The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 23, No. 2, pp.103-111, Apr. 30, 2023. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2023.23.2.103
JIIBC 2023-2-14

LSTM-GAN 기반 이상탐지 모델을 활용한 시계열 데이터의 동적 보정기법

A Dynamic Correction Technique of Time-Series Data using Anomaly Detection Model based on LSTM-GAN

정한석*, 김한준**

Hanseok Jeong*, Han-Joon Kim**

요 약 본 논문은 시계열 데이터에 존재하는 이상값을 정상값으로 변환하는 새로운 데이터 보정기법을 제안한다. 최근 IT기술의 발전으로 센서를 통해 방대한 시계열 데이터가 수집되고 있다. 하지만 센서의 고장, 비정상적 환경으로 인해, 대부분의 시계열 데이터는 다수의 이상값을 포함할 수 있다. 이상값이 포함된 원천 데이터를 그대로 사용하여 예측모델을 구축하는 경우, 고신뢰도의 예측 서비스가 실현되기 어렵다. 이에 본 논문은 LSTM-GAN 모델을 활용하여 원천 시계열 데이터에 존재하는 이상값을 탐지하고, DTW(Dynamic Time Warping) 및 GAN 기법을 결합하여 분할된 윈도우 단위로 이상값을 정상값으로 보정하는 기법을 제안한다. 기본 아이디어는 탐지된 이상값이 포함된 윈도우에 인접한 정상분포 데이터의 통계정보를 DTW에 적용하여 연쇄적으로 GAN 모델을 구축하여 정상적 시계열 데이터를 생성하는 것이다. 오픈 NAB 데이터를 활용한 실험을 통해, 우리는 제안 기법이 기존 2개의 보정기법보다 성능이 우수함을 보인다.

Abstract This paper proposes a new data correction technique that transforms anomalies in time series data into normal values. With the recent development of IT technology, a vast amount of time-series data is being collected through sensors. However, due to sensor failures and abnormal environments, most of time-series data contain a lot of anomalies. If we build a predictive model using original data containing anomalies as it is, we cannot expect highly reliable predictive performance. Therefore, we utilizes the LSTM-GAN model to detect anomalies in the original time series data, and combines DTW (Dynamic Time Warping) and GAN techniques to replace the anomaly data with normal data in partitioned window units. The basic idea is to construct a GAN model serially by applying the statistical information of the window with normal distribution data adjacent to the window containing the detected anomalies to the DTW so as to generate normal time-series data. Through experiments using open NAB data, we empirically prove that our proposed method outperforms the conventional two correction methods.

Key Words : anomaly detection, data quality, deep learning, LSTM, GAN, time-series data

*준회원, 서울시립대학교 전자전기컴퓨터공학과
**정회원, 서울시립대학교 전자전기컴퓨터공학부(교신저자)
접수일자 2023년 3월 23일, 수정완료 2023년 4월 3일
게재확정일자 2023년 4월 7일

Received: 23 March, 2023 / Revised: 3 April, 2023 /
Accepted: 7 April, 2023
*Corresponding Author: khj@uos.ac.kr
Dept. of Electrical and Computer Engineering, University of Seoul, Korea

Ⅰ. 서 론

최근 사물인터넷(IoT)의 확산으로 스마트 빌딩, 스마트 팩토리, 지능형 발전소 및 데이터 센터 등의 환경에서 네트워크로 연결된 센서를 통해 상당한 양의 시계열 데이터가 수집되고 있다[1][2]. 이 시계열 데이터는 특정 공간 및 환경을 지속적으로 모니터링하면서 이상 현상을 감지하여 사고 대응에 요긴하게 활용될 수 있다. 그런데 시계열 데이터 기반 모니터링은 지속적으로 육안 관찰해야 하는 부담이 크기 때문에, 최근 이를 자동화하기 위해 기계학습 기반 예측모델의 구축에 관한 연구가 활발하다[3][4]. 예를 들어, 최근 공장 설비의 진동, 속도, 온도 등의 변화를 감지하여 화재와 같은 긴급한 상황을 대응하기 위한 예측모델 구축에 관한 연구가 있다[5]. 최근 딥러닝 기술은 이상탐지 도메인에 적용되어 유의미한 변수값의 실시간 변화를 예측하는 예측모델을 통해, 고신뢰도 예측 서비스를 실현하는데 크게 기여하고 있다[3].

기본적으로 신뢰도가 높은 예측모델을 구축하기 위해서는 양질의 학습데이터가 필요하다[6][7]. 하지만 특정 환경에서 센서를 통해 수집된 데이터는 센서 자체의 오염, 오작동, 환경 요인 등 다양한 이유로 정상적이지 않은 이상값을 포함할 수 있다[8]. 이상값이 포함된 데이터를 정제 또는 보완하지 않고 그대로 이용하여 예측모델을 구축한다면, 유의미한 예측 서비스를 기대할 수 없다[9]. 따라서 시계열 데이터를 활용한 예측모델의 신뢰성을 확보하기 위해서, 최근 진행되고 있는 기계학습 기반 시계열 데이터의 품질 개선에 관한 연구는 그 가치와 활용성이 매우 높다[10][11][12].

