

외부 환경 데이터를 활용한 비지도 공정 이상치 탐지 모델의 성능 예측 연구

최유빈, 조재영, 권구영, 김선혁
국립공주대학교

202301893@smail.kongju.ac.kr, whwodud02@smail.kongju.ac.kr, gykwon@kongju.ac.kr,
seonh@kongju.ac.kr

A Study on the Performance Prediction of Anomaly Detection Model using External Environmental Data

Yu Bin Choi, Jae Yeong Jo, Gu-Young Kwon, Seon Hyeog Kim
Kongju National University

요 약

본 논문은 정답 레이블이 없는 실제 공장 시계열 데이터의 이상치를 탐지하고, 성능을 검증하는 방법론을 제안한다. 실제 산업 현장의 데이터는 센서별로 분포와 양이 불균등하여 이상치 탐지 시, 딥러닝 모델의 과적합 위험이 크다는 한계가 있다. 이러한 문제를 해결하기 위해 본 연구에서는 시계열 데이터의 시간적 특징을 학습하는 Long Short-Term Memory와 비지도 이상치 탐지 모델인 Deep Support Vector Data Description을 결합한 모델을 기본으로 설정하였다. 나아가 과적합을 방지하고 일반화 성능을 높이기 위해, 메타학습 기법을 적용한 모델을 제안하고 두 모델의 성능을 비교 분석하였다. 또한, 객관적인 검증 기준이 없는 문제를 해결하고자 공정 특성에 기반하여, 외부 환경 데이터를 준-정답으로 활용하여 탐지된 이상치의 정당성을 검증하였다.

I. 서 론

최근 스마트 팩토리의 대두와 함께, 공장 설비에서 수집되는 많은 양의 데이터를 분석하여 문제를 미리 예측하는 기술이 핵심으로 자리 잡고 있다. 공장의 설비 이상 징후를 조기에 발견하는 것은 갑작스러운 고장으로 인한 생산을 막는 핵심 과정으로, 정확한 이상치 탐지는 필수적이다. 따라서 본 연구에서는 딥러닝 모델을 활용하여 실제 공장 데이터의 이상치를 탐지하고자 한다. 이를 위해 시간적 패턴을 잘 학습하는 LSTM(Long Short-Term Memory)과 비지도 이상치 탐지에 특화된 Deep SVDD(Deep Support Vector Data Description)를 결합한 모델을 기본으로 사용한다. 하지만 적은 데이터만으로 학습해야 하는 경우에는 과적합에 빠질 위험이 있다. 따라서 메타학습 기법인 MAML(Model-Agnostic Meta-Learning)을 추가로 결합해 2 차 검증을 진행할 것이다. 최종적으로, 두 모델의 탐지 결과를 비교, 분석하고, 정답 레이블이 없는 데이터의 한계를 극복하기 위해 외부 환경 데이터를 활용하여 탐지된 이상치의 성능을 검증하는 것을 목표로 한다.

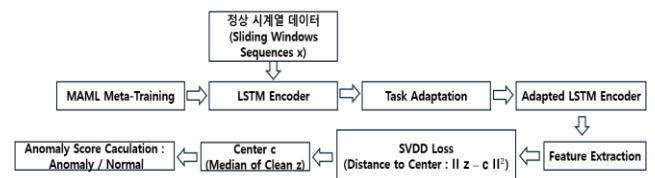
II. 본론

2.1 Deep SVDD-LSTM 모델

Deep SVDD와 LSTM을 결합해 이상치를 추출하는 모델이다. 먼저, LSTM은 시계열 데이터의 장기적 의존성 문제를 해결하는 순환 신경망(RNN)의 한 종류이다. 입력, 망각, 출력 게이트를 가지며, 이를 통해 과거의 중요한 정보는 기억하고, 불필요한 정보는 잊는 메커니즘을 갖는다.

Deep SVDD는 딥러닝 인코더를 학습시켜, 정상 데이터의 특징 벡터를 최소 부피의 초구(Hypersphere) 안에

모이도록 가중치를 최적화하는 비지도 이상치 탐지 기법이다. 학습이 완료된 후, 정상 데이터는 구 중심에 가깝게 매핑되는 반면, 이상 데이터는 구 경계 밖이나 중심에서 멀리 떨어진 곳에 매핑된다. 이 중심점으로부터의 거리가 바로 이상치 점수(Anomaly Score)이다. 이상치 탐지는 개별 데이터 포인트 또는 고정된 크기의 벡터를 기반으로 작동하는데, 이는 데이터의 시간적 순서를 고려하지 못한다는 한계를 가진다.



