https://doi.org/10.7840/kics.2022.47.9.1310

오토인코더 모델의 은닉층 정보를 활용한 네트워크 이상탐지 시스템

김 미 르•, 계 효 선*, 권 민 혜°

Network Anomaly Detection System Using Hidden Layer Information of Autoencoder

Miru Kim•, Hyoseon Kye*, Minhae Kwon°

요 약

최근 네트워크 사용량 증가에 따른 사이버 공격 건 수 증가로 네트워크 침입 탐지 시스템(Network Intrusion Detection System; NIDS)의 중요성이 더욱 강조되고 있다. NIDS 구축을 위한 기술적 방안으로 딥러닝 기반의 연구가 활발히 진행되고 있는데, 특히 여러 딥러닝 구조 중 오토인코더(Autoencoder)가 가장 대표적으로 활용되고 있다. 대부분의 오토인코더 기반 연구들은 오토인코더 모델의 입력층과 출력층만의 정보를 활용하여 침입 여부를 판단한다. 즉, 은닉층의 정보는 활용하지 않는다는 한계점을 가지고 있다. 이러한 제한적 정보의 활용은 침입 탐지 성능의 한계를 가져오기에 더욱 정밀한 침입 탐지를 위해서는 은닉층 정보를 반드시 고려해야 한다. 본 논문에서는 이와 같은 한계점을 극복하기 위하여 오토인코더 모델의 은닉층 정보를 활용하여 더욱 높은 탐지 성능을 가져오는 방안을 제안한다. 제안하는 시스템의 성능 검증을 위해 네트워크 이상탐지 기술 평가에 널리 사용되는 두 가지 네트워크 데이터 셋을 활용하였다. 모의실험 결과 제안하는 시스템이 모든 평가항목에서 기존 연구 결과들 대비 우수한 성능을 보이는 것을 정량적으로 확인하였다. 특히 Accuracy 항목과 F1-score 항목에서 기존 방식은 평균 80%의 성능을 보이는 것에 비해, 제안 방식은 평균 98%로 기존 방식 대비 확연히 높은 성능을 가지는 것을 확인하였다.

Key Words : Network Intrusion Detection System, Anomaly Detection, Dimensionality Reduction, Autoencoder

ABSTRACT

As Internet usage has been increasing, the importance of network intrusion detection systems (NIDS) has been highlighted. A promising solution for the NIDS is an autoencoder, a type of deep learning model. The conventional autoencoder uses only the input and output layers to detect intrusion, which draws a limitation. In this case, the information embedded in hidden layers would be ignored. The hidden layers of the autoencoder should be included in the detection process since they have information about the data. In order to overcome such limitations, we propose a novel anomaly detection solution that utilizes not only the input and output layers of the autoencoder but also hidden layers of the autoencoder to improve the detection performance. To

※ 본 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단(NRF-2020R1F1A1069182)과 정보통신기획평가원(IITP- 2021-0-00739)의 지원을 받아 수행된 연구임.
• First Author : Soongsil University School of Electronic Engineering, 48rlaalfm@soongsil.ac.kr, 학생회원
° Corresponding Author : Soongsil University School of Electronic Engineering, minhae@ssu.ac.kr, 종신회원
* Soongsil University School of Electronic Engineering, ksh8301000@soongsil.ac.kr, 학생회원
논문번호 : 202201-001-B-RN, Received December 30, 2021; Revised March 28, 2022; Accepted June 20, 2022

evaluate the detection performance of the proposed solution, we use two popular network intrusion data sets and compare our solution with existing state-of-the-art methods. As a result, we confirm that the proposed solution outperforms other comparison methods. Specifically, our solution shows as high as 98% Accuracy and F1-score on average, while the comparison method shows 80% Accuracy and F1-score on average.

Ⅰ. 서 론

최근 네트워크 기술의 급진적인 발전과 감염병 유행으로 비대면 문화가 확산함에 따라 개인, 기관, 기업 등 사회 전반에서 온라인 환경을 적극적으로 사용하고 있다. 이러한 온라인 환경 활용의 활성화로 인하여 네트워크상에서 공유되는 중요 개인 정보들의 양이 증가하였고, 이를 불법적으로 탈취하려는 네트워크 침입 시도 또한 급격하게 증가하고 있다. 이 문제를 해결하고 안전한 네트워크 환경을 제공하기 위해서 사이버 공격자의 침입을 빠르게 탐지하여 대응할 수 있는 네트워크 침입 탐지 시스템에 관한 연구가 큰 관심을 받고 있다. 특히, 인공지능 기술이 점점 발전함에 따라 이를 사용하는 네트워크 침입 탐지 시스템에 관한 연구가 활발히 진행되고 있다.

네트워크 침입 탐지 시스템은 침입 분석 기법에 따라 오용탐지와 이상탐지 방식으로 구분할 수 있다. 오용탐지는 과거의 네트워크 침입 데이터들을 바탕으로 침입을 탐지하는 방법으로, 새롭게 발생한 네트워크 트래픽이 이미 알려진 침입 패턴과 일치한다면 이를 침입으로 판단한다[1,2]. 그러므로 오용탐지는 이미 알려진 방식으로 공격자가 침입을 시도하는 경우라면 높은 정확도로 탐지할 수 있다. 하지만 새로운 방식으로 진화된 침입 시도가 생기는 경우에는 사전에 알려진 침입 패턴과 상이하기에 이를 탐지하지 못한다는 단점이 있다. 반면, 이상탐지는 침입 데이터가 아닌 정상적인 네트워크 트래픽 데이터를 이용하여 정상상태의 패턴을 학습하고, 새로운 네트워크 트래픽 패턴이 정상상태의 패턴과 상이하다면 침입으로 판단하는 방식이다. 따라서 이상탐지는 침입 패턴이 파악되지 않은 공격에 대해서도 정상적인 패턴과 비교를 통해 탐지가 가능하므로, 새로운 유형의 침입 방식이 지속해서 발생하는 환경에서 사용하기에 더욱 적합하다.

이상탐지에 활용되는 대표적인 기술로 stacked autoencoder (SAE) 모델이 있다. SAE 모델은 심층신경망 구조를 사용하는 딥러닝 모델 중 하나로, 다수의 은닉층을 통해 입력 데이터를 축소한 뒤, 다시 복원하는 구조로 되어 있다. 하지만 기존의 SAE 모델을 활용한 이상탐지는 은닉층의 정보를 활용하지 않고 입력과 출력층의 정보만을 사용한다는 한계점을 가지고 있다. 이러한 제한적인 정보의 활용은 탐지 성능의 한계를 가져오기에, 더욱 정밀한 탐지를 위해서는 은닉층의 정보의 고려가 반드시 필요하다. 이에 본 논문에서는 SAE 모델의 입력층과 출력층 정보와 더불어, 은닉층의 정보를 활용하는 이상탐지 시스템을 제안한다.

본 논문은 다음과 같이 구성되어 있다. Ⅱ장에서는 본 논문의 배경이 되는 선행연구를 소개하고, Ⅲ장에서는 SAE 모델의 은닉층의 정보를 활용한 이상탐지 방법에 대해서 제안한다. Ⅳ장에서는 실험을 통해 제안하는 방법의 성능을 평가하고, 기존 방법의 성능과 비교 및 분석을 진행한다. Ⅴ장에서는 제안하는 오토인코더의 은닉층의 정보를 활용한 네트워크 이상탐지 시스템의 연구에 대한 결론을 맺는다.

Ⅱ. 선행 연구

네트워크 트래픽 데이터는 다양한 feature를 포함하는 고차원의 데이터이다[3]. 이러한 고차원의 데이터를 직접적으로 침입 탐지에 활용하기에는 높은 계산 복잡도를 요구하기에 실용적인 시스템 구축에 어려움이 있다. 또한, 네트워크 데이터의 feature 간 중복성이 많이 존재하기에 네트워크 침입 탐지를 하는데 데이터를 저차원으로 표현하는 차원 축소 기술을 적극적으로 활용하고 있다. 대표적인 차원 축소 방식으로는 주성분 분석 (principal component analysis; PCA)과 오토인코더 (Autoencoder; AE) 모델이 있다.