본 연구는 시계열 데이터에 내재된 이상값을 탐지하여 이를 정상값으로 보정하는 기법을 제안한다. 여기서 이상값의 탐지는 기존의 LSTM-GAN 모델[13]을 활용한다. 제안 기법의 우수성을 보이기 위해서, 우리는 다양한 분야를 포괄하는 오픈 소스 데이터인 NAB 데이터셋[14]을 사용하였으며, 제안 기법을 통해 보정된 데이터로 구축한 예측모델의 성능이 향상됨을 실험적으로 보인다.

본 논문의 구조는 다음과 같다. II절은 이상탐지 및 데이터 보정과 관련된 기존 연구를 소개하고, III절은 시계열 이상값 데이터에 대한 동적 보정 기법을 소개한다. IV절은 NAB 데이터셋에 제안 기법을 적용하여 최종 구축한 예측모델의 성능 비교를 통해 제안 기법의 우수성을 서술하고, V절에서 결론 및 향후 연구를 서술한다.

Ⅱ. 관련 연구

1. 이상탐지 연구

이상탐지(anomaly detection)는 정상적 데이터의 분포에서 크게 벗어난 이상값을 식별하는 이진 분류 문제로 정의된다. 이를 위한 전통적인 기법으로서, PCA(Principal Component Analysis)[15]와 PLS(Partial Least Squares)[16]는 선형 방식의 차원 축소 기반 이상 감지 기법이다. 하지만 이는 주어진 데이터가 정규 분포로 생성됨을 가정하고 있고, 변수간 상관관계가 높은 데이터에만 효과적으로 작동한다. 또 다른 방식으로서 거리 기반 이상탐지 기법이 있는데, 흔히 k-Nearest Neighbor (k-NN)[17] 알고리즘이 사용된다. k-NN 기법은 클러스터링을 위한 이웃(neighbor) 개수를 사람이 지정해주어야 하는데, 이를 위해서 대상 데이터의 도메인에 대한 충분한 사전 지식을 요구한다.

최근 제안된 LSTM[18], 오토인코더(auto-encoder)[19]와 같은 딥러닝 기반 비지도 이상탐지 기법이 뛰어난 성능을 보임에 따라 많은 큰 관심을 받고 있다. 우리는 여기에 유사 데이터를 생성하는 GAN(Generative Adversarial Network)[20] 딥러닝 기술을 결합할 수 있다. GAN 딥러닝 모델은 생성 데이터 분포와 실제 데이터의 분포를 일치시키는 특징을 가지고 있어, 이를 이상탐지 목적에 활용할 수 있다. 최근에 제안된 LSTM과 GAN 기술을 결합한 LSTM-GAN (일명 TAnoGAN)모델은 시계열 데이터에서 이상값을 탐지하는데 효과적으로 사용되고 있다[13]. 이에 본 연구는 시계열 데이터에 포함되어 있는 이상값을 보정하는 것이 목적인 바, 이상탐지를 위해서 LSTM-GAN 모델을 활용한다.

2. 데이터 보정 연구

시계열 데이터 내 탐지된 이상값은 정상값으로 보정하는 연구는 이상탐지 관련 연구 대비 활발하지 않으나, 최근 기초 통계량 및 기계학습을 활용한 연구가 진행되었다. 이상값의 보정을 위한 전통적인 기법은 평균값, 중앙값, 최빈값 등의 기초 통계량을 활용하거나 과거 관측값과의 오차를 활용한 회귀분석 모형 등을 포함한다[21]. 하지만 이는 시간의 흐름에 따른 시계열 데이터의 특성을 효과적으로 반영하지 못하여 그 정확도가 높지 못하다. 또한 서포트벡터회귀(support vector regression)와 같은 기계학습 기법들을 이용하여 데이터 보정 과정을 통해 시계열 예측모델의 성능을 높였다[22]. 본 연구는 서포

트벡터회귀 기반 시계열 데이터 보정기법보다 우수한 성능을 성취하기 위해, GAN과 DTW(Dynamic Time Warping)를 결합한 방안을 제안한다.

Ⅲ. 제안 기법

이상값의 보정을 위해서는 우선 이상값을 탐지하는 것이 선행되어야 한다. 앞서 언급한 바와 같이, 본 연구는 기존 이상탐지 모델인 LSTM-GAN[13]을 이용하여 이상탐지를 수행한다. 그리고나서, 이상값을 정상값으로 대체하기 위해, Soft DTW 기법[23]을 반영한 손실함수를 정의하여 정상값과 유사한 데이터를 생성한다. 기본적으로 DTW 기법은 두 개의 시계열 데이터가 서로 얼마나 유사한지 평가할 때 자주 사용되는데, 이는 시계열 데이터의 시간축이 상이하더라도 시계열 분포 형태의 유사도를 측정할 수 있는 장점을 가진다[24]. 하지만 이는 시계열 값 비교를 위해 단 하나의 매핑 경로를 고려하여 시계열 간 거리를 계산하므로, 이를 가지고 미분이 가능한 손실함수(loss function)를 만들기 어렵다. 이를 개선한 Soft DTW 기법은 다수의 매핑 경로들을 동시에 고려하기 때문에 딥러닝 모델 학습 메커니즘 내의 손실함수의 인자로 활용이 가능하다[23]. 그림 1은 제안하는 보정기법의 전체적인 프로세스를 보여준다.

