

경사면의 안정성 모니터링 데이터의 품질관리를 위한 2 단계 접근방안

Two-Phase Approach for Data Quality Management for Slope Stability Monitoring

최준혁¹, 김용진², 조준휘³, 정우철³, 석송희³, 최 송³, 김용성⁴, 지봉준^{5*}

Junhyuk Choi¹, Yongjin Kim², Junhwi Cho³, Woocheol Jeong³, Songhee Suk³, Song Choi³, Yongseong Kim⁴, Bongjun Ji^{5*}

¹Member, Graduate Student, Department of Industrial Management Engineering, Pohang University of Science and Technology, 77 Cheongam-Ro, Pohang, Gyeongsangbuk-do 37673, Republic of Korea

²Non-Member, CEO, Smart Geotech, 1 Kangwondaehak-gil, Chuncheon-si, Gangwon-do 24341, Republic of Korea

³Non-Member, Graduate Student, Department of Regional Infrastructure Engineering, Kangwon National University, 1 Kangwondaehak-gil, Chuncheon-si, Gangwon-do 24341, Republic of Korea

⁴Member, Professor, Department of Regional Infrastructure Engineering, Kangwon National University, 1 Kangwondaehak-gil, Chuncheon-si, Gangwon-do 24341, Republic of Korea

⁵Member, Assistant Professor, Department of Regional Infrastructure Engineering, Kangwon National University, 1 Kangwondaehak-gil, Chuncheon-si, Gangwon-do 24341, Republic of Korea

ABSTRACT

In order to monitor the stability of slopes, research on data-based slope failure prediction and early warning is increasing. However, most papers overlook the quality of data. Poor data quality can cause problems such as false alarms. Therefore, this paper proposes a two-step hybrid approach consisting of rules and machine learning models for quality control of data collected from slopes. The rule-based has the advantage of high accuracy and intuitive interpretation, and the machine learning model has the advantage of being able to derive patterns that cannot be explicitly expressed. The hybrid approach was able to take both of these advantages. Through a case study, the performance of using the two methods alone and the case of using the hybrid approach was compared, and the hybrid method was judged to have high performance. Therefore, it is judged that using a hybrid method is more appropriate than using the two methods alone for data quality control.

요 지

경사면의 안정성을 모니터링 하기 위해 데이터 기반으로 사면의 붕괴를 예측, 경보를 하려는 연구가 증가하고 있다. 하지만 대부분의 논문에서는 데이터의 품질에 대해 간과하고 있다. 이는 오경보와 같은 문제를 발생시킬 수 있다. 이에 본 논문에서는 사면에서 수집된 데이터의 품질관리를 위한 규칙과 기계학습 모델로 구성된 2 단계의 접근 방안을 제안하였다. 규칙 기반은 높은 정확도와 직관적인 해석이 가능하다는 장점이 있으며 기계학습 모델은 명시적으로 표현할 수 없는 패턴을 도출할 수 있다는 장점이 있으며 2단계의 접근 방안은 이 두 장점을 모두 취할 수 있었다. 사례연구를 통해 두 방법을 단독으로 사용하였을 경우와 2단계의 접근 방안을 사용하였을 때의 성능을 비교하였고 2단계 접근 방안이 높은 성능을 보이는 것으로 판단되었다. 따라서 데이터의 품질관리를 위해 단독으로 두 방법을 사용하는 것보다 2단계 접근 방안 방법을 사용하는 것이 적절할 것으로 판단된다.

Keywords : Slope failure, Data quality, Machine learning, Anomaly detection, Rule-based method

1. 서 론

경사면의 안정성을 모니터링하는 것은 매우 중요하다. 경사면의 안정성 모니터링을 통한 경사면 붕괴의 조기 감

Received 23 Feb. 2023, Revised 16 Mar. 2023, Accepted 24 Mar. 2023

*Corresponding author

Tel: +82-33-250-6464; Fax: +82-33-242-7355

E-mail address: bjji@kangwon.ac.kr (B. Ji)