PCA는 대표적인 선형 차원 축소 알고리즘으로, 이를 사용한 이상탐지 시스템 연구가 활발히 진행되었다[4-6]. PCA는 데이터의 주성분 축을 사용하여 데이터의 손실을 최소화하면서 선형적으로 차원을 축소하는 기법이다. 하지만 PCA를 통한 선형 차원 축소는 비선형성을 지니는 네트워크 데이터의 패턴을 잘 나타내기 어렵다는 한계점이 존재한다. 이에 비선형 차원 축소 기법인 AE 모델을 활용한 이상탐지 방식이 제안되었으며, 이는 PCA보다 더 높은 성능을 보인다[7].

AE 모델은 딥러닝 방식을 적용한 비지도 학습 기법의 하나로 인코더와 디코더로 구성되어 있다[8-11]. 인코더는 입력 데이터의 차원을 비선형 방식으로 축

소하며, 디코더는 축소된 데이터를 입력 데이터와 동일한 차원으로 복원한다. AE 모델 기반 이상탐지는 정상 데이터를 이용하여 인코더의 입력값인 원본 데이터와 디코더의 출력값인 복원 데이터의 오차인 복원 오차를 최소화하는 방향으로 학습을 하게 되며, 이 과정에서 정상 데이터의 특징이 AE 모델에 학습된다. 따라서 학습을 마친 AE 모델은 정상 데이터에 대하여 작은 복원 오차를 가지고 비정상 데이터에 대하여 큰 복원 오차를 가진다. 이때, 복원 오차의 크기를 anomaly score로 정의하여 새로운 데이터의 anomaly score가 특정 임계값 보다 크면 비정상으로 판단한다.

AE 모델은 하나의 은닉층을 가지고 있다. 이처럼 하나의 은닉층만 사용하여 차원 축소를 진행할 경우, 한 번에 많은 차원을 축소하므로 데이터의 고유 특성 정보가 크게 손실될 수 있는 문제점이 존재한다. 이러한 문제점을 보완하기 위해 복수의 은닉층을 통해 여러 단계에 걸쳐 데이터의 차원을 축소하고 복원하는 SAE 모델이 제안되었다[11]. [11]의 연구 결과에 따르면 하나의 은닉층을 가진 AE 모델 보다 복수의 은닉층을 가진 SAE 모델을 기반으로 한 이상탐지 방식이 더 높은 성능을 보인다. 하지만 SAE 모델 기반의 이상탐지는 구조적으로 여러 은닉층을 가짐에도 불구하고, 입력층과 출력층의 출력 결과만을 사용하여 anomaly score를 측정하므로 학습된 모델의 정보를 모두 반영하지 못하는 한계점이 존재한다. 이러한 한계점을 보완하기 위해 SAE 모델의 은닉층의 정보를 활용하는 방법에 관한 연구가 진행되었으며, 기존의 SAE 모델을 사용한 이상탐지 방식에 대비하여 성능이 향상됨을 입증했다[12-14]. [12]의 연구는 입력 데이터에 대한 은닉층의 출력과 복원 데이터에 대한 은닉층의 출력의 오차인 '층별 복원 오차'를 은닉층의 정보로 활용하였다. [12]의 연구에서는 이러한 은닉층의 정보를 활용하여 MNIST, F-MNIST와 같은 이미지 데이터에 적합한 anomaly score를 구하기 위한 distance measure를 제안하였다. [13]의 연구는 네트워크 데이터 셋에 [12]의 연구에서 제안한 distance measure를 anomaly score 측정에 단순 적용하여 이상탐지를 진행하였다. 하지만 네트워크 데이터 셋은 이미지 데이터 셋과 다른 특성을 가지고 있으므로 네트워크 데이터 셋에 적합한 anomaly score를 도출하기 위한 연구가 필요하다. 이에 본 논문에서는 SAE 모델의 은닉층의 정보를 활용하는 네트워크 이상탐지 시스템을 위한 anomaly score 측정 방식을 제안하고자 한다.

Ⅲ. 제안하는 이상탐지 시스템

본 장에서는 제안하는 네트워크 이상탐지 시스템에 관하여 서술한다. 먼저 이상탐지 시스템에서 네트워크 데이터를 사용하기 위한 전처리 과정을 서술한 뒤, SAE 모델의 학습에 관하여 서술한다. 이후 학습된 SAE 모델의 은닉층의 정보를 활용하는 이상탐지 방식에 관해 설명한다.

3.1 데이터 전처리

네트워크 데이터는 다수의 feature로 구성되어 있다. 각 feature는 서로 다른 정보를 담고 있기 때문에 각각 다른 값의 범위(스케일)를 가진다. 이처럼 다른 스케일의 feature가 포함된 데이터를 전처리하지 않고 모델을 학습하게 되면, 학습된 모델이 큰 값을 가진 feature에 편향되어 생성될 수 있다는 문제점이 존재한다. 따라서 feature 별의 범위를 통일하기 위한 전처리 과정이 필요하다. 본 절에서는 네트워크 데이터 셋의 정의 및 전처리 방식에 대하여 서술한다.

네트워크 데이터 셋 $\boldsymbol{X}\in\mathbb{R}^{N\times M}$은 M개의 feature를 가진 N개의 데이터 샘플 x_n으로 구성되며 수식 (1)과 같이 정의한다.

$$\boldsymbol{X}=\left[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n,\cdots,\boldsymbol{x}_N\right]^T,\boldsymbol{x}_n\in\mathbb{R}^{M\times 1} \tag{1}$$

데이터 셋 $\boldsymbol{X}$는 정상 데이터와 네트워크 공격을 의미하는 비정상 데이터로 구성되어 있다. 데이터 셋 $\boldsymbol{X}$ 중 정상 데이터의 일부를 학습용 정상 데이터 셋으로 사용하며 이를 전처리의 기준으로 사용한다. 학습용 정상 데이터 셋 $\boldsymbol{X}_{train}\in\mathbb{R}^{N_{train}\times M}$ 은 N_{train}개의 데이터 샘플로 구성되며 수식 (2)와 같이 정의한다.

$$\boldsymbol{X}_{train}=\left[\boldsymbol{x}_{1_{train}},\cdots,\boldsymbol{x}_{n_{train}},\cdots,\boldsymbol{x}_{N_{train}}\right]^T,\boldsymbol{x}_{n_{train}}\in\mathbb{R}^{M\times 1} \tag{2}$$

수식 (2)에서 $\boldsymbol{x}_{n_{train}}$은 M개의 feature를 가진 n_{train}번째 데이터 샘플로 정의된다. 전처리 방식은 학습용 정상 데이터 셋 $\boldsymbol{X}_{train}$의 feature 별 평균 $\boldsymbol{\mu}\in\mathbb{R}^{M\times 1}$과 표준편차 $\boldsymbol{\sigma}\in\mathbb{R}^{M\times 1}$을 기반으로 standard scaler를 사용하며 수식 (3)과 같이 정의한다. 전처리 과정을 통과한 데이터 셋 $\overline{\boldsymbol{X}}\in\mathbb{R}^{N\times M}$은 수식 (4)와 같이 나타낼 수 있다.

$$\overline{\boldsymbol{x}}_n = \frac{\boldsymbol{x}_n - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \tag{3}$$

$$\overline{\boldsymbol{X}} = \left[\overline{\boldsymbol{x}}_1, \cdots, \overline{\boldsymbol{x}}_n, \cdots, \overline{\boldsymbol{x}}_N\right]^T, \overline{\boldsymbol{x}}_n \in \mathbb{R}^{M\times 1} \tag{4}$$

전처리가 완료된 학습용 정상 데이터 셋 $\overline{\boldsymbol{X}}_{train}$은 SAE 모델 학습에 사용하며, 전처리가 완료된 데이터 셋 $\overline{\boldsymbol{X}}$ 중 학습용 정상 데이터 셋을 제외한 데이터는 SAE 모델을 검증하기 위해 테스트 데이터로 사용한다.

