2025년도 한국통신학회 추계종합학술발표회

공장 설비 유형별 전력소모량 데이터의 특성 분석 및 최적 전처리 방법론 연구

조윤호, 조재영, 권구영, 김선혁
국립공주대학교

202102082@smail.kongju.ac.kr, whwodud02@smail.kongju.ac.kr, gykwon@kongju.ac.kr, seonh@kongju.ac.kr

A Study on the Characteristics and Optimal Preprocessing Methodology of Power Consumption Data by Factory Equipment Type

Yun Ho Jo, Jae Yeong Jo, Gu-Young Kwon, Seon Hyeog Kim
Kongju National University.

요 약

본 논문은 산업 현장의 시계열 데이터 품질과 유용성을 극대화하기 위한 데이터 유형별 최적 전처리 프레임워크를 제안한다. 이를 위해 실제 공정 데이터를 통계적 지표에 기반하여 연속 가동형과 간헐적 가동형으로 분류하고, 각 유형의 문제점을 해결하기 위한 복수의 방법론을 비교 실험하였다. 실험 결과, 연속 가동형 데이터는 2 단계 처리 방식이, 간헐적 가동형 데이터에서는 신호의 특성을 보존하는 방식이 가장 효과적임을 예측 성능 개선도를 통해 입증하였다. 본 연구는 MES 와 같은 부가 정보가 없는 환경에서도 데이터의 물리적 의미를 해석하였다는 점에서 의의를 가진다.

Ⅰ. 서 론

데이터 기반의 스마트 팩토리 전환이 가속화되면서, 현장에서 수집된 원본 데이터의 품질을 높이는 전처리 과정이 필수적이다. 하지만 공장 내 모든 설비 데이터에 통일된 기법을 일괄 적용하는 것은, 운전 특성이 다른 설비의 정상적인 신호를 노이즈로 오인하여 데이터의 의미를 훼손할 위험이 있다.

따라서 본 연구는 이러한 문제의식에서 출발하여, 통계적 지표를 통해 설비 데이터의 유형을 객관적으로 분류하고, 각 유형의 특성에 맞는 최적의 전처리 방법론을 비교 검증하여 제시하는 것을 목표로 한다.

Ⅱ. 본론

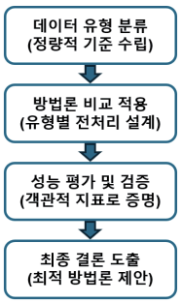
2.1. 실험 과정 설계

본 연구는 <그림 1>과 같이 데이터 유형 분류, 방법론 비교 적용, 성능 평가, 최종 결론 도출의 4 단계로 진행되었다.

분석에 사용된 5 개 설비의 원본 데이터는 <그림 2>와 같이 시각적으로 뚜렷하게 구분되는 두 가지 패턴을 보였다. 이러한 패턴 차이를 객관적으로 분류하기 위해, 설비의 비가동 시간을 나타내는 0 값 비율(%)과 데이터의 상대적 변동성을 나타내는 변동 계수(CV:Coefficient of Variation)를 핵심 지표로 사용하였다. 분석 결과, 0 값 비율이 1% 미만이고 변동 계수가 0.1 미만인 설비들은 연속 가동형으로, 0 값 비율이 40%를 초과하고 변동 계수가 2.0 이상인 설비들은 간헐적 가동형으로 명확히 분류되었다.

2.2. 연속 가동형 데이터 전처리

연속 가동형 데이터의 주요 문제인 일시적 데이터 강하(Dip)와 노이즈(Noise)를 해결하기 위해 두 가지 방법론을 비교했다. 방법 A 는 2 단계로 나눠 1 차로 Dip 구간을



<그림 1> 전체 실험 절차

선형 보간하고, 2 차로는 이동 평균 필터를 적용하여 Noise 를 평탄화 하는 방식이다. 방법 B 는 단일 필터 Savitzky-Golay 를 적용하는 방식이다.

<그림 3>은 연속형 데이터(D 공정)에 방법 A 를 적용한 결과이다.