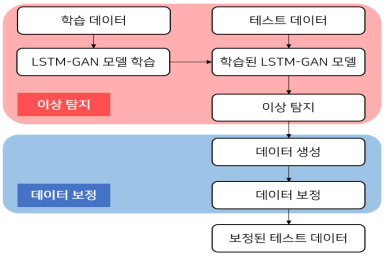

그림 1. 시계열 데이터에 대한 이상탐지 및 데이터 보정 과정
Fig. 1. A process of anomaly detection and correction for time series data

1. 이상탐지

LSTM-GAN 모델 기반 이상탐지 기법은 시계열 데이터 처리를 위해 생성자(generator)와 판별자(discriminator) 모두 LSTM 모델을 이용하며, 이들 상호 간의 적대적 학습을 통해 주어진 데이터셋의 분포를 학습한다. 기본적으로 LSTM-GAN 모델은 효과적인 학습을 위해 전체 시계열 데이터를 슬라이딩 윈도우(sliding window) 단위로 분할한다.

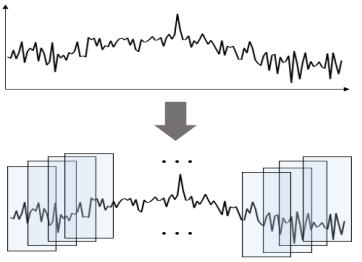

그림 2. LSTM-GAN 모델 학습을 위한 시계열 데이터에 대한 슬라이딩 윈도우 기반 분할
Fig. 2. A sliding window-based partitioning for time series data for LSTM-GAN model training

시계열 데이터에 대한 LSTM-GAN 모델의 목적함수 $V(D,G)$는 식(1)과 같다.

$$\min_G \max_D V(D,G) = E_{x\sim p_x}[logD(x)] + E_{z\sim p_z}[log(1-D(G(z))] \quad (1)$$

여기서, x는 실제 데이터(윈도우) 분포 P_x에서 가져온 하나의 데이터 값(윈도우)을 의미하고, 잠재 벡터공간 Z에서 선택된 노이즈 z는 데이터분포 p_z를 따르는 데이터(윈도우)를 의미한다. 실제 데이터에 대한 판별자 D의 결정이 $E_{x\sim p_x}[logD(x)]$을 최대화함으로써 그 결과가 정확하기를 기대하며, $z\sim p_z$의 분포로 생성된 가짜 데이터(윈도우) $G(z)$가 주어졌을 때, 판별자의 예측 결과는 $D(G(z))$가 된다. 식(1)의 구현을 위해서, 우리는 [13]에 근거하여 생성자(G)는 각각 32, 64, 128개의 은닉 유닛을 가지는 3개의 LSTM 레이어를 포함하고, 판별자(D)는 100개 은닉 유닛을 가지는 단일 LSTM 레이어를 포함한다.

그림 3은 시계열 데이터 내 이상값의 탐지를 위한 LSTM-GAN 모델 구조를 보여준다. LSTM-GAN 모델은 이상탐지를 수행하기 위해서, 실제 데이터(윈도우) x를 잠재 벡터공간에 역매핑(inverse mapping)하여 얻은 노이즈 데이터 z를 가지고 생성한 $G(z)$와 실제값을 포함하는 데이터(윈도우) x간의 잔차 손실(residual loss) $R(x,G(z))$를 계산하고, 또한 진위 판별에 따른

차별 손실(discrimination loss) $D(x, G(z))$를 계산한다. 여기서 학습을 위해 필요한 최종 손실함수는 2개 손실값의 합으로 정의된다. 또한 LSTM-GAN 모델이 산출한 잔차 손실과 차별 손실은 이상값 탐지를 위한 이상점수(anomaly score) $A(x, G(z))$를 정의하는데 활용된다. (식(2) 참조)

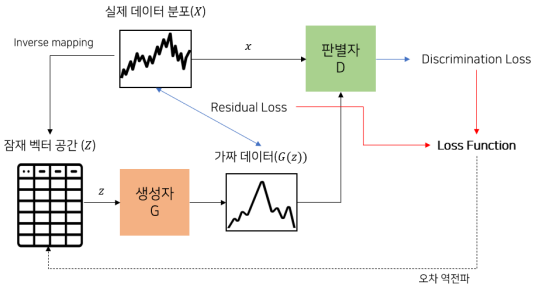

그림 3. 시계열 데이터 내 이상탐지를 위한 LSTM-GAN 모델
Fig. 3. LSTM-GAN model for time series-anomaly detection

$$A(x, G(z)) = (1-\gamma) \cdot R(x, G(z)) + \gamma \cdot D(x, G(z)) \quad (2)$$

여기서, γ는 $R(x, G(z))$와 $D(x, G(z))$간의 상대적 중요도를 결정하는 비중값이다. 이상점수가 임계값(α) 이상이면, 해당하는 시계열 값이 '이상'으로 탐지된다[13]. 일반적인 데이터 분포를 학습한 생성자와 판별자는 이상값이 포함된 실제 데이터(윈도우)에 대해서는 이상점수를 커지게 한다. 결과적으로, LSTM-GAN 모델은 주어진 임계값보다 큰 이상점수를 가지는 데이터(윈도우)를 '이상(Anomaly)' 클래스로 분류하게 된다.

