에너지 빅데이터를 활용한 머신러닝 기반의 생산 예측 모형 연구

강미영* · 김석

호남대학교

A Study on Production Prediction Model using a Energy Big Data based on Machine Learning

Mi-Young Kang* · Suk Kim

Honam University

E-mail : kmy2021@honam.ac.kr / ks2021@honam.ac.kr

요 약

전력망의 역할은 안정적인 전력공급이 최우선이다. 예고 없는 불안정한 상황에 대한 여러 가지 대비에 대한 방안이 필요하다. 기상 데이터를 활용하여 탐구적 데이터 분석을 통한 피처 간의 관계를 파악하여 머신러닝 기반의 에너지 생산 예측 모형을 모델링한다. 본 연구에서는 주성분분석을 사용하여 에너지 생산 예측 시 영향을 미치는 피처를 추출하였으며 머신러닝 모델에 적용함으로써 예측 신뢰도를 높였다. 제안한 모형을 사용하여 특정 기간을 대상으로 생산 에너지를 예측하고 해당 시점의 실제 생산 값과 비교함으로써 주성분분석을 적용한 에너지 생산 예측에 대한 성능을 확인하였다.

ABSTRACT

The role of the power grid is to ensure stable power supply. It is necessary to take various measures to prepare for unstable situations without notice. After identifying the relationship between features through exploratory data analysis using weather data, a machine learning based energy production prediction model is modeled. In this study, the prediction reliability was increased by extracting the features that affect energy production prediction using principal component analysis and then applying it to the machine learning model. By using the proposed model to predict the production energy for a specific period and compare it with the actual production value at that time, the performance of the energy production prediction applying the principal component analysis was confirmed.

키워드

Machine Learning, Energy Big Data, Energy Production Prediction Model, Principal Component Analysis

Ⅰ. 서 론

국내 대부분의 지역에서 매해 높은 기온의 기록을 경신하고 있다. 2021년 7월 서울·인제 35.9도 기록 경신[1], 2017년부터 2022년까지 연간 전국 평균 폭염 일수 또한 계속 증가하고 있다[2]. 이러한 기록적인 폭염에 전력수요 또한 급증하였다.

전력 수급에 안정화를 위해 국내뿐만 아니라 전 세계적으로 다양한 연구가 진행되고 있다. 머신러닝과 딥러닝에 대한 연구가 활발해지면서 데이터를 이용한 전력수요 예측 모형을 에너지 효율 분야에 적용하려는 연구가 진행되고 있다[3-6].

본 논문에서는 수요 에너지를 예측하는데 있어서 어떠한 기상변수가 중요한지 분석하여 기상변수 적용에 따른 모델의 성능을 향상 시킨다. 실험결과 기온, 기압, 강수량, 습도 등이 에너지 수요

* speaker

예측에 있어 중요 피처임을 확인하였다.

기상변화 데이터 중 중요 변수와 에너지 데이터를 이용하여 인공지능 학습 모델에 적용해 봄으로써 에너지 사용량을 예측한다.

제2장에서는 본 연구에 사용된 데이터 수집과 주성분분석 머신러닝 모형에 관한 관련연구에 대해 설명한다. 제3장에서는 기상변화 데이터에 주성분분석을 사용함으로써 중요 피처를 추출 하기위한 실험환경을 설명한다. 제4장에서는 본 연구에서 제안한 학습모형을 적용하여 예측에 대한 신뢰도를 확인한다.

II. 관련 연구

기상변화 데이터는 다수의 독립변수로 되어 있으며, 일반적으로 차원이 증가할수록 즉 피처(feature)가 많아질수록 예측 신뢰도가 떨어지고 과적합(overfitting)이 발생하고 개별 피처 간의 상관관계가 높을 가능성이 있다.

PCA(주성분분석, Principal Component Analysis)는 고차원의 데이터를 저차원의 데이터로 축소하는 차원 축소 알고리즘이다. 기상변화 데이터는 훈련 데이터의 피처가 많다. 그렇지만 모든 피처가 결과에 주요한 영향을 미치는 것은 아니다. 가장 중요한 피처가 있을 것이고 순서에 따라 중요한 피처들이 존재할 것이며 그중에는 쓸모없는 피처도 있을 것이다. 이런 피처들 중 가장 중요한 피처들만 추출하여 사용하는 것이 PCA이다[7].

본 연구를 진행하면서 기상변화 데이터 변수들 중 어떤 피처가 모델의 성능에 큰 영향을 줄지 파악했으며, 피처를 선택/가공하는 과정을 거쳤다. 아래에 기술된 관련 접근법 중 피처 추출법을 사용하였다.

