ISSN 1229-1889(Print)
ISSN 2287-9005(Online)
J Korean Soc Qual Manag Vol. 50, No.3: 459-471, September 2022
https://dx.doi.org/10.7469/JKSQM.2022.50.3.459

TadGAN 기반 시계열 이상 탐지를 활용한 전처리 프로세스 연구

이승훈* · 김용수**†

* 경기대학교 일반대학원 산업경영공학과
** 경기대학교 산업시스템공학과

A Pre-processing Process Using TadGAN-based Time-series Anomaly Detection

Lee, Seung Hoon* · Kim, Yong Soo**†

* Department of Industrial and Management Engineering, Kyonggi University Graduate School
** Department of Industrial System Engineering, Kyonggi University

ABSTRACT

Purpose: The purpose of this study was to increase prediction accuracy for an anomaly interval identified using an artificial intelligence-based time series anomaly detection technique by establishing a pre-processing process.

Methods: Significant variables were extracted by applying feature selection techniques, and anomalies were derived using the TadGAN time series anomaly detection algorithm. After applying machine learning and deep learning methodologies using normal section data (excluding anomaly sections), the explanatory power of the anomaly sections was demonstrated through performance comparison.

Results: The results of the machine learning methodology, the performance was the best when SHAP and TadGAN were applied, and the results in the deep learning, the performance was excellent when Chi-square Test and TadGAN were applied. Comparing each performance with the papers applied with a Conventional methodology using the same data, it can be seen that the performance of the MLR was significantly improved to 15%, Random Forest to 24%, XGBoost to 30%, Lasso Regression to 73%, LSTM to 17% and GRU to 19%.

Conclusion: Based on the proposed process, when detecting unsupervised learning anomalies of data that are not actually labeled in various fields such as cyber security, financial sector, behavior pattern field, SNS. It is expected to prove the accuracy and explanation of the anomaly detection section and improve the performance of the model.

Key Words: Pre-processing Process, Time-series Anomaly Detection, TadGAN, Unsupervised Learning

● Received 13 July 2022, 1st revised 2 August 2022, accepted 5 August 2022
† Corresponding Author (kimys@kgu.ac.kr)

* 본 연구는 경기도의 경기도 지역협력연구센터 사업의 일환으로 수행하였음[GRRC경기2020-B03, 산업통계 및 데이터마이닝 연구].

1. 서 론

최근 4차 산업혁명 및 센서 기술의 발전으로 인공지능 기술이 발전함에 따라 딥러닝 이상 탐지 방법이 대두되면서 각종 산업에서 딥러닝을 기반한 이상 탐지 기술을 도입하고 있다. 많은 기업에서 인공지능을 활용한 이상 탐지 방법을 활용해 다양한 문제를 해결하고자 한다. 이로 인해 센서 데이터를 학습하여 예측, 분류하는 인공지능의 활용이 중요해지고 있다. 특히, 제조업에서 공정 데이터를 분석하여 양품과 불량을 분석하거나 공정 장비의 고장을 찾아내는 이상 탐지 연구도 활발하게 진행되고 있다(Hwang J. H. et al., 2021). 제조업뿐 아니라, 사이버 보안, 금융 분야, 행동 패턴 분야, SNS 등 다양한 분야에서도 응용되고 있으며 그 가치가 커지고 있다. 또한, 센서 데이터를 통해 이상 진단을 통한 RUL 예측 연구(Kim H. S. et al., 2020), 정상 펌프와 고장 펌프를 구분하는 알고리즘 연구(Lee G. H. et al., 2020), 고장 모드 분류 진단(Park H. J., 2020) 그리고 센서 데이터에 기반한 고장진단 연구(Park H. J. et al., 2021) 등이 연구되고 있다.

기존에 연구된 이상 탐지 방법에는 특정 기준에 맞는 데이터를 다루는 방법에만 의존하고 있어 정확한 이상 구간에 대한 탐지에 어려움이 있다. 이로 인해, 통계적, 경험적 기반의 이상 탐지 기법의 한계점을 보완한 머신러닝, 딥러닝 기반의 이상 탐지 방법들이 연구되고 있다. 머신러닝, 딥러닝 기반의 이상 탐지 방법은 지도학습, 준지도학습, 비지도 학습을 사용하는 방법으로 나누어진다. 지도학습의 경우 학습데이터를 정상 데이터와 이상 데이터를 Labeling 할 수 있을 때 사용한다. 준지도학습은 학습데이터 중 정상 개체에만 Label 정보가 있고 Label 정보가 없는 자료에 대해 정상, 비정상 여부를 알 수 없는 경우에 사용하며 Label을 지정하고 결정된 데이터에 추가하여 기본 학습기를 재학습해 성능 향상시키는 것을 목표로 한다(Kim H. J., 2022). 비지도 학습은 Labeling이 되어있지 않거나 정상, 이상에 대한 Labeling이 어려울 것으로 예상할 때 적용하며 데이터에 이상치가 매우 적을 때 유용한 기법이다.

센서를 활용한 이상 탐지를 하기 위해서 시계열 데이터에 대한 딥러닝 기반 이상 탐지 방법이 필요하다. 공정의 경우 시간에 따라 이상 구간이 어느 시점에서 발생하는지 확인해야 하며, SNS, 사이버 보안 등 특정한 분야에서는 시간에 따른 이상 구간에 대한 정보가 필요하다. 따라서 비지도 학습을 기반으로 한 기존의 알고리즘에 시계열을 더한 시계열 이상 탐지 방법들이 개발되고 있다. 비지도 학습방법에는 AutoEncoder, GAN과 같은 알고리즘이 존재하는데, 시계열을 기반으로 한 LSTM-AutoEncoder(Srivastava, N. et al., 2015), TadGAN 등이 개발되었다. 각각의 알고리즘은 그 기반으로 하는 알고리즘이 다르므로 서로 다른 결과가 도출될 수 있다.

