조건부 랜덤 포레스트 기반의 설명 가능한 일사량 예측

문지훈, 황인준
고려대학교 전기전자공학과
johnny89@korea.ac.kr, ehwang04@korea.ac.kr

Explainable Solar Irradiation Forecasting Based on Conditional Random Forests

Jihoon Moon, Eenjun Hwang
School of Electrical Engineering, Korea University

요 약

태양광 발전은 이산화탄소 배출로 인한 기후 변화에 대응하는 주요 수단으로 인식되어 수요와 필요성이 급격하게 증가하고 있다. 최적의 태양광 발전 시스템의 운영을 위해서는 정교한 전력수요 및 태양광 발전량 예측 모델이 요구되며, 온도 및 일사량은 태양광 발전량 예측 모델의 필수적인 입력 변수이다. 하지만, 한국 기상청의 동네예보는 일사량에 관한 예측값을 제공하지 않아 정교한 태양광 발전량 예측 모델을 구축하는 것은 어렵다. 이를 위해 일사량 예측 기법에 관한 많은 연구 사례가 보고되고 있지만, 다수의 연구들은 충분한 데이터 셋을 이용하여 일사량 예측 모델을 개발하였다. 초기 태양광 발전 시스템 운영을 위해서는 불충분한 데이터 셋을 이용한 예측 모델 개발이 필요하나 이에 대한 사례는 불충분하다. 본 논문은 실제 태양광 발전 시스템에서 수집된 불충분한 데이터 셋을 이용한 단기 일사량 예측 기법을 제안한다. 먼저, 기상청 동네예보의 다양한 기상 요인들을 이용하여 일사량 예측 모델을 위한 입력 변수를 구성한다. 다음으로, 조건부 랜덤 포레스트를 이용하여 일사량 예측 모델을 구성하며, 설명 가능한 일사량 예측뿐만 아니라 더욱더 많은 데이터 셋을 학습하기 위해 시계열 교차검증을 수행한다. 실험 결과, 제안한 기법은 다른 예측 기법들보다 높은 예측 정확도를 보일 뿐만 아니라 설명 가능한 예측 결과를 제시할 수 있음을 보여준다.

1. 서 론

최근 기후 변화 및 에너지 부족 문제를 대비하기 위해 신재생 에너지(Renewable Energy)를 적극 활용한 스마트 그리드 기술의 관심이 커지고 있다[1]. 스마트 그리드(Smart Grid)는 정보통신기술(ICT: Information and Communication Technologies)을 기존의 전력망과 접목하여 에너지 효율을 최적화하는 기술이다[1,2]. 신재생 에너지는 스마트 그리드의 핵심 요소 중 하나이며, 태양광(PV: Photovoltaics), 풍력 등과 같은 천연 자원을 통해 목적에 따라 전기 생산이 가능하다[3]. 태양광 발전은 공간 제약 없이 설치할 수 있는 장점이 있어, 이와 관련된 기술이 빠르게 발전하고 있다[1,3].

태양광 발전 시스템은 다양한 기상 요인으로 인해 발전에 크게 영향을 받으며, 일사량(Solar Irradiation)은 태양광 발전의 중요한 요인이다[4]. 그러나 기상청의 동네예보는 기온, 습도 등과 같은 기상 요인의 예측값은 제공하지만 일사량의 예측값은 제공하지 않는다[5]. 따라서, 국내 태양광 발전 시스템의 운영을 위해서는 정확한 일사량 예측 모델이 필요하다. 그리하여, 인공지능(AI: Artificial Intelligence) 기술 기반의 일사량 예측 기법에 관한 많은 연구가 수행되었다[3-5].

다수의 연구들은 일정 기간 이상의 충분한 데이터 셋을 사용하여 예측 모델 구성 및 예측 성능 평가를 수행하였으나, 불충분한 데이터 셋을 사용하여 인공지능 기술을 기반으로 일사량을 예측한 사례는 미미하다. 따라서, 초기 태양광 발전 시스템의 효율적인 운영을 위해서는 불충분한 데이터 셋을 통해 정교한 예측 모델을 구성할 필요가 있다. 의사결정나무 기반 알고리즘들은 작은 데이터 셋에서도 만족스러운 예측 성능이 가능하다[6,7].

본 논문은 불충분한 데이터 셋을 이용하여 조건부 랜덤 포레스트(CRF: Conditional Random Forests) 기반의 일사량 예측 기법을 제안하고, 예측 성능을 다중선형 회귀(MLR: Multiple Linear Regression) 및 다양한 의사결정나무 기반의 알고리즘들과 비교한다. 본 논문의 주요 기여도는 아래와 같다.

