Regular paper
Journal of the KIECS. pp. 39-44, vol. 18, no. 1, Feb. 28. 2023, t. 117, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2023.18.1.39

기상 빅데이터를 활용한 신재생 에너지 발전량 예측 모형 연구

강미영*

Renewable Energy Generation Prediction Model using Meteorological Big Data

Mi-Young Kang*

요 약

태양광, 풍력 등의 신재생 에너지는 기상조건 및 환경변화에 민감한 자원이다. 설치위치 및 구조에 따른 설비의 발전량이 달라질 수 있기 때문에 정확한 발전량 예측은 중요하다. 기상 빅데이터를 활용하여 주성분 분석을 기반으로 데이터 전처리 과정을 진행하여 신재생 에너지 발전량 예측 시 영향을 미치는 피처간의 관계를 모니터링하였다. 또한, 본 연구에서는 영향을 미치는 민감도에 따라 데이터셋을 재구성하여 머신러닝 모델에 적용하여 예측도를 테스트하였다. 제안한 모형을 사용하여 신재생 에너지를 대상으로 기상환경에 따라 에너지 발전량을 예측하고 해당 시점의 실제 생산 값과 비교함으로써 랜덤 포레스트 회귀 분석을 적용한 에너지 발전량 예측에 대한 성능을 확인하였다.

ABSTRACT

Renewable energy such as solar and wind power is a resource that is sensitive to weather conditions and environmental changes. Since the amount of power generated by a facility can vary depending on the installation location and structure, it is important to accurately predict the amount of power generation. Using meteorological data, a data preprocessing process based on principal component analysis was conducted to monitor the relationship between features that affect energy production prediction. In addition, in this study, the prediction was tested by reconstructing the dataset according to the sensitivity and applying it to the machine learning model. Using the proposed model, the performance of energy production prediction using random forest regression was confirmed by predicting energy production according to the meteorological environment for new and renewable energy, and comparing it with the actual production value at that time.

키워드

Energy Generation, Meteorological Big Data, Random Forest Regression, Renewable Energy, PCA
에너지 발전량, 기상 빅데이터, 랜덤 포레스트 회귀, 신재생 에너지, 주성분 분석

* 교신저자 : 호남대학교 정보통신공학과
• 접 수 일 : 2022. 12. 24
• 수정완료일 : 2023. 01. 18
• 게재확정일 : 2023. 02. 17
• Received : Dec. 24, 2022, Revised : Jan. 18, 2023, Accepted : Feb. 17, 2023
• Corresponding Author : Mi-Young Kang
Dept. Information & Communication Engineering, Honam University,
Email : kmy2021@honam.ac.kr

Ⅰ. 서 론

최근 인공지능이 신재생 에너지 분야에서 핵심기술로 떠오르고 있다[1]. 신재생 에너지는 일사량, 풍속 등의 기상 환경적 요인 등 외부요인에 매우 민감한 발전 특성을 가지고 있다.

인공지능 기술은 신재생 에너지의 설계, 최적화, 진단, 관리 및 예측분야 등에 활용되고 있다[2]. 신재생 에너지 전원 중 대부분 풍력, 태양광 발전 부분에 인공지능 기술 적용이 되고 있으며, 과거 발전량 이력, 기상정보 및 기후예보모델 등을 입력 변수로 하는 머신러닝 및 딥러닝 모델 기반의 발전량 예측기술 연구가 활발히 진행되고 있다[3-4]. 풍력발전은 바람이 가진 운동에너지를 변환하여 전기 에너지를 생산하는 발전시스템이다. 육상에 설치된 풍력발전기를 육상풍력발전기, 해상에 설치된 풍력발전기를 해상풍력발전기라 분류하여 해상풍력발전기는 설치 형식에 따라 고정식과 부유식으로 분류된다.

