

디지털 트윈 환경에서의 이상치 탐지 알고리즘 비교분석

신 윤 종¹ · 이 남 연^{2*}¹경희대학교 인공지능학과 석사과정^{2*}한신대학교 IT경영학과 부교수

A Comparative Analysis of Anomaly Detection Algorithm for Digital Twin

Yoon-Jong Shin¹ · Namyeon Lee^{2*}¹Master's Course, Department of Artificial Intelligence, Kyung Hee University, Seoul 02447, Korea^{2*}Associate Professor, Department of IT Management, Hanshin University, Osan 18101, Korea

[요 약]

디지털 트윈은 현실 세계의 객체 및 시스템을 가상의 디지털 트윈으로 복제하고 분석 및 시뮬레이션을 통해 효과적인 운영과 예측을 가능하게 하는 기술이다. 이 때, 디지털 트윈을 구성하는 데이터의 품질이 보장되지 않는다면 결과물인 디지털 트윈 또한 정상적으로 구현되기 어렵기 때문에 디지털 트윈에서 사용되는 데이터의 모니터링을 위해 이상치 탐지 기법을 활용해왔다. 그러나 기존 포인트 이상치 탐지 기법 (PAD; Point anomaly detection) 기반의 이상치 탐지 기법들은 디지털 트윈 환경에서 수집되는 데이터들의 특성들로 인해 효과적인 이상치 탐지를 수행하는데 한계점을 가진다. 따라서 본 논문은 디지털 트윈 환경에서 수집되는 데이터의 특성과 유사한 데이터셋을 기반으로 이상치 탐지 알고리즘의 성능을 비교 분석하여 디지털 트윈 환경에서 가장 적합한 이상치 탐지 알고리즘을 제안하였고, 실험의 결과를 바탕으로 향후 디지털 트윈 환경에서의 이상치 탐지 연구에서의 주요하게 고려되어야 할 이슈들을 도출하였다.

[Abstract]

Digital twin is a technology that duplicates real-world objects and systems into a virtual twin, and analyzes and simulates them to enable effective operation and prediction. If the quality data is not guaranteed, the resulting digital twin is also difficult to be normally implemented. So, anomaly detection methods have been used to monitor the quality of data. However, existing anomaly detection methods are not enough for digital twins due to the data characteristics such as class imbalance problem of multivariate time series data, and the limitations of point anomaly detection (PAD) performed by existing methods. Therefore, this paper implements existing anomaly detection algorithms and compares their performance, and based on this, we present the issue of anomaly detection in digital twin environments and future research directions for contextual outlier detection (CAD).

색인어 : 디지털 트윈, 이상치 탐지, 시계열데이터, 딥러닝, 가상물리시스템**Keyword** : Digital twin, Anomaly detection, Time series data, Deep learning, Cyber physical system<http://dx.doi.org/10.9728/dcs.2022.23.12.2497>

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 14 November 2022; Revised 23 November 2022

Accepted 24 November 2022

*Corresponding Author; Namyeon Lee

Tel: +82-31-379-0595

E-mail: nylee@hs.ac.kr

I. 서론

오늘날 사물인터넷, AI, 5G 등 4차 산업혁명을 대표하는 핵심 기술의 집약체인 디지털 트윈은 스마트 시티, 스마트 팩토리, 지능형 교통체계 등 다양한 산업에 접목되고 있다. 디지털 트윈 기술은 2003년 Micheal Grieves가 처음 언급한 기술로 제품의 설계, 개발, 검증, 양산, 추적관리 및 폐기에 이르는 전주기적 관리를 수행하기 위해 실제 제품과 1:1로 대응하여 제품의 주요 정보를 지속적으로 수집, 가공 및 처리하는 정보 구조체로 정의한다[1]. 또한 글로벌 시장조사 전문기관인 가트너에서도 디지털 트윈 기술을 2017년부터 2019년까지 10대 전략 기술 중 하나로 언급하고 있으며, 실제 객체 또는 시스템에 대응되는 물리적 사물, 프로세스, 조직, 사람 또는 기타 추상적 정보를 소프트웨어 객체 혹은 컴퓨팅 모델로서 디지털화한 표현이라고 규정하였다. 디지털 트윈 기술은 다양한 산업 분야에서 활용 가능하며, 디지털 트윈 기술을 주도하는 대표적인 기업으로는 GE, SIEMENS, IBM 등이 있으며 초기 분석 및 모델링 기능 중심에서 3D모델 및 3D프린팅을 통한 생산 공정 설계 최적화 단계를 거쳐 최근에는 인공지능 기술을 적용한 시뮬레이션 및 예측 분야에 적용되고 있다.

디지털 트윈은 데이터의 품질이 가장 중요하며, 데이터의 품질을 담보할 수 없는 경우 디지털 트윈을 이용하여 소프트웨어 객체를 생성하고 시뮬레이션을 통해 도출된 결과물 역시 품질이 떨어질 수 있다[2]. 특히 디지털 트윈 서비스에서의 데이터가 가지는 특성은 다양한 웨어러블 디바이스 및 멀티모달 센서를 통해 수집되는 데이터로 이들 데이터는 다변량 시계열 데이터라는 특성이 있으며, 이러한 데이터의 특성에 맞는 데이터 처리 및 분석의 필요성이 증가하고 있다. 이러한 디지털 트윈을 위한 데이터는 많은 경우 현장에 설치된 센서로부터 수집되고 있지만, 실제 현장의 다양한 상황적인 요인으로 인해 저품질의 데이터를 수집하거나 데이터가 누락되는 상황이 발생할 수 있다[3]. 그리고 최근에는 서로 다른 디지털 트윈 간 상호작용을 가능케 하는 표준을 수립하기 위한 연구가 활발히 수행되고 있는 상황 속에서 통신에 따른 데이터 손실 방지에 대한 중요성이 증가하고 있다[4]. 따라서, 성공적인 디지털 트윈의 구축을 위해서는 데이터의 품질을 높일 수 있도록 오류 데이터를 적절하게 탐지하고 보정할 수 있는 기술 연구가 필요하다.

오늘날 수많은 데이터 품질 관리와 장애 이벤트 발생을 다루기 위한 기술로 이상치 탐지가 활발히 활용되고 있다[5]. 이상치 탐지는 주어진 데이터의 분포에서 눈에 띄게 벗어난 인스턴스를 탐지하는 프로세스로, 이상 거래 탐지, 고객 이탈, 설비 유지보수, 침입 탐지 시스템 등의 다양한 산업 분야에서 사용하고 있다. 이상치 탐지를 위해 쓰였던 통계적 모델링 또는 기계학습을 활용한 알고리즘들은 특성 공학을 위한 도메인 전문 지식이 필요했으나 딥러닝이 등장하면서 많은 양의 데이터를 손쉽게 학습할 수 있게 되었다. 그러나 딥러닝 방법론은 일반적으로 지도 학습의 경우 데이터의 라벨 정보가 필

요하므로 시간과 비용이 많이 든다는 한계가 있다[6]. 게다가 라벨링 해야 할 데이터가 실시간 데이터라면 대부분 정상과 비정상에 대한 라벨이 함께 생성되기 어려워진다. 특히 디지털 트윈 환경에서 생성되는 다변량 시계열 데이터를 위한 이상치 탐지는 지도 학습 패러다임으로 접근하기 더욱 어렵게 만든다. 혹여나 만약 데이터 라벨을 모두 준비하더라도 대부분 비정상 데이터보다 정상 데이터의 분포가 더 크기 때문에 비정상 데이터를 탐지하기 위한 모델 학습 시 탐지 정확도가 높지 않을 수 있다[8]. 이와 같이 디지털 트윈 맥락에서의 이상치 탐지는 기존 통상적인 이상치 탐지 방법론으로는 디지털 트윈 환경에서 생산되는 데이터의 품질을 개선하기 어렵다는 문제점이 있다.