<그림 1> Deep SVDD-LSTM 모델 블록 다이어그램

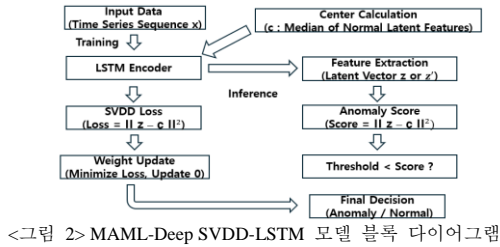
본 모델에서 LSTM은 Deep SVDD의 인코더로 활용되어, 시퀀스 길이의 센서 데이터, 파생 특성을 입력으로 받아 시간적 패턴을 압축한 단일 특징 벡터를 생성한다. 이를 통해 LSTM은 Deep SVDD가 시간적 흐름을 인식할 수 있도록 정제된 정보를 제공하며, Deep SVDD는 해당 특징을 기반으로 이상치를 효과적으로 판별한다.

2.2 MAML-Deep SVDD-LSTM 모델

이 모델은 앞서 구성한 Deep SVDD-LSTM 구조에 MAML을 결합한 형태이다. MAML은 소량의 데이터로도 빠른 학습을 할 수 있도록, 다양한 작업에서 공통적으로 유용한 초기 파라미터를 찾는 것이 MAML 목표이다.

본 연구에서는 MAML 을 센서 그룹 단위로 적용하여, 그룹 내 센서들의 공통된 정상 상태 특성을 학습하고 이를 각 개별 센서 모델의 초기값으로 활용하였다.

실제 공장 데이터는 센서별로 데이터 분포가 상이하고, 정상 상태 정의 기준이 다르다. 센서별 데이터 불균형이 심한 상황에서 각 센서 모델을 독립적으로 처음부터 학습시키는 기존 방식은 과적합 위험을 갖는다. MAML 은 이러한 문제를 완화하기 위해, 다양한 센서 그룹의 학습 경험을 통해 일반화된 초기 가중치를 Deep SVDD 에게 전달해준다. 이를 기반으로 각 개별 센서의 Deep SVDD-LSTM 모델은 무작위 가중치에서 시작하는 대신, 사전 훈련된 최적의 출발점에서 학습을 시작한다. 따라서 MAML 의 도입은 각 센서 모델이 과적합을 피하고, 높은 일반화 성능을 갖도록 유도한다.



<그림 2> MAML-Deep SVDD-LSTM 모델 블록 다이어그램

3.1 모델 학습

연구에 사용된 데이터는 특정 공정의 운영을 위해 저온 유지가 필수적인 생산 시설의 시계열 데이터로, 6 개월 간 유량계, 압력계, 온도계, 전력량계의 센서 값을 포함한다. 해당 공정은 외부 기온 상승 시, 온도를 낮게 유지하기 위해 냉수 공급량을 높이는 운영 특성을 가진다. 따라서 냉각수 유량 및 관련 설비의 전력 사용량이 증가하게 될 것이라고 가정하였다.

수집된 원본 데이터를 전처리 과정을 수행하였다. 데이터 결측치 보간은 누적 데이터의 시간별 변화량을 계산하여 진행하였으며, 롤링 윈도우 기법을 사용해, 생성한 이동 평균과 이동 표준 편차를 변수로 설정하였다.

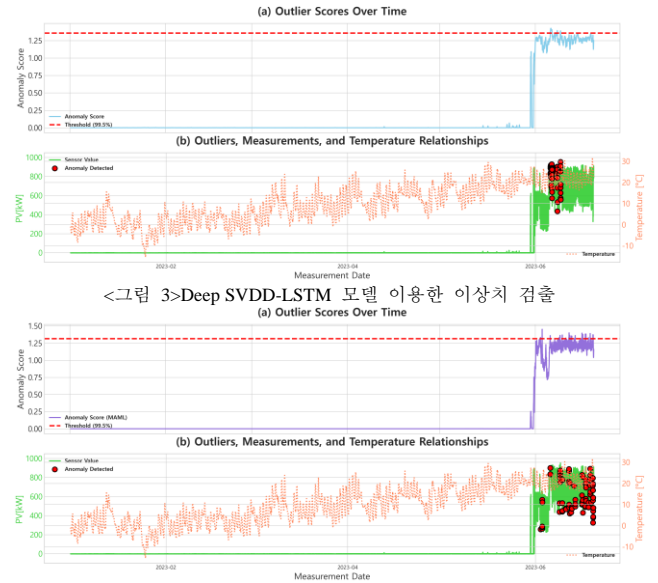
본 연구에서는 센서 유형별 특성을 고려해 학습 데이터를 구성하였다. 온도 센서는 전체 기간의 데이터를 사용했지만, 그 외 센서인 유량계, 압력계, 전력량계는 유의미하게 가동되는 상태를 학습시키기 위해, 센서 값이 0 보다 큰 양수 값 만을 대상으로 학습데이터와 훈련데이터를 구성했으며, 이를 토대로 학습을 진행하였다.