본 연구에서는 모델기반의 발전량 예측 데이터와 실제 발전한 데이터를 비교분석하여 신재생 에너지의 정확한 발전량 예측기술에 적용하고자 한다. 제2장에서는 본 연구에 사용된 데이터 수집과 주성분 분석 머신러닝 모형에 관한 관련 연구에 대해 설명한다. 제3장에서는 기상 빅데이터를 사용하여 주성분 분석을 적용함으로써 중요 피처를 추출하기 위한 실험환경을 설명한다. 제4장에서는 본 연구에서 제안한 모형을 사용하여 신재생 에너지를 대상으로 기상환경에 따라 에너지 생산량을 예측하고 신뢰도를 확인한다.

Ⅱ. 신재생 에너지 발전량 예측 모델

신재생 에너지 발전량 예측 목적은 두 가지로 분류할 수 있다. 첫째는 신재생 에너지 경제성 분석을 위한 연간, 월간 발전량 예측이 있으며, 두 번째는 전력계통 안정화 및 신재생 에너지 거래를 위한 하루 전, 한 시간 전 등의 짧은 주기의 발전량 예측기술이다. 본 논문에서는 발전량 예측 관련 연구에 필요한 기술을 살펴본다.

2.1 회귀분석

회귀분석은 변수들 사이의 관계를 모델링하는 기법이다. 여러 분야에서 가장 광범위하게 사용되는 통계기법 중 하나이다. 회귀분석의 목적은 관심이 있는 종속변수들에 영향을 주는 독립변수들을 찾고, 독립변수들과 종속변수들의 관계를 나타내는 모델을 만드는 것이다. 신재생 에너지 발전량 예측은 다양한 변수들을 고려해야 하므로, 다중회귀분석을 기반으로 발전량을 예측한다.

신재생 에너지의 기온, 습도, 일사량 등의 다수개의 독립변수와 예측하려는 발전량인 종속변수 사이의 관계를 수식으로 구성하여 예측모형을 구성하며, 최소제곱법 또는 손실함수 등의 알고리즘 구성이 쉽다는 장점을 가지고 있다[5-6]. 그림 1에서는 종속변수(Y)에 영향을 미치는 독립변수가 여러 개인(X1, X2) 다중회귀분석 모델을 보여주고 있다.

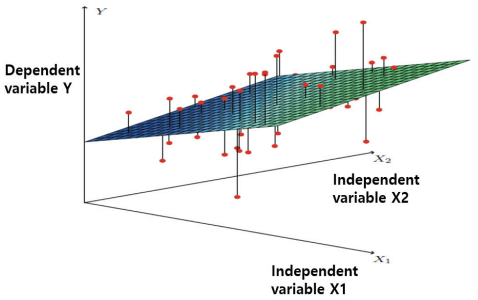

그림 1. X1, X2 독립변수에 대한 다중회귀분석
Fig. 1 Multiple regression analysis (X1, X2)

2.2 ARIMA (Autoregressive integrated moving average)

ARIMA 모형은 과거의 관측 값과 오차를 사용해서 현재의 시계열 값을 설명하는 ARMA 모델을 일반화 한 것으로 분기/반기/연간 단위로 다음 지표를 예측한다거나 주간/월간 단위로 지표를 리뷰하며 트렌드에 이상치가 없는지를 모니터링 하는 데 사용되는 분석 기법이다[7-9]. 태양광, 풍력 등의 신재생 에너지 발전량 예측을 위해서 ARIMA 모형이 활발하게 사용되었다.

2.3 랜덤 포레스트(Random Forest)

랜덤 포레스트는 다수의 결정트리(Decision Tree)

로부터 예측된 값의 평균 또는 가중 평균을 출력하는 앙상블 기법들 중 하나이다[10].