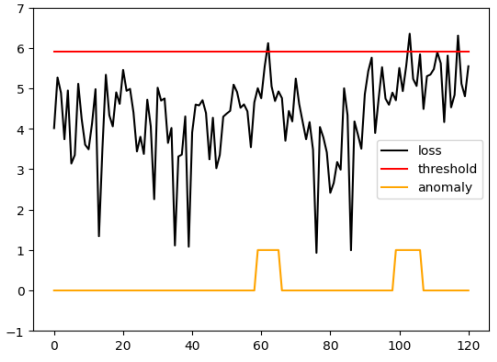

그림 4. 이상점수를 이용하여 이상탐지를 수행한 예
Fig. 4. Result of anomaly detection using anomaly score

그림 4는 식(2)의 이상점수로 이용하여 이상탐지를 수행한 예시를 보여준다. 검정선은 이상점수를 의미하고, 적색선은 임계값을 의미한다. 최하단 주황선은 정답(ground-truth)에 해당하는 데이터를 표시한 것이며, 정상값을 0으로, 실제 이상값을 1로 표시하여 이상 유무를 나타낸다.

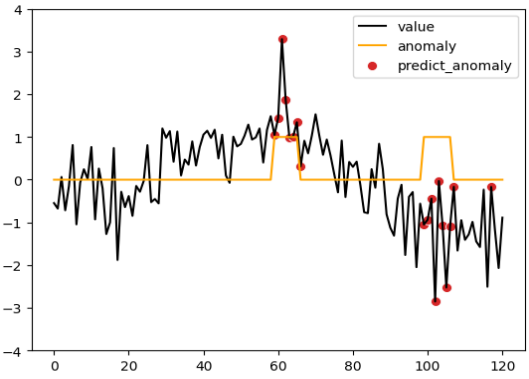

그림 5. 이상점수로 탐지한 이상값을 데이터에 매핑 결과
Fig. 5. Result of mapping the anomalies detected by the anomaly score to the data

그림 5는 이상점수를 통해 이상탐지한 결과를 실제 데이터에 중첩한 결과를 보여준다. 검정선은 이상값을 포함하는 실제 데이터를 의미하고, 적색점은 LSTM-GAN 모델이 탐지한 이상값임을 표시한다. 우리는 이러한 이상 유무를 표시한 정답선과의 비교를 통해, LSTM-GAN 모델이 이상값을 효과적으로 탐지할 수 있음을 확인하였다.

2. 데이터 생성 및 보정

탐지된 이상값에 대한 정상값으로의 보정은 정상값으로 간주되는 대체값을 생성함으로써 이루어진다. 시계열 데이터는 시간 연속성을 가지기 때문에, 특정 시점의 데이터값은 가장 가까운 시점의 데이터값들의 영향을 많이 받을 수밖에 없다. 따라서 정상값과 유사한 값을 생성하기 위해서, 우리는 이상값이 포함된 윈도우와 바로 인접한 정상 분포 데이터를 포함하는 윈도우를 활용한다. 여기서 윈도우의 크기는 샘플 개수로 정의하며, 기본값은 60이다. 그리고 이상값 보정은 윈도우 단위로 이루어진다.

그림 6은 시계열 데이터 내 이상값의 보정을 위해 인접한 정상 분포를 가지는 데이터(윈도우)를 활용한 방안을 보여준다. 이상값이 포함된 윈도우의 보정을 위해 우리는 3가지 경우를 고려한다. 첫째는 이상값이 포함된

윈도우가 인접하지 않고 홀로 존재하는 경우이다. 이상값이 탐지된 7번째 윈도우 보정을 위해 이전 시점의 정상 분포 데이터를 가지는 6번째 윈도우의 통계값을 반영하여 실제 데이터와 유사한 데이터를 생성·보정한다. 둘째는 이상값이 포함된 2개의 윈도우가 연속하는 경우이다. 그림 7에서, 16번째와 17번째 윈도우가 이상값을 가질 때, 16번째 윈도우 보정을 위해 15번째 윈도우의 정보를 활용하고, 17번째 윈도우는 가장 가까운 18번째 윈도우의 정보를 기반으로 보정된다. 셋째는 이상값이 포함된 3개 윈도우가 연속하는 경우이다. 그림 7에서, 26번째, 27번째, 28번째 윈도우에 이상값이 연속해서 탐지된 경우에는 2단계에 걸쳐 데이터 보정이 이루어진다. 우선 26번째 윈도우 보정을 위해 25번째 윈도우 정보를 활용하고, 28번째 윈도우 보정을 위해 29번째 윈도우 정보를 활용한다. 그리고나서 27번째 윈도우의 보정을 위해 사전 보정된 26번째 윈도우의 정보를 기반으로 대체 데이터가 생성된다. 이상값이 포함된 윈도우가 4개 이상 존재하는 경우에도, 셋째 경우와 같이 인접한 정상 분포 윈도우를 점진적으로 탐색하여 데이터 보정이 수행될 수 있다.