지도학습은 비지도학습에 비해 학습된 모델의 정확도가 높다는 장점이 있지만, 제조, 설비, 마케팅과 같은 데이터의 경우 비정상 Sample을 구하기 어렵다. 또한, Label이 없거나 Labeling 오류의 경우 학습이 어렵고 새로운 이상 패턴이 발생하게 되면 새로 학습을 진행해야 한다는 단점이 존재한다. 이러한 한계점으로 인해 Labeling이 없어도 이상 탐지가 가능한 비지도 학습이 활용되고 있지만, 일반적으로 지도학습보다 정확도가 낮으며 비지도 학습으로 생성된 이상 구간에 대한 정확성을 입증하기 어렵다.

따라서, 본 연구에서는 시계열 이상 탐지 비지도 학습의 한계점을 보완하기 위해, 이상 구간에 대한 정확성을 입증하고 성능도 높일 수 있는 전처리 프로세스를 제안하고자 한다.

2. 이론적 배경 및 선행연구

관련 문헌 연구는 실제 이상 탐지 활용사례에 관한 연구와 시계열 데이터 이상 탐지 방법론 연구로 구분하여 진행하였다.

최근 AI 기술의 발전으로 사람이 고려하지 못하는 이상 부분에 대한 탐지 기술이 발전함에 따라 실제 이상 탐지 활용사례에 관한 연구가 활발하게 진행되고 있다. Cook A. A. et al. 는 IoT 이상 탐지를 위한 응용 사례와 함께 IoT 데이터에 이상 탐지 기술을 적용할 때 직면할 수 있는 과제에 대한 배경을 제공하였다. Carletti M. et al.은 딥러닝 기반 이상 탐지 방식을 사용하여 이상 점수를 각 압력 프로파일에 연결하여 생산단계에서 더 자세한 검사, 수행한 작업을 최적화하였다. Preuveneers D. et al.은 실제 사이버 침입에 대해 탐지하는 기술이 향상되고 있으며 연구에서 제안하는 블록체인 기반 연합 학습 솔루션은 일반화되어 더 정교한 신경망 구조를 이루고 다양한 사례에 적용하였다. Jiang W. Q. et al. 는 불균형 시계열에서 Class 불균형 문제에 대한 높은 분류 정확도를 얻기 위해 GAN(Generative Adversarial network)을 기반으로 한 새로운 이상 탐지 접근법을 제안하였다. Hwang J. H. et al.은 생산 설비의 이상 탐지를 위해 제조데이터를 이상 탐지 딥러닝 모델인 AutoEncoder, Variational AutoEncoder, Adversarial AutoEncoder에 적용하였다. Ramotsoela. D. et al.은 기계학습 기반 이상 탐지 방법을 사용하여 ISWN이 배포되는 환경의 비정상적인 변화를 감지하여 보안 문제를 향상시켰다. Meyer P. et al.은 시간 민감 네트워킹(TSN)의 스트림 별 필터링 및 폴리싱이 자동차에서 잘못된 동작과 트래픽 흐름을 식별하기 위한 핵심기술로 네트워크 이상 탐지기로의 역할을 하였다. Song B. et al.은 텍스트 정보를 중심으로 사고에 대한 보고서 문서를 사용해 비정상적인 상태를 탐지하기 위한 텍스트 마이닝 기반 로컬 이상치 인자(LOF) 알고리즘을 제안하였다. Rezapour M. et al. 는 비지도 학습 중 AutoEncoder, SVM, 마할라노비스 특이치 탐지를 활용하여 사기꾼들의 고객에 대한 행동과 패턴에 따라 접근하는 방식을 고려하여 이상 탐지를 하였다.

빅데이터 시대에 실시간으로 데이터가 수집되면서 시계열 데이터에 대한 분석이 활발하게 진행되고 있으며 그중에서도 시계열을 활용한 이상 탐지 알고리즘 연구가 중요해지고 있다. Oh M. J. et al.은 유압 시스템에 부착된 다중 센서 데이터를 기반으로 장비의 고장 예측과 이상 발생 시점 예측을 결합해 제조 설비 이상 탐지를 위한 지도학습 및 비지도 학습을 설계하였다. Jiao, Y. et al.은 새로운 자기 지도학습의 대조 손실을 기반으로, 다변량 시계열 데이터(TimeAutoAD)를 위한 자율 이상 감지 기술을 제안하였다. Nguyen H. D. et al.은 다변량 시계열 데이터를 예측하기 위한 LSTM-AutoEncoder 방법과 시계열 데이터의 이상을 탐지하기 위한 하이브리드 알고리즘의 하이퍼 파라미터 최적화 방법을 제안하였다. Geiger A. et al.은 시계열 데이터에서 이상을 감지하는 것은 라벨 부족 및 복잡한 시간 상관관계로 어렵기 때문에 GAN을 기반으로 한 비지도 변칙 탐지 방식은 TadGAN을 제안하였다. Chang K. B. et al.은 이상 감지에 대한 비지도 ML 접근방식인 TadGAN으로 정상/이상 조건의 분리 가능성에 대해 제안하였다. 또한, Oh S. et al.은 한 지점이 아닌 여러 시간 단위에 걸쳐 기록된 이상치를 TadGAN 알고리즘을 적용하여 이상징후를 탐지하였다. 그럴 뿐만 아니라 GAN 기반의 이상 탐지 알고리즘의 경우 생성기에 최신 유행하는 기법인 Transformer를 추가한 연구들도 진행되고 있다. Xu J. et al.은 서비스 모니터링, 우주 및 지구탐사, 수처리 등의 데이터 적용에 대해 우수한 결과를 달성한 비지도 학습 시계열 이상 탐지 방법인 Anomaly Transformer를 제안하였다. Li Y. et al.은 모드 붕괴 문제를 완화하기 위해 여러 개의 생성기와 단일 판별기를 사용하는 Dilated Convolutional Transformer based GAN(DCT-GAN)을 제안하였다. Tipirneni S. et al. 는 세밀한 행렬 표현 대신 시계열을 3개의 세트로 처리하여 손실을 줄이는 Self-Supervised Transformer for time-series(STraTS)를 제안하였다.