지는 정보를 통해 경제적, 환경적 피해는 물론이고 인명 피해 또한 방지할 수 있는 기능을 한다. 또한 경사면 파손이 발생할 가능성을 적시에 예측할 수 있도록 도와 취약한 부분에 보강 조치를 취하도록 도울 수 있다는 측면과 사람이 직접 관찰이 어려운 산간 오지에 존재하는 경사면 등을 저렴한 운영비용으로 지속적으로 모니터링할 수 있다는 측면에서 유지관리에 큰 도움이 되고 있다. 하지만 IoT 장비 혹은 센서에서 수집하는 데이터는 태생적 한계로 다양한 데이터 품질 문제가 발생한다(Zhang et al., 2010). 다양한 원인 중 센서 오작동, 환경 요인, 네트워크 오류는 가장 대표적인 예시이다(Teh et al., 2020). 이 중 센서 오작동은 센서 데이터의 이상치 발생 중 가장 일반적인 원인이다. 이는 센서 데이터의 노화, 캘리브레이션(Calibration)의 실패, 고장 등의 이유를 포함한다(Rabatel et al., 2011). 온도, 습도 또는 전자기 간섭이나 실외에 위치한 센서의 경우 동,식물, 바람 등에 의한 간섭 또한 센서에 영향을 미칠 수 있다(Jesus et al., 2017). 또한 경우에 따라 여러 센서가 있는 복잡한 시스템의 경우 센서 간의 혼선으로 인해 데이터에 이상치가 측정될 수 있다(Zhang et al., 2018). 이러한 한계로 센서 데이터는 불안정한 수 밖에 없지만, 불안정한 데이터를 분석에 그대로 이용할 시 왜곡된 해석을 낳을 수 있다(Li and Parker et al., 2014). 예를 들어 뾰족한 점(Spike)과 같이 정상범위를 완전히 벗어난 값의 경우 그 수가 적더라도 평균값과 분산에 미치는 영향을 무시할 수 없다(Bosman et al., 2017; Dereszynski and Dietterich, 2011; Rassam et al., 2014). 특히 사면에서 수집된 데이터를 잘못 해석하는 경우 잘못된 정보와 불필요한 개입을 야기할 수 있으므로 데이터의 품질 관리는 중요하다. 사면의 붕괴는 지질학적, 수문학적, 기계적 요인을 포함한 다양한 요인이 복합적으로 영향을 미친다. 단층과 균열의 존재 뿐만 아니라 토양, 암석의 특성과 같은 지질학적 요인은 경사면의 안정성을 결정하는데 중요한 역할을 하며, 폭우, 동결, 눈이 녹는 현상과 같은 수문학적 요인은 수분 침투와 간극 압력을 증가시켜 사면 붕괴를 야기하기도 한다(Mori et al., 2017). 기상 조건에 따라 변하는 온도와 습도는 토양 수분 함량에 영향을 미치는 인자로 사면 붕괴에 간접적으로 영향을 미친다(Zhang et al., 2011). 따라서 본 연구에서는 사면에서 수집되는 정보, 특히 온도 데이터의 품질 관리를 위한 접근방안을 제공하고자 한다. 변위와 같이 사면의 붕괴를 직접적으로 모니터링할 수 있는 데이터의 경우, 이상치 감지는 사면 붕괴를 모니터링하기 위한

지표로 사용될 수 있기에 이상치 감지는 데이터 품질관리 보다는 사면의 정보에 활용하는 것이 적합하다. 반면 온도의 경우, 이상치를 제거하여 고품질의 데이터를 유지함으로써 이들이 변위에 어떻게 영향을 미치는 지 파악하는 기초 자료로 활용될 수 있다. 본 논문에서 데이터 품질 관리를 위한 2단계 접근 방식을 제안한다. 접근 방식의 첫 번째 단계는 규칙 기반이며 전문 지식과 미리 정의된 규칙을 사용하여 명시적으로 알려진 데이터 품질 문제를 식별하고 제거하도록 설계하였다. 두 번째 단계는 기계학습 기반이며 훈련된 모델을 사용하여 데이터의 이상을 감지하도록 설계하였다. 기계학습 기반의 이상치 감지 모델은 Long Short-Term Memory(LSTM) Autoencoder 기반(Liu et al., 2022)의 이상치 감지 모델을 사용하였다.

2. 관련 연구

본 섹션에서는 이상치 감지에 관해 널리 사용되는 세 가지 접근 방법인 규칙 기반 이상치 감지, 통계적 방법론 기반 이상치 감지, 기계학습 기반의 이상치 감지에 대해 설명한다.

2.1 규칙 기반 이상치 감지

규칙 기반 이상치 감지는 사전에 정의된 규칙을 데이터에 적용하는 이상치 감지 방법(Duffield et al., 2009)으로 여러 분야에 걸쳐 널리 사용되는 방법이다. 규칙 기반 이상치 감지는 규칙을 명시적으로 정의할 수 있을 때 효과적이다. 규칙을 정의하기 위해 도메인 전문가는 시스템이 정상적인 동작을 수행하는 조건을 식별하고, 이를 임계값, 빈도, 상승/하강 비율과 같은 패턴으로 추출하여 정상 데이터의 범위의 경계를 정의하는데 사용한다. 규칙이 정의된 후에는 정의된 범위를 벗어나는 모든 데이터 포인트에 대해 이상치로 판별한다. 규칙 기반 이상치 감지는 모든 유형의 이상치를 감지할 수 있는 규칙을 정의하기 어렵다는 단점이 있지만 직관적이고 강력한 이상치 감지 방법이다(Ilgun, 1992).