기존 이상치 탐지 알고리즘들은 디지털 트윈 환경에서 생산되는 데이터를 다루는데 어려움을 겪고 있기 때문에 본 연구에서는 성공적인 디지털 트윈의 구축을 위한 이상치 데이터 탐지 및 보정 기술을 위한 이상치 탐지 이슈를 도출하였다. 2장 문헌 연구에서는 기존 이상치 탐지를 방법론에 따라 살펴보고 방법론 별 특징과 장단점을 살펴본다. 특히 디지털 트윈 맥락에 적용했을 때 적합하다고 판단되는 딥러닝 기반 이상치 탐지 방법론을 리뷰 한다. 3장에서는 디지털 트윈 문헌 및 기존 이상치 탐지 방법들에 대한 문헌 조사 결과를 바탕으로 디지털 트윈 환경에서 발생하는 데이터의 특성과 이에 따른 디지털 트윈에 적용할 수 있는 이상치 탐지 알고리즘을 구현한다. 실제 다변량 시계열 데이터인 기상청의 기후 관측 데이터를 기반으로 학습하여 기존 전통적인 이상치 탐지 알고리즘과 딥러닝 기법을 비교 분석한다. 그리고 4장에서는 실험 결과를 통해 도출된 디지털 트윈 환경에서의 이상치 탐지 주요 이슈를 설명한다.

II. 문헌 연구

2-1 전통적 이상치 탐지 방법론

기존에 오랫동안 널리 쓰인 이상치 탐지 방법론으로는 데이터의 분포 정보를 활용하는 통계적 접근 기반 알고리즘이 있다. 비정상적인 데이터는 주된 데이터 분포에서 벗어난 개체이므로 데이터를 설명하는 통계치를 바탕으로 임계치를 설정하고 이를 기준으로 이상치를 분류하게 된다. 먼저 기본적인 접근법으로 그럽스 검정(Grubb's test)이 있는데, 그럽스 검정은 정규분포를 만족하는 단변량 자료에서 이상치를 검정하는 방법으로 t-분포에 근거한 임계치를 산출하여 검정 통계량이 임계치보다 큰 경우 이상치로 분류하는 알고리즘이다[7]. 또 다른 통계적 접근 기반 알고리즘으로는 사분위수 범위를 활용하여 이상치를 분류하는 사분위수 범위 방법론(IQR; interquartile range)이 있다. 데이터의 분포는 최소값, 최대값, 제 1사분위수, 제 2사분위수, 제 3사분위수와 같은 통계치를 시각적으로 요약할 수 있는데, 이를 상자 그림이

라고 한다. 상자 그림에서 표현되는 최소값과 최대값은 이상치를 제외한 데이터의 최대값과 최소값을 의미하며 사분위수 범위는 제 1사분위수와 제 3사분위수의 차이로 정의되며, 사분위수 범위의 1.5배를 초과하는 관측치는 약한 이상치로, 3배를 초과하는 관측치는 강한 이상치로 정의한다.

이상치 탐지는 전통적으로 다양한 도메인에서 활용되어 왔으며 특히 품질 관리를 위한 다양한 연구가 진행되어왔다. 그 중 슈하르트 관리도 (Shewhart control chart)는 이상치 탐지를 품질 관리에 적용한 모델이다. 관리도란 품질 관리에 사용되는 지표로 품질이 어떤 값을 중심으로 어느 선까지 달라지는 것은 우연변동으로 취급하나, 그 이상을 벗어나면 이상 원인에 의한 변동으로 간주한다[8]. 슈하르트 관리도는 데이터의 중심으로부터 양쪽으로 표준편차의 세 배에 임계치를 설정하며 이 때 표준편차는 우연변동의 척도로 간주하여 근사적으로 99.7%의 관리한계 내에 품질이 보장될 것을 의미한다. 슈하르트 관리도의 종류로는 관리 모수와 관리 통계량에 따라 다양한 관리도가 존재하는데, 관리 모수는 평균과 분산 등과 같이 관리하고자 하는 통계량을 의미한다.

전통적 이상치 탐지 방법론 중 또 하나로 데이터 개체 간 거리를 활용한 접근법이 있다. 정상 데이터 개체 간에는 서로 거리가 가깝고, 이상치 개체는 정상 데이터로부터 떨어져 있다는 점을 활용하는 것이다. 따라서 서로 거리가 가까운, 즉 유사한 데이터들을 같은 군집으로 묶고 정상 범주에 속한 군집에서 떨어진 개체를 이상치로 분류하게 된다. 가장 대표적인 개체 간 거리 기반 이상치 탐지 알고리즘으로는 k-평균 군집화를 이용한 이상치 탐지 알고리즘이 있다. k-평균 군집화는 주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로, 각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작한다. 이 알고리즘은 자율 학습의 일종으로 레이블 정보가 존재하지 않는 입력 데이터에 레이블을 부여할 수 있다. 그러나 k-평균 군집화 알고리즘은 손실 함수 최적화 과정에서 전역 최솟값이 아닌 지역 최솟값으로 수렴할 가능성이 있으며 이상치에 민감하고 하이퍼 파라미터인 클러스터 개수를 지정해야 한다는 한계점이 있다.

k-평균 군집화의 대안으로 쓰이는 밀도 기반 군집화는 유사한 데이터 포인트들이 그룹화되는 방식으로 데이터 포인트 세트를 그룹화하는 방법이다. 따라서 군집화 알고리즘은 데이터 포인트 간의 유사성 또는 비 유사성을 찾게 된다. DBSCAN (Density based spatial clustering of applications with noise)은 노이즈가 있는 밀도 기반 공간 군집화를 대표하는 알고리즘이며 특정 군집이나 이상치가 존재하는 군집을 찾을 수 있다[9]. DBSCAN의 장점은 하이퍼 파라미터인 군집의 개수를 정할 필요가 없고 군집의 밀도에 따라 군집을 서로 연결하기 때문에 기하학적인 모양을 가질 수 있으므로 기존 선형 군집화 알고리즘보다 유연하다. 군집화를 수행하는 동시에 이상치도 분류할 수 있으므로 이상치에 의해 군집화 성능이 하락하는 현상을 완화할 수 있다. 그러나 거리 기반 알고리즘의 경우 각 개체 간 거리를 계산하는

만큼 데이터가 많아질수록 연산 시간복잡도가 기하급수적으로 증가한다는 한계점이 있다.

기존 전통적 이상치 알고리즘들을 종합하여 보면 다음과 같은 한계점이 존재한다. 정상 데이터의 분포에서 크게 벗어나는 값을 이상치로 분류하는 통계적 접근 기반 방법론이나 거리 기반 이상치 탐지 방법론과 같은 기존 방법론들의 문제점은 정상치를 모두 고려해야 하므로 계산 부하가 크다는 점과 매번 정상치를 프로파일링하여 이상치를 계산하므로 모델을 생성하고 재활용하기 어렵다는 한계점이 있었다. 트리 기반 이상치 탐지는 앞선 기존 방법론들의 문제점을 데이터를 분할하면서 이상치를 분류하는 방식으로 보완한다. 대표적으로 Isolation forest는 관측치 사이의 거리 또는 밀도에 의존하지 않고 의사결정 나무를 이용하여 이상치를 탐지하는 기법이다[10]. Isolation forest는 의사결정 나무 기법으로 분류 모형을 생성하여 모든 관측치를 고립시키면서 분할 횟수로 이상치를 분류하는데, 모형에서 적은 횟수로 잎 노드에 도달하는 관측치일수록 이상치라고 볼 수 있다. 즉, 일반적으로 정상 데이터의 경우 더 많은 재귀 이진 분할이 필요하지만 이상치의 경우 정상 데이터보다 이진 분할이 덜 필요하다는 점을 활용하게 된다.