3.2 SAE 모델 학습

SAE 모델은 각각 l개의 은닉층을 가진 인코더 g와 디코더 f로 구성되어 있다. 인코더 g는 데이터의 차원을 축소하는 역할을 하며, 디코더 f는 축소된 데이터를 입력 데이터와 동일한 차원으로 복원하는 역할을 한다. 학습용 정상 데이터 샘플 $\overline{\boldsymbol{x}}_{n_{train}}$에 대하여 인코더 g의 i번째 $(1 \le i \le l)$ 은닉층의 출력값을 수식 (5)와 같이 정의하며, 각 층의 계산 과정은 수식 (6)과 같이 나타낼 수 있다.

$$g_{:i}(\overline{\boldsymbol{x}}_{n_{train}}) = g_i \circ g_{i-1} \circ \cdots \circ g_1(\overline{\boldsymbol{x}}_{n_{train}}) \tag{5}$$

$$\begin{aligned} &g_i(g_{:i-1}(\overline{\boldsymbol{x}}_{n_{train}})) \\ &= \phi(\boldsymbol{W}_{g_i} \bullet g_{:i-1}(\overline{\boldsymbol{x}}_{n_{train}}) + b_{g_i}) \end{aligned} \tag{6}$$

수식 (6)의 ϕ는 비선형 활성 함수로, 입력 데이터를 비선형적으로 변환하여 출력값을 생성하는 함수이다. 여기서 $i=1$인 경우, $g_{:0}(\overline{\boldsymbol{x}}_{n_{train}}) = \overline{\boldsymbol{x}}_{n_{train}}$으로 정의한다. $\boldsymbol{W}_{g_i}$는 i번째 인코더 층의 weight 행렬을 의미하며 인코더 층 g_i의 입력 데이터의 차원 $K_{g_{i-1}}$와 축소된 데이터의 차원 K_{g_i}에 대하여 $\boldsymbol{W}_{g_i} \in \mathbb{R}^{K_{g_{i-1}} \times K_{g_i}}$을 만족한다. b_{g_i}는 i번째 인코더 층의 bias 값을 의미한다. 학습용 정상 데이터 샘플 $\overline{\boldsymbol{x}}_{n_{train}}$이 모든 인코더 층을 통과하여 최종적으로 축소된 결과는 수식 (5)에 의해 $g_{:l}(\overline{\boldsymbol{x}}_{n_{train}}) = \boldsymbol{h}_{n_{train},l}$로 정의한다.

인코더를 통해 축소된 데이터 $\boldsymbol{h}_{n_{train},l}$은 디코더 f의 입력으로 사용한다. 디코더 f의 i번째 $(1 \le i \le l)$ 은닉층의 결과를 수식 (7)과 같이 정의하며, 각 층의 계산 과정은 수식 (8)과 같이 나타낼 수 있다.

$$f_{:i}(\boldsymbol{h}_{n_{train},l}) = f_i \circ f_{i-1} \circ \cdots \circ f_1(\boldsymbol{h}_{n_{train},l}) \tag{7}$$

$$\begin{aligned} &f_i(f_{:i-1}(\boldsymbol{h}_{n_{train},l})) \\ &= \phi(\boldsymbol{W}_{f_i} \bullet f_{:i-1}(\boldsymbol{h}_{n_{train},l}) + b_{f_i}) \end{aligned} \tag{8}$$

수식 (8)에서 ϕ는 인코더와 동일하게 비선형 활성 함수를 의미한다. 여기서 $i=1$인 경우, $f_{:0}(\boldsymbol{h}_{n_{train},l}) = \boldsymbol{h}_{n_{train},l}$로 정의한다. $\boldsymbol{W}_{f_i}$는 i번째 디코더 층의 weight 행렬을 의미하며 디코더 층 f_i의 입력 데이터의 차원 $K_{f_{i-1}}$와 복원 데이터 차원 K_{f_i}에 대하여 $\boldsymbol{W}_{f_i} \in \mathbb{R}^{K_{f_{i-1}} \times K_{f_i}}$를 만족한다. b_{f_i}는 i번째 디코더 층의 bias 값을 의미한다. 이렇게 정의된 디코더 f를 통해 복원된 데이터는 $\hat{\boldsymbol{x}}_{n_{train}}$으로 정의한다.

SAE 모델은 학습 과정에서 입력 데이터 $\overline{\boldsymbol{x}}_{n_{train}}$와 복원 데이터 $\hat{\boldsymbol{x}}_{n_{train}}$ 사이의 오차인 복원 오차를 최소화하는 방향으로 학습을 진행한다. 학습을 진행하기 위한 모델의 손실 함수로 평균 제곱 오차인 Mean Squared Error (MSE)를 사용하며 수식 (9)와 같이 나타낼 수 있다.

$$\begin{aligned} J(\boldsymbol{W},\boldsymbol{b}) &= \frac{1}{N_{train}} \sum_{n_{train}=1}^{N_{train}} (\overline{\boldsymbol{x}}_{n_{train}} - \hat{\boldsymbol{x}}_{n_{train}})^2 \\ \boldsymbol{W} &= \left\{\boldsymbol{W}_{g_1}, \boldsymbol{W}_{g_2}, \cdots, \boldsymbol{W}_{g_l}, \boldsymbol{W}_{f_1}, \boldsymbol{W}_{f_2}, \cdots, \boldsymbol{W}_{f_l}\right\} \\ \boldsymbol{b} &= \left\{b_{g_1}, b_{g_2}, \cdots, b_{g_l}, b_{f_1}, b_{f_2}, \cdots, b_{f_l}\right\} \end{aligned} \tag{9}$$

수식 (9)의 $\boldsymbol{W}$는 각 은닉층의 weight 행렬 $2l$개를 원소로 가지는 집합이며, $\boldsymbol{b}$는 각 은닉층의 bias 값 $2l$개를 원소로 가지는 집합이다. SAE 모델은 손실 함수 $J(\boldsymbol{W},\boldsymbol{b})$를 최소화하는 방향으로 $\boldsymbol{W}$와 $\boldsymbol{b}$를 최적화한다.

3.3 Anomaly score 설계

SAE 모델 기반의 이상탐지 시스템은 학습된 모델에 전처리 된 네트워크 데이터를 입력하여 도출한 anomaly score가 임계값을 초과하는 경우에 해당 데이터를 비정상으로 탐지한다. 본 절에서는 은닉층의 정보를 활용한 기존의 anomaly score 측정 방식에 대하여 설명한 뒤, 기존의 방식이 feature 별 정규화 과정을 거치지 않기 때문에 가지는 문제점을 설명한다.

이후 기존 방식이 가지는 문제점을 해결하기 위해 정규화 과정을 거치는 anomaly score 측정 방식을 새롭게 제안한다.

3.3.1 은닉층의 정보의 활용

anomaly score에 은닉층의 정보를 반영하기 위해 입력 데이터 $\overline{\boldsymbol{x}}_n$의 i번째 은닉층의 결과 $\boldsymbol{h}_{n,i}$와 복원 데이터 $\hat{\boldsymbol{x}}_n$의 i번째 은닉층의 결과 $\hat{\boldsymbol{h}}_{n,i}$를 수식 (10)과 수식 (11)과 같이 정의한다.

$$\boldsymbol{h}_{n,i} = g_{:i}(\overline{\boldsymbol{x}}_n) \tag{10}$$

$$\hat{\boldsymbol{h}}_{n,i} = g_{:i}(\hat{\boldsymbol{x}}_n) \tag{11}$$

입력 데이터 $\overline{\boldsymbol{x}}_n$의 i번째 은닉층의 결과 $\boldsymbol{h}_{n,i}$와 복원 데이터 $\hat{\boldsymbol{x}}_n$의 i번째 은닉층의 결과 $\hat{\boldsymbol{h}}_{n,i}$의 오차를 $\boldsymbol{d}_i(\overline{\boldsymbol{x}}_n)$로 정의하여 수식 (12)와 같이 표현하고, 이를 층별 복원 오차라고 명명한다.

