Isolation Forest 와 Autoencoder 하이브리드 이상치 탐지 기법을 활용한 산업 전력 데이터 분석 및 최적화

권우현 [1], 박준성 [1], 구남석 [1], 이충호 [2], 허태욱 [2], *이상금 [1]
*국립한밭대학교 [1], 한국전자통신연구원 [2]

{mfireon0520, js0309, kns2931}@gmail.com, {leech, htw398}@etri.re.kr,
sangkeum@hanbat.ac.kr

Industrial Power Data Analysis and Optimization Using a Hybrid Anomaly Detection Technique Combining Isolation Forest and Autoencoder

Woohyeon Kwon[1], Junseong Park[1], Namseok Koo[1], Chungho Lee[2], Taewook Heo[2], and *Sangkeum Lee[1]
*Hanbat National University[1], Electronics and Telecommunications Research Institute[2]

요 약

본 논문은 2022 년 국내 식품 산업의 전력 소비량도 꾸준히 증가함에 따라 데이터의 이상치도 증가하고 있다. K-Means, Isolation Forest, Autoencoder 를 결합한 하이브리드 이상치 탐지 기법을 활용해 식품 산업 전력 데이터를 분석한다. K-Means 와 Isolation Forest 로 정상치와 이상치를 분리한 후, Autoencoder 를 통해 재구성 손실 값을 기반으로 이상치를 탐지한다. 두 모델 모두 우수한 성능을 보였으나, Isolation Forest-Autoencoder 모델이 더 높은 AUC(Area Under the Curve)값과 강건성을 가지며, 비지도 학습 방식으로 이상치 탐지의 초기 단계를 수행하고, Contamination 값을 최적화하여 더 강건하고 뛰어난 이상치 탐지 성능을 나타냈다. 특히, Isolation Forest 는 고차원 데이터에서도 빠르고 효율적으로 이상치를 분리할 수 있다. 이를 통해 에너지 효율성과 비용 절감을 위한 효과적인 이상치 처리 방안을 제안했다. 향후 연구로는 Autoencoder 및 다양한 이상치 탐지 알고리즘의 결합을 모색한다.

Ⅰ. 서 론

2022 년 식품산업통계정보의 국내 식품 시장 규모가 전년 대비 15.73%로, 이에 따라 식품 산업의 전력 소비량도 꾸준히 증가하는 추세이다. 산업체 전력 데이터에서 이상치는 에너지 사용의 오류나 비효율적인 특징을 나타낸다. 따라서, 이상치 처리는 식품 산업의 지속적인 성장과 효율적인 전력 사용을 위해 필수적이다. 그린 버튼 플랫폼을 통해 에너지 절약 및 효율성을 높이고 비용 절감에 기여할 수 있다[1]. 그린 버튼 플랫폼과 연계하여 식품 산업 전력 데이터를 분석하고, K-Means, Isolation Forest, Autoencoder 를 활용한 하이브리드 기법의 이상치 탐지를 통해 전력 데이터 이상치 처리 방안을 제안한다. K-Means 는 데이터를 지정된 개수의 클러스터로 군집화 하는 거리 기반 알고리즘이다. Isolation Forest 는 높은 밀도로 분포된 정상 값과는 달리, 낮은 밀도로 분포된 낮은 이상치는 빠르게 분리되어 이상치를 검출한다[2]. 고차원의 데이터나 복잡한 패턴에서 정확도가 떨어진다. 정상 데이터로 학습된 Autoencoder 는 이상치가 포함된 데이터의 재구성 손실 값을 통해 이상치 탐지가 가능하다[3]. 고차원 데이터의 패턴을 학습하여 복잡한 이상치 탐지에 적합하므로, Isolation Forest 의 약점을 보완할 수 있다. 두 알고리즘은 강건함과 정확성을 바탕으로 데이터의 전반적인 이상치를 탐지하는 데 높은 성능을 제공한다.

본 연구에서는 K-Means 와 Isolation Forest 로 정상치와 이상치 데이터를 분리하고, Autoencoder 모델의 출력에 기반하여 파라미터 비율에 따른 이상치 비율을 비교함으로써 에너지 효율성을 높일 수 있는 방안을 제시한다.

Ⅱ. 본론

2.1 데이터 설명

그림 1 은 15 분 간격으로 수집된 식품 산업 전력 데이터로, 주말보다 주중에 상대적으로 많은 전력 사용량을 보이는 분포를 가진다. 이에 대해 [0, 1]으로 변환하는 정규화 기법을 적용하여, 데이터 간의 차이를 줄인다.