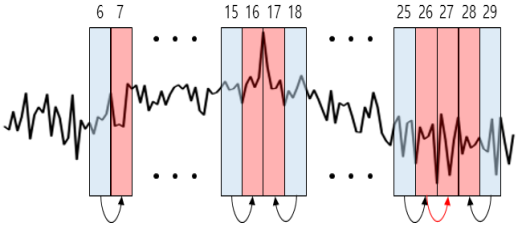

그림 6. 데이터 보정을 위해 정상 데이터를 가지는 인접 윈도우의 활용 방법

Fig. 6. How to use adjacent windows with normal data for data correction

우리는 탐지된 이상값을 정상 분포를 따르는 데이터로 대체하기 위해서, LSTM-GAN 모델에서 얻은 이상점수 $A(x, G(z))$ (식(2) 참조)에 시계열 데이터 내 값들 간의 유사도를 계산하는 Soft DTW 점수인 $softDTW(x', G(z))$를 추가하여 새로운 보정점수(correction score)인 $C(x', G(z))$를 식(3)과 같이 정의한다.

$$C(x', G(z)) = A(x', G(z)) + softDTW(x', G(z)) \quad (3)$$

여기서, 식(2)의 x는 주어진 시계열 데이터를 세분화한 각 윈도우를 의미하며, 식(3)의 x'은 이상값을 포함한 윈도우 x와 인접한 윈도우에 해당하는 변수이다. 이 보정점수를 통해 $G(z)$는 정상 분포를 가지는 윈도우 x'와 가까워지고 이상 점수도 낮아지도록 학습이 수행된다. 최종적으로 이상값이 포함된 윈도우 x는 정상 분포와 유사하도록 학습된 생성자에 의해 $G(z)$로 대체되어 보정된다.

표 1은 LSTM-GAN 모델 기반 이상탐지 연산과 제안하는 데이터 보정 연산을 보여주는 의사코드이다. 우리는 이상탐지를 위해서 [13]에서 제시한 adversarialTrain 함수, anomalyScore 함수와 [23]에서 제시한 SoftDTW 함수를 활용하여 손실함수를 구성한다. 우선 adversarialTrain 함수를 통해 이상탐지를 위한 LSTM-GAN 모델을 구축하였고, 학습된 생성자 G와 판별자 D를 이용하여 anomalyScore 함수를 통해 실제 데이터에 대한 이상점수를 계산한다. 이 이상점수가 임계값(α)보다 큰 경우, 해당 윈도우는 '이상' 클래스로 분류되며, '이상' 윈도우는 GAN 기반 DataGenerate 함수를 통해 생성된 새로운 '정상' 윈도우로 대체된다.

앞서 언급한 바와 같이, 시계열 데이터의 이상값 보정을 위해 이상값이 포함된 윈도우와 가장 가까운 정상 윈도우의 통계량을 활용하여 새로운 정상 분포를 가지는 윈도우를 생성한다. 이 때 SplitIndex 함수는 '정상' 윈도우와 '이상' 윈도우를 구분하는 인덱스를 부여한다. 이 윈도우 인덱스 정보를 이용하여 이상값이 포함된 윈도우와 인접한 '정상' 윈도우 탐색을 위해 NearestIndexDetect함수가 호출된다. 그 결과, '이상' 윈도우의 인덱스 i는 인접한 '정상' 윈도우의 인덱스인 i'로 대체되어, 윈도우 $X[i']$가 x'에 할당된다. 이어서 anomalyScore 함수를 통해 대체된 x'와 학습된 생성자 G와 판별자 D를 이용하여 잠재벡터공간으로부터 추출된 z를 가지고 생성된 데이터 $G(z^{i'})$에 대한 이상점수와 SoftDTW 함수를 통해 x'와 생성된 데이터에 대한 거리 정보를 이용하여 보정점수 C를 계산한다. 이 보정점수를 바탕으로 이상점수 A와 거리값 softDTW가 적어지도드록 $z^{i'}$를 반복적으로 갱신하게 된다. 최종적으로 얻어진 $z^{i'}$로 생성된 데이터 $G(z^{i'})$는 이상값이 포함된 윈도우를 대체하게 된다.

그림 7은 표 1의 데이터 보정 과정을 통해, 이상값에 해당하는 적색선이 정상 분포에 근사한 검정선으로 보정됨을 보여준다. 그리고 그림 8은 보정된 데이터에 대하여 이상탐지 모델로 이상탐지를 수행한 결과, 이상값이 모두 정상값으로 탐지되었음을 보여준다.

표 1. 이상탐지 및 데이터 보정 과정
Tabel 1. Process of anomaly detection and data correction

A:lgorithm 1: Algorithm for Data Correction

Input: A set of real time series windows X
Output: A set of corrected time series windows X
Function NearestIndexDetect(IdxList, Idx):
    $NW = \arg\min_i(|IdxList - Idx|)$
  return NW

Function SplitIndex(x, G, D):
  A = anomalyScore(x, G, D)
  for i in 1 to number_of_windows **do**
    if $A[i] > \alpha$
      AnomalyIdx ← i
    else if $A[i] < \alpha$
      NormalIdx ← i
  return AnomalyIdx, NormalIdx

Function DataGenerate(x, G, D):
  A = anomalyScore(x, G, D)
  AnomalyIdx, NormalIdx = SplitIndex(x, G, D)
  for i in AnomalyIdx **do**
    i' = NearestIndexDetect(NormalIdx, i)
    x' = $X[i']$
    Sample a noise vector $z^{i'}$ from $p_g(z^{i'})$.
    for λ in 1 to Γ **do**
      Generate a fake data vector $G(z^{i'})$ from $z^{i'}$.
      Calculate $A(G(z^{i'}))$ for $x^{i'}$ utilizing G and D, and Calculate $SoftDTW(x^{i'}, G(z^{i'}))$
      $C = A(G(z^{i'})) + SoftDTW(x^{i'}, G(z^{i'}))$
      Update $z^{i'}$ using gradient descent.
  return $G(z^i)$