$$\begin{aligned}\boldsymbol{d}_i(\overline{\boldsymbol{x}}_n) &= \boldsymbol{h}_{n,i} - \hat{\boldsymbol{h}}_{n,i} \\ \boldsymbol{d}_0(\overline{\boldsymbol{x}}_n) &= \overline{\boldsymbol{x}}_n - \hat{\boldsymbol{x}}_n\end{aligned} \tag{12}$$

수식 (12)에서 $i=0$일 때, $\boldsymbol{d}_0(\overline{\boldsymbol{x}}_n)$는 입력 데이터 $\overline{\boldsymbol{x}}_n$와 복원 데이터 $\hat{\boldsymbol{x}}_n$의 복원 오차로 정의한다. 은닉층의 정보를 사용하여 이상탐지를 진행하는 [13]의 anomaly score 측정 방식은 층별 복원 오차 $\boldsymbol{d}_i(\overline{\boldsymbol{x}}_n)$를 단순 누적한 L1 norm을 적용한다. 이는 수식 (13)과 같이 나타낸다.

$$s_{L1} = \sum_{i=0}^{l} \| \boldsymbol{d}_i(\overline{\boldsymbol{x}}_n) \|_1 \tag{13}$$

[13]의 anomaly score s_{L1}은 은닉층의 정보를 활용한다는 장점이 있지만, 층별 복원 오차에 정규화 과정을 적용하지 않기 때문에 다음과 같은 두 가지 문제점을 가지고 있다.

첫 번째 문제는 s_{L1}이 정상 데이터들의 복원 오차 평균으로부터 떨어진 거리가 아닌 원점으로부터의 거리를 측정한다는 점이다. 이 경우 정상 데이터 복원 오차들의 평균과 거리가 멀지만, 원점과 거리가 가까운 비정상 데이터나, 정상 데이터 복원 오차들의 평균과의 거리가 가깝지만, 원점과 거리가 먼 정상 데이터 등이 존재할 경우, 높은 정확도의 이상탐지 결과를 기대할 수 없다. 두 번째 문제는 각 은닉층마다 층별 복원 오차 $\boldsymbol{d}_i(\overline{\boldsymbol{x}}_n)$의 범위가 상이하여, 범위가 큰 은닉층의 정보가 anomaly score에 결정적인 영향을 주게 된다는 점이다. 따라서 anomaly score가 모든 은닉층의 정보를 공정하게 반영하지 못한다는 한계점이 있다. 따라서 이러한 문제점을 해결하기 위해 본 논문에서는 층별 복원 오차를 정규화한 후, L1 norm을 적용하는 Normalized L1 norm (N-L1) 방식을 제안한다.

3.3.2 제안하는 Normalize L1 norm (N-L1) 방식

제안하는 N-L1 방식은 층별 복원 오차의 평균을 0으로 이동시키는 단계와 각 층별 복원 오차의 공분산을 $\boldsymbol{I}$로 통일시키는 단계를 통해 정규화를 진행한다. 각 단계는 학습용 정상 데이터 셋의 i번째 층의 복원 오차 행렬 $\boldsymbol{D}_i \in \mathbb{R}^{N_{train} \times K_{g_i}}$를 사용하며, 다음과 같이 수식 (14)로 정의한다.

$$\boldsymbol{D}_i = [\boldsymbol{d}_i(\overline{\boldsymbol{x}}_{1_{train}}), \boldsymbol{d}_i(\overline{\boldsymbol{x}}_{2_{train}}), \cdots, \boldsymbol{d}_i(\overline{\boldsymbol{x}}_{N_{train}})]^T \tag{14}$$

층별 복원 오차의 정규화 과정에서 i번째 층의 복원 오차 행렬 $\boldsymbol{D}_i$의 평균과 특잇값 분해 (singular value decomposition; SVD)를 사용하며, 이는 수식 (15), (16)을 통하여 구할 수 있다.

$$\boldsymbol{m}_i = \mathbb{E}(\boldsymbol{D}_i) \tag{15}$$

$$\overline{\boldsymbol{D}}_i = \boldsymbol{D}_i - \mathbf{1}_{N_{train} \times 1} \times \boldsymbol{m}_i = \boldsymbol{U}_i \boldsymbol{\Sigma}_i \boldsymbol{V}_i^T \tag{16}$$

수식 (15)의 $\boldsymbol{m}_i \in \mathbb{R}^{1 \times K_{g_i}}$는 $\boldsymbol{D}_i$의 열(column)평균을 의미하며, 수식 (16)의 $\overline{\boldsymbol{D}}_i$는 $\boldsymbol{m}_i$를 이용하여 $\boldsymbol{D}_i$의 평균을 원점으로 이동시킨 행렬이다. 수식 (16)은 SVD를 통해 $\overline{\boldsymbol{D}}_i$를 $\boldsymbol{U}_i \in \mathbb{R}^{N_{train} \times N_{train}}$, $\boldsymbol{V}_i \in \mathbb{R}^{K_{g_i} \times K_{g_i}}$, $\boldsymbol{\Sigma}_i \in \mathbb{R}^{N_{train} \times K_{g_i}}$의 세 행렬의 곱으로 나타낸 것으로, 세 행렬은 각각 좌특이 벡터, 우특이 벡터, 특이값을 의미한다[15]. 제안하는 N-L1 방식은 수식 (15)를 통해 구한 $\boldsymbol{m}_i$와 수식 (16)에서 구한 $\boldsymbol{V}_i, \boldsymbol{\Sigma}_i$를 이용하여 층별 복원 오차의 평균은 원점으로, 공분산은 $\boldsymbol{I}$로 정규화한다. 이에 대한 증명은 다음과 같다.

Property 1. $\boldsymbol{A}=(\boldsymbol{D}_i-\mathbf{1}_{N_{train}\times 1}\times \boldsymbol{m}_i)\boldsymbol{V}_i\boldsymbol{\Sigma}_i^{-1}$ ***라면,*** $\mathbb{E}(\boldsymbol{A})=\mathbf{0}$, $Cov(\boldsymbol{A})=\boldsymbol{I}$***을 만족한다.***

Proof 1. 먼저 $\boldsymbol{A}=(\boldsymbol{D}_i-\mathbf{1}_{N_{train}\times 1}\times \boldsymbol{m}_i)\boldsymbol{V}_i\boldsymbol{\Sigma}_i^{-1}$인 경우, $\mathbb{E}(A)=\mathbf{0}$임을 증명한다. 행렬 $\boldsymbol{A}$는 수식 (16)의 $\overline{\boldsymbol{D}}_i$ 정의에 따라 $\boldsymbol{A}=\overline{\boldsymbol{D}}_i\boldsymbol{V}_i\boldsymbol{\Sigma}_i^{-1}$로 표현할 수 있다. 여기서 $\boldsymbol{A}$를 구성하는 행렬의 (i,j)번째 원소를 $\overline{\boldsymbol{D}}_i$는 y_{ij}, $\boldsymbol{V}_i$는 v_{ij}, $\boldsymbol{\Sigma}_i^{-1}$는 s_{ij}로 정의한다. 이를 통해 $\overline{\boldsymbol{D}}_i$의 k번째 열을 $\mathbb{Y}_k=[y_{1k},y_{2k},\cdots,y_{N_{train}k}]^T$라고 정의하면, $\mathbb{Y}_k$와 $\boldsymbol{V}_i$의 원소 v_{ij}를 통해 행렬 곱 $\overline{\boldsymbol{D}}_i\boldsymbol{V}_i$의 k번째 열 $\mathbb{V}_k$는 다음과 같이 표현할 수 있다.

$$\mathbb{V}_k=\mathbb{Y}_1v_{1k}+\mathbb{Y}_2v_{2k}+\cdots+\mathbb{Y}_{K_{y_i}}v_{K_{y_i}k}=\sum_{i=1}^{K_{y_i}}v_{ik}\mathbb{Y}_i$$

다음으로, 행렬 곱 $\overline{\boldsymbol{D}}_i\boldsymbol{V}_i\boldsymbol{\Sigma}_i^{-1}$로 표현되는 $\boldsymbol{A}$의 k번째 열 $\mathbb{S}_k$는 $\mathbb{V}_k$와 $\boldsymbol{\Sigma}_i^{-1}$의 원소 s_{ij}를 사용하여 다음과 같이 표현할 수 있다.