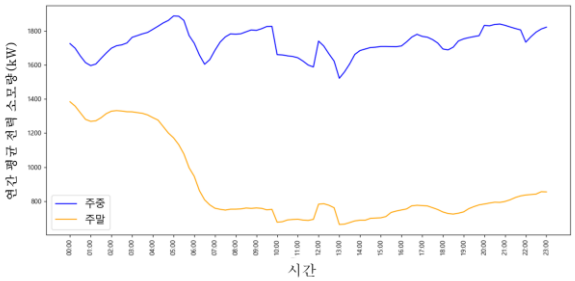
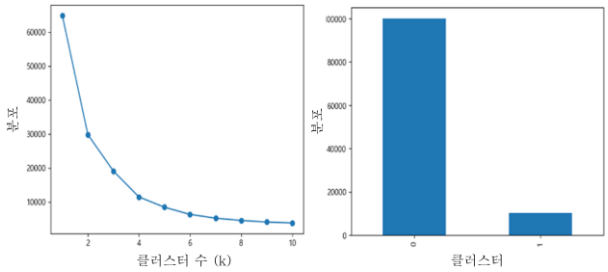


그림 1. 식품 산업의 주중, 주말 연 평균 전력 사용량

2.2 군집화를 이용한 이상치 데이터 셋 구성



(a) (b)

그림 2. Elbow 기법으로 얻은 클러스터 개수와 정상치를 0 으로 이상치는 1 로 구분한 그래프

본 연구에서는 이상치를 탐지하기 위해 군집화를 이용해서 데이터의 정상치와 이상치를 분리하였다. 이상치 데이터 셋 구성을 위하여 Elbow 기법을 사용하여 최적의 클러스터 개수를 2 개로 지정하였고, 정상치와 이상치의 비율을 90.58% 대 9.414%로 군집화하였다. 그림 2 은 K-Means 를 통해 얻은 클러스터 결과를 시각화한 것이다.

Isolation Forest 는 트리 기반의 비지도 학습 알고리즘으로, 설정한 이상치 비율(Contamination) 값에 따라 이상치와 정상치를 분리하였다. 이상치가 상대적으로 낮은 밀도로 분포된 특성을 이용하여 빠르고 효율적으로 탐지한다.

2.3 하이브리드 이상치 탐지 모델 설계

2.3.1 K-means-Autoencoder 모델

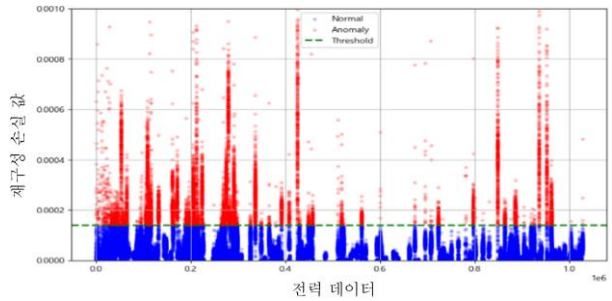


그림 3. 재구성 손실(Reconstruction Loss) 값을 시각화하고 이상치 여부를 구분한 결과

K-means 로 군집화된 정상치 데이터를 Autoencoder 에 학습 데이터로 활용하였다. 학습된 Autoencoder 모델은 정상 데이터를 재구성할 수 있도록 설계되었으며, 이후 전체 데이터를 입력하여 재구성 손실 값을 계산하였다. 재구성 손실 값이 사전 설정된 임계값(Threshold)을 초과하는 데이터는 이상치로 간주하였다. 그림 3 은 K-means-Autoencoder 모델의 재구성 손실 값을 시각화한 결과이며, AUC(Area Under the Curve)값은 0.9628 이다. K-Means 로 정상치를 분리한 후 Autoencoder 로 이상치를 탐지하는 하이브리드 모델은 안정적인 재구성과 높은 정확도를 통해 효과적으로 이상치를 탐지할 수 있다.

2.3.2 Isolation Forest-Autoencoder 모델

설정한 Contamination 비율에 따라 정상치 데이터를 분리 후, Autoencoder 에 학습 데이터로 활용하였다. 그림 4 는 Isolation Forest-Autoencoder 모델의 ROC(Receiver Operating Characteristic) 곡선, 표 1 은 훈련된 Autoencoder 모델에 전체 데이터를 입력하여 재구성 손실 값과 ROC 곡선을 나타낸다. Isolation Forest 로 1 차 이상치를 제거한 후 Autoencoder 를 결합한 모델은 이상치 탐지에서 우수한 성능과 강건성을 제공, 특히 Contamination 값을 조정하여 모델 성능을 최적화할 수 있다.

