

기상 데이터를 활용한 기계학습 기반 난방 에너지 소비 이상 탐지

윤혜주, 권혜영, 전창재*
세종대학교

i3yyj5676@gmail.com, hy913149@sju.ac.kr, *cchun@sejong.ac.kr

Machine Learning-Based Anomaly Detection for Heating Energy Consumption Based on Weather Data

HyeJu Yun, HyeYoung Gwon, Chang-Jae Chun*
Sejong University

요 약

난방 에너지 공급 시스템은 매우 규모가 크고, 소비하는 대상 인원이 매우 많은 주요 에너지 공급/소비 시스템 중 하나이다. 이러한 난방 에너지 시스템에서 이상이 발생하면 그로 인한 파급효과가 매우 심각하다. 따라서 본 논문에서는 난방 에너지 소비 데이터와 기상 데이터를 활용하여 기계학습 기반 이상 탐지 방법을 제안한다. 시뮬레이션을 통해 기본적인 이상 탐지 모델인 임계점 기반 이상 탐지 방법보다 이상 상황의 정도에 따라 10~20% 성능이 향상된 것을 확인했다. 또한 여러 기계학습 모델 및 신경망 모델의 성능 비교를 통해 Extra Tree 알고리즘을 활용한 모델의 성능이 가장 높은 것을 확인했다.

I. 서 론

아파트에서는 난방을 목적으로 톤(ton) 단위의 가스가 쓰인다. 난방은 수많은 사람이 사용하기 때문에 사용에 문제가 생기게 되면 여러 가지 불편함이 발생하며, 특히 가스가 이동하는 배관에서 누출, 침습 등 이상이 발생하게 되면 재산 및 인명 피해가 발생할 수도 있다. 이러한 상황들을 미연에 방지하기 위해선 철저한 사전, 사후 관리 필요하다. 그러나 사람이 직접 전 세대 난방 사용량을 관리하기에는 많은 인적, 시간적 자원이 필요하므로 현실적으로는 어려움이 있다. 따라서 최근에는 기계학습을 이러한 이상 탐지에 활용하고 있다 [1]. 기계 학습 알고리즘을 이용하기 위해서는 사전에 정상과 이상 데이터가 균일하게 잘 분포된 데이터셋이 필요하다. 그러나 실제 상황에서 이상 데이터가 발생하는 빈도가 매우 작으므로 현장에서 충분한 양의 이상 데이터를 확보하기에는 어려움이 있다. 이러한 문제를 해결하기 위해 연구 [2]에서는 오픈 소스 시뮬레이터를 사용하여 인공 데이터셋을 생성하였고 연구 [3]에서는 실제 값을 임의의 값으로 수정하거나 특정 값을 더해주는 방식으로 이상치를 생성하였다. 본 논문에서는 실제 난방 사용량에 일정 값을 곱하여 이상 데이터를 생성하였고, 모델 성능향상을 위해 난방사용량뿐만 아니라 기상 데이터 (외기온도 및 습도 데이터)를 사용하였다.

II. 본론

본론에서는 임계점 기반 모델, 기계학습 기반 모델 및 신경망 기반 모델에 대해 상세 설명을 하였다.

2.1 기계학습 기반 모델

본 연구에서 사용한 기계 학습 모델은 Logistic Regression, Decision Tree, Extra Tree, LightGBM 이다.

기계학습 기반 모델에 사용한 데이터셋은 전처리한 데이터셋에서 랜덤으로 데이터를 골라 기존 난방 사용량에서 1.2, 1.5, 2, 3 을 곱해주어 인공 이상 데이터를 생성하였다. 실험에 따라 외기 온도, 습도 데이터를 추가한 후 표준화하여 훈련에 사용하였다.

2.2 임계점(Threshold) 기반 모델

임계점 기반 모델은 특정 임계점을 기준으로 정상과 이상을 구별해내는 모델이다. 해당 모델은 기계학습과 동일한 데이터셋을 적용하여 훈련하였다. 해당 모델은 정확한 비교를 위해 신경망 기반 모델의 FNR(False Negative Rate)값과 가능한 가장 비슷하면서도 훈련했을 때 총 10 개의 정확도가 0 이나 NULL 이 아닌 경우를 만족하는 조건으로 FNR 을 제한하여 정확도를 산정하였다.