Function Correction(x):
  G, D = adversarialTrain(x)
  A = anomalyScore(x, G, D)
  for i in 1 to number_of_windows **do**
    if A[i] $> \alpha$
      $X[i]$ ← DataGenerate(x, G, D)
  return X

* For the adversarialTrain function and anomalyScore function, refer to Bashar's paper[10]
* For the SoftDTW function, refer to Cuturi's paper[20]

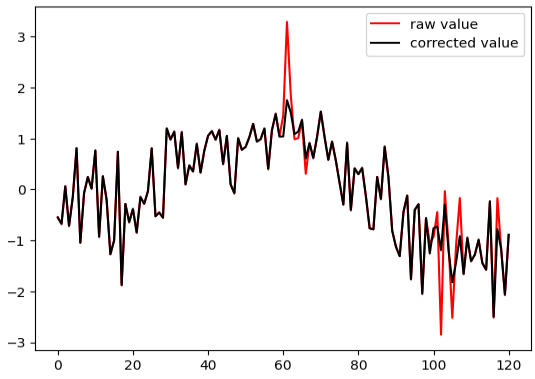

그림 7. 이상값이 포함된 원천 데이터와 보정된 데이터 비교
Fig. 7. Comparison of raw data with anomalies and corrected data

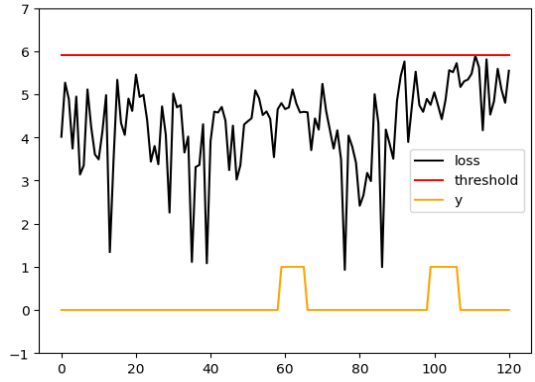

그림 8. 보정된 데이터에 대한 이상탐지 결과
Fig. 8. Anomaly detection results for corrected data

Ⅳ. 실험 및 결과

본 실험은 시계열 데이터의 일반 분포를 학습한 LSTM-GAN 모델을 이용하여 이상탐지를 수행하고, 탐지된 이상값을 정상값으로 보정한 효과를 시계열 예측모델의 성능 평가를 통해 간접적으로 확인하고자 한다.

1. 데이터

우리는 제안 기법의 효능을 실험을 위해 오픈소스 데이터인 NAB 데이터셋을 활용하였다. NAB (Numenta Anomaly Benchmark) 데이터셋[14]은 Kaggle (http://www.kaggle.com)에 개방된 시계열 데이터 모임으로서, 이상값에 대한 정답 정보를 포함하며, 본 실험은 이 NAB 데이터셋 내 이상탐지 관련 연구에 자주 사

표 2. 원천데이터와 보정데이터에 대한 예측모델의 정확도 비교
Table 2. Comparison of accuracy of prediction model for original data and correction data

Data		LSTM		GRU		ARIMA	
		MAE	MSE	MAE	MSE	MAE	MSE
Ambient temperature system failure	Raw data	0.554	0.060	0.610	0.054	0.813	1.022
	mean-based correction	0.852	0.077	0.849	0.071	0.799	0.996
	SVR-based correction	0.481	0.054	0.467	0.052	0.844	1.072
	Proposed correction	**0.450**	**0.055**	**0.454**	**0.055**	**0.798**	**0.979**
CPU utilization asg misconfiguration	Raw data	0.617	0.368	1.066	0.282	0.976	1.744
	mean-based correction	0.673	0.316	0.965	0.242	0.947	1.576
	SVR-based correction	0.620	0.272	0.936	0.239	0.938	1.563
	Proposed correction	**0.573**	**0.224**	**0.853**	**0.227**	**0.702**	**1.115**

용되는 'Ambient temperature system failure'와 'CPU utilization asg misconfiguration' 데이터를 활용한다. 그리고 이 시계열 데이터에 대한 딥러닝 모델 구축을 위해 전처리 작업이 수행될 필요가 있으며, 이는 식 (4)에 따라 전체 데이터의 값(x_i)들을 평균(μ_i)이 0, 표준편차(σ_i)가 1이 되도록 하는 정규화 작업을 수반한다.

$$\hat{x_i} = \frac{x_i - \mu_i}{\sigma_i} \tag{4}$$

앞서 언급한 바와 같이, 본 연구에서는 이상값 보정은 윈도우 단위로 이루어지므로, 전체 시계열 데이터가 적당한 샘플 개수를 가지는 세분화된 시계열 데이터(윈도우)로 분할된다. 여기서 각 데이터 윈도우에 포함된 샘플 개수는 60으로 설정하였으며, 윈도우 크기에 따라 제안 기법의 성능이 거의 영향받지 않음을 실험적으로 확인하였다.