$$\begin{aligned}\mathbb{S}_k&=\mathbb{V}_1s_{1k}+\mathbb{V}_2s_{2k}+\cdots+\mathbb{V}_{K_{y_i}}s_{K_{y_i}k}\\&=\sum_{j=1}^{K_{y_i}}s_{jk}\mathbb{V}_j=\sum_{j=1}^{K_{y_i}}s_{jk}\sum_{i=1}^{K_{y_i}}v_{ij}\mathbb{Y}_i\end{aligned}$$

여기서 $\frac{1}{N_{train}}\sum_{n_{train}=1}^{N_{train}}y_{n_{train}k}=0$이므로, $\mathbb{E}(\mathbb{S}_k)$는 다음과 같다.

$$\begin{aligned}\mathbb{E}(\mathbb{S}_k)&=\frac{1}{N_{train}}\sum_{j=1}^{K_{y_i}}s_{jk}\sum_{i=1}^{K_{y_i}}v_{ij}\sum_{n_{train}=1}^{N_{train}}y_{n_{train}i}\\&=\sum_{j=1}^{K_{y_i}}s_{jk}\sum_{i=1}^{K_{y_i}}v_{ij}\left(\frac{1}{N_{train}}\sum_{n_{train}=1}^{N_{train}}y_{n_{train}i}\right)\\&=0\end{aligned}$$

즉, $\boldsymbol{A}$의 k번째 열의 평균 $\mathbb{E}(\mathbb{S}_k)=0$을 만족한다. 따라서 $\boldsymbol{A}$의 열 평균 벡터 $\mathbb{E}(\boldsymbol{A})=\mathbf{0}$을 만족한다.

다음으로, $Cov(\boldsymbol{A})=\boldsymbol{I}$의 증명은 다음과 같다. 특이값 분해의 좌특이 벡터 $\boldsymbol{U}_i$와 우특이 벡터 $\boldsymbol{V}_i$는 직교 행렬이므로 $\boldsymbol{V}_i^T\boldsymbol{V}_i=\boldsymbol{I}$, $\boldsymbol{U}_i^T\boldsymbol{U}_i=\boldsymbol{I}$를 만족한다. 따라서 $Cov(A)=\boldsymbol{I}$임을 다음과 같이 증명할 수 있다.

$$\begin{aligned}Cov(\boldsymbol{A})&=\boldsymbol{A}^T\boldsymbol{A}\\&=(\boldsymbol{D}_i\boldsymbol{V}_i\boldsymbol{\Sigma}_i^{-1})^T(\boldsymbol{D}_i\boldsymbol{V}_i\boldsymbol{\Sigma}_i^{-1})\\&=((\boldsymbol{U}_i\boldsymbol{\Sigma}_i\boldsymbol{V}_i^T)\boldsymbol{V}_i\boldsymbol{\Sigma}_i^{-1})^T((\boldsymbol{U}_i\boldsymbol{\Sigma}_i\boldsymbol{V}_i^T)\boldsymbol{V}_i\boldsymbol{\Sigma}_i^{-1})\\&=((\boldsymbol{U}_i\boldsymbol{\Sigma}_i(\boldsymbol{V}_i^T\boldsymbol{V}_i)\boldsymbol{\Sigma}_i^{-1})^T((\boldsymbol{U}_i\boldsymbol{\Sigma}_i(\boldsymbol{V}_i^T\boldsymbol{V}_i)\boldsymbol{\Sigma}_i^{-1})\\&=(\boldsymbol{U}_i\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_i^{-1})^T(\boldsymbol{U}_i\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_i^{-1})\\&=\boldsymbol{U}_i^T\boldsymbol{U}_i\\&=\boldsymbol{I}\end{aligned}$$

이에 따라 $\boldsymbol{A}=(\boldsymbol{D}_i-\mathbf{1}_{N_{train}\times 1}\times \boldsymbol{m}_i)\boldsymbol{V}_i\boldsymbol{\Sigma}_i^{-1}$이면 $\mathbb{E}(\boldsymbol{A})=\mathbf{0}$, $Cov(\boldsymbol{A})=\boldsymbol{I}$을 만족한다. □

제안하는 N-L1 방식은 층별 복원 오차 $\boldsymbol{d}_i(\overline{\boldsymbol{x}}_n)$를 정규화한 뒤, L1 norm을 적용하여 anomaly score를 측정하며, 이는 수식 (17)과 같이 표현된다.

$$s_{N-L1}=\sum_{i=0}^{l}\left\|\left(\boldsymbol{d}_i(\overline{\boldsymbol{x}}_n)-\boldsymbol{m}_i^T\right)^T\boldsymbol{V}_i\boldsymbol{\Sigma}_i^{-1}\right\|_1 \qquad (17)$$

따라서 제안하는 N-L1 방식의 anomaly score인 s_{N-L1}은 층별 복원 오차의 정보를 공정하게 반영할 수 있다.

Ⅳ. 성능 평가

본 장에서는 제안하는 방식의 우수성을 입증하기 위해 네트워크 데이터를 사용하여 이상탐지 성능 평가 실험을 진행한다. 본 실험은 전처리 단계, 학습용 정상 데이터를 통한 SAE 모델 학습 단계, 학습된 모델을 이용한 이상탐지 테스트 단계로 구성된다.

실험의 첫 번째 단계인 전처리 단계에서는 정상 데이터를 8:2로 랜덤 샘플링하여 학습용 정상 데이터와 테스트용 정상 데이터로 분류한다. 테스트용 비정상 데이터는 공정한 성능 평가를 위해 정상 데이터와 동일한 개수로 랜덤 샘플링하여 구성한다. 데이터의 전처리는 학습용 정상 데이터의 평균과 표준편차를 사용하여 standard scaler 방식으로 진행한다. 실험의 두 번째 단계인 SAE 모델 학습 단계에서는 전처리 된 학습용 정상 데이터를 사용하여 모델의 학습을 진행한다. 모델은 정상 데이터의 복원 오차를 최소화하는 방향으로 학습을 진행한다. 따라서 학습된 모델은 정상 데이터에 대해서는 작은 anomaly score를 가지며, 비정상 데이터에 대해서는 큰 anomaly score를 가지게 된다. 실험의 세 번째 단계인 이상탐지 단계에서는 학습된 SAE 모델에 테스트 데이터를 입력하여 anomaly

score를 측정한다. 측정된 anomaly score가 임계값 이하일 경우 해당 데이터를 정상으로 판단하고, 임계값을 초과할 경우 비정상으로 판단하여 탐지를 진행한다.

위의 실험 단계에 따라 성능 평가 실험을 진행한다. 또한, 제안하는 방식이 특정 데이터 셋에서만 우수한 성능을 보이는 것이 아님을 확인하기 위하여 두 가지의 네트워크 트래픽 데이터 셋을 사용해 실험을 진행하고, 그 성능을 확인하였다.

4.1 사용한 데이터 셋

본 실험에서는 NSL-KDD 데이터 셋과 CSE-CIC-IDS 2018 데이터 셋을 사용하였다. 먼저, NSL-KDD 데이터 셋은 41개의 feature로 구성되어 있는 데이터로 침입 탐지 시스템 평가를 위해 보편적으로 사용된다[8,13,14,16-18]. NSL-KDD 데이터 셋은 67,343개의 정상 데이터와 58,630개의 비정상 데이터로 구성되어 있다. 다음으로 CSE-CIC-IDS 2018 데이터 셋은 캐나다 사이버보안 연구소 (CIC)와 캐나다 통신 보안기구 (CSE)에서 실제 네트워크 환경과 유사하게 만들어진 데이터 셋이다[19]. CSE-CIC-IDS 2018 데이터 셋은 83개의 feature를 가지고 있으며, 13,484,708개의 정상 데이터와 2,748,235개의 비정상 데이터로 구성되어 있다.