본 연구를 진행하면서 신재생 에너지 발전량을 예측하는 머신러닝 학습 모델로 랜덤 포레스트 기법을 사용하여 적용하였다. 랜덤 포레스트는 1) 기존 배깅의 이점을 살리고 2) 변수를 랜덤으로 선택하는 과정을 추가함으로써 결정 트리들의 상관성을 줄여서 예측력을 향상한 앙상블 모형이다. 이때 배깅은 트리를 만들 때 training set의 부분집합을 활용하여 형성하는 것을 말한다.

1) 배깅을 사용하는 앙상블 모형

훈련 과정에서 구성한 다수의 결정 트리로부터 분류 또는 평균 예측치(회귀 분석)을 출력함으로써 동작한다. 배깅의 장점인 편의(Bias)를 유지하면서 분산을 낮추는 효과를 가져온다.

그림 2는 부트스트랩을 통해 resampling한 데이터로 학습된 트리들을 집계하는 방법론인 랜덤포레스트 배깅 과정이다.

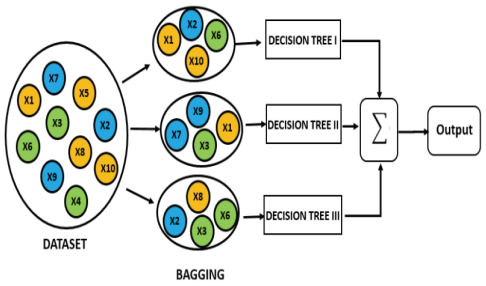

그림 2. 랜덤 포레스트에서 배깅 과정
Fig. 2 Bagging process in random forest

2) 변수 선택하는 과정을 통해 결정 트리들의 상관성을 줄여 예측력 향상

랜덤포레스트에 주요한 파라미터는 랜덤 포레스트를 구성하는 결정 트리의 개수와 트리의 최대 깊이가 있다. 결정 트리의 개수를 늘리면 연산량이 늘어나서 속도가 느려지지만, 주어진 데이터에 대한 과적합을 피할 수 있다. 한편 트리의 최대 깊이를 줄이면 데이터에 대한 과소적합(underfitting)이 발생하게 된다. 따라서 본 실험에서는 깊이의 튜닝을 통해 가장 적절한 값을 찾아내는 것이 중요하다.

2.4 주성분 분석

기상 빅데이터는 다수의 독립변수로 되어 있다. 독립변수가 많아질수록 예측 신뢰도가 떨어지고 과적합이 발생하고 개별 피처(feature)간의 상관관계가 높을 가능성이 있다. 본 연구에서는 주성분 분석을 기반으로 데이터 전처리 과정을 진행하였다.

주성분 분석은 고차원의 데이터를 저차원의 데이터로 축소하는 차원 축소 알고리즘이다. 기상 빅데이터는 훈련 데이터의 피처가 많다. 그렇지만 모든 피처가 결과에 주요한 영향을 미치는 것은 아니다. 영향을 가장 많이 미치는 피처가 존재할 것이며 그중에는 영향을 미치는 정도가 미미한 피처가 있을 것이다. 이런 피처들 중 가장 영향을 많이 미치는 피처들만 추출하여 사용하는 것이 PCA이다[11].

본 실험에서는 다음과 같은 단계로 전처리 과정이 진행된다.

1) 주성분 추출

표준 행렬 분해 기술인 SVD(Singular Value Decomposition)을 이용해 훈련데이터를 행렬의 점 곱으로 분해한다. 이때 찾고자하는 모든 주성분은 V에 담겨 있다.

$$SVD = U \cdot \Sigma \cdot V^T \quad (1)$$

2) d차원으로 투영

d개의 주성분으로 정의한 초평면에 투영하여 데이터셋의 차원을 d차원으로 축소한다. 이때의 초평면은 분산을 가능한 최대로 보존한 투영이다. 분산을 최대로 보존할 수 있는 축을 선택하는 것이 정보를 가장 손실을 적게 할 수 있다. 분산이 커야 데이터들 사이의 차이점이 명확해질 테고 결국 본 실험의 모델을 더욱 좋은 방향으로 만들 수 있다.