2. 평가 척도

제안한 시계열 데이터의 보정기법은 주어진 데이터를 사용하여 구축된 예측모델의 성능 평가를 통해 간접적으로 평가되는 것이 일반적이다. 우리는 예측모델의 구축을 위해 LSTM, GRU, ARIMA 총 3개의 알고리즘을 사용하였다. LSTM과 GRU 모델은 각각 100개의 은닉 유닛을 가지는 단일 레이어 모델로 구현되었으며, ARIMA 모델은 ARIMA(1,1,1) 모델을 이용하였다. 그리고 이상값이 포함된 원천 데이터와 보정된 데이터에 대한 예측모델의 정량적인 평가를 위해 식(5)의 평균 절대오차(MAE)와 식 (6)의 평균 제곱오차(MSE)를 이용하며, 이 값이 작을수록 해당 예측모델의 성능이 우수함을 의미한다. 여기서 Y_i는 정답 수치, $\widehat{Y_i}$는 예측 수치를 의미한다.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\widehat{Y_i} - Y_i|, \tag{5}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{Y_i} - Y_i)^2 \tag{6}$$

3. 실험 결과

제안 기법의 효능을 평가하기 위해 시계열 내 이상값을 평균값(mean)으로 보정한 데이터, 서포트벡터회귀(SVR)로 보정한 데이터, 그리고 제안 기법으로 보정한 데이터에 대하여 3가지 예측모델을 구축하고 이것의 성능을 비교하였다.

표 2는 NAB데이터셋 내 'Ambient temperature system failure'와 'CPU utilization asg misconfiguration' 시계열 데이터에 대한 예측모델 성능을 MAE 및 MSE 척도로 보여준다. 제안 기법으로 보정된 데이터에 대한 예측모델의 성능이 2개 기존 기법으로 보정된 데이터에 대한 예측모델의 성능보다 우수함을 확인하였다. 그리고 서포트벡터회귀 기반 보정기법이 평균값 기반 보정기법보다 그 성능이 우세하였다. 구체적으로 서포트벡터회귀 기반 보정기법 대비 제안 기법으로 보정된 데이터의 예측모델이 'Ambient temperature system failure' 데이터의 경우에는 약 4.9%, 'CPU utilization asg misconfiguration' 데이터의 경우에는 약 15.5%의 개선 효과를 보였다. 또한 제안 기법으로 보정 연산을 거친 시계열 데이터에 대해 LSTM-GAN 모델로 이상탐지를 수행한 결과, 모든 이상값이 정상값으로 보정됨을 확인하였다.

V. 결 론

본 논문은 시계열 데이터의 이상값에 대한 보정을 위해 GAN 딥러닝 모델과 Soft-DTW 기법을 결합한 방안을 제안하였다. 이상값 보정을 위해 LSTM-GAN 기반 이상탐지 모델[13]을 활용하여 이상탐지를 먼저 수행하였고, 탐지된 이상값은 인접한 정상 분포를 가지는 슬라이딩 윈도우 데이터에 근사한 새로운 값으로 대체된다. 제안 기법의 데이터 보정 효과를 정량적으로 평가하기 위해, 원천 데이터와 보정 데이터에 대한 예측모델을 구축하여 MAE와 MSE를 비교함으로써 제안 보정 기법이 기존 서포트벡터회귀 기반 보정기법보다 우수함으로 실험적으로 증명하였다.

본 논문에서 제안된 기법은 단변량(univariate) 시계열 데이터에 내재된 이상값에 대한 보정에 적용된다. 향후 우리는 다변량(multivariate) 시계열 데이터의 보정 문제를 풀기 위하여, 데이터 보정 과정에서 변수 간 상관 및 의존 관계를 학습하는 딥러닝 모델을 추가함으로써 일반화된 시계열 데이터 보정기법을 개발할 예정이다.

References

[1] M. Ge, H. Bangui, and B. Buhnova, 'Big Data for Internet of Things: A Survey', Future Gener. Comput. Syst., vol. 87, pp. 601–614, Oct. 2018.
DOI: https://doi.org/10.1016/j.future.2018.04.053

[2] J. Choi and Y. Shin, 'Design of Efficient Big Data Collection Method based on Mass IoT devices', *J. Korea Inst. Inf. Electron. Commun. Technol.*, vol. 14, no. 4, pp. 300–306, Aug. 2021.
DOI: https://doi.org/10.17661/JKIIECT.2021.14.4.300

[3] R. Chalapathy and S. Chawla, 'Deep Learning for Anomaly Detection: A Survey', 2019
DOI: https://doi.org/10.48550/ARXIV.1901.03407

[4] T. Kim and J. Park, 'Image Segmentation for Fire Prediction using Deep Learning', The Journal of The Institute of Internet, Broadcasting and Communication, vol. 23, no. 1, pp. 65–70, Feb. 2023
DOI: https://doi.org/10.7236/JIIBC.2023.23.1.65

[5] W. Mao, W. Wang, Z. Dou, and Y. Li, 'Fire Recognition Based On Multi-Channel Convolutional Neural Network', Fire Technol., vol. 54, no. 3, pp. 809–809, May 2018.
DOI: https://doi.org/10.1007/s10694-018-0705-3

[6] M. Douglas C, Introduction to statistical quality control. John Wiley & Sons, 2020.