3. 방법론 소개

3.1 Feature Selection 방법

광업 센서 데이터에 대한 전처리 과정을 거쳐 3가지 Feature Selection 방법을 통해 중요인자를 선정하고 TadGAN으로 이상 구간을 탐지하는 과정으로 진행된다. 전처리 과정으로는 기존 데이터의 180개씩 반복된 데이터의 평균값, 중앙값 중 더 나은 결과를 보인 평균값을 이용하여 시간별 1개의 데이터가 나올 수 있게 하였다. 또한, 데이터 재구성 방식을 채택하여 전처리 과정에 활용하였다(Lee, S. H. et al., 2020). 전처리한 데이터에 Feature Selection을 하기 위해서 SHAP(SHapley Additive exPlanations)를 사용하여 어떻게 중요변수를 선정했는지에 대한 Force Plot 2가지와 Summary Plot을 이용하여 중요변수를 선정하였다.

Figure 1. Force Plot(1)

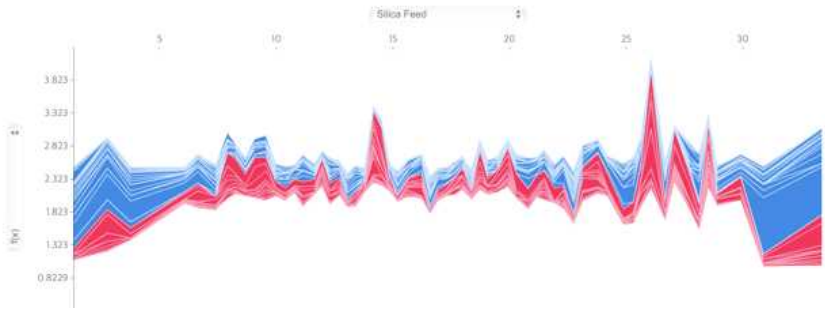

Figure 2. Force Plot(2)

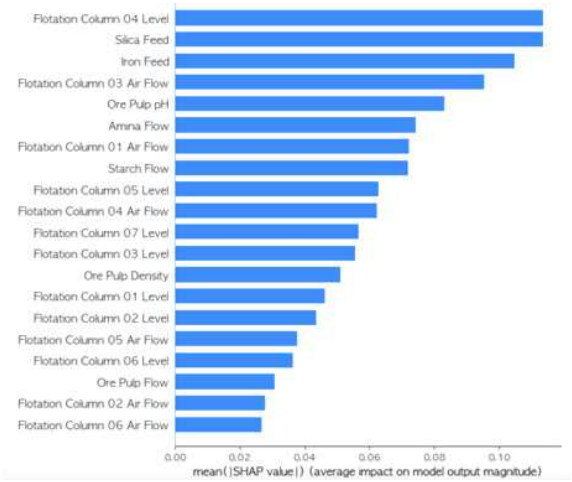

Figure 3. Summary Plot

또한, RFE(Recursive Feature Elimination)는 SHAP와 달리 시각적으로 나타나지 않고 Feature Selection에 가장 단순한 방법으로서, 모든 Feature를 하나하나 제거하면서 원하는 개수의 Feature가 남을 때까지 반복한다. Feature importance가 가장 낮은 Feature부터 하나씩 제거하면서 가장 최적의 중요변수를 추출하였다. 추가로 sklearn 패키지의 Feature Selection에 주로 사용하는 Chi-square Test도 SHAP처럼 시각적으로 나타나지 않았지만, Chi-square를 활용한 계산을 통해 중요변수를 추출하였다. 앞서 TadGAN 설명에 언급하였던 rolling window기법을 적용하여 예측된 신호를 재구성하고 unroll하기 위해 중앙값으로 선택하여 재구성 그래프를 구성하였다. 기존의 값과 재구성한 데이터를 시계열에 따라 나타낸 후, 신호 재구성이 잘되었는지를 파악하여 두 신호 사이의 불일치 오류의 계산에 사용한다. 이를 활용해 재구성한 값과 기존 값의 오류 절댓값에 대한 그래프를 나타내고 오류의 정도에 대한 그래프를 추가하여 어느 정도를 이상 구간으로 선정할지에 대한 Threshold를 설정한다. TadGAN을 적용하여 선별된 이상 구간에 대한 정보를 기존 데이터에서 제외한 후, 정상구간으로 판별된 데이터만 연결하여 새로운 데이터를 선정한다.

3.2 머신러닝 기법

3.2.1 다중회귀분석(Multi Linear Regression)

MLR은 다중선형회귀로 여러 개의 독립변수를 통해 종속변수를 예측하기 위한 회귀모형이다. 즉, 여러 독립변수 중 하나의 독립변수만 변화한다고 가정했을 때 종속변수가 얼마나 변화하는지를 측정한다는 특징을 가진다. 다중 회귀분석의 일반식은 아래와 같다.

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots b_n x_n \qquad \text{Equation(1)}$$

위 식의 b_0는 절편, $b_1, b_2, \cdots, b_n$ 은 편 회귀 계수(Partial Regression Coefficient)를 나타내며 계수를 통해서 설명변수 x_n의 종속변수에 대한 영향력을 보여준다. 단순회귀분석과 마찬가지로 최소제곱법을 편미분하여 계수를 추정할 수 있다. 독립변수들에 따라 설명변수에 얼마나 영향을 미치는지 여부를 파악할 수 있으며, 종속변수에 대한 영향력과 차이가 있을 수 있어 다중공선성을 제거하는 것이 중요하다.