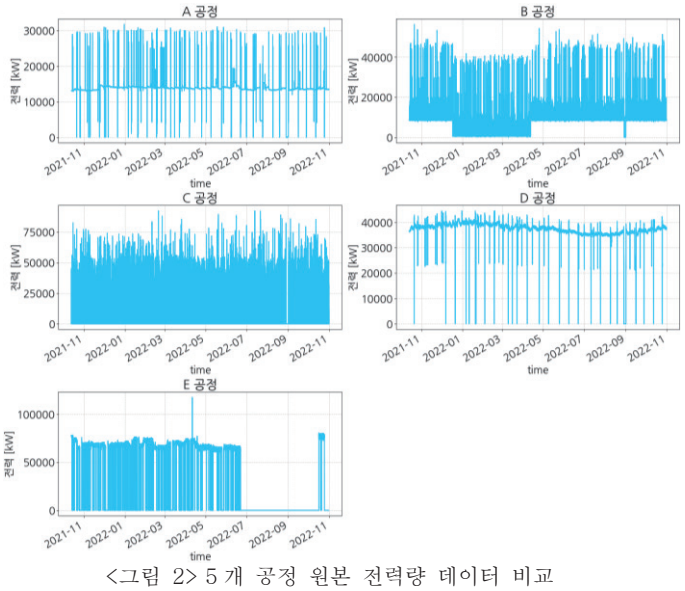
2.3. 간헐적 가동형 데이터 전처리

간헐적 가동형 데이터의 주요 문제인 돌입 전류 스파이크(Spike)를 해결하기 위해 두 가지 방법론을 비교했다.(0 값은 정상 신호이므로 처리 대상에서 제외함)

간헐적 가동형 데이터의 2 가지 전처리 방법 중 첫번째 방법 C 는 데이터 분포의 상위 0.1%를 초과하는 값들의 크기만 임계 값으로 강제 조정(Clipping)하는 방식인 백분위수 클리핑 기법을 사용하였다. 마지막 방법 D 로는 사분위수(IQR:Interquartile Range)를 이용해 통계적 이상치를 탐지하고, 해당 값을 데이터의 중앙값으로 대체하는 방식이다.

2.4. 성능 평가 및 최적 방법론 입증

각 전처리 방법론의 우수성을 입증하기 위해, 전처리 전후 데이터로 동일한 RandomForest 예측 모델을.



<그림 2> 5 개 공정 원본 전력량 데이터 비교

학습시켜 예측 오차(MAE:Mean Absolute Error) 개선도를 측정하였다

<그림 4>는 각 공정 데이터에 두 가지 방법론을 적용한 결과를 시각적으로 보여준다. 연속 가동형 데이터의 경우, 2 단계 접근법인 방법 A 가 압도적인 성능을 보였다. 이 방법은 A, B, D 공정 데이터에 적용 시 예측 성능을 각각 74.77%, 68.48%, 91.00% 만큼 크게 향상시켰다. 이는 Dip 이라는 큰 오류를 먼저 보간하고 Noise 라는 미세한 오류를 나중에 다듬는 2 단계 접근법이 매우 효과적이었음을 의미한다. 간헐적 가동형 데이터에서는 매우 중요한 결과가 도출되었다. 방법 D 를 적용하자 예측 성능이 오히려 C, E 공정에서 각각 -2.36%, -13.58%로 크게 악화되었다. 이는 돌입 전류 스파이크가 단순한 이상치가 아니라, 설비의 가동 시작을 알리는 중요한 신호임을 의미한다. IQR 기반 대체 방식은 이 중요한 신호를 훼손시켜 데이터의 가치를 훼손했다. 반면, 스파이크의 크기만 제어하고 신호 정보는 유지한 방법 C 는 예측 성능을 소폭 개선시켜, 간헐형 데이터에 더 적합한 방법론임이 확인되었다.