4.2 성능 평가 지표

본 논문에서는 이상탐지를 위해 학습용 정상 데이터 셋의 anomaly score를 기반으로 임계값을 설정하며, 입력 데이터의 anomaly score가 임계값 이하이면 정상으로 예측하고 임계값을 초과하면 비정상으로 예측한다. 예측의 성능을 평가하기 위한 성능 평가 지표는 오차 행렬을 기반으로 연산한다. 오차 행렬은 실제값과 예측값을 바탕으로 데이터를 4가지의 유형으로 분류하며, 표 1과 같이 나타낼 수 있다. 실제 정상 데이터에 대한 올바른 예측이면 True Negative (TN), 실제 정상 데이터에 대한 틀린 예측이면 False Positive (FP), 실제 비정상 데이터에 대한 올바른 예측이면 True Positive (TP), 실제 비정상 데이터에 대한 틀린 예측이면 False Negative (FN)로 분류한다. 이렇게 분류된 오차 행렬의 TP, FP, TN, FN를 이용하여 대표적인 8가지의 성능 평가 지표를 통해 본 실험의 성능을 종합적으로 분석한다.

표 1. 오차 행렬
Table 1. Confusion matrix

Confusion Matrix		예측값	
		비정상	정상
실제값	비정상	True Positive (TP)	False Negative (FN)
	정상	False Positive (FP)	True Negative (TN)

• Precision

Precision은 예측값이 비정상인 데이터 중에서 실제값이 비정상인 데이터로 올바르게 예측한 비율이다. 이 지표는 1에 가까울수록 네트워크 침입 탐지 시스템에서 공격이라고 예측한 값이 공격일 확률이 높아진다.

$$Precision = \frac{TP}{TP+FP}$$

• Recall

Recall은 실제값이 비정상인 데이터 중에서 예측값이 비정상인 데이터로 올바르게 예측한 비율이다. 이 지표는 1에 가까울수록 네트워크 침입 탐지 시스템에서 실제 공격을 탐지할 비율이 높아진다.

$$Recall = \frac{TP}{TP+FN}$$

• F1-score

F1-score는 Precision과 Recall의 조화 평균이다. Precision과 Recall은 trade-off 관계를 가지고 있다. Precision을 최대한 높이기 위해서는 비정상으로 예측을 보수적으로 하도록 임계값을 설정할 수 있다. 이 경우에 정상 데이터와 분포가 겹쳐 있는 비정상 데이터들에 대하여 정상으로 판단할 가능성이 높다는 문제점이 생긴다. 반면, Recall을 최대한 높이기 위해서는 조금이라도 정상 범위에서 벗어나는 경우를 모두 비정상으로 예측하도록 임계값을 설정할 것이다. 이 경우에는 정상 데이터를 비정상 데이터라고 예측하는 Fall-out이 높아진다는 문제점이 생긴다. 이러한 Precision과 Recall의 trade-off를 적절히 고려한 F1-score는 이상탐지 시스템을 구축할 수 있도록 하는 성능 지표이다.

$$F1-score = \frac{2(precision \times recall)}{precision + recall}$$

• Accuracy

Accuracy는 예측값이 실제값과 일치하는 정도를 나타내는 지표이다. 이 지표는 1에 가까울수록 네트워크 침입 탐지 시스템의 예측이 실제와 일치한다.

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP}$$

• Fall-out

Fall-out은 실제값이 정상인 데이터 중에서 예측값이 비정상인 데이터로 잘못 예측한 값의 비율이며, 오경보율이라고도 불린다. 이 지표는 0과 가까울수록 네트워크 침입 탐지 시스템의 오경보율이 감소한다.

$$Fall-out = \frac{FP}{TN+FP}$$

• Specificity

Specificity는 실제값이 정상인 데이터 중에서 예측값이 정상인 데이터로 올바르게 예측한 비율이다. 이 지표는 1에 가까울수록 오경보율인 Fall-out이 감소한다.

$$Specificity = \frac{TN}{TN+FP}$$

• Metthews Correlation Coefficient (MCC)

MCC는 예측값과 실제값의 상관계수이다. 이 지표는 -1과 1 사이의 값을 가지며 1에 가까울수록 예측값과 실제값이 일치하는 완벽한 예측, -1에 가까울수록 예측값과 실제값이 상반되는 예측, 0에 가까울수록 예측값과 실제값이 연관성을 가지지 않는 무작위 예측을 의미한다[20].

$$MCC = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

• Area Under Curve (AUC)

AUC는 Fall-out 대 Recall의 그래프인 Receiver Operating Characteristic (ROC) 그래프의 곡선 아래 영역이다. AUC는 0.5와 1 사이의 값을 가지며, 1에 가까울수록 작은 Fall-out에 대하여 높은 Recall을 가지는 것을 의미한다. 이는 정상 데이터를 공격 데이터로 예측하지 않으면서, 공격 데이터를 탐지해 내는 비율이다[21].

4.3 성능 비교 평가

제안하는 방식의 성능 평가를 위해 4.2절에서 설명한 8가지의 성능 지표를 사용한 실험을 진행한다. 본 실험은 일반화된 우수한 성능임을 보여주기 위하여 10개의 랜덤시드에 대한 각 성능 지표의 평균과 표준편차를 제시하였다. 또한 제안하는 방식과 다음의 두 선행연구의 성능 비교를 수행하였다.

- **SAE 방식**[11]: 학습된 SAE 모델의 입력층과 출력층의 복원 오차만을 사용하여 anomaly score를 구하는 방식이다. 이때, 은닉층의 정보를 포함하지 않기 때문에 학습된 모델의 모든 정보를 반영하지 못한다.
- **L1 방식**[13]: 인코더의 층별 복원 오차를 은닉층의 정보로 이용하며, 각 층의 복원 오차에 L1 norm을 적용하여 anomaly score를 구한다. 이때, 층별 복원 오차에 정규화 과정은 포함되지 않는다. 따라서 층별 복원 오차의 범위가 상이하므로 특정 은닉층의 복원 오차가 집중적으로 반영될 수 있다는 한계점이 존재한다.

표 2는 NSL-KDD 데이터 셋에 관한 SAE 방식, L1 방식, 그리고 제안하는 N-L1 방식의 실험 결과를 나타낸다. L1의 방식은 SAE 방식에 대비하여 은닉층의 정보를 추가적으로 활용하였음에도 불구하고 성능에 큰 차이가 없는 것을 확인할 수 있다. 그러나 SAE 방식과 은닉층의 정보를 정규화하여 사용하는 제안 방식을 비교하였을 때, 제안 방식이 SAE 방식 대비 Precision, Specificity, Recall, Accuracy, 그리고 F1-score의 항목에서 성능이 약 8~9% 향상되었다. MCC는 21.4%로 성능이 큰 폭으로 향상되었는데, 이는 SAE 방식에 비해 제안하는 방식이 예측값과 실제값의 연관성을 더 잘 나타냄을 의미한다. 특히, Fall-out은 74.5%의 가장 높은 성능 향상률을 보였다. 이는 오경보율을 대폭 축소하여 이상탐지 시스템 경보의 신뢰성을 높일 수 있음을 나타낸다. 결론적으로 각 은닉층의 정보를 활용을 통한 이상탐지 성능의 향상에 정규화 과정이 큰 역할을 한다는 것을 확인하였다.