2-2 딥러닝 기반 이상치 탐지 방법론

오늘날 딥러닝 연구가 활발히 진행되면서 이상치 탐지 분야에서도 딥러닝을 활용한 연구가 수행되고 있는데, 대표적으로 순환 신경망을 활용한 이상치 탐지 연구가 활발히 진행되고 있다[11]. 순환신경망은 입력과 출력을 시퀀스 단위로 처리하는 모델로써, 디지털 트윈 환경에서 발생하는 데이터와 같은 시계열 또는 다변량 데이터를 학습하는 데 특화된 모형이다. 더 나아가 장단기 메모리 모델은 기존 순환신경망이 입력 데이터 시퀀스가 길어질수록 기울기 소실이 발생하는 문제를 셀 상태를 통해 이전 단계에서 학습 시킨 정보를 그대로 보존하며 학습을 진행함으로써 보완하였다. 이러한 이점으로 인해 최근 이상치 탐지를 위한 딥러닝 연구에서 순환신경망 형태의 모델 아키텍처를 기반으로 이상치 탐지 연구를 수행하고 있다. 특히 LSTM(Long Short-Term Memory) 또는 GRU(Gate Recurrent Unit) 알고리즘을 기반으로 한 이상치 탐지 연구가 활발하게 진행되어 왔다.

LSTM을 기반으로 한 이상치 탐지 연구들을 살펴보면, Malhotra 외(2016)의 연구에서는 여러 층의 장단기 메모리를 학습하여 오류 벡터를 계산하고 오류의 다변량 가우스 분포에서 임계치를 넘는 오류에 대해서는 이상치로 분류하였다[12]. 이들 연구에서는 순환신경망 모델의 기울기 소실 문제를 극복하기 위해 장단기 메모리 모델을 사용하여 순환신경망보다 더 높은 예측 성능을 보임을 확인하였다. 또한 LSTM 기반의 이상치 탐지 성능을 높이기 위해서 오토인코더를 활용하기도 하였다. 오토인코더를 활용한 연구의 경우에는 일정 기간의 시계열 데이터를 입력받아 향후 데이터를 예측하고

이를 원본 데이터와의 차이를 비교하여 값이 큰 데이터를 이상치로 분류한다. 그리고 LSTM 자체로 이상치 탐지에 활용하기보다 LSTM과 오토인코더를 앙상블하여 사용하는 연구 또한 활발히 수행되고 있다[13], [14]. LSTM-오토인코더를 학습할 때 인코더 부분은 데이터를 압축하여 데이터의 중요한 특징으로 변환하고, 디코더는 인코더에서 학습한 정보를 바탕으로 데이터를 재구성한다. 그리고 원본 데이터와 디코더를 통해 출력된 복원된 데이터의 차이인 복원 손실 함수가 작아 지도록 모델을 학습시킨다. 오토인코더 모델이 정상 데이터로만 학습하므로 재구성 오류가 큰 데이터일 경우 이상치라고 판단할 수 있다. 재구성 오류는 사용자가 설정한 임계치보다 값이 커지면 이상치로 판단하거나 정상 데이터의 분포를 적용하는 등 이상치를 판단하는데 활용하는 데이터로써 사용된다.

LSTM-오토인코더를 활용한 이상치 탐지 연구 사례를 살펴보면, Maleki 외(2021)의 연구에서는 산업용 가스터빈의 오작동 탐지를 위해 산업용 가스터빈의 온도 측정 데이터, CPU 사용량 데이터를 사용하여 이상치 탐지를 시도하였다 [15]. 해당 연구는 학습 전, 표준 점수를 임계치로 활용하여 훈련 데이터로 사용할 정상 데이터를 필터링하였다는 특징이 있다. Nguyen 외(2021)는 데이터의 분포 가정을 사용하지 않기 위해 단일 부류 서포트 벡터 머신을 사용하여 정상과 비정상 데이터에 대해 구분하여 공급망 관리를 위한 판매 시점 관리 데이터에서 이상 탐지를 위한 알고리즘을 제안하였다 [16]. 한편 다차원 시계열 데이터에서의 이상치 탐지를 위한 연구들도 함께 진행되었다. Kieu 외(2018)는 다차원 시계열 데이터를 기반으로 이상치를 탐지하기 위해 입력 시계열 벡터의 통계적 특징을 사전에 계산하고 이를 장단기 메모리 오토인코더 모델에 사용하였다[17]. Zhang 외(2018)는 다변량 시계열 간의 상관관계를 고려하기 위해서 시계열 데이터의 상관관계 및 시간 정보를 인코딩한 후 학습하는 합성곱 순환 인코더-디코더 모델을 제안하였으며 발전소 시스템에서 발생하는 시계열 데이터로부터 이상치를 탐지하였다[18].

LSTM과 더불어 LSTM을 변형하여 간소화 시킨 게이트 순환 유닛 (GRU: gated recurrent unit) 또한 딥러닝 이상치 탐지 모델로써 활발히 연구되고 있다. GRU는 망각 게이트와 입력 게이트를 하나의 업데이트 게이트로 합치고 셀 상태와 은닉 상태를 하나로 합쳐 변형함으로써 LSTM보다 단순한 구조를 가지도록 하였다. LSTM과 GRU는 성능상 큰 차이는 없지만, 연구 목적에 따라 LSTM이 더 좋을 수도, GRU가 더 좋은 모델이 될 수도 있다[19]. GRU는 LSTM에 비해 단순해서 연산량이 적으므로 큰 네트워크 구축 시 효율적이다. 또한 GRU는 모델 구현 시 고려해야 할 파라미터의 수가 더 적어서 사용자 편의성 측면에서 빠르게 네트워크를 구축할 수 있다. 반면 LSTM은 각 메모리와 결과값을 제어할 수 있지만 GRU는 메모리와 결과치의 제어가 불가능하다는 측면에서 정교한 모델 구축 시에는 LSTM이 더 좋은 선택이 될 수 있다. GRU도 LSTM과 마찬가지로 오토인코더를 앙상블하여 구현하는 방식으로 이상치 탐지 연구가 활발히 수행되고 있다. 여

기서는 게이트 순환 유닛 모델을 단편적으로 사용하지 않고, 정상 데이터만을 학습하기 위한 전처리 방법이나 다른 알고리즘과 결합한 모델을 통해 입력 데이터의 이상행동을 감지한다. 게이트 순환 유닛 오토인코더를 통해 입력 벡터의 특징들을 학습하고, 디코더를 통해 재구성 오류를 출력하여 이상치 판단 기준인 가우시안 분포나 임계치를 통해 이상치 탐지를 수행한다[19]-[21].

III. 이상치 탐지 알고리즘 성능 비교

본 연구에서는 앞서 설명한 이상치 탐지를 위한 알고리즘들이 과연 디지털 트윈 환경에서도 적합한 알고리즘인가를 확인하기 위해서 기상데이터를 기반으로 각 알고리즘의 성능을 비교 평가하였다.