3.2.2 Random Forest

Random Forest는 의사 결정 나무의 앙상블 기반 모형으로 Breiman에 의해 제안되어 배깅의 원리와 임의적 특성을 더한 형태이다. 빠른 학습속도, 많은 양의 데이터 처리 능력 그리고 이상치에 대한 영향이 크지 않다는 장점이 있다(Breiman, L. 2001). Decision Tree 분류기를 여러 개 훈련하며 앙상블 방식을 사용해 과대 적합 되거나 과소적합이 될 경우, 해당하는 오차가 상대적으로 무시되는 경향을 보이며 일반적으로 Decision Tree를 하나 사용할 때보다 더 좋은 모델 성능을 보여준다. Random Forest는 분류 및 회귀 문제에 모두 사용이 가능하며 결측치를 다루기 쉽다는 장점이 있다. 또한, 대용량 데이터 처리에 우수하며 분류 모델에서 상대적으로 중요한 변수를 선정하고 순위를 정할 수 있고 Overfitting문제를 회피하여 모델 정확도를 향상시킬 수 있다.

3.2.3 XGBoost

XGBoost는 CART(Classification and Regression Tree)를 기반으로 하여 Random Forest와 같이 Decision Tree를 조합해서 사용하는 앙상블 알고리즘이다. CART는 여러 가지 Decision Tree를 통한 방법론으로 CART의 원리는 Additive Learning로 정의되며 아래와 같은 수식으로 표현할 수 있다.

$$Y' = a^{*}treeA + b^{*}treeB + c^{*}treeC + \cdots \quad \text{Equation(2)}$$

Y'는 Y에 대한 예측값인 $a, b, c, \cdots$는 각 트리 $A, B, C, \cdots$에서 나온 가중치를 의미한다. 위의 개념을 XGBoost와 Gradient Boosting Tree에 적용시키면 아래의 식들로 표현할 수 있다(Chen, T. et al., 2016).

$$y_i' = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad \text{Equation(3)}$$

$$obj = \sum_{i=1}^{n} l(y_i, y_i') + \sum_{k=1}^{K} \Omega(f_k) \quad \text{Equation(4)}$$

첫 번째 식의 y_i'는 XGBoost에 적용한 예측값을 의미하며, f_k는 k번째 Decision Tree를 의미한다. 두 번째 식의 l은 손실함수를 의미하며, Ω는 정규화 기간을 의미한다. 따라서 여러 개의 Decision Tree 모델들을 학습시켜 예측값을 더한 것으로 선택한다. 또한, 더해진 예측 점수들로 결론을 내려 과적합이 일어나는 경우와 기존 모델이 잘 설명하지 못하는 부분에 대해서 보완할 수 있다(Choi, S. H. et al., 2020).

3.2.4 Lasso Regression

기존 선형회귀에서 MSE 값이 최소가 되는 가중치와 편향을 찾아내며, 가중치들의 절댓값의 합을 최소로 만드는 회귀모델이다. Lasso Regression은 L1-norm 페널티를 가진 선형회귀 방법으로 가중치의 모든 원소가 0에 가깝거나 0이 되어야 한다. 식은 아래와 같이 나타낼 수 있다.

$$\begin{aligned} & MSE + Penalty \\ &= MSE + \alpha \cdot L_1 - norm \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{m} |w_j| \end{aligned} \quad \text{Equation(5)}$$

α는 페널티의 효과를 조절해주는 하이퍼 파라미터이며, m은 가중치의 개수 즉, 특징의 개수를 나타낸다. α의 값이 커지게 되면 페널티 항의 영향력이 커지기 때문에 전체적으로 작아져 0에 가까워지면 선형회귀와 같아진다. Lasso Regression의 목적은 MSE와 페널티 항의 합이 최소가 될 수 있는 w와 b를 찾는 것이다. Lasso Regression의 장점은 두 가지로 볼 수 있다. 첫 번째는 제약 조건으로 인해 일반화된 모형을 찾을 수 있다는 것이다. 두 번째는 가중치들이 0이 되게 하여 그에 해당하는 특성들을 제외시킨다. 이로 인해, 모델에서 가장 중요한 특성이 어떤 것인지 알게되는 효과를 주며 모델의 해석력이 높아진다.

3.3 딥러닝 기법

3.3.1 LSTM(Long Short-Term Memory)

LSTM은 Hochreiter and schmidhuber(1997)에 의해 제안되어 순환신경망(RNN, Recurrent Neural Network) 기반 알고리즘으로, 순환신경망의 단점으로 알려진 긴 의존 기간에 대한 학습을 수행할 수 있으며 순차적인 시계열 패턴의 학습 그리고 예측에 많이 이용된다. 앞서 순환신경망의 단점으로 시간이 증가하면서 나타나는 경사(Gradient)가 '0'에 수렴하는 문제로 없어지거나 무한대로 수렴하는 경우가 발생하는데, 이를 메모리블록과 게이트 프로세스를 사용하여 문제를 해결하고자 한 것이다. LSTM은 이전의 상태들은 더 잘 기억할 수 있게 개선한 신경망을 셀 스테이트(Cell State)에 정제된 구조를 가진 게이트라는 요소를 활용하여 정보를 더하거나 제거하는 기능을 가진다. LSTM은 구성하는 대표적인 유닛으로 입력, 망각, 출력 3가지 게이트가 존재한다. 각 게이트는 시그모이드 함수를 통해 입력할 값, 망각할 값, 출력할 값을 정한다. 위 방식을 통해 시계열 자료 학습 시 경사 값이 사라지는 장기 의존성 문제를 해결한다.