2.2 통계적 방법론 기반 이상치 감지

통계적 접근 방법 또한 이상치 감지에 널리 사용되는 방법이다. 가장 널리 알려진 통계적 이상치 감지 기법 중

하나는 햄펠(Hampel) 필터로 데이터의 중간 절대 편차 (Median Absolute Deviation;MAD)를 이용하는 방법이다 (Liu et al., 2004). 햄펠 필터는 시계열 데이터에서 이상치를 감지하는데 특히 유용하다. 이 방법은 각 데이터 포인트를 인접한 데이터 포인트 창(Window)의 중앙값과 비교한다. 중앙값에서 데이터 포인트의 절대 편차가 지정된 임계값보다 크면 이상치로 감지한다. 이상치 감지에 널리 사용되는 또 다른 통계적 방법은 Mahalanobis 거리(Mahalanobis, 1936)이다. 마할라노비스(Mahalanobis) 거리는 변수 간의 공분산을 고려하여 데이터 포인트와 분포 중심 사이의 거리를 측정하는 방법으로 아래의 수식과 같이 계산된다. x 는 각 데이터 포인트, μ 는 분포의 중심을 뜻하며 S 는 공분산 행렬을 나타낸다.

$$d_M = \sqrt{(x - \mu)^T S^{-1} (x - \mu)} \quad (1)$$

비정상적인 데이터 포인트는 분포 중심에서 큰 마할라노비스 거리를 가지며 마할라노비스 거리를 지정된 임계값과 비교하여 감지할 수 있다. 통계적 방법론은 널리 사용되는 방법론으로 높은 성능을 보여주지만 데이터의 기

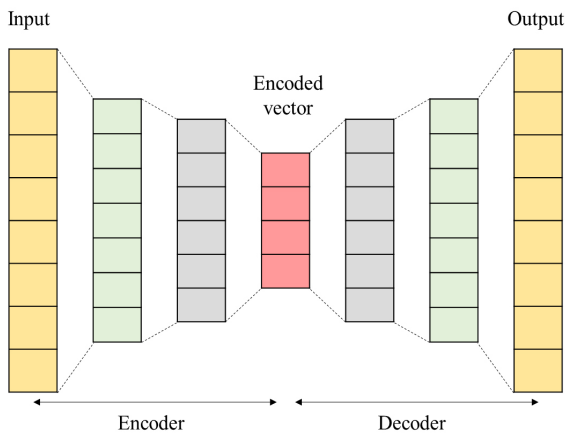


Fig. 1. Conceptualized architecture for Autoencoder

본 분포에 대한 가정을 필요로 한다. 또한 통계적 방법은 적절한 매개 변수를 선택하기 위해 전문 지식과 경험이 필요하다라는 단점이 있다.

2.3 기계학습 기반 이상치 감지

기계학습 기반의 이상치 감지는 데이터의 기본 분포에 대한 가정을 요구하지 않으며 데이터 자체에서 직접 데이터의 패턴, 관계를 추출해낼 수 있어 최근 들어 널리 사용되기 시작했다(Omar et al., 2013; Lane and Brodley, 1997; Nassif et al., 2021). 가장 널리 사용되는 시계열 데이터의 이상치 감지 방법은 Autoencoder 기반의 이상치 감지 방법이다(Chen et al., 2018). Autoencoder는 입력 데이터를 압축된 차원의 벡터로 표현하는 신경망 아키텍처이다. 압축된 벡터를 재구성하여 입력 데이터와 차이를 계산한 후 이를 미리 정의된 임계값과 비교하여 이상 감지에 사용한다. 센서 데이터, 음성 신호, 금융 데이터와 같은 다양한 시계열 데이터에 성공적으로 적용되었으며 다양한 변형이 제안되고 있다. Autoencoder는 Fig. 1과 같이 입력 데이터를 더 작은 차원으로 압축하는 Encoder와 압축된 벡터를 복원하는 Decoder로 구성되어 있다. 예를 들어 Input의 차원이 100이라면 Encoder를 통해 차원이 100 미만인 30차원으로 압축하는 부분이 Encoder, 30차원으로 압축된 벡터를 다시 100차원으로 복원하는 부분인 Decoder이다.

시계열 데이터에서 이상 징후를 감지하기 위한 또 다른 기계학습 기반의 접근방법은 Recurrent Neural Network(RNN) (Rumelhart et al., 1986)와 Long Short Term Memory(LSTM) (Hochreiter and Schmidhuber, 1997)과 같은 순서가 있는 데이터에 특화된 모델을 활용하는 방법이다. RNN 및 LSTM은 데이터의 장기적인 종속성을 모델링 할 수 있다(Fig. 2). 기계학습 기반의 접근 방식은 점점 대중화되고 있으며 적용 도메인 또한 넓어지고 있다.