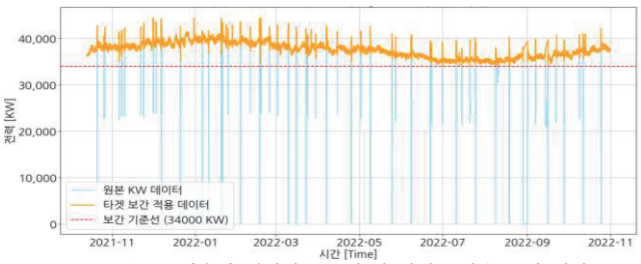
Ⅲ. 결론

본 연구는 공정 설비의 원본 전력 데이터를 분석하여 유형별로 최적의 전처리 방법론을 제시하는 것을 목표로 하였다. 실험 결과, 데이터의 특성을 고려하지 않고 통계적 기법을 일괄적으로 적용하는 것은 오히려 데이터의 가치를 훼손할 수 있다. 연속 가동형 데이터의 경우 큰 오류와 작은 노이즈를 순차적으로 제거하는 2 단계 방식이 효과적이었으며, 간헐적 가동형 데이터 같은 경우에는 중요한 신호를 보존하면서 극단적인 크기만 제어하는 방식이 성능을 향상시켰다.

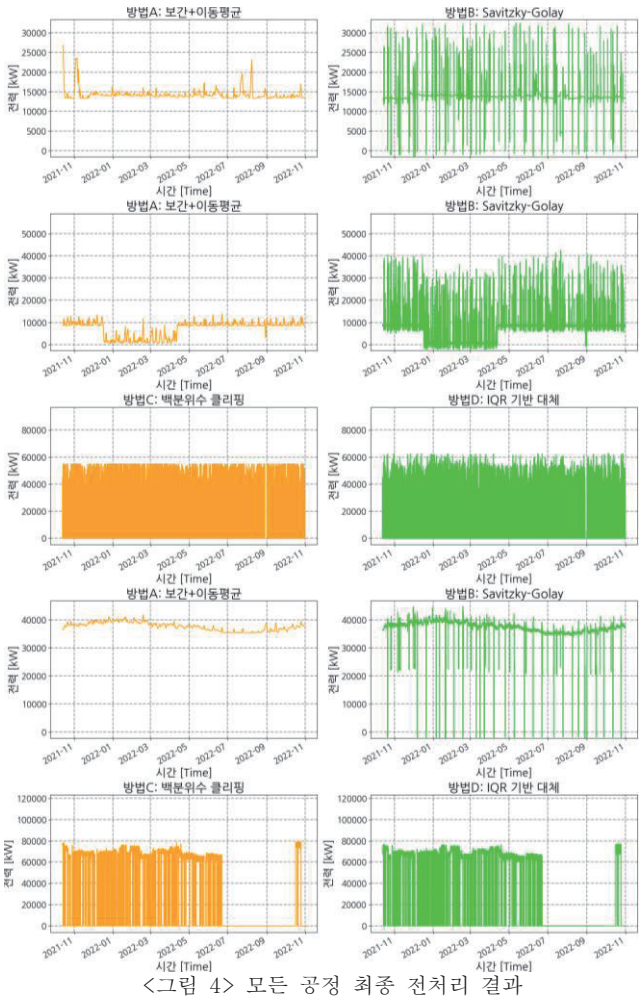
본 연구에서 제안하는 데이터 유형 분류 및 맞춤형 전처리 접근법은, 향후 다양한 산업 데이터 분석의 정확도를 높이는 실용적인 가이드라인이 될 수 있을 것으로 기대된다.

ACKNOWLEDGMENT

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



<그림 3> 연속형 데이터(D 공정)에 방법 A 적용 보건 결과



<그림 4> 모든 공정 최종 전처리 결과

참 고 문 헌

[1] 강민지, 김서림, 김선혁, 허태욱, 권구영. (2022). “LSTM-VAE 기반 공장 이상 모니터링 기술 개발”, *한국통신학회 학술대회 논문집.*

[2] 장민영, 이정일, 정남준, 고영준. (2023)."LSTM 과 GRU 를 활용한 업종별 전력사용량 데이터 보간 방법", *대한전기학회 논문지*, 72(3). 413-418.

[3] 권혁록. (2022). “딥러닝 기반 전력 계량데이터 결측 보정 모델에 대한 연구”, *조선대학교 박사학위논문.*

[4] 김서림, 강민지, 김선혁, 허태욱, 권구영. (2025). “시계열 데이터 이상치 탐지를 위한 LSTM-USAD 모델 개발”. *한국통신학회 동계종합학술발표회 논문집.*