1) 피처 선택(Feature Selection) : 특정 피처에 종속성이 강한 불필요한 피처는 아예 제거한다. 장점은 선택한 피처의 해석이 용이하다는 점이고 단점은 피처 간 상관관계를 고려하기 어렵다는 점이다.
2) 피처 추출(Feature Extraction) : 더 작은 차원으로 피처들을 맵핑한다. 장점은 상관관계를 고려하기 용이하고 피처의 개수를 많이 줄일 수 있다는 점이고 단점은 추출된 변수의 해석이 어렵다는 점이다.
3) 피처 생성 (Feature Engineering) : 데이터 테이블에서 피처가 부족한 상황일 때 적용하는 기법으로 해당 데이터와 만들고자 하는 머신러닝 모델의 기능 활용 목적에 따라 새로운 피처들을 생성해 내는 것이다.

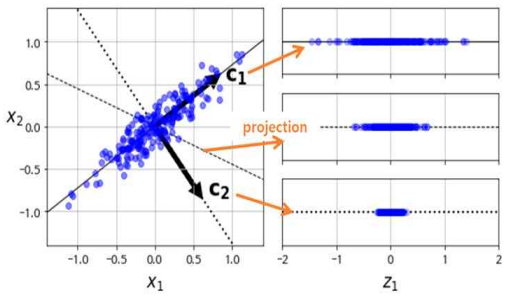


그림 1. 주성분분석(피처 추출)

그림 1에서 보는 것처럼 PCA는 데이터의 분산이 최대가 되는 축을 찾게 된다. 기상변화 데이터의 모든 차원을 살리면서 차원을 축소할 수는 없다. 모든 특성을 살릴 수는 없지만, 최대한 특성을 살리며 차원을 낮춰주는 방법을 사용한다.

에너지 사용량을 예측하는 학습 모델로 본 연구에서는 앙상블 알고리즘 중 랜덤 포레스트 모델을 사용한다.

랜덤 포레스트는 앙상블 기법들 중에서도 같은 알고리즘(결정트리)으로서 여러 개의 분류기를 만들어 예측하는 배깅 방식의 대표적인 알고리즘이다. 비교적 빠른 수행 속도를 가지고 있으며 다양한 영역에서 높은 성능을 보이며 부스팅 방식과 마찬가지로 기반 알고리즘인 결정트리의 쉽고 직관적인 장점을 가지고 있다[8-9].

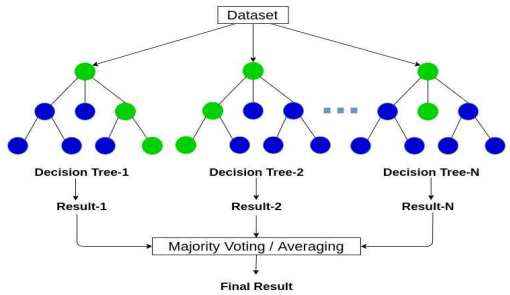


그림 2. 랜덤 포레스트 알고리즘

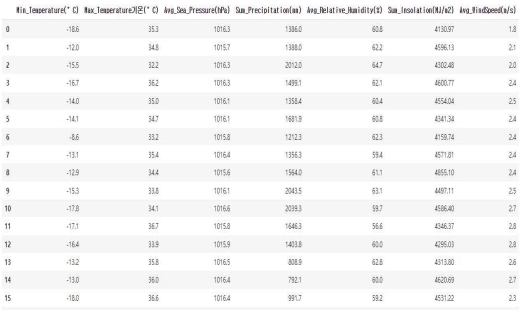
III. 실험 환경 및 결과

기상변화 데이터는 공공데이터 포털 사이트에서 데이터를 수집하고 전력 사용량 데이터를 수집하여 실험을 진행하였다.

일차적으로 기상변화 데이터를 수집하여 머신러닝 모델인 주성분분석을 적용하였다.

1단계 : 기상변화 데이터의 공분산 행렬을 구한다.
2단계 : 공분산 행렬에서 고유 벡터와 고윳값을 구한다.
3단계 : 고윳값이 큰 것부터 순서대로 정렬한다.

4단계 : 기상변화에 영향을 미치는 고유 벡터를 추출함으로써 차원 축소를 한다.
5단계 : 추출된 고유 벡터를 축으로 하여 데이터의 차원을 줄인다.