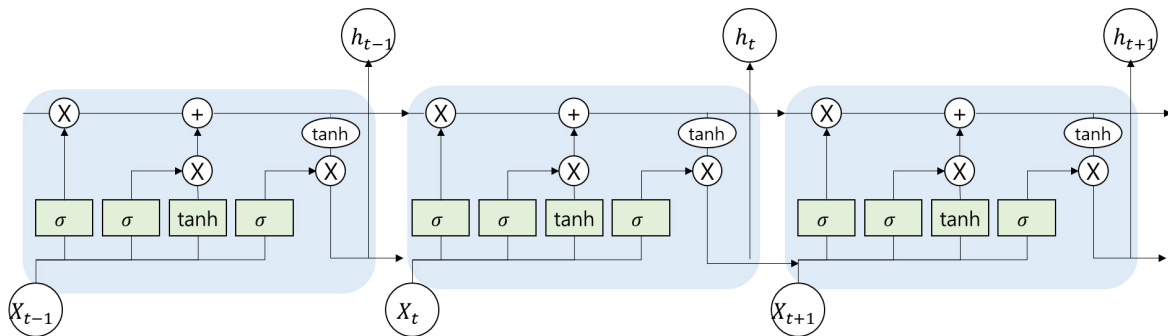


Fig. 2. Conceptualized architecture for LSTM

3. 연구 방법 및 절차

3.1 이상치 감지 모델

본 연구에서는 규칙 기반 및 기계학습 기반의 이상치 탐지 방법을 모두 사용하는 2단계 접근 방안을 사용한다. 2단계 접근 방식은 규칙 기반 또는 기계학습 기반의 접근 방식을 단일로 사용했을 때보다 더 강력한 성능을 보인다. 규칙 기반 방법은 명시적으로 정의된 규칙에 대해서는 잘 작동하나 정의되지 않은 복잡한 패턴의 이상치 데이터는 처리할 수 없다. 반면 기계학습 기반의 접근 방법은 복잡한 패턴의 이상치를 감지하는데 높은 성능을 보이기 때문에 상호 보완적으로 활용가능하다. 또한 2단계 접근 방안은 많은 부분에서 해석 가능한 이상치 감지가 가능하기 때문에 도메인 전문가들에게 더욱 선호된다. 기계학습 기반의 이상치 감지는 복잡한 수학적 연산에 의존하기 때문에 해석이 어려운 단점이 있다. 따라서 두 가지 방법을 결합함으로써 이상치 감지 시스템을 사용하는 사용자로 하여금 결과를 이해하고 신뢰할 수 있게 만들 수 있다. 마지막으로 규칙 기반의 방법을 통해 많은 양의 데이터를 일차적으로 필터링 하기 때문에 연산 속도가 빠르고 효과적, 효

율적인 이상치 감지가 가능하다. 아래의 그림은 본 연구에서 제안하는 2 단계의 이상치 감지 기법이다. 첫 번째 단계는 규칙을 기반으로 이상치를 감지하고 두 번째 단계에서는 기계학습 모델을 사용하여 이상치를 감지한다.

3.2 데이터 준비

온도와 습도는 사면의 안정성에 영향을 미치는 주요 인자 중 하나이다. 연구에서 사용된 데이터는 Fig. 3과 같이 Sensirion 제조사의 SHT-31 온습도 센서를 통해 수집되었다. 센서는 실내에 위치한 토사 표면 10cm 아래에 설치하였고 무선으로 데이터를 수신받기 위해 LPWAN(Low Power Wide Area Network)통신 프로토콜을 사용하였다.

저전력 장거리 통신에 적합한 LoRa(Long Range) 네트워크를 구축하였으며 Fig. 4에서 보이는 바와 같이 2023년 1월 4일부터 2023년 1월 16일 까지 2초 간격으로 데이터를 수집하여 451,459개의 데이터가 수집되었다. 데이터 측정 결과 1월 9일 이후로 전, 후 데이터와 값의 차이가 비정상적으로 큰 Spike 형태의 데이터 이상치가 감지되었다. 이는 센서의 오류와 전파 간섭으로 인해 발생하는 전형적인 이상치의 형태이다.



Fig. 3. Sensor used for collecting data

4. 이상치 감지 결과 및 분석

4.1 규칙 기반 이상치 감지 모델의 규칙 추출

본 데이터가 수집된 위치는 강원도 춘천시로 겨울철 평년 최고 기온은 섭씨 17.2도이다. 온도 센서가 설치된 위치는 실내이며, 설정가능한 최대 난방 온도가 28도 임을 감안하면 28도를 초과하는 온도가 관측되는 것은 불가능하다는 규칙을 생성할 수 있었다. 이에 28도를 초과하는 데이터는 이상치로 판단하고 아래와 같이 이상치를 제거

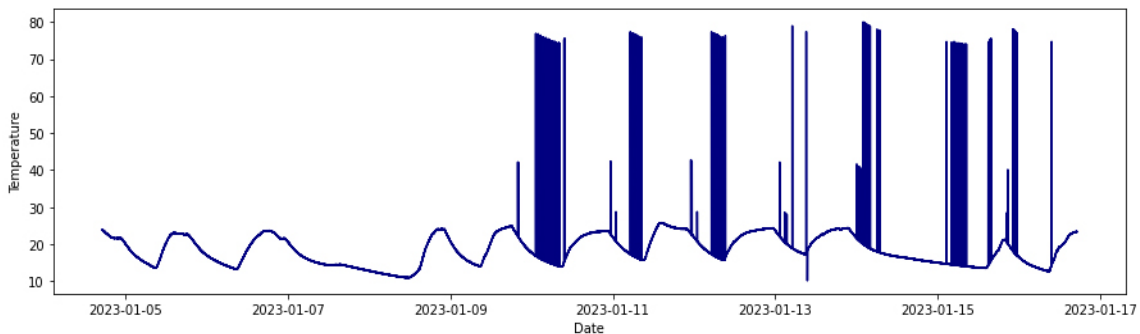


Fig. 4. Visualized collected data (Temperature)