초평면에 훈련 세트를 투영하기 위해서는 행렬 X와 d개의 주성분을 담은 행렬 Wd를 점 곱한다.

$$X_{d-proj} = X \cdot W_d \quad (2)$$

그림 3에서 보는 것처럼 PCA는 데이터의 분산이 최대가 되는 축을 찾게 된다. 기상 빅데이터의 모든 차원을 살리면서 차원을 축소 할 수는 없다. 모든 특

성을 살릴 수는 없지만 최대한 특성을 살리며 차원을 낮춰주는 방법을 사용한다.

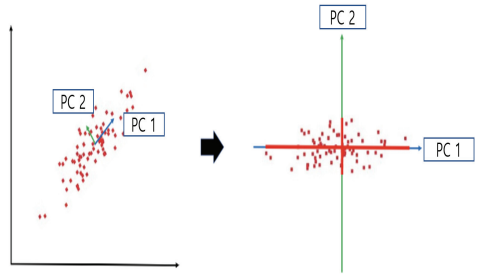

그림 3. 주성분 벡터로의 투영
Fig. 3 Projection into PCV

III. 실험 환경 및 결과

기상 빅데이터는 공공데이터 포털 사이트에서 데이터를 수집하고 신재생 에너지 발전량 데이터를 수집하여 실험을 진행하였다.

표 1. 기상 빅데이터
Table 1. The meteorological Big Data

wind speed	air pressure	temp	water temp	wave height
8.0	1026.7	7.4	16.5	2.0
7.7	1016.5	15.7	16.7	1.7
11.1	1016.0	11.9	15.3	3.5
1.8	1015.5	14.0	16.0	0.7
5.0	1008.0	17.5	16.5	2.1
3.0	1008.7	21.3	21.5	1.9
5.0	1006.4	21.5	22.5	1.4
5.3	1010.5	28.7	28.8	1.6
4.2	1012.2	24.5	25.2	0.8
…	…	…	…	…

1단계 데이터셋은 제주지역 2020년1월~2022년 7월까지(2년7개월) 기상 빅데이터 정보를 수집하였으며 데이터 피처는 5개를 선택 수집하였다.

2단계 신재생 에너지 발전량 데이터는 제주지역 풍력발전량으로 2020년1월~2022년 7월까지(2년7개월) 데이터를 기반으로 하였다.

표 2. 신재생 에너지 풍력 발전량
Table 2. Renewable energy wind power

Date	Generation
20200115	1714560
20200215	782487
20200315	3162054
20210115	155418
20210215	2243926
20210315	2571625
20220115	416096
20220215	4175214
20220315	1237441
…	…

표 1의 훈련 데이터셋 피처 중 Scree Plot 그래프를 통해 고윳값 크기를 기반으로 영향을 미치는 차원의 수를 고려하였다.

3.1 데이터셋 전처리

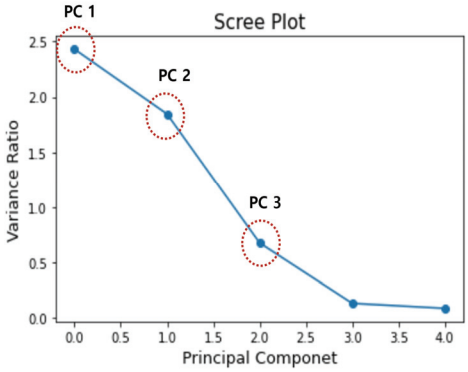

그림 4. 주성분 분석 Scree Plot
Fig. 4 PCA Scree Plot

주성분 분석 후 주성분 수를 선정하기 위해 고윳값-주성분의 분산 변화를 보는 그래프를 모니터링 하였다. 그림 4의 분석을 통해 고윳값 변화율이 완만해지는 부분을 필요한 주성분의 수로 결정할 수 있다. 기상 빅데이터를 몇 퍼센트 정보를 설명하고자 하는 지에 따라 선택할 수 있다. 본 실험에서는 주성분 수 3개를 선택할 때 아래와 같은 결과를 얻을 수 있었다.