3-1 실험 데이터

디지털 트윈 환경에서 수집되는 대부분의 데이터들은 데이터의 이상치 레이블 정보가 없는 시계열 데이터이다. 이러한 관점에서 본 연구에서 디지털 트윈 환경에서 적합한 이상치 탐지 알고리즘을 확인하기 위해서 기후데이터를 선택하였다. 기후데이터는 시계열로 수집되는 데이터이며 데이터 자체에 이상치를 판단하는 레이블 정보가 포함되어 있지 않다. 따라서 본 연구에서는 기상청에서 제공하는 기후 관측 데이터를 이용하여 이상치로 고려되는 이상기온을 탐지하고자하였다. 실험에 사용된 데이터는 그림 1과 같이, 1961년부터 2021년까지 서울의 일자별 평균기온, 최저기온, 최고기온으로 구성된 기온 다변량 데이터 셋이며 총 22,280건으로 구성되어 있다. 본 실험에 사용된 데이터는 일평균기온, 일최저기온, 일최고기온으로 구성되어 있다.

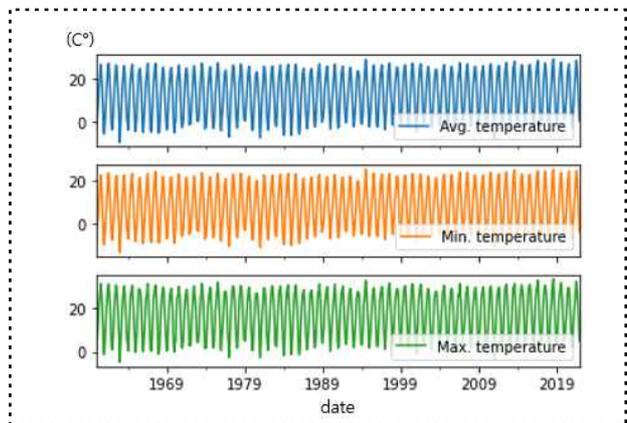


그림 1. 다변량 기온 데이터 시각화. 상단부터 일간 평균 기온, 일간 최저 기온, 일간 최고 기온

Fig. 1. Visualization of multivariate temperature data. Average daily temperature from top, minimum daily temperature, maximum daily temperature

3-2 실험 방법

본 연구에서는 문헌연구를 통해 확인한 전통적 이상치 탐지 알고리즘 중 Shewart control, IQR, Isolation forest, DBSCAN 알고리즘과 LSTM, GRU 기반의 이상치 탐지 알고리즘의 성능을 비교 분석하였다. 각 알고리즘은 파이썬을 통해 구현되었으며 특히, 딥러닝 모델인 LSTM과 GRU는 Nvidia Tesla K80 그래픽카드를 이용해 학습을 수행하였다. 구현한 이상치 탐지 알고리즘 중 Shewart control은 전체 데이터 평균으로부터 표준편차의 세 배만큼 상한선과 하한선을 도출하여 구현하였다. 둘 째로 IQR은 하한선을 1분위수로부터 사분위수 범위의 1.5배만큼 낮은 지점을, 상한선을 3분위수로부터 사분위수 범위의 1.5배만큼 높은 지점으로 설정하였다. 전통적 이상치 탐지 알고리즘 중 Isolation forest와 DBSCAN은 파이썬 라이브러리 사이킷런(Scikit-learn)으로 구현하였다. 이들 모델은 학습하는 과정에서 정상 데이터 분포에서 벗어난 노이즈를 동시에 분류해내는 비지도학습 방식으로 이상치 탐지를 수행한다. 즉, 라벨이 주어지지 않더라도 학습할 수 있지만 새로운 데이터가 입력될 경우 처음부터 새로 학습을 다시 해야한다.

LSTM과 GRU는 텐서플로 (Tensorflow) 프레임워크로 구현되었다. 두 모델 모두 두 개의 층을 쌓았으며 마지막 출력 시 선형 인공신경망 층을 통해 회귀를 수행한다. 이때 회귀를 수행한다는 것은 자기회귀 모델로써 라벨을 자기 자신의 데이터를 활용하는 기법이다. 시계열 데이터의 과거값을 이용해 미래의 시계열을 예측하도록 학습하고, 다변량 가우스 분포를 이용해 오차 벡터를 계산한다. 만약 예측값과 실제값의 오차가 다변량 가우스 분포의 특정 임계치를 벗어난다면 일반적으로 예측될 수 있는 값의 범위를 벗어나므로 이상치라 판단한다. 이상치 판정 시 가우스 분포 외에도 슈하르트 검정이나 사분위수 범위를 벗어난 데이터를 이상치로 판단할 수 있다. 전체적인 이상치 탐지 알고리즘 성능평가 프로세스는 그림 2와 같다.

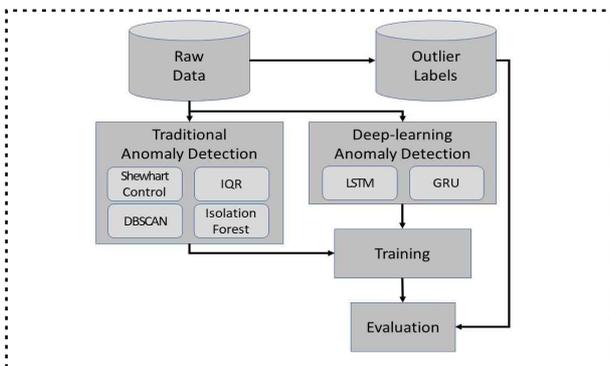


그림 2. 이상치 탐지 알고리즘 성능 평가 프로세스
 Fig. 2. Anomaly detection algorithm performance evaluation process

수집한 기온 데이터에는 매일 연속적으로 수집되어 데이터 규모가 크고 이상기온을 나타내는 레이블 정보가 포함되어 있지 않아, 알고리즘의 이상치 탐지 성능을 평가하기 위해서 라벨 정보가 필요하다. 따라서 본 연구에서는 기상청에서 정의한 이상기온을 기반으로 각 일자별 기온의 이상유무를 나타내는 레이블을 포함시켜 성능평가에 활용하였다. 세계기상기구에서 정의한 이상기온은 기상학적 측면에서 기온, 강수량 등의 기후요소가 30년 동안의 기록에 비해 현저히 높거나 낮은 수치를 나타내는 극한 현상으로 표준편차의 두 배 범위로 정의하고 있다[22]. 그림 3과 같이, 기상청에 따르면 통계적으로 평균 영하 7.78도보다 낮을 때, 그리고 평균 영상 33.64도보다 높을 때 이상치로 분류할 수 있다. 이 수치를 이용하여 30년 기상 데이터에서 발생한 이상기온으로 판단되는 일수는 총 169일로 나타났으며, 향후 실험에서는 각 이상치 탐지 알고리즘이 이 이상기온 데이터를 얼마나 정확하게 예측하는지를 바탕으로 성능을 검증하였다.

