

# 산업 전기데이터 무결성을 위한 LSTM 네트워크를 이용한 결측치 처리 정보보호 플랫폼

정범희<sup>1</sup>, 박선아<sup>1</sup>, 김주령<sup>1</sup>, 이충호<sup>2</sup>, 이상금<sup>1</sup>

국립한밭대학교<sup>1</sup>, 한국전자통신연구원<sup>2</sup>

qjagmlck@gmail.com, 20222024@edu.hanbat.ac.kr, juryeong1@gmail.com, leech@etri.re.kr, sangkeum@hanbat.ac.kr

## Missing Value Processing based Information Protection Platform by LSTM network for Industrial Electrical Data Integrity

Beomhee Jung<sup>1</sup>, Suna Park<sup>1</sup>, Juryeong Kim<sup>1</sup>, Chungho Lee<sup>2</sup>, Sangkeum Lee<sup>1</sup>

Hanbat National University<sup>1</sup>

Electronics and Telecommunications Research Institute (ETRI)<sup>2</sup>

### 요약

본 논문은 탄소중립과 지속 가능한 운영을 위한 산업 전기 데이터의 더욱 중요해진 정확성과 신뢰성을 유지하기 위해 산업체의 전기 사용량 데이터를 기반으로 최적의 하이퍼파라미터를 적용한 LightGBM 분류 모델을 이용한다. 산업체를 7 가지 유형으로 세분화하여 각 산업체의 전력 사용 패턴과 상관관계를 파악하고, 산업체의 일 별 전기 사용량을 분석한다. 이를 통해 데이터의 정확성을 저해시킬 수 있는 요인인 데이터의 결측치를 확인하며, 산업 전기 데이터의 무결성을 유지하기 위해 딥러닝 기반 LSTM 네트워크를 활용해 결측된 8 월 데이터를 7 월과 9 월 데이터를 기반으로 전기 사용량을 예측하고 대체하는 방안을 제안한다. 또한 예측된 전기 사용량 시각화로 모델의 성능을 평가하여 실제 값과 비교해 모델의 정확성을 확인한다. 연구 결과를 통해 데이터의 정확성과 일관성을 높여 지속 가능한 산업 운영에 기여할 수 있는 가능성을 제시하며, 그린 버튼 플랫폼과 같은 지속 가능한 에너지 사용 및 절약을 촉진하는 시스템 구축에 있어 데이터의 무결성과 신뢰성을 보장하여 데이터를 보호 관리하고 유지하는데 기여하고자 한다.

### I. 서론

산업체의 전기 데이터는 운영과 생산 프로세스의 핵심적인 구성 요소로 생산라인의 원활한 운영과 생산 과정의 효율성을 측정하고 개선하는 데 필수적인 정보를 제공한다. 특히, 탄소 중립으로 나아가는 과정에서 전기 데이터의 정확성과 신뢰성은 더욱 중요해졌다[1]. 이와 관련해 소비자들이 에너지 사용량을 관리하고 효율적으로 활용할 수 있도록 돕는 시스템인 그린 버튼 플랫폼은 에너지 절약 및 지속 가능한 에너지 사용을 촉진하는 데 기여한다는 점에서 이산화탄소 저감을 위한 탄소중립 및 환경 문제에 대한 해결책으로 제시된다. 이를 산업체 전기 데이터에 적용하면 산업체의 전기 데이터를 사용하여 운영 효율성을 개선하는 데 도움이 된다. 데이터 분석을 통한 에너지 효율성과 생산 프로세스의 최적화는 내외부적인 환경 영향을 줄이는 데에도 직접적인 영향을 미치므로, 전기 데이터의 정확성과 신뢰성은 지속 가능한 산업 운영을 위해 중요한 역할을 한다. 그러나 때로는 데이터에 결측치가 발생하여 데이터 플랫폼의 무결성과 신뢰성이 저해되는 경우가 있다. 이를 극복하기 위해 딥러닝 기반 LSTM (Long Short-Term Memory Network) 네트워크와 같은 기술을 활용하여 전기 데이터 측정에서 발생하는 결측치를 예측하고, 보완하여 데이터의 무결성을 유지하고자 한다. LSTM 네트워크는 시계열 데이터 분석에서 강력한 성능을 보이며, 데이터의 시간적 특성을

잘 파악하여 예측하는 데 적합하다[2]. 따라서 본 연구에서는 LSTM 네트워크를 활용하여 전기 데이터 측정에서 발생하는 결측치를 예측하여 데이터의 무결성을 유지하고 데이터를 보호하는 방안을 모색하고자 하며 이를 통해 데이터 저장 및 공유를 통한 그린 버튼 플랫폼 구축에서의 데이터 정확성과 일관성을 높여 신뢰할 수 있는 데이터를 보호하고 유지 관리하는데 기여할 것으로 기대된다.

