계학과

요 약

통계적 공정관리에서 프로파일 관리도란 다수의 품질 특성치 간 함수관계의 변화를 탐지하는 것을 말한다. 두 변수 간 선형의 관계가 있는 경우, 선형 프로파일을 가정하고 절편과 기울기가 일정한지 모니터링한다. 이때 선형 프로파일에 관한 대부분의 기존 연구에서는 모든 프로파일에서 설명변수의 관측치가 동일하다고 가정한다. 그러나 프로파일마다 설명변수의 값이 랜덤하게 관측되는 경우도 존재한다. 본 논문에서는 단순 선형 프로파일 모니터링에서 설명변수가 프로파일마다 랜덤하게 관측된다는 가정하에 기존의 방법을 확장 적용하고자 한다. 모의실험을 통해 제안한 방법의 탐지 성능을 확인하고 네트워크 침입 탐지 알고리즘 성능을 비교하기 위한 NSL-KDD 데이터를 이용하여 제안된 침입 탐지 결과를 비교해 보았다.

주요용어: 프로파일 모니터링, 단순선형 프로파일, 침입 탐지, NSL-KDD 데이터

이 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 진행된 논문입니다 (No.2019R1A2C1003257).

[1]교신저자: (16890) 경기도 용인시 수지구 죽전로 153, 단국대학교 응용통계학과. E-mail: silee@dankook.ac.kr