3.3.2 GRU(Gated Recurrent Unit)

GRU는 Cho, K. et al.(2014)에 의해 처음 제안되었고 RNN에서 Encoder와 Decoder를 최초로 제시하며 LSTM을 보다 효율적으로 바꿨다. LSTM은 기존 Vanilla RNN의 한계점으로 대두된 Long-Term Dependency에 강인하게 설계되어 셀 스테이트와 입력, 망각, 출력이라는 3가지 게이트를 적용해 장기간 정보 손실문제를 해결하였다. GRU는 LSTM의 구조에 영감을 받아 만들어졌으며, 기존 게이트의 중복성을 제거하고 보다 효율적인 처리방식을 위해 간단한 구조를 제안하였다. LSTM의 핵심인 셀 스테이트 개념을 없애고 다시 Hidden state 단일 방식을 사용하면서, Long-Term Dependency 문제를 효과적으로 해결할 수 있는 방식을 제안했다. 또한, 기존 LSTM의 망각 게이트와 같은 역할을 하는 Reset 게이트와 입력 게이트와 망각 게이트 개념을 합친 Update 게이트로 이루어져 있다.

3.4 이상 탐지 알고리즘

3.4.1 GAN(Generative Adversarial Network)

GAN은 Goodfellow, I. et al. (2014)에 의해 제안되어 적대적 생성 신경망으로도 불리며, 실제에 가까운 이미지 또는 사람이 쓴 글 등 여러 개의 가짜 데이터를 생성하는 모델이다. GAN은 서로 다른 2개의 네트워크인 생성자(Generator)와 판별자(Discriminator)를 적대적으로 학습시켜 실제 데이터와 비슷한 데이터를 생성하는 모델이며, 생성된 데이터에 대한 Label 값이 없어 비지도 학습 기반 생성모델로 분류된다. GAN의 서로 다른 네트워크 중 생성자(G)는 생성된 노이즈(z)를 받아 실제 데이터와 비슷한 데이터를 만들어내도록 학습한다. 그러나 판별자(D)는 실제 데이터와 생성자가 생성한 가짜 데이터를 구별하도록 학습한다.

3.4.2 AutoEncoder

AutoEncoder는 단순히 입력을 출력으로 복사하는 신경망이지만 네트워크에 여러 가지 방법의 제약을 주어 어려운 신경망을 만든다. 아래 그림처럼 은닉층의 뉴런 수를 입력층보다 작게 하여 데이터 차원 축소를 하거나, 입력데이터에 노이즈를 추가하여 원본 입력을 복원하는 네트워크를 학습시키는 등 다양한 AutoEncoder 알고리즘이 있다.

이러한 AutoEncoder에 주어진 제약들은 단순히 입력을 출력으로 복사하지 못하게 방지하고, 데이터를 효율적으로 표현하는 방법을 학습할 수 있도록 제어하는 역할을 한다.

3.4.3 TadGAN(Time-series Anomaly Detecion GAN)

TadGAN은 AutoEncoder와 GAN의 단점을 보완한 시계열 데이터를 분석하여 이상탐지하는 새로운 알고리즘으로 Geiger, A. et al. (2020)에 의해 제안되었다. AutoEncoder의 경우 L2목적함수를 사용할 때, 재구성한 데이터가 너무 정확하게 나와 비정상 데이터까지 Fitting한다는 문제가 있다. GAN의 경우, 생성자가 데이터의 숨은 분포를 완전히 포착하게끔 학습시키는 것에 비효율적이고 False Alarm을 일으키며 생성자와 판별자 간 학습 불균형으로 판별자의 성능이 뛰어나 생성자가 어떤 이미지 생성하더라도 구분할 수 있게 하는 모드 붕괴 문제를 일으킨다. TadGAN은 이 두 가지 단점을 보완하여 성능을 높였다.

4. 연구 프로세스

4.1 데이터 소개

본 연구에 사용된 데이터는 부유선광의 실리카 농축액의 비율을 실시간으로 추정하기 위해 만든 데이터를 시계열 분석으로 전처리한 데이터를 사용하여 분석하였다(Lee, S. H. et al., 2020). 데이터는 2017년 3월 10일 1시부터 2017년 3월 16일 5시, 2017년 3월 29일 12시부터 2017년 9월 9일 23시까지로 구성되어있으며 총 737,453개로 이루어져 있다. 본 논문에서 3월 16일 5시 이전의 데이터가 2017년 9월 9일 23시 이후의 데이터에 영향을 충분히 줄 수 있다고 판단하여 중단된 일자는 없으며 바로 연결된다고 가정하였다. 데이터는 날짜변수(date)와 부유선광에 투입하는 Input 변수(% Silica Feed, % Iron Feed) 2가지와 Output 변수 1가지 그리고 공정 흐름이 발생하면서 나타나는 프로세스 변수 22가지로 이루어져 있다. 해당 데이터는 1시간마다 입력변수, 출력변수 데이터는 180개의 값이 같지만, 프로세스 데이터는 서로 다른 180개의 데이터로 이루어져 있어 이를 평균값, 중앙값, 최빈값으로 전처리한 후, 가장 성능이 좋았던 평균값을 선정하였고 1시간 간격으로 4094개의 데이터로 재구성하였으며 데이터 구성은 아래 Figure 1. 와 같다.