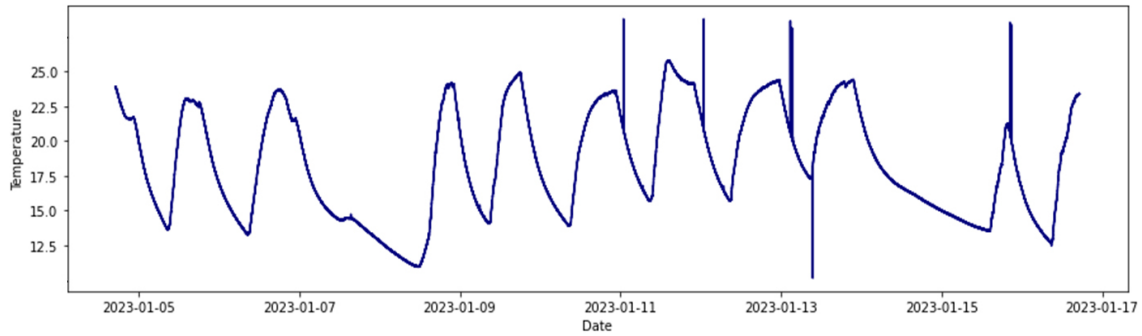


Fig. 5. Data which anomaly is deleted usign rule-based model

하였다(Fig. 5).

많은 비율의 이상치가 제거 되었지만 1월 13일 10도를 가리키는 음의 방향으로의 Spike나 28도 미만의 값을 갖는 이상치는 제거할 수 없는 것을 알 수 있다.

4.2 이상치 감지 모델 학습

단변량 시계열 데이터에서 이상치 감지를 위해 규칙 기반의 이상치 감지 방법과 기계학습 기법의 이상치 감지 기법을 2 Phase로 적용하였다. 본 연구에서 사용한 기계학습 모델은 LSTM Autoencoder이다. LSTM Autoencoder는 시계열 데이터에서 이상 감지에 사용되는 Neural Network 아키텍처로, 아래의 그림과 같이 인코더와 디코더 네트워크로 구성된다. 인코더 네트워크는 입력되는 시계열 데이터를 저차원인 잠재 공간(Latent Space)으로 매핑하는데 사용된다. 디코더 네트워크는 잠재 공간에서 원래의 시계열 데이터를 재구성하는데 사용한다. LSTM Autoencoder 모델은 인코더와 디코더 모두에서 LSTM셀을 활용한다. LSTM 셀은 시계열 데이터와 같은 순차 데이터를 처리하는데 적합한 셀로, 내부 메모리에 정보를 선택적으로 저장하고 검색하여 데이터의 장기 종속성을 추출할 수 있다. LSTM Autoencoder의 학습은 입력 시계열 데이터와 재구성된 시계열 데이터 사이의 차이를 최소화 하는 방향으로 이루어진다. 입력값과 재구성 값 간의 차이는 Mean Squared Error를 사용하여 계산된다(식 (2)). 특히 LSTM Autoencoder는 대표적인 비지도학습(Unsupervised learning) 방법으로 이상치에 대한 Labeling을 필요로 하지 않는다. 즉, 정상치의 비율만 높다면 추가로 어떤 데이터가 이상치인지에 대해 사람이 정해주지 않아도 이상치 감지가 가능하다. 본 연구에서도 이상치가 포함된 데이터를 학습시켜 재구성된 값과 비교를 통해 이상치 판별을 하고자 한다. Fig. 6는 본 연구에서 활용된 LSTM Autoencoder의 구조이다.