2.3 신경망 기반 모델

신경망 기반 모델에 사용한 데이터셋은 전처리한 데이터셋을 24 시간 주기로, 한 시간의 간격으로 잘라서 가장 마지막 시간의 난방 사용량에 1.2, 1.5, 2, 3 을 곱해주어 인공 이상 데이터를 생성하였다. 이는 현재 시점에서의 난방 사용량이 이상인지 정상인지 판단하게 하도록 만든 환경이다. 이상 데이터를 생성한 데이터셋을 '평균', '분산', '왜도', '첨도', '최대값', '최소값', '중앙값'의 7 개 특징값을 갖는 데이터 세트로 재구성하였다. 생성된 데이터 세트의 개수는 총 1824 개(정상 912 개, 이상 912 개)이다. 신경망 기반 이상 탐지를 위해 활용한 모델은 Feedforward Neural Network (FFNN)이고 해당 모델을 입력층과 7 개의 은닉층, 출력층으로 구성하였다.

각 층이 가지는 노드의 개수는 다음과 같다: 입력층은 온도와 습도 특성 추가에 따라 7, 14, 21 개이다. 은닉층은 순서대로 50, 80, 120, 120, 120, 80, 50 개이며 출력층은 2 개이다. 또한 활성화함수는 Relu, 출력함수는

Softmax, 손실함수는 Categorical Cross entropy 를 사용하였으며 최적화를 위해 Adam 알고리즘을 사용하였다. 학습 횟수는 150 epoch 이고 학습률은 0.0001 로 설정하였다.

III. 실험 결과

3.1 실험 환경 설정

본 논문에서는 한국에너지기술연구원에서 수집한 수원시 아파트 단지 난방 사용량 데이터셋을 활용하였다. 해당 데이터셋은 1 년 치 데이터이지만, 여름철은 난방사용량이 매우 낮으므로 이상 탐지 대상에서 제외했으며, 평균 난방 사용량이 높은 겨울철 데이터를 활용했다. 좀 더 구체적으로, 10 월부터 12 월 까지의 날짜, 난방 사용량, 외기 온도, 습도 데이터를 이상 탐지 모델 구축에 활용했다.

기계학습 기반 모델과 신경망 기반 모델 모두 모델 훈련을 한 번 시행할 때마다 데이터셋을 무작위로 섞어 주었고 이를 총 10 번 반복한 정확도의 평균을 계산하여 최종 결과값으로 산정하였다. 또한 이상 데이터와 정상 데이터의 비율은 5:5 로, 훈련 데이터와 테스트 데이터의 비율은 8:2 로 설정하였다. 전체 모델 학습 및 테스트는 Python 3.8.13, TensorFlow 2.7.0 환경에서 수행하였다.

3.2 모델에 따른 이상 탐지 정확도

정확한 비교를 위해 난방 사용량, 외기 온도, 습도 특성을 모두 추가하여 훈련한 전체 모델 간 성능 비교 결과를 그림 1 에 나타냈다. 비교한 모델은 Feedforward Neural Network (FFNN), Threshold-based, Logistic Regression, Decision Tree, Extra Tree, LightGBM 을 사용했다. 그림 1 에서와 같이 Extra Tree 기반 이상 탐지 모델이 가장 정확하게 이상을 감지함을 확인하였다. 구체적인 결과값은 표 1 에 기재했다.

3.3 온도, 습도 특성 추가에 따른 이상 탐지 정확도

이상 탐지 정확도가 가장 높은 모델인 Extra Tree 기반 모델에 대해 난방 사용량만 사용한 모델 ExtraTree(1), 난방 사용량과 외기 온도 특성을 사용한 모델 ExtraTree(2), 난방 사용량과 외기 온도, 습도 특성을 사용한 모델 ExtraTree(3)간 성능 비교 결과, 그림 2 에서와 같이 난방 사용량, 외기 온도, 습도 특성을 모두 사용한 모델인 ExtraTree(3)이 그렇지 않은 경우보다 큰 차이로 높은 정확도를 기록했다.