	Input		process					output
date	X1	X2	X3	···	X21	X22	X23	X24	Yt
2017-03-10 1:00	55.2	16.98	3316.754		446.37	1.81	2.83	2.83	1.5
2017-03-10 2:00	55.2	16.98	3687.332		498.075	2.83	2.83	1.5	1.94
2017-03-10 3:00	55.2	16.98	4497.958		458.567	2.83	1.5	1.94	1.09
2017-03-10 4:00	55.2	16.98	4707.07		427.669	1.5	1.94	1.09	1.51
				⋮					

Figure 4. Data Configuration

4.2 TadGAN기반 전처리 프로세스 제안

본 연구에서 제안하는 연구 프로세스는 아래 Figure 2. 와 같다. 광업 센서 데이터를 일반적인 전처리 과정으로 데이터를 사용하기 쉽게 만든 후, Feature Selection 방법인 SHAP, RFE, Chi-square Test를 통해 각각에 대한 중요 인자를 선정하였고 TadGAN으로 이상 구간을 탐지하는 과정으로 진행된다. 앞선 단계에서 TadGAN을 통해 이상 구간을 선별하고 그 구간을 제외한다면 정상구간에 대한 시계열 특징을 볼 수 있다. 따라서 남아있는 정상구간 데이터를 연결하여 하나의 데이터로 만든다. 재구성한 데이터를 머신러닝 방법 중 다중회귀분석(MLR), Random Forest, XGBoost, Lasso Regression 알고리즘에 적용하여 기존 같은 데이터로 연구된 다른 논문과의 성능 비교를 시행한다. 또한, 딥러닝 방법 중 시계열 데이터에 우수하다고 평가받는 LSTM, GRU모델에 적용한 후, Lee, S. H. et al. (2021) 연구에 언급된 베이지안 기반 하이퍼 파라미터 최적화를 하여 머신러닝과 같이 더 나은 성능을 가지는지에 대해 평가한다.

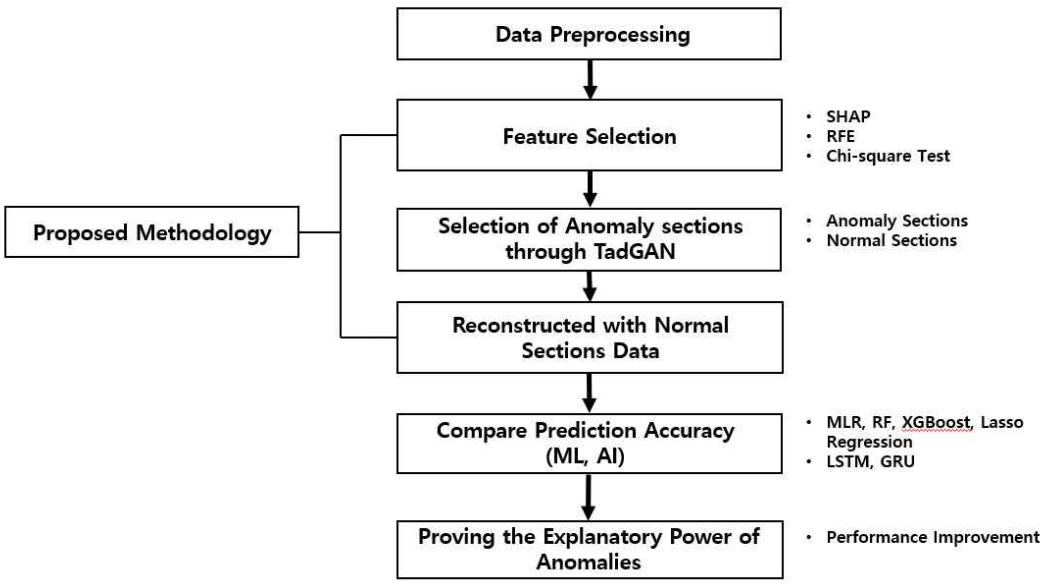

Figure 5. Research Process

4.3 머신러닝, 딥러닝 적용 및 성능 비교

TadGAN을 통해 이상 구간을 선별하고 그 구간을 제외한 다른 정상구간 데이터를 연결하여 만든 데이터를 사용한다. 기존 데이터를 머신러닝 방법 중 MLR(다중회귀분석), Random Forest, XGBoost 알고리즘을 적용한다. 기존 전처리 전 데이터를 활용한 다른 방법론들과 비교하여 성능이 우수한지 혹은 부족한지에 대해 평가해본다. 또한, 머신러닝 알고리즘만이 아닌 딥러닝 방법 중 시계열 데이터에 성능이 우수하다고 평가받는 LSTM, GRU 알고리즘을 적용하고 하이퍼 파라미터 최적화를 활용해 더 나은 성능을 가지는지에 대해 평가한다.

5. 실험결과 및 분석

5.1 Feature Selection 결과

먼저 데이터를 전처리한 후, 해당 센서 데이터를 이용하여 Feature Selection을 하기 위해 SHAP, RFE, Chi-Square Test 방법을 적용하였다. Feature Selection에는 Python 3.8.12 버전, Tensorflow 2.7.0 버전을 사용하였다. SHAP를 통한 Feature Selection에는 중요변수 선정을 SHAP 중요도가 0.05이상 되는 값들을 선별하였다. RFE의 경우 scikit-learn 패키지를 이용하여 클래스 선언, fit, transform 과정을 거치며 Feature importance 속성을 지원하는 의사 결정 나무 기반의 모델을 사용한다. 이를 활용하여 Feature importance가 가장 낮은 Feature부터 하나씩 제거하여 선별하였다. Chi-square Test를 통한 중요변수 선정에는 Chi-square 검정 즉, Chi-square 분포에 기초한 통계적 방법을 적용해 관찰된 빈도가 기대 빈도와 통계적으로 다른지를 판단한다. Chi-square 분포는 독립성 검정, 동질성 검정으로 나뉘게 되는데 변수 선택에서는 독립성 검정을 사용한다.