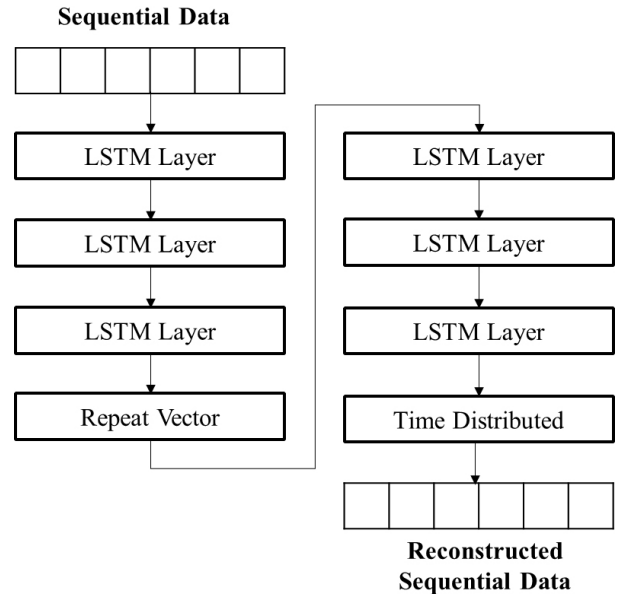


Fig. 6. Architecture of the machine learning model used in this study

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (2)$$

모델이 학습된 후에는 시계열 데이터를 다시 입력하여 이상치를 감지하는데 사용할 수 있다. 이상 감지 프로세스는 입력된 시계열 데이터와 재구성된 시계열 데이터간의 Error를 계산한 후, Error의 값이 특정 임계값 높으면 입력된 시계열의 값이 정상적인 동작에서 벗어났다고 판단할 수 있다. 학습 때와 마찬가지로 Error를 계산하기 위해서 Mean Squared Error를 사용하였으며 이때 임계값은 학습 데이터의 평균에 2 sigma 또는 3 sigma를 더한 값으로 설정하였다. 이상 감지에서 2 sigma를 더한 값은 일반적으로 널리 통용되는 값으로, 정규분포를 따를 경우 95%의 데이터는 2 sigma 내에 존재한다. 3 sigma는 99%의 데이터가 존재하는 경우로, 본 연구에서는 데이터의 수는 풍부한 반면, 이상치 값으로 인해 발생할 수 있는 비용은 크기 때문

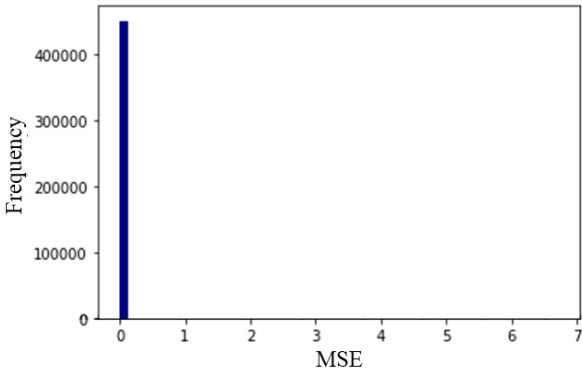


Fig. 7. Distribution of the reconstructed error

Table 1. Statistics of MSE in train and test data

Statistics	MSE value
Mean	0,004
Standard deviation	0,010
Max value	6,704

에 보수적으로 이상치 감지를 수행하고자 2 sigma를 사용하였다. 아래의 Fig. 7과 같이 대부분의 MSE가 0에 가까운 값을 가지는 것으로 나타났으나 일부 MSE의 경우 최대 6.7의 값을 갖는 것으로 나타났다.

전체 데이터에서의 입력 데이터와 재구성된 예측값 간의 차이인 MSE에 대한 주요 통계치는 아래의 Table에서 확인할 수 있다.

4.3 이상치 감지 모델 성능 평가

본 연구에서는 이상치 감지 모델의 성능 평가를 위한 지표로 Precision를 사용하였다. 이상치 감지 모델의 성능 평가를 위해서 널리 사용되는 지표는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-score 등이 있다. 정확도는 총 예측 수에 대한 참값의 비율로 모델 예측의 전반적인 정확성을 측정하는 간단하고 일반적인 성능지표이다. 그러나 이상치와 같이 클래스가 불균형할 때 즉, 대부분의 값이 정상치 일 때 정확도는 오해의 소지가 있을 수 있다.

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (3)$$

정밀도는 모델이 수행한 긍정적 예측(Positive, 본 연구의 경우 이상치로 판단)의 총 수 중에 실제로 긍정으로 밝혀진 비율을 측정하는 성능지표이다. 이상 감지에서는 예

측된 이상치 중 실제 이상치 비율을 나타낸다.

$$Precision = \frac{\text{True positive}}{\text{True positive} + \text{False positive}} \quad (4)$$

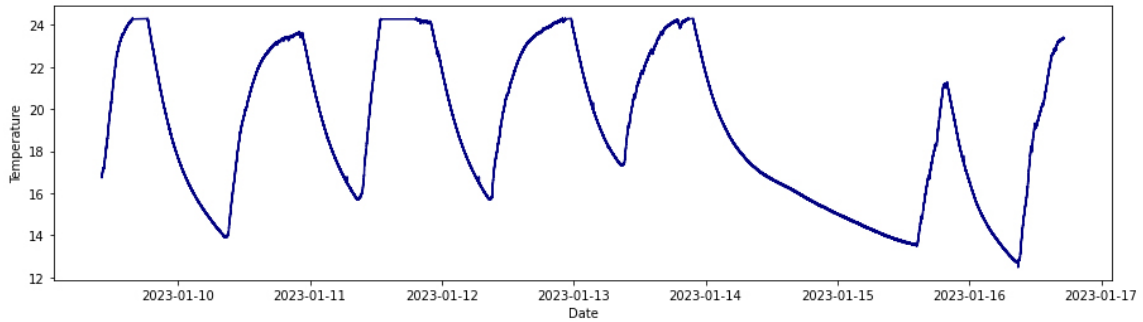
재현성은 데이터 셋의 전체 이상치 중에서 실제로 감지한 이상치 비율을 측정하는 성능지표이다. 재현율이 높다는 것은 이상치를 실제로 이상하다고 식별할 수 있는 성능이 높다는 것을 뜻한다.