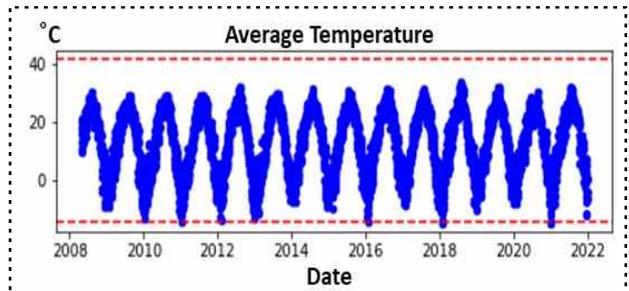


그림 3. 평균 기온 데이터와 기상청에서 정의한 이상치 기준(점선)
 Fig. 3. Anomaly detection algorithm performance evaluation process

3-3 실험 결과

이상치 탐지 알고리즘의 성능 비교 결과 딥러닝 기반의 알고리즘들이 전통적 알고리즘에 비해 성능이 우수하게 나타났다. 실험에 사용된 모든 모델의 이상치 탐지 분류 결과는 그림 4와 같이 나타났다. 우선 슈하르트 검정의 경우 전체 데이터에서 이상치를 탐지를 못하였다. 즉, 우리나라 평균 기온은 6시그마 내에 모두 분포하고 있다고 볼 수 있다. 기상청에서 정의한 이상치의 기준은 전체 평균 기온으로부터 표준편차 두 배에 해당하는 임계치이지만, 슈하르트 검정의 경우 표준편차 세 배를 기준으로 이상치를 분류한다. 슈하르트 검정에 따른 이상치 하한선이 영하 18.13도, 상한선은 43.97도로 나타났으며, 실제 기상청 데이터에서 이 범위 밖에서 나타나는 기온 데이터는 존재하지 않았다. IQR의 경우에는 전체 이상기온을 나타낸 169일 중 34%인 57일 만을 이상기온으로 판단하였고, 이때 이상치를 탐지하는 기준은 하한선이 영하 24도, 상한선 49.4도로 나타났다. 통계적 접근법인 IQR 이상치 탐지 기법과 슈하르트 검정 기법은 데이터의 분포를 통해 이

상치를 분류할 고정된 임계치를 산출하여 이상치를 탐지하기 때문에, 시계열의 특성을 반영하거나 분포의 변화를 고려하지는 못하는 한계점을 가진다. DBSCAN과 Isolation forest 알고리즘 기반의 이상치 탐지 방법은 비지도 학습 기반의 이상치 탐지 기계학습 알고리즘이다. DBSCAN 기법의 경우 전체 데이터들 사이의 유클리디안 거리를 계산하여 거리가 가까운 데이터들끼리의 클러스터를 도출한 후 클러스터에 포함되지 않는 데이터를 이상치로 판단하였다. 분석 결과 총 180개의 데이터가 클러스터에 포함되지 못하여 이상치로 판단되었고, 이상치에 대한 검출 정확성은 91%로 나타났다. Isolation forest은 모델 학습 결과 총 496개의 데이터를 이상치로 분류하였으며 실제 이상치 검출 정확도는 90%이다. DBSCAN과 Isolation forest 알고리즘을 비교하면 DBSCAN이 Isolation forest보다 정확도 측면에서 다소 높게 나타났으나, 정밀도와 재현율을 동시에 고려하는 성능지표인 F1-score의 경우에는 Isolation forest 알고리즘이 0.22로 0.17인 DBSCAN에 비해 높게 나타났다. 즉, Isolation forest가 양성 이상치 데이터를 좀 더 잘 분류한다고 볼 수 있다.

딥러닝 알고리즘인 LSTM과 GRU 알고리즘은 이전의 통계적 접근법과는 다른 방식으로 이상치 탐지를 수행한다. 이들은 직전 시계열 데이터를 입력받아 미래의 값을 예측하는 방식으로 작동한다. 학습 때에는 미래의 값과 모델이 예측한 값이 같아지도록 모델을 학습하게 된다. 학습 결과, LSTM 알고리즘은 검증 데이터셋의 평균 손실 함수 값은 0.05, 정확도는 95%, F1-score는 0.26으로 나타났다. GRU는 검증 데이터셋은 LSTM보다 전체적으로 높은 성능을 보였으며 검증 데이터셋 평균 손실 함수 0.03, 정확도는 96%였다. 결과적으로 GRU가 이상치 탐지를 위한 알고리즘에 대한 성능 실험 결과 정확도, 정밀도, 재현율, F1-score의 4가지의 성능 지표 모두에서 가장 좋은 성능을 보이는 것으로 나타났다. 특히 LSTM과 GRU와 같은 딥러닝 알고리즘은 분류 성능이 전통적 이상치 탐지 알고리즘에 비해 높았다. 또한, 디지털트윈에서의 동시에 온라인 예측이 가능하다는 점에서 실제 활용도에서 장점을 가진다. 통계 기반의 알고리즘은 분석 대상 데이터를 모두 활용하여 계산하기 때문에 실시간 이상치 탐지 예측이 어렵다는 한계점이 있지만, 딥러닝 모델은 과거 데이터로 모델을 학습하고 배포하여 실시간 입력되는 데이터를 즉시 처리할 수 있다. 따라서 이상치 탐지 알고리즘에서는 딥러닝 기반의 이상치 탐지 알고리즘을 사용하는 것이 효과적이다. 하지만 딥러닝 기반의 알고리즘이 가지고 있는 한계점도 존재한다. 본 실험에서는 기존 데이터만을 바탕으로 성능을 비교평가 하였으나, 기후 데이터는 기온, 습도, CO2 등 다양한 변수로 설명되어야 한다. 하지만 기존의 알고리즘들에서는 이러한 다변량 데이터를 기반으로 이상치인 이상기후를 예측하는 것에 한계가 있었다. 다음 4절에서는 이상의 3가지 한계점에 대해 디지털 트윈 환경에서의 이상치 탐지 알고리즘이 고려해야 할 부분에 대해 보다 자세히 기술하였다.

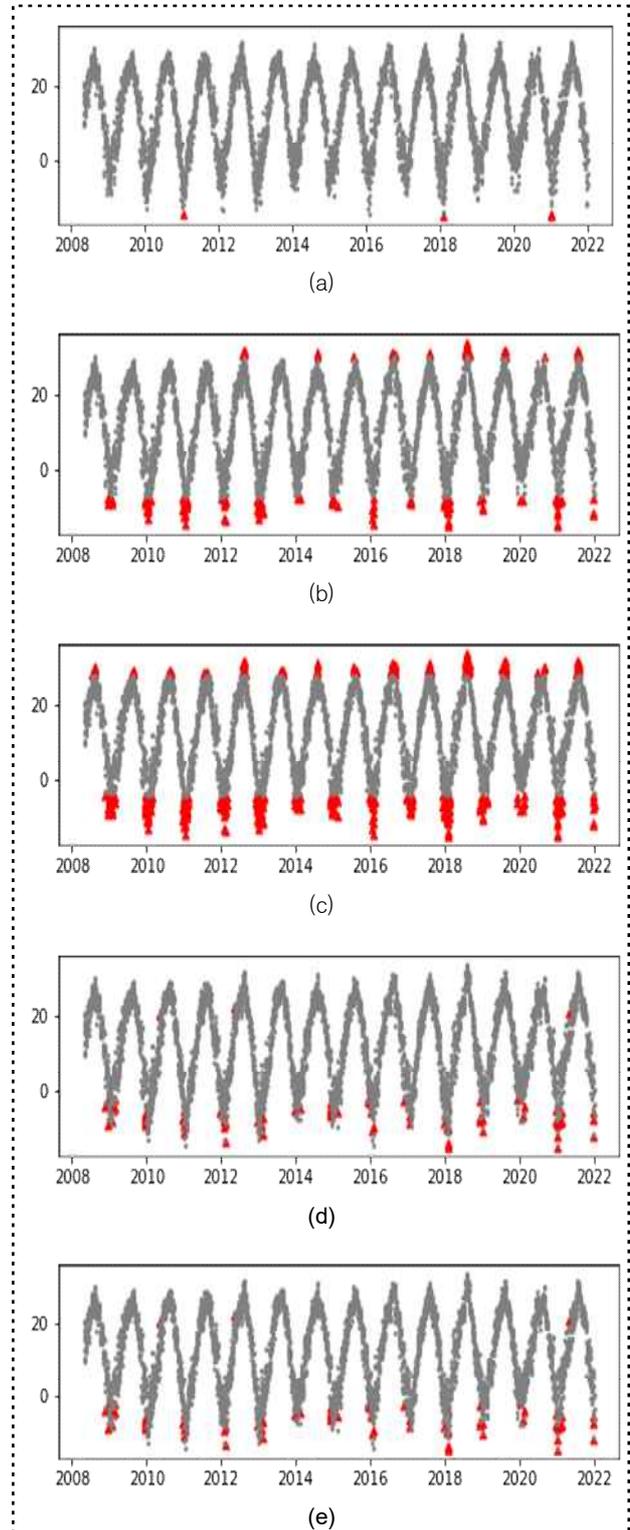


그림 4. 2008년부터 2022년까지의 (a) IQR (b) DBSCAN (c) Isolation forest (d) LSTM (e) GRU의 이상치 탐지 분류 결과