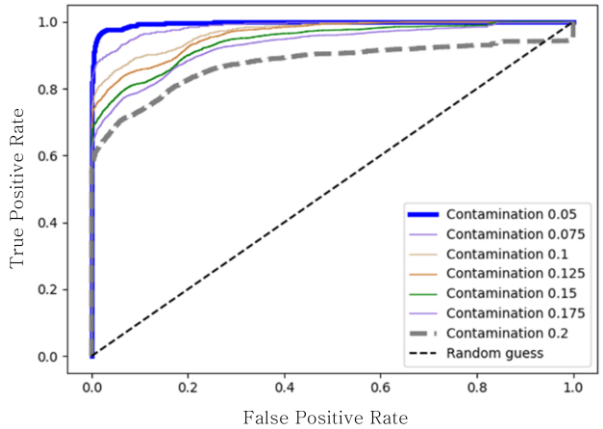


그림 4. Isolation Forest 와 Autoencoder 하이브리드 모델로 이상치를 탐지한 ROC 곡선

Contamination	AUC	재구성 손실 값
0.050	0.9964	0.0055
0.075	0.9873	0.0148
0.100	0.9703	0.0380
0.125	0.9603	0.0286
0.150	0.9419	0.0300
0.175	0.9292	0.0282
0.200	0.8732	0.0212

표 1. Contamination 에 따른 AUC 및 재구성 손실 값

2.4 하이브리드 모델 비교

K-Means 와 Isolation Forest 를 각각 Autoencoder 와 결합하여 하이브리드 이상치 탐지 모델을 구축한 결과, 두 모델 모두 높은 성능을 보였다. 특히, Contamination 값을 최적화한 Isolation Forest-Autoencoder 모델이 더 강건한 탐지 성능과 높은 AUC 값을 기록하며 우수성을 입증했다. 이를 통해 식품 산업 전력 데이터의 이상치 탐지에 가장 적합한 접근 방안을 도출할 수 있었다.

Ⅲ. 결론

본 논문에서는 식품 산업의 전력 데이터를 분석하여, 두 하이브리드 이상치 탐지 기법의 성능을 비교한다. K-Means 와 Autoencoder 하이브리드 모델은 군집화를 기반으로 데이터의 분포를 효과적으로 분석하고, 정상치 데이터로 Autoencoder 를 학습하여 안정적인 성능과 높은 재현성을 보였다. 반면, Isolation Forest 와 Autoencoder 하이브리드 모델은 비지도 학습 방식으로 이상치 탐지의 초기 단계를 수행하고, Contamination 값을 최적화하여 더 강건하고 뛰어난 이상치 탐지 성능을 나타냈다. 특히, Isolation Forest 는 고차원 데이터에서도 빠르고 효율적으로 이상치를 분리할 수 있어 Autoencoder 와의 결합에서 더 높은 AUC 값을 기록하며 우수한 모델 결합으로 평가되었다.
향후 연구에서는 다른 산업 분야의 전력 데이터를 적용하거나, LSTM-Autoencoder 를 활용해 시계열 데이터의 이상치 탐지를 더욱 향상시킬 계획이다[4]. 또한, 다양한 이상치 탐지 알고리즘을 분석하고 다양한 모델을 하이브리드로 결합할 방안을 모색하고자 한다. 이를 통해 그린 버튼 플랫폼과 같은 스마트 에너지 관리 시스템의 효율성과 안정성을 높이는 데 기여할 수 있을 것으로 기대한다. 이러한 연구는 에너지 소비의 최적화를 이끌며 지속 가능한 에너지 사용과 비용 절감을 제공하려 한다.

ACKNOWLEDGMENT

본 연구는 산업통상자원부(MOTIE)와 한국에너지기술평가원 (KETEP)의 지원을 받아 수행한 연구 과제입니다. (No. RS2023-00237018)

참 고 문 헌

[1] 최경호. (2019). 4 차 산업혁명 핵심기술을 활용한 기후변화 대응 및 관련 법제 연구 – 빅데이터를 중심으로 –. 강원법학, 58, 779-815. 10.18215/kwlr.2019.58..779

[2] 정광훈 외. (2024-06-19). "Isolation Forest 을 활용한 축산 온·습도 이상치 탐지에 관한 연구", *한국통신학회 학술대회논문집*, 제주

[3] 박노진. (2023). "시계열 이상치 탐지: 오토인코딩을 활용한 사례 분석", *한국데이터정보과학회지*, 34(4), 649-657. 10.7465/jkdi.2023.34.4.649

[4] 전승현 외. (2023)." LSTM 오토인코더를 이용한 이상 탐지의 임계치 결정 방법", *한국정보기술학회논문지*, 21(4), 21-30. 10.14801/jkiit.2023.21.4.21