$$Recall = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \quad (5)$$

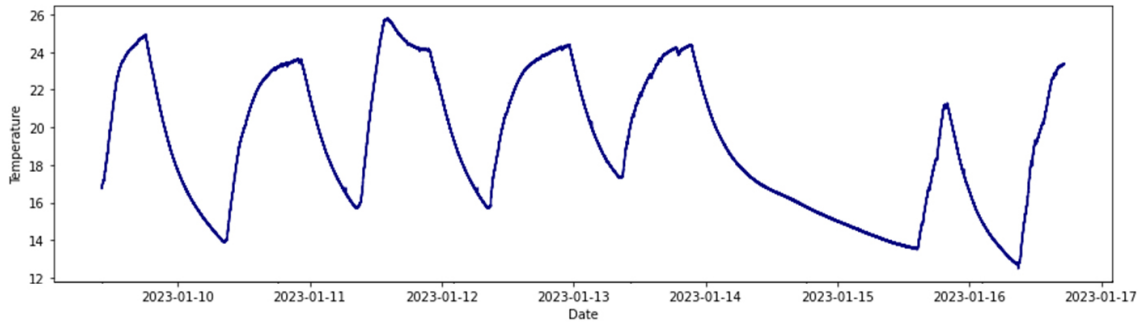
F1점수는 정밀도와 재현율을 모두를 고려하여 단일 성능 지표로 사용하는 성능평가 지표로 아래와 같이 계산된다. F1 점수의 값은 0에서 1 사이의 값을 지니며 1에 가까울수록 높은 성능을 보인다.

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} \quad (6)$$

본 논문에서는 이상치를 이상치라고 판별하는 것이 중요하다. 즉 이상치가 아님에도 이상치라고 판별할 때의 비용보다, 이상치를 이상치가 아니라고 판별할 때의 비용이 더 크다. 이런 경우 Recall, Precision 순으로 가장 중요한 성능 지표가 된다. 규칙 기반 방법만 적용하였을 경우에는 이상치를 감지하지 못한 경우가 다수 존재했다. 반면 LSTM Autoencoder 기반의 이상치 감지는 Spike 형태의 이상치 감지에 높은 성능을 보였는데 Fig. 8의 (a)와 (b) 모두 LSTM Autoencoder를 이용하여 이상치를 제거한 결과이다. (a)와 (b) 모두에서 Spike인 이상치가 성공적으로 감지되었음을 알 수 있다. 즉 두 접근 방법 모두 높은 Recall을 보였다. 반면 Fig. 8의 (a)와 (b)는 24도 이상의 데이터에 대해 다른 결과를 보인다. 1월 9일과 10일 사이의 데이터, 1월 12일의 데이터의 경우 이 현상이 두드러지는데 규칙기반의 이상치 감지 없이 기계학습기반의 이상치 감지만을 수행한 경우, 24도 이상의 데이터 포인트는 많은 경우 이상치로 판단, 제거하는 경향을 보인다. 반면 규칙 기반 이상치 감지를 통해 28도를 초과하는 데이터를 이상치로 판단 제거한 후 학습한 경우, 24도 이상의 데이터 포인트를 이상치로 감지하지 않고, Spike만 제거하는 모습을 보였다. 즉 2 단계 접근 방법을 적용한 경우가 단순 기계학습의



(a) Without rule-based deletion of anomaly



(b) After applying rule-based anomaly deletion

Fig. 8. Result of removing values that has greater MES than threshold

접근방법을 적용한 경우보다 높은 Precision을 보임을 알 수 있었다.

5. 결론

본 연구에서는 사면 안정성 모니터링에서의 데이터 품질 관리를 위해 2단계 접근방안을 제안하여 이상치를 제거하는 연구를 수행하였으며 그 결과를 요약하면 다음과 같다.

- (1) 제안된 접근 방안은 규칙 기반 이상 감지와 LSTM Autoencoder 기반 이상 감지를 결합하여 이상 데이터 포인트를 효과적으로 식별할 수 있었다.
- (2) 규칙 기반 이상 감지 방법은 미리 정의된 규칙과 임계값을 기반으로 이상치를 감지하는데 사용되는 반면, LSTM Autoencoder 기반 방법은 시계열의 입력 데이터와 재구성된 출력데이터의 편차를 기반으로 이상치를 감지하는데 사용
- (3) 실제 온도 데이터에 대한 실험 결과는 제안된 2 단계의 접근 방안이 이상 감지 정확도 및 오경보율 측면에서 개별 방법보다 성능이 우수함을 보였다.
- (4) 규칙 기반 방법은 높은 확률로 큰 편차를 잡아내는데 성공한 반면, LSTM Autoencoder는 규칙 기반 방법으

로 포착하기 어려운 상대적으로 미묘한 이상 징후를 감지할 수 있었다.

- (5) 이는 둘 중 하나의 방법만을 사용했을 때 얻을 수 없는 결과로 2단계 접근 방안을 활용하여 두 방법의 장점을 결합하고 보다 강력하고 정확한 이상 감지 시스템을 만들 수 있었다.