Fig. 4. Anomaly detection classification results of (a) IQR (b) DBSCAN (c) Isolation forest (d) LSTM (e) GRU from 2008 to 2022

표 1. 이상치 탐지 알고리즘 성능

Table 1. Performance of anomaly detection algorithms

Model	Accuracy	Precision	Recall	F1-score
Shewhart control	0.0	0.0	0.0	0.0
IQR	0.34	0.05	0.02	0.03
DBSCAN	0.91	0.01	0.42	0.17
Isolation forest	0.90	0.02	0.34	0.22
LSTM	0.95	0.17	0.47	0.26
GRU	0.96	0.18	0.47	0.27

IV. 디지털 트윈 이상치 탐지 주요 이슈

4-1 컨텍스트 이상치 탐지

기존의 경우에는 대부분 포인트 이상치 탐지 기법을 사용하고 있다. 포인트 이상치 탐지 기법은 과거 데이터를 기반으로 현재 시점 값이 이상치인지 분류하는 기법이다. 포인트 검출 후 즉시 이상치인지 여부를 판단하더라도 이는 디지털 트윈의 과거 이벤트이다. 컨텍스트 이상치 탐지 기법은 지속적으로 데이터의 흐름이 비정상적으로 판단되는 것을 의미한다. 기존의 포인트 이상치 탐지 기법은 레이블이 정확히 지정되어 있는 데이터에서 좋은 성능을 보일 수 있지만, 많은 환경이 이러한 데이터를 제공해주지 못한다. 따라서 비정상적인 상황을 유연하게 감지할 수 있는 컨텍스트 이상치 탐지 모델의 연구가 필요하다. 디지털 트윈은 센서, 건물관리, 제조 설비 등 데이터가 실시간으로 생성되는 동적 데이터이므로 디지털 트윈의 가치를 극대화하기 위해서는 과거의 데이터에 대해 이상치를 판단하는 알고리즘보다 맥락적으로 비정상적인 상황을 감지할 수 있는 디지털 트윈 기반의 이상 탐지 시스템이 필요하다. 포인트 이상치 탐지 기법 주로 데이터 포인트 그 자체의 이상치로 분류될 확률을 계산한다. 전통적인 포인트 기반 이상치 탐지 방법은 정확하게 레이블이 지정된 데이터에서 잘 수행될 수 있지만, 많은 환경에서는 이러한 데이터를 수집하기 쉽지 않다. 그러나 컨텍스트 이상치 탐지 기법은 데이터의 분포 또는 시계열의 확률 분포가 변화하는 전환점을 고려하여 이상치 탐지를 수행할 수 있다.

포인트 이상치 탐지 기법은 과거 데이터를 바탕으로 이상치를 판별하는 사후방식으로, 비정상 신호를 사전에 검출하기 어렵다. 한편, 컨텍스트 이상치 탐지 기법은 데이터의 전체적인 흐름을 파악하여 이상 징후를 탐지할 수 있으므로, 사전에 예측할 수 있다는 장점이 있다. Hayes & Capretz [23]의 연구에서, 그들은 다변량 군집화 방법을 기반으로 컨텍스트 이상치 탐지 모델을 구성하여 대규모 센서 데이터에서 실시간으로 이상을 식별했다. Lu 외 (2020)의 연구에서는 컨텍스트 이상치 탐지를 데이터 분포의 전환점을 고려하는 베이지

안 방법과 접목하여 수행하였다[24]. 문맥 이상 검출에 대한 정확한 방법론은 없지만, 디지털 트윈 데이터의 특성상 실시간으로 생성되며 데이터 라벨링이 쉽지 않다는 점을 고려할 때 실시간으로 적용할 수 있는 문맥 이상 검출 모델이 필요하다.

실제로 전통적 이상치 탐지 기법과 딥러닝 이상치 탐지 기법의 기후 데이터 실험 결과를 살펴보면 문맥을 고려해야 할 필요성을 발견할 수 있다. 기후 데이터를 장단기 메모리로 학습한 후 이상치로 분류한 결과가 그림 5와 같이 나타나고 있다. 세로로 그려진 직선이 전환점으로, 이는 데이터 분포가 바뀐 시점이며 'X'자로 그려진 지점이 이상치로 분류한 지점이다. 결과를 살펴보면 대부분의 이상치가 분포가 바뀐 전환점 주변에 머물러 있으며 일부 데이터 부분에 집중되어 나타나고 있다. 그림 5에서 2016년과 2020년의 두 모델 결과에서 차이가 확연히 나타나고 있다. 두 구간 모두 전환점이 나타난, 즉 기후변화가 일어난 상황에서 기후 환경 자체가 바뀐 상태를 고려한다면 오히려 전통적 이상치 탐지 방법론이 2016년도에 예측한 이상치는 정상일 가능성도 있다. 반대로 기상청이 제시하는 이상치 기준에 따르면 2020년도 겨울에는 기후변화가 아닌데 맥락상 이상치로 분류되어야 하는 데이터가 나타났다. 데이터 분포가 계절에 따라 바뀌는 기후와 디지털 트윈은 기상청의 기준처럼 이상치를 분류하는 임계점을 고정하고 분류하기보다 동적으로 모델이 판단할수 있어야 한다. 실험결과와 같이 이상치로 탐지되는 경우가 있더라도 전체 분포를 볼 때 사실은 이상치가 아닌 정상 값이 될 수 있다. 이는 포인트 이상치 탐지 알고리즘이 하나의 포인트만을 고려하여 이상치를 분류하기 때문에 이러한 현상이 나타나는 것이다. 따라서 맥락을 고려할 수 있는 컨텍스트 이상치 탐지 기법을 적용해야 한다.