3.2 결과 및 성능 예측

두 모델 모두 외부 기온이 상승하는 6 월 초부터 이상치 점수가 급등하는 패턴을 성공적으로 탐지한 것을 확인할 수 있다. 이는 두 모델이 기온 변화에 따른 공정 상태 변화를 의미 있는 이상 신호로 인식했음을 보여준다. 두 결과를 비교했을 때, <그림 3>(a)는 비정상 구간의 초입에서 이상치 점수가 임계값(Threshold)에서 불안정하게 변동하는 모습을 보이지만, <그림 4>(a)는 같은 구간에서 이상치 점수가 훨씬 더 급격하게 임계값을 넘었고, 비정상 구간 내에서도 안정적으로 정상과 비정상의 경계를 뚜렷하게 구분하였다. 이는 MAML 의 메타 학습을 수행한 모델이 더 우수한 성능을 띠는 것을 보여준다.

또한 객관적인 수치로 뒷받침하기 위해, 각 모델이 산출한 이상치 점수와 기온 간의 피어슨 상관계수를 계산하였다. 상관 계수는 -1 부터 1 사이의 값을 가지며, 1 에 가까울수록 두 데이터가 동일한 방향으로 움직이는 강한 양의 선형 관계를 의미한다. 두 모델 모두 0.65 를 상회하는 높은 상관 계수를 보여, 두 모델의 이상치 점수가

기온 변화라는 물리적 현상과 매우 밀접하게 연관되어 있음을 확인하였다. 이를 통해 두 모델이 산출한 Threshold 가 논리적으로 정당하다고 해석하였으며, MAML 을 적용한 모델의 상관 계수가 0.69 로 더 높게 나타난 것은 보다 정확한 패턴 분석을 수행하였다는 것을 뒷받침하는 결과로 바라본다.



<그림 3> Deep SVDD-LSTM 모델 이용한 이상치 검출

III. 결론

본 연구에서는 정답 레이블이 없는 실제 공장 시계열 데이터에 대해 딥러닝 기반의 비지도 이상치 탐지 모델을 구현하고, 공정에 직접적인 원인을 미치는 외부 데이터인 기온을 활용하여 그 성능을 비교 검증하였다. 실험 결과, 외부 기온 상승에 따라 이상치가 나타나는 것을 확인했다. 이는 저온 유지가 필수적인 공정 특성상 외부 기온이 상승하면 냉각 시스템의 가동률이 높아져 관련 설비의 전력 사용량이 증가할 것이라는 가설과 부합하는 결과이다. 따라서 객관적인 평가 기준이 없는 비지도 학습 환경에서도 공정의 물리적 특성과 관계가 있는 외부 데이터를 준-정답(Quasi-ground truth)으로 활용하여 모델의 성능을 실증적으로 예측하고 검증할 수 있다는 핵심 가설을 성공적으로 입증하였다. 또한 센서별 데이터 양과 분포가 다른 실제 산업 환경에서 MAML 의 그룹별 사전 훈련 방식이 과적합을 방지하고 모델의 일반화 성능을 높이는 데 효과적이라는 것을 보여준다.

ACKNOWLEDGMENT

본 연구는 산업통상자원부(MOTIE)와 한국에너지기술평가원(KETEP)의 지원 (No. RS- 2023-00237018) , 그리고 국립공주대학교 학술연구지원사업의 의하여 연구되었음.

참 고 문 헌

- [1] 김서림, 강민지, 김선혁, 허태욱, 권구영. (2025). “시계열 데이터 이상치 탐지를 위한 LSTM-USAD 모델 개발”, *한국통신학회 학술대회논문집*.
- [2] 강민지, 김서림, 김선혁, 허태욱, 권구영. (2025). “LSTM-VAE 기반 공장 이상 모니터링 기술 개발”, *한국통신학회 학술대회논문집*.