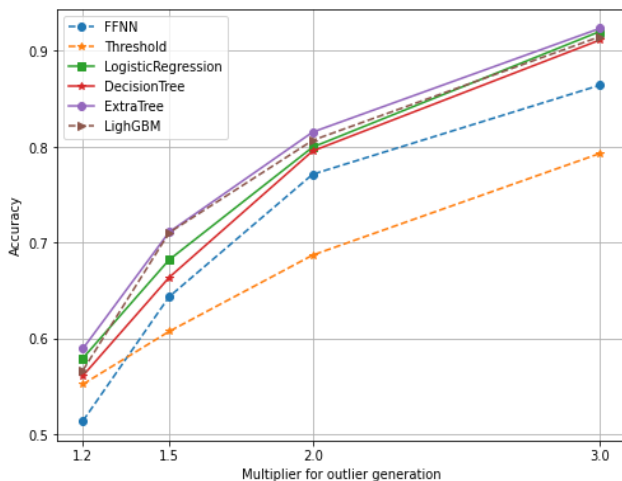


그림 1. 이상 탐지 정확도 결과비교

표 1. 모델간 결과값 비교 (정확도(표준편차))

	*1.2	*1.5	*2	*3
FFNN	0.5143 (0.0345)	0.6437 (0.0396)	0.7714 (0.0198)	0.8643 (0.0114)
Threshold-based	0.5525 (0.0088)	0.6075 (0.0080)	0.6871 (0.0092)	0.7931 (0.0073)
Logistic Regression	0.5797 (0.0257)	0.6822 (0.0209)	0.7997 (0.0187)	0.9203 (0.0069)
Decision Tree	0.5614 (0.0180)	0.6638 (0.0278)	0.7957 (0.0276)	0.9114 (0.0139)
Extra Tree	0.5897 (0.0220)	0.7111 (0.0172)	0.8154 (0.0150)	0.9235 (0.0122)
LightGBM	0.5670 (0.0214)	0.7105 (0.0110)	0.8068 (0.0160)	0.9149 (0.0150)

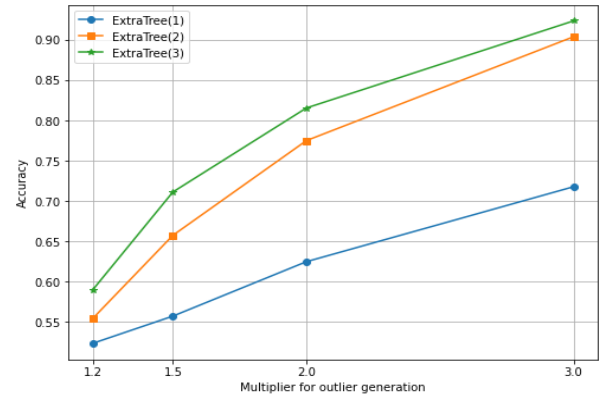


그림 2. 엑스트라 트리 기반 모델 비교

IV. 결론

본 논문에서는 난방 에너지 소비 데이터와 기상 데이터를 활용하여 기계학습 기반 이상 탐지 방법을 제안했다. 임계점 기반 모델, 기계학습 모델, 및 신경망 모델의 성능 비교를 통해 Extra Tree 알고리즘을 활용한 모델의 성능이 가장 높은 것을 시뮬레이션을 통해 확인했다. 또한, 외기온도 및 습도 데이터를 고려했을 때 좀 더 성능이 향상되는 것을 확인하였다.

ACKNOWLEDGMENT

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00106, 복합 도메인의 에너지 수요 빅데이터를 활용한 설명 가능한 인공지능 기반 진단 및 분석 프레임워크 개발). 이 논문은 산업통상자원부 "열공급망 수용가 빅데이터 기반 에너지관리 기술개발 및 실증" 사업의 연구비지원(20212020900150)에 의해 수행됨. 이 논문은 2021 년도 세종대학교 교내연구비 지원에 의한 논문입니다.

참 고 문 헌

- [1] A. B. Nassif, M. A. Talib, Q. Nasir, and F. M. Dakalbab, "Machine learning for anomaly detection: A systematic review," IEEE Access, vol. 9, pp. 78658– 78700, 2021.
- [2] Abdullatif A., Rovetta S., and Masulli F., "Layered ensemble model for short-term traffic flow forecasting with outlier detection," IEEE 2nd Int. Forum Research and Technologies for Society and Industry Leveraging a better tomorrow, pp. 1– 6, 2016.
- [3] M. De Nadai and M. van Someren, "Short-term anomaly detection in gas consumption through ARIMA and artificial neural network forecast," IEEE workshop on Environmental, energy and structural monitoring systems, pp. 250–255, 2015.