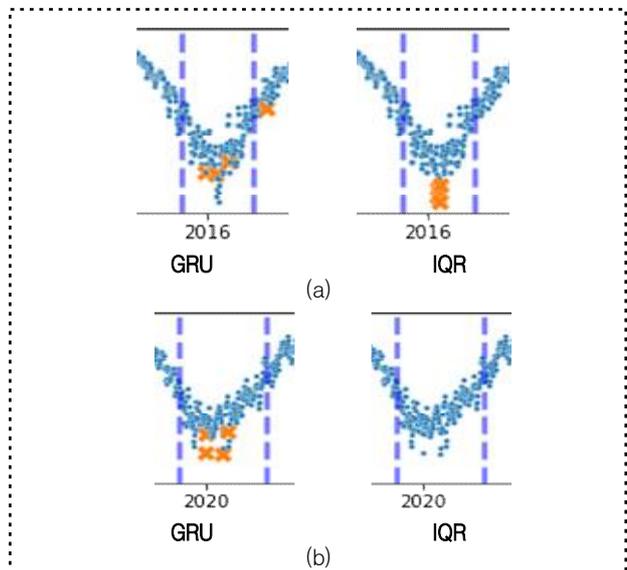


그림 5. (a) 2016년과 (b) 2020년 GRU와 IQR 이상치 탐지 분류 결과

Fig. 5. (a) 2016 and (b) 2020 GRU and IQR anomaly detection classification results

4-2 다변량 시계열 데이터

오늘날 스마트 팩토리, 발전소, 웨어러블 기기 등의 다양한 디지털 트윈 환경에서 다변량 시계열 데이터가 생산되고 있으며 이러한 디지털 트윈의 수는 점차 증가하고 있는 추세이다. 실제 디지털 트윈은 복잡계로 구성되어있기 때문에, 다양한 원인으로 인하여 비정상 데이터가 다변량 시계열 데이터를 오염시켜 전체적인 데이터와 시스템의 품질을 저하시킬 수 있다. 이러한 상황에서는 데이터의 품질 관리 기법의 일종인 이상치 탐지를 디지털 트윈에 적용하기 위해 평균, 이동평균, k-최근접 이웃, 그리고 슈하르트 차트와 같은 전통적인 통계 기법을 사용할 수 있다. 그러나 이러한 기법들은 시계열 데이터 간 상관관계를 고려하지 못한다는 한계점이 있다 [25]. 대부분의 디지털 트윈에서 생산되는 다변량 시계열 데이터로부터 서로 다른 시계열의 관계를 적절히 반영할 수 있다면 이상치 탐지의 성능을 높일 수 있다. 그러나 여전히 디지털 트윈 환경에서의 다변량 시계열 특성을 고려하여 변수 간 연관성 또는 시간 의존성을 반영한 이상치 탐지 연구가 활발히 이루어지지 않고 있다.

실제로 시계열 데이터의 변수 개수에 따라 모델이 예측하는 결과에 어떤 차이가 나타나는지 비교하고자 그림 6과 같이 평균 기온을 예측하는 장단기 메모리에 전체 평균, 최저, 최고 기온과 같은 다변량을 입력했을 때와 평균 기온만을 입력한 경우를 비교하였다. 다변량 모델의 경우 과거의 모든 다변량 데이터를 입력받아 미래의 평균 기온을 예측하며 단변량 모델은 과거의 평균 기온을 입력받아 미래의 평균 기온을 예측한다. 만약 다변량이든 단변량이든 딥러닝 모델이 변수 간 상관 관계를 고려하지 못한다면 결과는 같아야 하지만, 실제로는 그림 6와 같이 결과가 상이한 것으로 나타났다. 정량적인 성능 비교를 하면 다변량 모델의 손실 함수는 0.492이며 단변량 모델은 0.0527로 다변량 모델의 성능이 뛰어났다. 정량적으로 판단해보았을 때, 2021년과 2020년의 결과를 살펴보면 다변량 모델이 눈에 띄게 실제로 이상치로 보이는 값을 이상치로 분류해냈으므로 종합적으로 판단해보았을 때 디지털 트윈 환경에서 생산되는 다변량 시계열 데이터를 위한 이상치 탐지에 있어서 변수의 종합적인 상관 관계를 다룰 수 있는 방법론이 필요하다고 볼 수 있다.

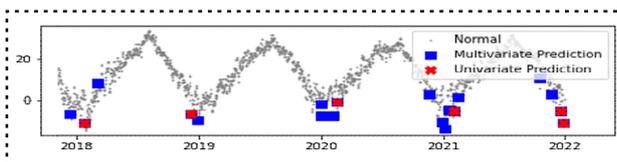


그림 6. 다변량 데이터로 학습한 GRU 모델과 단변량 데이터로 학습한 GRU 모델의 평균 기온 이상치 분류 예측 결과 차이
Fig. 6. Differences in average temperature anomaly classification prediction results between GRU learned from multivariate data and GRU learned from univariate data

4-3 클래스 불균형 문제

대부분의 경우 비정상 데이터보다 정상 데이터의 분포가 더 많이 발생 되기 때문에 비정상 데이터를 탐지하기 위한 모델 학습시 탐지정확도가 높지 않을 수 있다. 비정상 값은 전체 관측값의 분포에서 벗어난 수치로 정의하기 때문에 학습에 사용할 수 있는 정보가 절대적으로 부족해진다. 특히 데이터를 바탕으로 학습하는 기계학습 및 딥러닝 분야에서는 클래스 불균형 문제가 중요한 이슈로 다뤄지고 있다. 기계학습에서 클래스 불균형이 문제가 되는 이유는 소수 클래스의 특징을 충분히 활용할 기회가 적게 주어지기 때문이며, 모델은 주어진 목적함수를 기준으로 손실을 최소화하는 간편한 방법으로 대다수 클래스로 치우치게 예측할 수 있다. 기존 클래스 불균형을 다루는 방법론으로는 오버샘플링과 언더 샘플링 기법이 있다. 기존 디지털 트윈 데이터에서 클래스 불균형 문제를 해결하기 위해 다양한 시도가 있었다. Bai 외(2020)의 연구에서는 교통 디지털 트윈으로부터 오버샘플링의 일종인 스모트 기법을 통해 비정상 교통량을 증대시켰다[26]. 기존에 기상청은 이상 기온을 지난 30년간의 기온 분포로부터 표준편차의 두 배를 벗어난, 즉 전체 분포의 95%를 벗어나는 관측치를 이상 기온으로 정의하고 있다. 이러한 기준으로 라벨을 부여하고 이상치 탐지 분석을 수행한다면 기온 데이터 95%는 정상 수치, 나머지 5%는 이상치로 간주하고 모델을 학습시킨다. 이는 클래스 불균형 문제로 이어질 수 있으므로 샘플링 기법을 적용할 수 있어야 한다.

V. 결 론

디지털 트윈의 구축을 위해서는 무엇보다 데이터의 품질이 중요하며 이를 위해 실시간보다 빠른 시뮬레이션으로 데이터를 관리 및 모니터링할 수 있어야 한다. 이상치 탐지는 데이터 속 다른 관측값들과는 다른 양상을 보이는 이상치를 발견해내는 데이터 분석기법으로 보안, 의료진단, 제조공정에서의 모니터링 등 다양한 산업 분야에 적용되고 있다. 특히 오늘날 딥러닝을 활용한 이상치 탐지 알고리즘에 대한 연구가 활발히 수행되고 있으며 특히 장단기 메모리와 게이트 순환 유닛 등의 순환신경망 계열 모델 구조와 오토인코더를 활용하고 있다. 그러나 아직까지는 디지털 트윈과 같은 특별한 환경에서 이상치 탐지를 수행한 기존 연구가 부족하기 때문에 본 연구에서는 기상청의 기온 데이터를 학습한 이상치 탐지 모델로부터 도출된 결과를 토대로 디지털 트윈 이상치 탐지 이슈를 도출하였다. 디지털 트윈 이상치 탐지 주요 이슈로는 클래스 불균형 문제가 발생하는 다변량 시계열 데이터를 컨텍스트 이상치 탐지 기법으로 해결해야 한다고 정리할 수 있다. 따라서 성공적인 디지털 트윈 구축을 위한 이상치 탐지 알고리즘 연구가 필요하며 특히 향후 연구로 데이터의 문맥을 파

악하고 데이터 분포의 전환점을 활용하는 새로운 컨텍스트 이상치 탐지 알고리즘을 개발해야 한다.