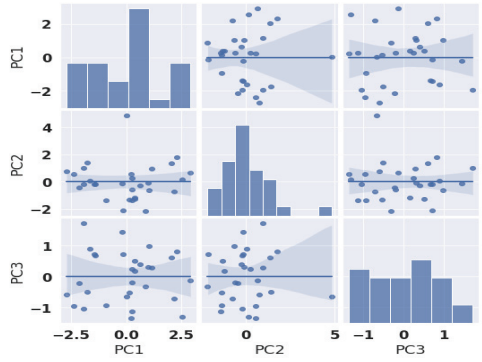

그림 5. 기상 빅데이터 변수 간 분포
Fig. 5 Distribution between meteorological Big Data

그림5에서는 기상 빅데이터를 전처리하여 신재생 에너지 발전량에 영향을 미치는 중요 피처를 이용하여 주성분 분석은 데이터의 분산이 최대가 되는 축을 찾게 된다. 전처리 과정을 거친 데이터셋을 머신러닝 학습 모형에 적용함으로써 신재생 에너지 발전량을 예측한다.

3.2 에너지 발전량 예측

신재생 에너지 풍력 발전량을 예측하기 위한 데이터로 제주지역 2020년 1월부터 2022년 7월까지 발전량 데이터를 수집하였으며 그림6에서는 연도에 따른 월별 데이터(2년7개월) 최대 풍력 발전량에 대한 실측 데이터를 보여주고 있다. 실험은 파이썬 기반으로 실제 데이터를 제안한 예측 모형에 적용하여 진행하였다.

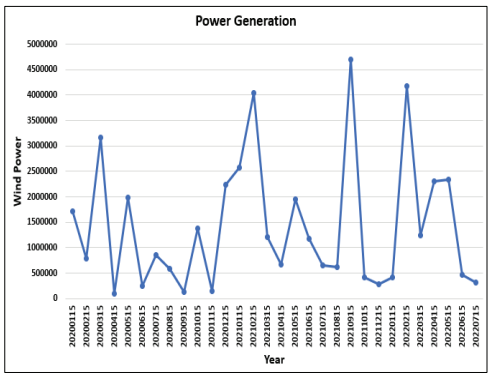

그림 6. 연도별 신재생 풍력 발전량
Fig. 6 Renewable Wind Power Generation by Year

신재생 에너지 풍력 발전량 예측의 정확성은 에너지 수요 예측 오차율을 통해서도 확인할 수 있다.

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

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

본 연구에서는 신재생 에너지 발전량을 예측하는 머신러닝 학습 모델로 랜덤 포레스트 기법을 사용하여 적용하였다.

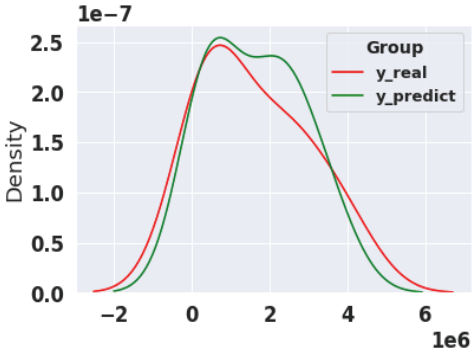

그림 7. 제안한 모형 적용한 예측 결과 비교
Fig. 7 Comparison of prediction results

Ⅳ. 결 론

공공데이터에서 기상 빅데이터를 일차적으로 수집하여 데이터를 정제하였다. 제주지역 풍력 발전량 데이터를 수집하여 이차적으로 데이터에 대한 정제화 과정을 거쳤다.

본 연구에서는 수집한 기상 빅데이터를 주성분 분석 알고리즘을 사용하여 영향도가 높은 피처들을 추출하여 머신러닝 모형에 적용하여 신재생 에너지 풍력 발전량을 예측하였다.