### II. 전기 사용량에 따른 산업체 분류 모델 활용

산업체별 전기 사용량을 분석하기 위해 2020 년 1 월 ~ 12 월, 1 년간 산업체에서 발생하는 일별 전기 사용량(kW)을 센서로 15 분 단위 데이터를 수집했다. 전기 사용량에 따른 산업체 분류 모델을 활용하기 위해 원본 데이터셋에서 산업체와 시간별 데이터를 제외한 부분을 데이터 전처리했다. 산업체의 문자열 값을 LabelEncoder 를 사용하여 숫자로 변환하였고, 모델 성능을 평가하기 위해 80%를 훈련 데이터로 나머지 20%를 테스트 데이터로 사용한다.

#### 2-1. LightGBM 분류 모델

LightGBM(Light Gradient Boosting Machine)은 Gradient Boosting 알고리즘을 기반으로 한 트리 기반 학습 알고리즘으로서, 계산 속도와 메모리 소비 측면에서

기존의 GBDT(Gradient Boosting Decision Tree) 알고리즘보다 높은 성능을 발휘한다[3]. 본 연구에서는 반복실험을 통해 그림 1 과 같은 모델이 가장 우수한 성능을 달성하는 최적의 하이퍼파라미터를 선정했다.

Parameter	Value
Learning Rate	0.2
Max_depth	10
Min_Child_Samples	30
N_Estimators	300

그림 1. LightGBM 최적의 하이퍼파라미터

## 2- 2. 산업체 데이터 상관관계 분석

그림 1 의 LightGBM 하이퍼파라미터를 적용한 분류 모델을 학습하여 90.29%의 정확도를 얻었으며, 산업체별 전기 사용 패턴을 분석하고 상관관계를 파악했다. 그림 2 를 통해 전기 사용량에 따라 산업체가 7 개로 분류된다는 것을 확인할 수 있다. 따라서 본 연구에서는 그림 3 을 통해 국내 7 종류의 산업체 중 비교적 패턴이 단조로운 것이 금속 산업체라는 것을 확인하고, 이에 대해 연구하고자 한다.

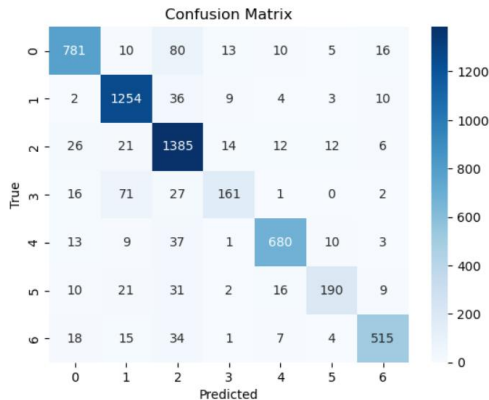


그림 2. 산업체 데이터 상관관계 분석을 위한 Confusion Matrix 히트맵

## III. LSTM 네트워크를 이용한 전기 사용량 예측

### 3-1. 예측 모델에 활용하기 위한 데이터 전처리

데이터 분석을 통해 NAN 값이 나오는 데이터, 일 평균 전기 사용량과 근사치가 0 인 데이터들을 이상치로 판단해 데이터를 제거했다. 산업체 전기 데이터의 일 별 전기 사용량을 분석했을 때 모든 산업체에서 8 월 데이터가 결측되어 있는 것을 확인했다. LSTM 네트워크 모델을 사용할 때, 예측하고자 하는 값의 실제 값이 있어야 하므로 결측된 8 월 데이터는 계절 관련성이 높은 7 월과 9 월 데이터의 일 평균 값으로 대체했다. 전처리한 데이터를 바탕으로 예측 모델 성능을 평가하기 위해 7 월 1일부터 31 일까지의 각 00:00 시부터 23:45 시까지 15 분 간격의 시간 별 평균 데이터를 입력 데이터로, 7 월 2일부터 8 월 1 일까지 15 분 간격의 시간 별 평균 데이터를 출력 데이터로 설정해 하루 데이터를 예측했다.

### 3-2. 딥러닝 기반 LSTM 네트워크 모델

딥러닝 기반 LSTM 네트워크는 RNN(Recurrent Neural Network)의 한 종류로, 시계열 데이터를 처리하고 학습하는 데 사용되는 신경망 아키텍처이다. 장기적인

의존성을 갖는 시계열 데이터를 효과적으로 학습할 수 있는 장점이 있어 시계열 데이터의 결측치를 예측하는데 활용할 수 있다. 8 월 1일부터 31 일까지의 데이터를 예측하기 위해 LSTM 레이어 4 개를 추가하고 Dropout 을 0.2 로 설정해 복잡한 주기의 변동성에 적응하도록 했다. 또한, Callback 을 활용하여 모델 체크포인트를 저장하고 과적합을 방지해 최적의 모델을 선정해 MSE(Mean Squared Error) Loss 0.02 값을 얻었다.