5.2 성능 비교 결과

평가 지표는 RMSE를 사용하였으며 제안한 프로세스를 적용한 결과는 아래 Table 1. 과 같으며 KwameO, E.'s. 의 경우 다중회귀분석과 랜덤포레스트의 결과만 있어 나머지 알고리즘 결과는 명시되어있지 않다. 제안한 방법의 머신러닝 결과는 SHAP와 TadGAN을 적용했을 때 성능이 가장 우수하였고 딥러닝을 적용했을 때의 결과는 Chi-square Test와 TadGAN을 적용했을 때 성능이 우수함을 보였다. 각각의 성능을 같은 데이터를 활용해 적용한 논문들과 비교해보았을 때, 머신러닝의 경우, 대부분 알고리즘에서 15% 이상 성능이 향상되었음을 알 수 있었고 딥러닝의 경우, 17~19% 성능이 향상되었음을 보여준다.

Table 1. Result of Modeling

Method / Algorithm	Proposed Method(RMSE)			Conventional Study(RMSE)	
	SHAP+ TadGAN	RFE+ TadGAN	Chi-square Test+ TadGAN	KwameO, E.'s result	Raw data Result
Multi Linear Regression	0.7057*	2.4683	2.4406	1.0315	0.8382
Random Forest	0.7082*	0.7237	0.7313	0.9081	0.9363
XGBoost	0.7083*	0.7842	0.7958	-	1.0070
Lasso Regression	0.6998*	0.7815	0.7879	-	2.5791
LSTM	0.6924	0.6744	0.6640*	-	0.7992
GRU	0.6947	0.6723	0.6646*	-	0.8180

6. 결 론

비지도 학습을 기반으로 하는 이상 탐지 방법에 대해 정확성을 입증하는 것에는 어려움이 많다. 이를 위해 비지도 학습 기반 이상 탐지 방법의 평가 방식에 관한 연구가 활발하게 진행되고 있다. 그러나 대부분 연구의 경우 지도학습을 기반으로 한 연구에 접목하여 그 평가 방법을 제시하거나, 이상 구간에 대한 데이터가 충분하지 않은 경우 평가를 하지 못하는 어려움이 발생할 수 있다.

본 연구에서는, 이상 구간에 대한 정보가 없는 데이터를 대상으로 시계열 이상 탐지 기법인 TadGAN을 적용하여 이상 구간을 선정한 후, 이상 구간에 대한 전처리 프로세스를 구축하여 예측 모델링 정확도를 높이며 이상 구간에 대한 설명력을 높이는 것을 목적으로 하였다. 제안한 프로세스는 총 2단계의 과정으로 구성되어 있으며, 첫 번째 단계에서는 광업 센서 데이터에 대한 전처리 과정을 거쳐 3가지 Feature Selection 방법인 SHAP, RFE, Chi-square Test를 통해 중요인자를 선정하고 TadGAN으로 이상 구간을 탐지하는 과정으로 진행된다. 두 번째 단계에서는, 앞서 첫 번째 단계에서 TadGAN을 통해 이상 구간을 선별하고 그 구간을 제외한 다른 정상구간 데이터를 연결하여 만든 데이터를 사용한다. 기존 데이터를 머신러닝 방법 중 MLR, Random Forest, XGBoost 알고리즘에 적용하며 성능이 우수한지 평가하고 추가로, 시계열 딥러닝 방법에서 우수하다고 평가받는 LSTM, GRU 알고리즘을 적용해 더 나은 성능을 가지는지 평가한다.

제안한 프로세스를 광업 센서 데이터에 적용하여 3가지 Feature Selection 방법으로 추출된 중요 인자들을 추출하였고, 추출한 데이터를 TadGAN에 적용하여 방법별 이상 구간을 선정하였다. 각 이상 구간을 제외하고 정상구간 데이터를 선별하여 재조합한 데이터를 이용하여 머신러닝, 딥러닝 알고리즘을 적용해 본 결과, 머신러닝 방법론 적용에서는 SHAP와 TadGAN을 적용했을 때의 성능이 가장 우수하였고 딥러닝 적용에서는 Chi-square Test와 TadGAN을 적용했을 때 성능이 우수하였다. 각각의 성능을 같은 데이터를 활용해 일반적인 방법론으로 적용한 논문들과 비교해보았을 때, 머신러닝 방법인 MLR의 경우 15%, Random Forest는 24%, XGBoost는 30% 그리고 Lasso Regression은 73%로 크게 성능이 향상하였으며 LSTM 17%, GRU 19%로 딥러닝에서도 성능이 향상했음을 알 수 있었다. 이와 같은 결과를 통해 이상 구간을 제외한 정상구간에 대한 성능이 향상했기 때문에, 제외했던 이상 구간에 대한 설명력을 입증할 수 있다.

본 연구에서 제안한 프로세스를 기반으로 제조업뿐 아닌, 사이버 보안, 금융 분야, 행동 패턴 분야, SNS 등 다양한 분야에서의 실제 Labeling이 되지 않은 데이터의 비지도 학습 이상 구간탐지를 시행하였을 시, 이상 탐지 구간에 대한 정확성 및 설명력을 입증하고 모델에 대한 성능 향상 효과를 줄 수 있을 것으로 사료된다.

REFERENCES

Breiman, L. 2001. Random forests. Machine learning, 45(1):5-32.

Carletti, M., Masiero, C., Beghi, A., & Susto, G. A. 2019. A deep learning approach for anomaly detection with industrial time series data: a refrigerators manufacturing case study. Procedia Manufacturing 38:233-240.

Chang, K. B. G. N. A. Learning Anomaly Detection for Generating Predictive Maintenance Models from LBS-AUV Mission Data.

Chen, T. & Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining pp. 785-794.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv: 1406.1078.

Choi, S. H. & Hur, J. 2020. Optimized-XG boost learner based bagging model for photovoltaic power forecasting. Transactions of the Korean Institute of Electrical Engineers 69(7):978-984.

Cook, A. A., Mısırlı, G., & Fan, Z. 2019. Anomaly detection for IoT time-series data: A survey. IEEE Internet of Things Journal 7(7):6481-6494.