본 논문에서는 랜덤포레스트를 이용한 신재생 에너지 발전량 예측 모형을 제안하였다. 차원 축소된 피처 데이터를 이용하여 머신러닝 모델을 적용할 경우 발전량을 예측하는 데 있어 적은 수의 피처만으로 특정 현상을 확인하고 모델 성능 향상에 기여함을 확인할 수 있었다.

우리나라는 신재생 에너지 보급에 대한 정책 방향은 수립되었으나, 발전량 예측을 위한 계량 데이터 취득시스템이 구축 되어 있지 않은 경우가 더 많은 현실이다. 향후 신재생 에너지와 상관관계가 높은 다른 변수들을 고려하여 데이터 수집 및 빅데이터 분석에 따른 발전량 예측 기술 향상을 통한 연구를 진행하고자 한다. 또한 다양한 지역을 가지고 제안한 모형을 통해 범용성 여부를 확인하고자 한다.

본 제(결과물)는 2022년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다.(2021RIS-002)

This results was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-002)

References

[1] J. Lee, "Optimal Operation of Energy Storage Devices based on Artificial Intelligence," *Journal of the Korean Soloar Energy Society*, vol. 42, no. 1, Feb. 2022, pp. 155-175.

[2] J. T. Dellosa and E. C. Palconit, "Artificial Intelligence(AI) in Renewable Energy Systems: A Condensed Review of its Applications and Techniques," *IEEE International Conference(EEEIC/I&CPS Europe)*, Bari, Italy, Sept. 2021.

[3] J. Choi and H. Choi, "Prediction of Wind Power Generation using Deep Learning," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 16, no. 02, Apr. 2021, pp. 329-338.

[4] S. Lee, S. Jung and J. Koh, "Recurrent Neural Network based Prediction System of Agricultural Photovoltaic Power Generation," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 17, no. 05, Oct. 2022, pp. 825-832.

[5] M. Chaikumbung, "Institutions and consumer preferences for renewable energy: A meta-regression analysis," *Renewable and Sustainable Energy Reviews*, vol. 146, Aug. 2021, pp. 1-24.

[6] J. Lee and I. Lee, "Trends in ICT-based renewable energy generation prediction technology," *The Journal of Korean Institute of Communications and Information Sciences*, vol. 36, no. 11, Oct. 2019, pp. 3-8.

[7] D. Shin and C. Kim, "Short Term Forecast Model for Solar Power Generation using RNN-LSTM," *Journal of Advanced Navigation Technology*, vol. 22 no. 3, 2018, pp. 233-239.

[8] L. Martin, L. F. Zarzalejo, J. Polo, A. Navarro, R. Marchante and M. Cony, "Prediction of global solar irradiance based on time series analysis: Application to solar thermal power plants energy production panning," *Solar Energy*, vol. 84, no. 10, 2010, pp. 1772-1781.

[9] S. Park and S. Kim, "A study on short-term wind power forecasting using time series models," *The Korean Journal of Applied Statistics*, vol. 29, no. 7, 2016, pp. 1373-1383.

[10] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, 2001, pp. 5-32.

[11] B. M. S. Hasan and A. M. Abdulazeez, "A Review of Principal Component Analysis Algorithm for Dimensionality Reduction," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, 2021, pp. 20-30.

저자 소개

강미영(Mi-Young Kang)

2003년 전남대학교 대학원 정보통신공학과 졸업(공학석사)
2008년 전남대학교 대학원 컴퓨터정보통신공학과 졸업(공학박사)
2008년~2010년 전남대학교 PostDoc.
2021년~현재 호남대학교 정보통신공학과 교수
2021년~현재 광주시 스마트도시사업협의회 자문위원
※ 관심분야 : 임베디드 시스템, 인공지능, 신재생 에너지