본 연구 결과는 추후 이상치를 복원하는데 기초자료로 활용되거나, 안정성 모니터링에서 잘못된 경보로 발생하는 비용을 감소시킬 수 있을 것으로 기대된다.

Acknowledgement

This research was supported by “Ministry of the Interior and Safety” R&D program(RS-2022-00155667).

References

1. Bosman, H. H., Iacca, G., Tejada, A., Wörtche, H. J. and Liotta, A. (2017), “Spatial anomaly detection in sensor networks using neighborhood information”, *Information Fusion*, Vol.33, pp.41-56.
2. Chen, Z., Yeo, C. K., Lee, B. S. and Lau, C. T. (2018), “Autoencoder-based network anomaly detection”, *In 2018*

- Wireless telecommunications symposium (WTS)*, Arizona, pp.1-5.
3. Dereszynski, E. W. and Dietterich, T. G. (2011), "Spatiotemporal models for data-anomaly detection in dynamic environmental monitoring campaigns", *ACM Transactions on Sensor Networks*, Vol.9, No.1, pp.1-36.
 4. Duffield, N., Haffner, P., Krishnamurthy, B. and Ringberg, H. (2009), "Rule-based anomaly detection on IP flows", *In IEEE INFOCOM 2009*, Rio de Janeiro, pp.424-432.
 5. Hochreiter, S. and Schmidhuber, J. (1997), "Long short-term memory", *Neural computation*, Vol.9, No.8, pp.1735-1780.
 6. Ilgun, K. (1992), *A real-time intrusion detection system for unix*, Master Thesis, University of California Santa Barbara.
 7. Jesus, G., Casimiro, A. and Oliveira, A. (2017), "A survey on data quality for dependable monitoring in wireless sensor networks", *Sensors*, Vol.17, No.9, pp.2010.
 8. Lane, T. and Brodley, C. E. (1997), "An application of machine learning to anomaly detection", *In Proc. of the 20th national information systems security conference*, Baltimore, Vol.377, pp.366-380.
 9. Li, Y. and Parker, L. E. (2014), "Nearest neighbor imputation using spatial-temporal correlations in wireless sensor networks", *Information Fusion*, Vol.15, pp.64-79.
 10. Liu, H., Shah, S. and Jiang, W. (2004), "On-line outlier detection and data cleaning", *Computers & Chemical Engineering*, Vol.28, No.9, pp.1635-1647.
 11. Liu, P., Sun, X., Han, Y., He, Z., Zhang, W. and Wu, C. (2022), "Arrhythmia classification of LSTM autoencoder based on time series anomaly detection", *Biomedical Signal Processing and Control*, Vol.71, Part B, 103228.
 12. Mahalanobis, P. C. (1936), "On the generalised distance in statistics", *In Proc. of the National Institute of Science of India*, Kolkata, Vol.12, pp.49-55.
 13. Mori, A., Subramanian, S. S., Ishikawa, T. and Komatsu, M. (2017), "A case study of a cut slope failure influenced by snowmelt and rainfall", *Procedia engineering*, Vol.189, pp.533-538.
 14. Nassif, A. B., Talib, M. A., Nasir, Q. and Dakalbab, F. M. (2021), "Machine learning for anomaly detection: A systematic review", *Ieee Access*, Vol.9, pp.78658-78700.
 15. Omar, S., Ngadi, A. and Jebur, H. H. (2013), "Machine learning techniques for anomaly detection: an overview", *International Journal of Computer Applications*, Vol.79, No.2.
 16. Rabatel, J., Bringay, S. and Poncelet, P. (2011), "Anomaly detection in monitoring sensor data for preventive maintenance", *Expert Systems with Applications*, Vol.38, No.6, pp.7003-7015.
 17. Rassam, M. A., Maarof, M. A. and Zainal, A. (2014), "Adaptive and online data anomaly detection for wireless sensor systems", *Knowledge-Based Systems*, Vol.60, pp.44-57.
 18. Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986), "Learning representations by back-propagating errors", *Nature*, Vol.323, No.6088, pp.533-536.
 19. Teh, H. Y., Kempa-Liehr, A. W. and Wang, K. I. K. (2020), "Sensor data quality: A systematic review", *Journal of Big Data*, Vol.7, No.1, pp.1-49.
 20. Zhang, H., Liu, J. and Pang, A. C. (2018), "A Bayesian network model for data losses and faults in medical body sensor networks", *Computer Networks*, Vol.143, pp.166-175.
 21. Zhang, L. L., Zhang, J., Zhang, L. M. and Tang, W. H. (2011), "Stability analysis of rainfall-induced slope failure: a review", *Proc. of the Institution of Civil Engineers-Geotechnical Engineering*, Vol.164, No.5, pp.299-316.
 22. Zhang, Y., Meratnia, N. and Havinga, P. J. (2010), "Ensuring high sensor data quality through use of online outlier detection techniques", *International Journal of Sensor Networks*, Vol.7, No.3, pp.141-151.