그림 1, 2, 3은 NSL-KDD 데이터 셋에 대하여 각각 SAE 방식, L1 방식, 제안 방식으로 구한 anomaly score를 데이터 타입에 따라 히스토그램으로 표현한 그래프이다. 그래프의 파란색 히스토그램은 정상 데이터를 의미하며, 붉은색 히스토그램은 비정상 데이터를

표 2. NSL-KDD 데이터 셋 이상탐지 성능 평가 결과
Table 2. The performance of anomaly detection for NSL-KDD data set

		Precision	Fall-out	Specificity	Recall	Accuracy	F1-score	MCC	AUC
제안 방식	mean	**0.9679**	**0.0315**	**0.9685**	**0.9612**	**0.9648**	**0.9642**	**0.9302**	**0.9705**
	std	±0.0192	±0.0204	±0.0198	±0.0172	±0.0131	±0.0131	±0.0262	±0.0036
SAE 방식	mean	0.8828	0.1235	0.8865	0.8846	0.8806	0.8809	0.7661	0.9627
	std	±0.0472	±0.0614	±0.0615	±0.0525	±0.0066	±0.0052	±0.0122	±0.0035
L1 방식	mean	0.8721	0.1495	0.8605	0.9094	0.8799	0.8835	0.7643	0.9534
	std	±0.0423	±0.0543	±0.0551	±0.0381	±0.0121	±0.0091	±0.0202	±0.0077

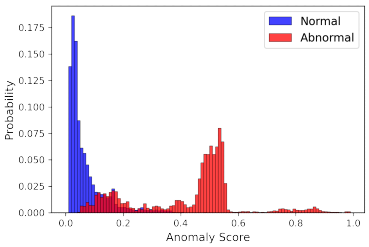


그림 1. SAE 방식에 따른 데이터 타입별 anomaly score의 히스토그램 (NSL-KDD 데이터 셋)
Fig. 1. Histogram of anomaly score based on the SAE method (NSL-KDD data set)

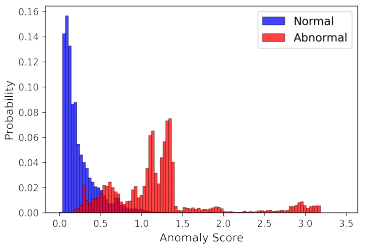


그림 2. L1 방식에 따른 데이터 타입별 anomaly score의 히스토그램 (NSL-KDD 데이터 셋)
Fig. 2. Histogram of anomaly score based on the L1 method (NSL-KDD data set)

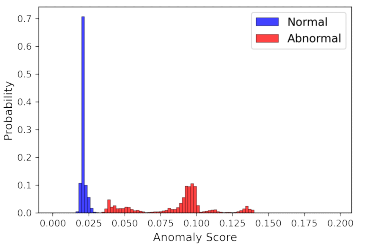


그림 3. 제안 방식에 따른 데이터 타입별 anomaly score의 히스토그램 (NSL-KDD 데이터 셋)
Fig. 3. Histogram of anomaly score based on the proposed solution (NSL-KDD data set)

의미한다. 그림 1, 2를 통해 SAE 방식과 L1 방식은 정상 데이터와 비정상 데이터의 anomaly score의 분포가 상당 부분 겹쳐 있는 것을 확인할 수 있다. 이처럼 정상과 비정상 데이터의 anomaly score 분포가 겹치게 되면, 비정상 판단의 기준이 되는 임계값 설정에 어려움이 생기게 되어 이상탐지 시스템의 높은 탐지 성능을 기대하기 어렵다. 반면, 제안 방식 결과에서는 정상과 비정상 데이터의 anomaly score 분포가 겹치지 않는 것을 그림 3에서 확인할 수 있다. 이에 따라 제안하는 방식을 활용하면 임계값 설정이 용이하여 더 좋은 탐지 성능을 가질 수 있음을 확인하였다.

표 3은 CSE-CIC-IDS 2018 데이터 셋에 관한 SAE 방식, L1 방식, 그리고 제안하는 방식의 실험 결과이다. L1의 방식은 SAE 방식에 대비하여 Precision과 Fall-out, 그리고 Specificity 항목에서 소폭 향상된 것을 확인할 수 있다. 그러나 Recall, Accuracy, F1-score, MCC, AUC 항목에서 모두 성능이 하락하였다. 반면에 제안 방식은 SAE 방식에 대비하여 모든 항목에서 성능 향상을 보이는 것을 확인할 수 있다.

이 중 Precision과 Specificity는 각각 41%와 53.25%의 높은 성능 향상률을 보였다. 특히 Fall-out과 MCC에서 각각 96.3%, 99%의 비약적인 성능 향상을 보였다.

결론적으로 CSE-CIC-IDS 2018 데이터 셋 또한 각 은닉층의 정보를 활용을 통한 이상탐지 성능의 향상에 정규화 과정이 큰 역할을 하는 것을 입증하였다.

그림 4, 5, 6은 CSE-CIC-IDS 2018 데이터 셋에 대하여 각각 SAE 방식, L1 방식, 제안 방식으로 구한 anomaly score를 데이터 타입에 따라 표현한 그래프이다. 그림 4, 5를 통해 SAE 방식과 L1 방식은 정상 데이터와 비정상 데이터의 anomaly score의 분포가 상당 부분 겹쳐 있음을 확인하였다. 따라서 CSE-CIC-IDS 2018 데이터 셋 또한 정상과 비정상 데이터의 anomaly score 분포가 겹쳐 있기에 임계값 설정에 어려움이 생기게 되어 이상탐지 시스템의 높은 성능을 기대하기 어렵다. 반면, 그림 6을 통해 제안 방식은 정상과 비정상 데이터의 anomaly score 분포가 겹치지 않는 것을 확인할 수 있다. 이를 통해 제안하는 N-L1 방식이 다른 방식에 비해 임계값 설정이 용이하며 더 좋은 성능을 가질 수 있음을 시각적으로도 확인하였다.

표 3. CSE-CIC-IDS 2018 데이터 셋 이상탐지 성능 평가 결과
Table 3. The performance of anomaly detection for CSE-CIC-IDS 2018 data set

		Precision	Fall-out	Specificity	Recall	Accuracy	F1-score	MCC	AUC
제안 방식	mean	**0.9864**	**0.0137**	**0.9862**	**0.9985**	**0.9923**	**0.9924**	**0.9850**	**0.9957**
	std	±0.0159	±0.0194	±0.0193	±0.0056	±0.0097	±0.0087	±0.0181	±0.0068
SAE 방식	mean	0.6997	0.3764	0.6435	0.8553	0.7394	0.7676	0.4949	0.7763
	std	±0.0833	v0.1846	±0.1846	±0.1029	±0.0535	±0.0286	±0.0934	±0.0258
L1 방식	mean	0.7437	0.3263	0.6737	0.7577	0.7199	0.7174	0.4703	0.6956
	std	±0.1097	±0.2554	±0.2554	±0.1625	±0.0603	±0.0364	±0.0824	±0.0644

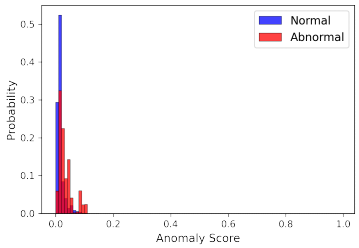


그림 4. SAE 방식에 따른 데이터 타입별 anomaly score의 히스토그램 (CSE-CIC- IDS 2018 데이터 셋)
Fig. 4. Histogram of anomaly score based on the SAE method (CSE-CIC-IDS-2018 data set)

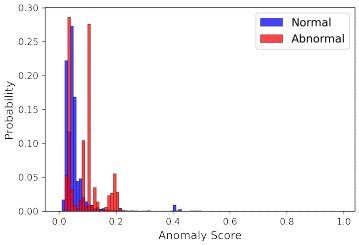


그림 5. L1 방식에 따른 데이터 타입별 anomaly score의 히스토그램 (CSE-CIC-IDS 2018 데이터 셋)
Fig .5. Histogram of anomaly score based on the L1 method (CSE-CIC-IDS-2018 data set)

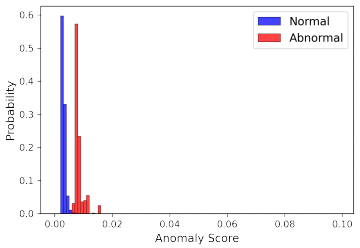


그림 6. 제안 방식에 따른 데이터 타입별 anomaly score의 히스토그램 (CSE-CIC-IDS 2018 데이터 셋)
Fig. 6. Histogram of anomaly score based on the proposed solution (CSE-CIC-IDS-2018 data set)

V. 결 론

본 논문은 SAE 모델의 은닉층의 정보를 반영하는 anomaly score를 구하고 이를 기반으로 하는 이상탐지 시스템을 제안하였다. 제안한 이상탐지 시스템은 복원된 데이터를 모델에 다시 입력하여 얻은 은닉층의 출력값을 원본 데이터의 은닉층의 출력값과 비교하여, 그 오차인 층별 복원 오차를 구한다. 이후 층별 복원 오차에 정규화 과정을 진행하는 N-L1 방식으로 anomaly score를 측정하고, 측정한 anomaly score 값이 임계값을 초과하면 해당 데이터를 비정상으로 탐지한다. 제안하는 네트워크 이상탐지 시스템의 성능을 확인하기 위해, 두 가지의 네트워크 데이터 셋을 바탕으로 두 가지의 기존 방식들과의 성능을 비교하여 제안하는 방식이 기존의 다른 방식들에 비해 가장 우수한 성능을 보임을 확인하였다.