[7] S. Lee, L. Juyoung, and J. Yun, 'Development of Quality Management Technology for Data Measured by Smart Water Meters', KAIS, vol. 24, no. 1, pp. 570–580, Jan. 2023.
DOI: https://doi.org/10.5762/KAIS.2023.24.1.570

[8] S. Zhanwei and L. Zenghui, 'Abnormal detection method of industrial control system based on behavior model', Comput. Secur., vol. 84, pp. 166–178, Jul. 2019.
DOI: https://doi.org/10.1016/j.cose.2019.03.009

[9] J. Zhu, Z. Ge, Z. Song, and F. Gao, 'Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data', Annu. Rev. Control, vol. 46, pp. 107–133, 2018.
DOI: https://doi.org/10.1016/j.arcontrol.2018.09.003

[10] Choh Man Teng, 'Polishing blemishes: issues in data correction', IEEE Intell. Syst., vol. 19, no. 2, pp. 34–39, Mar. 2004.
DOI: https://doi.org/10.1109/MIS.2004.1274909

[11] M.-K. Lee, S.-H. Moon, Y.-H. Kim, and B.-R. Moon, 'Correcting abnormalities in meteorological data by machine learning', in 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), San Diego, CA, USA, Oct. 2014, pp. 888–893.
DOI: https://doi.org/10.1109/SMC.2014.6974024

[12] S. Kim, E. Lee, and S. Lee, 'Method of Data Quality Evaluation for Improving Data Quality of DSEM-Trajectory', jkiit, vol. 20, no. 8, pp. 7–18, Aug. 2022.
DOI: https://doi.org/10.14801/jkiit.2022.20.8.7

[13] M. A. Bashar and R. Nayak, 'TAnoGAN: Time Series Anomaly Detection with Generative Adversarial Networks', in 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, ACT, Australia, Dec. 2020, pp. 1778–1785.
DOI: https://doi.org/10.1109/SSCI47803.2020.9308512

[14] A. Lavin and S. Ahmad, 'Evaluating Real-Time Anomaly Detection Algorithms -- The Numenta Anomaly Benchmark', in 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, Dec. 2015, pp. 38–44.
DOI: https://doi.org/10.1109/ICMLA.2015.141

[15] S. Li and J. Wen, 'A model-based fault detection and diagnostic methodology based on PCA method and wavelet transform', Energy Build., vol. 68, pp. 63–71, Jan. 2014.
DOI: https://doi.org/10.1016/j.enbuild.2013.08.044

[16] P. Geladi and B. R. Kowalski, 'Partial least-squares regression: a tutorial', Anal. Chim. Acta, vol. 185, pp. 1–17, 1986.
DOI: https://doi.org/10.1016/0003-2670(86)80028-9

[17] F. Angiulli and C. Pizzuti, 'Fast Outlier Detection in High Dimensional Spaces', in Principles of Data Mining and Knowledge Discovery, vol. 2431, T. Elomaa, H. Mannila, and H. Toivonen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 15–27.

DOI: https://doi.org/10.1007/3-540-45681-3_2

[18] T. Ergen and S. S. Kozat, 'Unsupervised Anomaly Detection With LSTM Neural Networks', IEEE Trans. Neural Netw. Learning Syst., vol. 31, no. 8, pp. 3127-3141, Aug. 2020.
DOI: https://doi.org/10.1109/TNNLS.2019.2935975

[19] C. Zhou and R. C. Paffenroth, 'Anomaly Detection with Robust Deep Autoencoders', in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax NS Canada, Aug. 2017, pp. 665-674.
DOI: https://doi.org/10.1145/3097983.3098052

[20] I. J. Goodfellow et al., 'Generative Adversarial Networks'. Advances in neural information processing systems, pp. 2672-2680, 2014.
DOI: https://doi.org/10.48550/arXiv.1406.2661

[21] S. K. Kwak and J. H. Kim, 'Statistical data preparation: management of missing values and outliers', Korean J. Anesthesiol., vol. 70, no. 4, p. 407, 2017.
DOI: https://doi.org/10.4097/kjae.2017.70.4.407

[22] Y.-H. Kim et al., 'Improved Correction of Atmospheric Pressure Data Obtained by Smartphones through Machine Learning', Comput. Intell. Neurosci., vol. 2016, pp. 1-12, 2016.
DOI: https://doi.org/10.1155/2016/9467878

[23] M. Cuturi and M. Blondel, 'Soft-DTW: a Differentiable Loss Function for Time-Series', Proc. int. Conf. Mach. Learn., pp. 894-903, 2017.
DOI : https://doi.org/10.48550/arXiv.1703.01541

[24] E. Keogh and C. A. Ratanamahatana, 'Exact indexing of dynamic time warping', Knowl. Inf. Syst., vol. 7, no. 3, pp. 358-386, Mar. 2005.
DOI: https://doi.org/10.1007/s10115-004-0154-9

저 자 소 개

정 한 석(준회원)

- 2021년 ~ 현재 : 서울시립대학교 전자전기컴퓨터공학과 석사과정
- 관심분야 : 머신러닝, 데이터마이닝, 데이터베이스, 딥러닝, 빅데이터

김 한 준(정회원)

- 1994년 : 서울대학교 계산통계학과(이학사)
- 1996년 : 서울대학교 전산과학과(이학석사)
- 2002년 : 서울대학교 컴퓨터공학부(공학박사)
- 2002년 ~ 현재 : 서울시립대학교 전자전기컴퓨터공학부 정교수
- 관심분야 : 데이터사이언스, 머신러닝, 텍스트마이닝, 데이터베이스, 정보검색

※ 본 논문는 2021년도 서울시립대학교 연구년교수 연구비에 의하여 연구되었음.