### 3-3. 전기 사용량 예측 분석

본 연구에서 8 월 결측 데이터를 7 월과 9 월 데이터의 일 평균 값으로 대체해 실제 값과 예측 값 차이의 표준편차를 이용한 95% 신뢰구간으로 예측 값 주변의 경계선을 설정하여 그림 3 처럼 출력 데이터의 불확실성 범위를 시각화 했다.

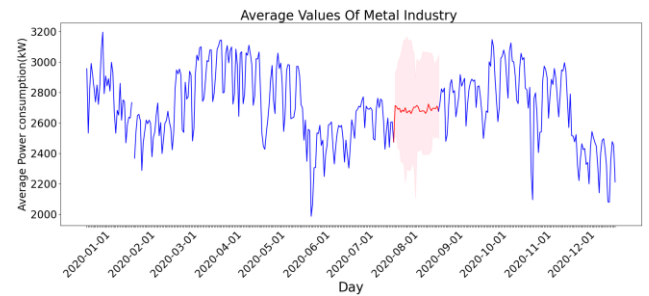


그림 3. LSTM 네트워크를 이용한 금속 산업 데이터 예측

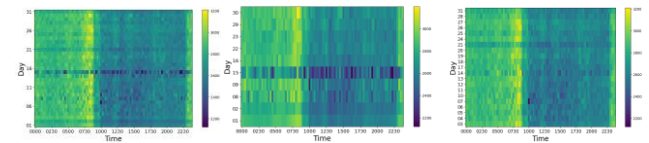


그림 4. 8 월 전체, 주말, 주중 전기 사용량 히트맵

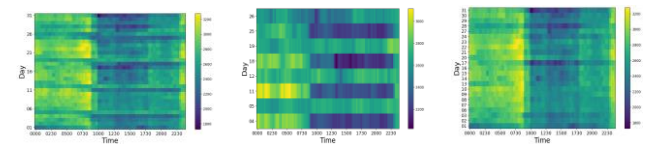


그림 5. 7 월 전체, 주말, 주중 전기 사용량 히트맵

다각도로 모델 성능을 평가하기 위해 같은 계절인 7 월 실제 전기 사용량과 8 월 예측 전기 사용량을 날짜와 시간 별로 그림 4, 5 처럼 출력 데이터를 비교 분석했다. 8 월 15 일 공휴일에 낮은 전기 사용량을 예측했고, 오후보다는 오전에 전기 사용량이 대체로 많다는 것과 제안된 모델이 주말보다 주중 예측 정확도가 높다는 것을 확인할 수 있다.

## IV. 결론

본 연구에서는 공장 에너지 소비량을 분석하기 위해 1 년간 산업체에서 발생하는 일 별 전기 사용량을 15 분 단위로 수집했다. 산업체 간 전력 사용 패턴 분석을 위해 산업체와 시간별 전기 사용량 데이터를 LightGBM 분류 모델을 최적의 하이퍼파라미터를 적용해 학습한 결과, 전기 사용량에 따라 산업체가 7 개로 분류됨을 확인했다. 8 월 데이터가 결측되어 있는 산업체 중 금속 산업체 데이터를 분석하고 이상치를 제거한 뒤, 금속 산업체의 8 월 전기 사용량을 예측하기 위해 7 월과 8 월 데이터를 딥러닝 기반 LSTM 네트워크 모델에 사용했다. 학습된 모델의 예측 성능을 평가하기 위해 MSE Loss 와 다양한

시각화를 통해 모델이 예측한 값과 실제 값 간의 차이를 수치화 했다. 이를 통해 데이터 저장 및 공유를 통한 그린 버튼 플랫폼 구축에서의 데이터 무결성과 신뢰성을 높여 데이터를 보호하고 유지 관리하는데 기여할 수 있다.

#### ACKNOWLEDGMENT

본 연구는 산업통상자원부(MOTIE)와 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구 과제입니다. (No. RS-2023-00237018)

#### 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] 이관영, 김선혁, 이상금, 허태욱, 이일우. (2023). 연속공정에서의 공장 에너지 수요 분석 및 예측 모델. 한국통신학회 학술대회논문집
- [3] KE, Guolin, et al, "Lightgbm: A highly efficient gradient boosting decision tree," Advances in neural information processing systems 30, pp.3149-3157, Dec 2017. DOI: 10.5555/3294996.3295074