References

[1] E. Aghaei and G. Serpen, “Ensemble classifier for misuse detection using N-gram feature vectors through operating system call traces,” *Int. J. Hybrid Intell. Syst.*, vol. 14, no. 3, pp. 141-154, 2017. (https://doi.org/10.3233/HIS-170247)

[2] S. Nielebock, R. Heumüller, K. M. Schott, and F. Ortmeier, “Guided pattern mining for API misuse detection by change-based code analysis,” *Automated Softw. Eng.*, vol. 28, no. 2, pp. 1-48, 2021. (https://doi.org/10.1007/s10515-021-00294-x)

[3] A. Sivanathan, H. H. Gharakheili, F. Loi, A. Radford, C. Wijenayake, A. Vishwanath, and V. Sivaraman, “Classifying IoT devices in smart environment using network traffic characteristics,” *IEEE Trans. Mobile Comput.*, vol. 18, no. 8, pp. 1745-1759, Aug. 2018. (https://doi.org/10.1109/TMC.2018.2866249)

[4] C. Callegari, L. Donatini, S. Giordano, and M. Pagano, “Improving stability of PCA-based network anomaly detection by means of kernel-PCA,” *Int. J. Computational Sci. and Eng.*, vol. 16, no. 1, pp. 9-16, Feb. 2018. (https://doi.org/10.1504/IJCSE.2018.089573)

[5] I. K. Savvas, A. V. Chernov, M. A. Butakova, and C. Chaikalis, "Increasing the quality and performance of N-Dimensional point anomaly detection in traffic using PCA and DBSCAN," *TELFOR*, Belgrade, Serbia, Nov. 2018. (https://doi.org/10.1109/TELFOR.2018.8611947)

[6] H. Kye and M. Kwon, "PCA-based low-complexity anomaly detection," *J. KICS*, vol. 46, no. 6, pp. 941-955, Jun. 2021. (https://doi.org/10.7840/kics.2021.46.6.941)

[7] K. Siwek and S. Osowski, "Autoencoder versus PCA in face recognition," *IEEE Int. Conf. CPEE*, Kutna Hora, Czech, Sep. 2017. (https://doi.org/10.1109/CPEE.2017.8093043)

[8] Z. Chen, C. K. Yeo, B. S. Lee, and C. T. Lau, "Autoencoder-based network anomaly detection," *IEEE WTS*, Phoenix, AZ, USA, Jan. 2018. (https://doi.org/10.1109/WTS.2018.8363930)

[9] H. Xu, W. Chen, N. Zhao, Z. Li, J. Bu, Z. Li, Y. Liu, Y. Zhao, D. Pei, Y. Feng, J. Chen, Z. Wang, and H. Qiao, "Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications," *WWW Conf.*, Lyon, France, Apr. 2018. (https://doi.org/10.1145/3178876.3185996)

[10] Q. P. Nguyen, K. W. Lim, D. M. Divakaran, K. H. Low, and M. C. Chan, "GEE: A gradient-based explainable variational autoencoder for network anomaly detection," *IEEE Conf. CNS*, Washington, D.C., USA, Jun. 2019. (https://doi.org/10.1109/CNS.2019.8802833)

[11] A. M. Vartouni, S. S. Kashi, and M. Teshnehlab, "An anomaly detection method to detect web attacks using stacked auto-encoder," *2018 6th Iranian Joint CFIS*, Kerman, Iran, Feb. 2018. (https://doi.org/10.1109/CFIS.2018.8336654)

[12] K. H. Kim, S. Shim, Y. Lim, J. Jeon, J. Choi, B. Kim, and A. S. Yoon, "RaPP: Novelty detection with reconstruction along projection pathway," *Int. Conf. Learn. Representations*, New Orleans, LA, USA, May 2019.

[13] M. Kim, H. Kye, and M. Kwon, "RaPP-based network anomaly detection systems," *Korea Artificial Intell. Conf.*, Jeju Island, Korea, Sep. 2021.

[14] H. Kye, M. Kim, and M. Kwon, "Hierarchical autoencoder for network intrusion detection," *IEEE ICC*, Seoul, Korea, May 2022. (https://doi.org/10.1109/ICC45855.2022.9839056)

[15] Z. He, M. Ng, and C. Zeng, "Generalized singular value decomposition for tensors and their applications," *Numerical Mathematics: Theory, Methods & Applications*, vol. 14, no. 3, pp. 692-713, Aug. 2021. (https://doi.org/10.4208/nmtma.OA-2020-0132)

[16] F. Z. Belgrana, N. Benamrane, M. A. Hamaida, A. M. Chaabani, and A. Taleb-Ahmed, "Network intrusion detection system using neural network and condensed nearest neighbors with selection of NSL-KDD influencing features," *IEEE Int. Conf. IoTaIS*, Bali, Indonesia, Jan. 2020. (https://doi.org/10.1109/IoTaIS50849.2021.9359689)

[17] B. Mohammed and E. K. Gbashi, "Intrusion detection system for NSL-KDD dataset based on deep learning and recursive feature elimination," *Eng. and Technol. J.*, vol 39, no. 7, pp. 1069-1079, 2021. (https://doi.org/10.30684/etj.v39i7.1695)

[18] H. Ao, "Using machine learning models to detect different intrusion on NSL-KDD," *IEEE Int. Conf. CSAIEE*, Aug. 2021. (https://doi.org/10.1109/CSAIEE54046.2021.9543241)

[19] https://www.unb.ca/cic/datasets/ids-2018.html

[20] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 6, Jan. 2020. (https://doi.org/10.1186/s12864-019-6413-7)

[21] S. Narkhede, "Understanding auc-roc curve," *Towards Data Sci.*, vol. 26, no. 1, pp. 220-227, Jun. 2018.

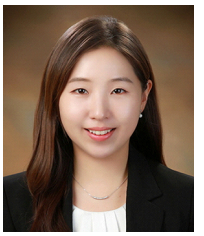
김 미 르 (Miru Kim)

2019년 3월~8월 : 숭실대학교 전자정보공학부 IT융합전공
2022년 9월~현재 : 숭실대학교 정보통신공학과 석사과정
<관심분야> 이상탐지기술, 인공지능, 연합학습
[ORCID:0000-0002-5394-4780]

계 효 선 (Hyoseon Kye)

2018년 3월~2021년 8월 : 숭실대학교 전자정보공학부 IT융합전공
2021년 9월~현재 : 숭실대학교 정보통신공학과 석사과정
<관심분야> 모바일 네트워크, 이상탐지기술, 인공지능, 연합학습
[ORCID:0000-0001-7808-0387]

권 민 혜 (Minhae Kwon)



2011년 8월 : 이화여자대학교 전자정보통신공학과 학사
2013년 8월 : 이화여자대학교 전자공학과 석사
2017년 8월 : 이화여자대학교 전자전기공학과 박사
2017년 9월~2018년 8월 : 이화여자대학교 전자전기공학과 박사 후 연구원
2018년 9월~2020년 2월 : 미국 Rice University, Electrical and Computer Engineering, Postdoctoral Researcher
2018년 9월~2020년 2월 : 미국 Baylor College of Medicine, Center for Neuroscience and Artificial Intelligence, Postdoctoral Researcher
2020년 3월~현재 : 숭실대학교 전자정보공학부 IT융합전공 조교수
<관심분야> 모바일네트워크, 이상탐지기술, 인공지능, 강화학습, 자율주행
[ORCID:0000-0002-8807-3719]