	Min_Temperature(° C)	Max_Temperature(°C)	Avg_Sea_Pressure(hPa)	Sum_Precipitation(mm)	Avg_Relative_Humidity(%)	Sum_Insolation(MJ/m2)	Avg_WindSpeed(m/s)
0	-18.6	35.3	1016.3	1386.0	60.8	4130.97	1.8
1	-12.0	34.8	1015.7	1388.0	62.2	4596.13	2.1
2	-15.5	32.2	1016.3	2012.0	64.7	4302.48	2.0
3	-16.7	36.2	1016.3	1499.1	62.1	4600.77	2.4
4	-14.0	35.0	1016.1	1358.4	60.4	4554.04	2.5
5	-14.1	34.7	1016.1	1691.9	60.8	4341.34	2.4
6	-8.6	33.2	1015.8	1212.3	62.3	4159.74	2.4
7	-13.1	35.4	1016.4	1356.3	59.4	4571.81	2.4
8	-12.9	34.4	1015.6	1564.0	61.1	4655.10	2.4
9	-15.3	33.8	1016.1	2043.5	63.1	4497.11	2.5
10	-17.8	34.1	1016.6	2039.3	59.7	4586.40	2.7
11	-17.1	36.7	1015.8	1646.3	56.6	4546.37	2.8
12	-16.4	33.9	1015.9	1403.8	60.0	4595.03	2.8
13	-13.2	35.8	1016.5	808.9	62.8	4313.80	2.6
14	-13.0	36.0	1016.4	792.1	60.0	4620.69	2.7
15	-18.0	36.6	1016.4	991.7	59.2	4531.22	2.3

그림 3. 기상 데이터

그림 3에서 보는 것처럼 기상변화 데이터의 피처는 7개를 선택 수집 하였으며 2011년부터 2021까지의 연간 데이터를 기반으로 한다.

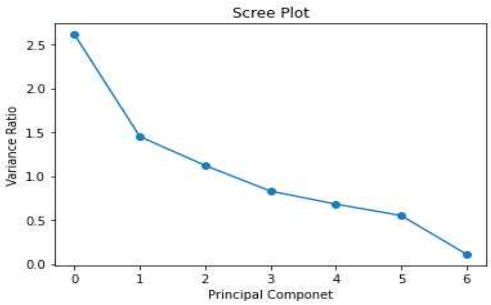


그림 4. Scree Plot

그림 4를 통해서 기상변수 데이터 피처 중 Scree plot 그래프를 통해 고윳값 크기를 기반으로 차원의 수를 3개로 결정하였다.

차원을 축소한 중요 피처를 이용하여 PCA는 데이터의 분산이 최대가 되는 축을 찾게 된다. 이를 이용해 머신러닝 학습 모형에 적용함으로써 에너지 사용량을 예측한다.

Ⅳ. 제안한 예측 모형

공공데이터에서 기상변화 데이터를 일차적으로 수집하여 데이터를 정제하였다. 전력 사용량 데이터를 수집하여 이차적으로 데이터에 대한 정제화 과정을 거쳤다.

본 연구에서는 기상변화 데이터를 수집하여 주성분분석 알고리즘을 사용하여 7차원 데이터를 머신러닝 모형에 적용하여 에너지 사용량을 예측하였다. 같은 데이터를 사용하여 3차원으로 차원 축소한 후 머신러닝 모형을 적용하여 에너지 사용량을 예측하였다.

에너지 사용량을 예측하는 데 있어 피처가 많을 때 주성분분석 알고리즘을 사용하여 유사한 성능을 유지하면서 빠르게 데이터를 분석할 수 있도록 하였다.

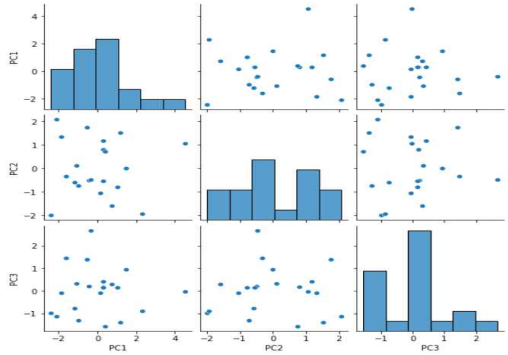


그림 5. 기상 데이터 변수 간 분포

표준화된 피처 데이터를 주성분 데이터로 변환하고 기상변화 데이터의 피처 사이의 전체적인 분포를 파악하였다. PC1의 분포가 가장 큰 것을 그림 5를 통해서 확인할 수 있다.