감사의 글

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2018-0-00387, 축산질병 예방 및 통제 관리를 위한 ICT 기반의 지능형 스마트 안전 축사 기술 개발)

참고문헌

- [1] M. Grieves, and J. Vickers, Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems, in *Transdisciplinary perspectives on complex systems*, Springer, pp. 85-113, 2017.
- [2] T. H. J. Uhlemann, C. Lehmann, and R. Steinhilper, "The digital twin: Realizing the cyber-physical production system for industry 4.0." *Procedia Cirp*, Vol. 61, pp. 335-340, 2017. <https://doi.org/10.1016/j.procir.2016.11.152>
- [3] M. Liu, S. Fang, H. Dong, and C. Xu, "Review of digital twin about concepts, technologies, and industrial applications." *Journal of Manufacturing Systems*, Vol. 58, pp. 346-361, March 2021. <https://doi.org/10.1016/j.jmsy.2020.06.017>.
- [4] Y. Lu, C. Liu, I. Kevin, K. Wang, H. Huang, and X. Xu, "Digital Twin-driven smart manufacturing: Connotation, reference model, applications and research issues." *Robotics and Computer-Integrated Manufacturing*, Vol. 61, 101837, August 2019. <https://doi.org/10.1016/j.rcim.2019.101837>
- [5] R. A. A. Habeeb, F. Nasaruddin, A. Gani, I. A. T. Hashem, E. Ahmed, and M. Imran, "Real-time big data processing for anomaly detection: A survey." *International Journal of Information Management*, Vol. 45, pp. 289-307, March 2019. <https://doi.org/10.1016/j.ijinfomgt.2018.08.006>
- [6] Y. Luo, Y. Xiao, L. Cheng, G. Peng, and D. Yao, "Deep learning-based anomaly detection in cyber-physical systems: Progress and opportunities." *ACM Computing Surveys (CSUR)*, Vol. 54, No. 5, pp.1-36, June 2022. <https://doi.org/10.1145/3453155>
- [7] F. E. Grubbs, "Procedures for detecting outlying observations in samples." *Technometrics*, Vol. 11, No. 1, pp. 1-21, 1969. <https://doi.org/10.1080/00401706.1969.10490657>
- [8] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey." *ACM computing surveys (CSUR)*, Vol. 41, No. 3, pp. 1-58, July 2009. <https://doi.org/10.1145/1541880.1541882>
- [9] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." *kdd*, Vol. 96, No. 34, pp. 226-231, August 1996.
- [10] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest." in *2008 eighth IEEE international conference on data mining*, NW Washington, DC, pp. 413-422, December 2008. <https://doi.org/10.1109/ICDM.2008.17>
- [11] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed and H. Arshad, "State-of-the-art in artificial neural network applications: A survey." *Heliyon*, Vol. 4, No. 11, e00938, November 2018. <https://doi.org/10.1016/j.heliyon.2018.e00938>
- [12] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "LSTM-based encoder-decoder for multi-sensor anomaly detection." *arXiv preprint arXiv:1607.00148*, July 2016.
- [13] H. D. Nguyen, K. P. Tran, S. Thomassey, and M. Hamad, "Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management." *International Journal of Information Management*, Vol. 57, 102282, February 2021. <https://doi.org/10.1016/j.ijinfomgt.2020.102282>
- [14] P. Liu, X. Sun, Y. Han, Z. He, W. Zhang, and C. Wu, "Arrhythmia classification of LSTM autoencoder based on time series anomaly detection." *Biomedical Signal Processing and Control*, Vol. 71, 103228, October 2021. <https://doi.org/10.1016/j.bspc.2021.103228>
- [15] S. Maleki, S. Maleki, and N. R. Jennings, "Unsupervised anomaly detection with LSTM autoencoders using statistical data-filtering." *Applied Soft Computing*, Vol. 108, 107443, April 2021. <https://doi.org/10.1016/j.asoc.2021.107443>
- [16] H. D. Nguyen, K. P. Tran, S. Thomassey, and M. Hamad, "Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management." *International Journal of Information Management*, Vol. 57, 102282, February 2021. <https://doi.org/10.1016/j.ijinfomgt.2020.102282>
- [17] T. Kieu, B. Yang, and C. S. Jensen, "Outlier detection for multidimensional time series using deep neural networks." in *2018 19th IEEE International Conference on Mobile Data Management (MDM)*, Aalborg, pp. 125-134. June 2018. <https://doi.org/10.1109/MDM.2018.00029>

- [18] Y. He, N. Zhang, and A. Wang, "Digital twin process and simulation operation control technology for intelligent manufacturing unit." *IOP Conference Series: Materials Science and Engineering*, Vol. 836, No. 1, 12010, April 2020. <https://doi.org/10.1088/1757-899X/836/1/012010>
- [19] K. Cho, B. V. Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches." *arXiv preprint arXiv:1409.1259*, October 2014.
- [20] S. Huang, G. Wang, Y. Yan, and X. Fang, "Blockchain-based data management for digital twin of product." *Journal of Manufacturing Systems*, Vol. 54, pp. 361-371, 2020. <https://doi.org/10.1016/j.jmsy.2020.01.009>
- [21] Z. Qu, L. Su, X. Wang, S. Zheng, X. Song, and X. Song, "A unsupervised learning method of anomaly detection using GRU." in *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Shanghai, pp. 685-688. January 2018. <https://doi.org/10.1109/BigComp.2018.00126>
- [22] J. I. Yun, S. O. Kim, K. M. Shim, and Y. S. Shin, "A field-specific service for agrometeorological hazards Early Warning." *Korean Society of Agricultural and Forest Meteorology*, Vol. 19, No. 3, pp. 180-194, September 2017. <https://10.5532/KJAFM.2017.19.3.180>.
- [23] M. A. Hayes, and M. A. Capretz, "Contextual anomaly detection framework for big sensor data." *Journal of Big Data*, Vol. 2, No. 1, pp. 1-22, February 2016. <https://doi.org/10.1186/s40537-014-0011-y>
- [24] Q. Lu, X. Xie, A. K. Parlikad, and J. M. Schooling, "Digital twin-enabled anomaly detection for built asset monitoring in operation and maintenance." *Automation in Construction*, Vol. 118, 103277, May 2020. <https://doi.org/10.1016/j.autcon.2020.103277>
- [25] Y. Guo, W. Liao, Q. Wang, L. Yu, T. Ji, and P. Li, "Multidimensional time series anomaly detection: A gru-based gaussian mixture variational autoencoder approach." in *Asian Conference on Machine Learning*, Beijing, pp. 97-112, November 2018.
- [26] S. Bai, Z. He, Y. Lei, W. Wu, C. Zhu, M., Sun and J. Yan, "Traffic anomaly detection via perspective map based on spatial-temporal information matrix." in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Long Beach, pp. 117-124, June 2019.



신윤종(Yoon-Jong Shin)

2022년 : 한신대학교 e-비즈니스 학과
(경영학사)

2015년~2022년: 한신대학교 e-비즈니스학과
2022년~현 재: 경희대학교 인공지능학과 석사과정
※ 관심분야 : 기계학습(Machine Learning), 이상치 탐지
(Anomaly Detection) 등



이남연(Namyeon Lee)

2008년 : 경희대학교 일반대학원
(경영학석사-경영정보시스템)
2013년 : 경희대학교 일반대학원
(경영학박사-경영정보시스템)

2014년~2017년: 성결대학교 파이테이아학부 조교수
2017년~현 재: 한신대학교 IT경영학과 부교수
※ 관심분야 : 텍스트마이닝, 기계학습, 디지털 트윈, 이상치
탐지 등