Geiger, A., Liu, D., Alnegheimish, S., Cuesta-Infante, A., & Veeramachaneni, K. 2020. TadGAN: Time series anomaly detection using generative adversarial networks. In 2020 IEEE International Conference on Big Data (Big Data) pp. 33-43. IEEE.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. 2014. Generative adversarial nets. Advances in neural information processing systems, 27.

Hochreiter, S. & Schmidhuber, J. 1997. Long short-term memory. Neural computation 9(8):1735-1780.

Hwang, J. H. & Jin, K. H. 2021. Anomaly Detection and Performance Analysis using Deep Learning. In Proceedings of the Korean Institute of Information and Commucation Sciences Conference pp. 78-81. The Korea Institute of Information and Commucation Engineering.

Jiang, W., Hong, Y., Zhou, B., He, X., & Cheng, C. 2019. A GAN-based anomaly detection approach for imbalanced industrial time series. IEEE Access 7:143608-143619.

Jiao, Y., Yang, K., Song, D., & Tao, D. 2022. TimeAutoAD: Autonomous Anomaly Detection With Self-Supervised Contrastive Loss for Multivariate Time Series. IEEE Transactions on Network Science and Engineering 9(3):1604-1619.

Kim, H. J. 2022. Semi-Supervised Learning to Predict Default Risk for P2P Lending. Journal of Digital Convergence 20(4):185-192.

Kim, H. S. & Choi, J. H. 2020. Distribution and Validation of RUL Prediction Parameters Considering Life Distribution. Journal of Applied Reliability 20(2):145-153.

Lee, G. H., Shin, B. C., & Hur, J. W. 2020. Fault Classification of Gear Pumps Using SVM. Journal of Applied Reliability, 20(2):187-196.

Lee, S. H. & Kim, Y. S. 2021. A Study on the Optimization of Long Short-Term Memory Hyperparameters Using the Taguchi Design of Experiments. Journal of Applied Reliability 21(3):238-245.

Lee, S. H., Yoon, Y. A., Jung, J. H., Chang, T. W., & Kim, Y. S. 2020. A Machine Learning Model for Predicting Silica Concentrations through Time Series Analysis of Mining Data. Journal of the Korean Society for Quality Management 48(3):511-520.

Li, Y., Peng, X., Zhang, J., Li, Z., & Wen, M. 2021. DCT-GAN: Dilated Convolutional Transformer-based GAN for Time Series Anomaly Detection. IEEE Transactions on Knowledge and Data Engineering.

Meyer, P., Häckel, T., Reider, S., Korf, F., & Schmidt, T. C. 2021. Network Anomaly Detection in Cars: A Case for Time-Sensitive Stream Filtering and Policing. arXiv preprint arXiv:2112.11109.

Nguyen, H. D., Tran, K. P., Thomassey, S., & Hamad, M. 2021. Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management. International Journal of Information Management, 57, 102282.

Oh, M. J., Choi, E. S., Roh, K. W., Kim, J. S., & Cho, W. S. 2021. A Study on the design of supervised and unsupervised learning models for fault and anomaly detection in manufacturing facilities. The Journal of Bigdata, 6(1), 23-35.

Oh, S. & Islam, M. R. 2021. Application TadGAN to Detect Collective Anomaly in Power Usage Data. The Journal of Contents Computing 3(1):297-306.

Park, H. J., Cho, S. H., Jang, K. H., Seol, J. W., Kwon, B. G., Kwon, J. Y. & Choi, J. H. 2020. Study on Fault Diagnosis of Planetary Gearbox in Unmanned Aerial Vehicle Using Multi sensor Data. Journal of Applied Reliability 20(4):332-342.

Park, H. J., Sim, J. W., Jang, J. W., Jang, K. H., Seol, J. W., Kwon, J. Y. & Choi, J. H. 2021. Study on Fault Severity Diagnosis of Planetary Gearbox in Unmanned Aerial Vehicle using Artificial Neural Network. Journal of Applied Reliability 21(4):329-340.

Preuveneers, D., Rimmer, V., Tsingenopoulos, I., Spooren, J., Joosen, W., & Ilie-Zudor, E. 2018. Chained anomaly detection models for federated learning: An intrusion detection case study. Applied Sciences 8(12):2663.

Ramotsoela, D., Abu-Mahfouz, A., & Hancke, G. 2018. A survey of anomaly detection in industrial wireless sensor networks with critical water system infrastructure as a case study. Sensors 18(8):2491.

Rezapour, M. 2019. Anomaly detection using unsupervised methods: credit card fraud case study. International Journal of Advanced Computer Science and Applications 10(11).

Song, B. & Suh, Y. 2019. Narrative texts-based anomaly detection using accident report documents: The case of chemical process safety. Journal of Loss Prevention in the Process Industries 57:47-54.

Srivastava, N., Mansimov, E., & Salakhudinov, R. 2015. Unsupervised learning of video representations using lstms. In International conference on machine learning pp. 843-852. PMLR.

TIPIRNENI, S. & REDDY, C. K. 2022. Self-Supervised Transformer for Sparse and Irregularly Sampled Multivariate Clinical Time-Series. ACM Trans. Knowl. Discov. Data, 1(1).

Xu, J., Wu, H., Wang, J., & Long, M. 2021. Anomaly transformer: Time series anomaly detection with association discrepancy. arXiv preprint arXiv:2110.02642.

저자소개

이승훈 경기대학교 산업경영공학과를 졸업하고, 동 대학원 데이터공학연구실에서 석사과정에 재학 중이다. 주요 관심 분야는 신뢰성공학 및 데이터 분석 등이다

김용수 KAIST 산업공학과에서 학사, 석사, 박사를 취득한 후 SK텔레콤에서 근무하였다. 현재 경기대학교 산업경영공학과 정교수로 재직 중이며, 품질 및 신뢰성, 기능안전, 통계 및 데이터마이닝 분야를 연구하고 있다.