그림 6. 연도별 최대전력 사용량

에너지 사용량을 예측하기 위한 데이터로 전력 사용량 데이터를 수집하였으며 그림 6에서는 연도별 최대전력 사용량에 대한 실측 데이터를 보여주고 있다. 실험은 파이썬 기반으로 실제 데이터를 제안한 예측 모형에 적용하여 진행하였다.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{\text{실제전력수요} - \text{예측전력수요}}{\text{실제전력수요}}\right|$$

에너지 사용량 예측의 정확성은 에너지 수요 예측 오차율을 통해서도 확인할 수 있다.

MAPE는 Mean Absolute Percentage Error의 약어이며 평균 절대 백분율 오차를 의미한다.

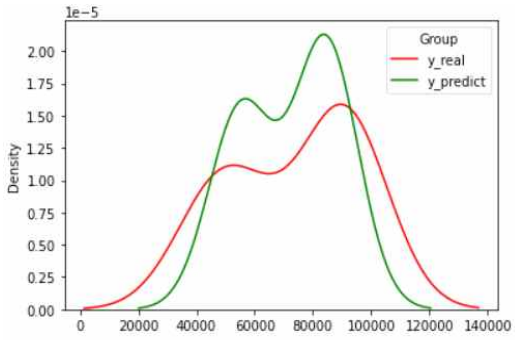


그림 7. 제안한 모형 적용한 예측 결과 비교

본 연구에서는 에너지 사용량을 예측하는 머신러닝 학습 모델로 랜덤 포레스트 기법을 사용하여 적용하였다. 차원이 큰 변수 간의 관계를 주성분분석 알고리즘을 사용하여 차원 감소를 시도 하였으며 랜덤포레스트 기법으로 모델링하여 결과를 예측하였다. 실제 데이터와 에너지 사용량 예측에 대한 비교 결과를 그림 7에서 보여주고 있다.

Ⅴ. 결 론

본 논문에서는 주성분분석을 이용한 에너지 생산 예측 모형을 제안하였다. 차원 축소된 피처 데이터를 이용하여 머신러닝 모델을 적용할 경우 에너지 사용량을 예측하는 데 있어 적은 수의 피처만으로 특정 현상을 확인하고 모델 성능 향상에 기여함을 확인 할 수 있다.

향후 전력수요와 상관관계가 높은 다른 변수들을 고려하여 연구를 진행하고자 한다. 또한 다양한 지역을 가지고 제안한 모형을 통해 범용성 여부를 확인하고자 한다.

Acknowledgement

이 과제(결과물)는 2022년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다.(2022RIS-002)

References

[1] G. Y. Lee, “Hankyoreh,” [Internet]. Available : http://www.hani.co.kr/arti/society/environment/1004705.html

[2] S. H. Han, “Maeil Economic Daily,” [Internet]. Available : http://www.mk.co.kr/news/society/view/2022/08/688334

[3] S Park, S Park, M Choi, S Lee, T Lee, K Cho. “Reinforcement LearningBased BEMS Architecture for Energy Usage Optimization.” Sensors 20, no. 17: 4918, 2020.

[4] A.M. Castro Martinez., S.H. Mallidi & B.T. Meyer. (2017). “On the relevance of auditory-based Gabor features for deep learning in robust speech recognition,” in *Computer Speech & Language, 45, 21-38.*

[5] Kim, S., & Park, S. (2011). “Multi-class classification of database workloads using PCA-SVM classifier,” in *The Korean Institute of information Scientists and Engineers, 38(1), 1-8.*

[6] Liang, C., Li, H., Lei, M., & Du, Q.(2018). “Dongting lake water level forcast and its relationship with the three gorges dam based on a long short-term memory network,” in *Water, 10(10), 1389.*

[7] Han, Jiawei, Micheline Kamber, andJian Pei. “Data mining: concepts and techniques,” in *Morgan kaufmann, 2012.*
http://doi.org/10.1088/1742-6596/971/1/012037

[8] M. F., Adiwijaya, & Al-faraby, S.(2017). “Text categorization on Hadith Sahih Al-Bukhari using Random Forest,” in *International Conference on data and Information Science, IOP Conference Series: Journal of Physics: Conf. Series 971.*
http://doi.org/10.1088/1742-6596/971/1/012037

[9] Kapsu Kim. “A Development of AI Education Model Using Weather and Power Usage Data” in *The Korea Society of Energy and Climate Change Education, 10(3), 237-246, 2020.*
http://doi.org/10.22368/ksecce.2020.10.3.237