- 조건부 랜덤 포레스트를 기반으로 예측 모델을 구축하여 다중 단기 일사량 예측을 수행한다.
- 국내 태양광 발전 시스템의 적용 가능성을 위해

기상청의 동네예보에 포함된 7 가지 기상 요인을 모두 고려한다.

- 예측 모델의 변수 중요도(Variable Importance)와 시계열 교차검증을 이용하여 예측값을 도출하는 과정을 설명한다.

본 논문의 나머지 부분은 아래와 같다. 2 장에서는 일사량 예측 모델을 위한 입력 변수 및 모델 구성에 대해 자세히 기술한다. 3 장에서는 예측 모델의 예측 성능을 비교 및 평가하기 위한 실험 과정을 기술하고 이를 논의한다. 4 장에서는 결론과 향후 연구 방향을 제시함으로 본 논문의 끝을 맺는다.

2. 일사량 예측 모델 구성

불충분한 데이터 셋을 이용하여 예측 모델을 구성하기 위해, 실제 대전에 있는 태양광 발전 시스템의 일사량 데이터를 수집하였다. 수집된 데이터 기간은 2018 년 6 월로 1 달 치의 데이터이며, 오전 6 시부터 저녁 8 시까지의 실측값으로 구성되어 있다. 표 1 에 수집된 데이터의 통계적 분석 결과를 나타내었다.

<표 1> 수집된 데이터의 통계적 분석 결과(단위: W/m^2)

기술 통계법	통계량 값
평균	330.07
표준 오차	12.93
중앙값	254.50
표준 편차	274.35
범위	805
최소값	0
최대값	805
합	148529.40
관측수	450

2.1. 입력 변수 구성

예측 모델을 위한 입력 변수를 구성하기 위해 기상청의 동네예보에서 제공하는 강수형태, 습도, 강수량, 하늘상태, 기온, 풍향, 풍속의 실측값을 기상자료개방포털에서 수집하였다. 여기서, 강수형태와 하늘상태는 범주형 데이터로 이루어지며, 강수형태는 비가 왔을 때에는 1, 그렇지 않으면 0 인 명목형 데이터로 구성되어 있다. 하늘상태는 맑음, 구름 조금, 구름 많음, 흐림을 각 1 부터 4 까지로 표기한 순서형 데이터로 구성되어 있다. 나머지 실측값인 습도, 강수량, 기온, 풍향, 풍속은 연속형 데이터의 특징을 가지고 있다.

시간 정보를 반영하기 위해, 6 시부터 20 시까지 총 15 시간 간격에 대해 명목 척도로 데이터 셋을 구성하였다. 이는 아침과 저녁에는 일사량이 적고 오후 시간대에는 일사량이 많은 일사량의 특정 시간대의 특징을 더욱 효과적으로 반영할 수 있다.

이뿐만 아니라, 과거 일사량 패턴 및 추세를 반영하기 위해, 예측 시점에서 과거 2 일의 일사량과 습도, 강수량, 하늘상태, 기온, 풍속, 풍향으로 총 14 개의 입력 변수를 구성하였다. 본 논문에서 고려한 입력 변수는 총 36 개이며 표 2 에 기술하였다.

<표 2> 예측 모델의 입력 변수 구성 및 특징

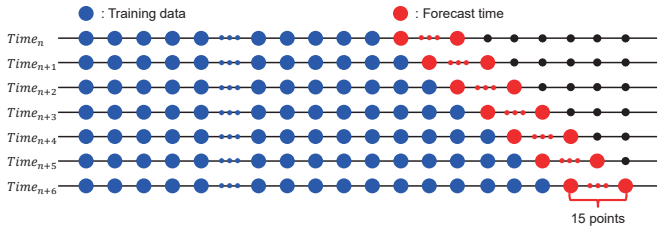
IV #	입력 변수 (특징)	IV #	입력 변수 (특징)
IV01	6 시 (명목형)	IV19	2 일 전 기온 (연속형)
IV02	7 시 (명목형)	IV20	2 일 전 풍향 (연속형)
IV03	8 시 (명목형)	IV21	2 일 전 풍속 (연속형)
IV04	9 시 (명목형)	IV22	2 일 전 일사량 (연속형)
IV05	10 시 (명목형)	IV23	1 일 전 습도 (연속형)
IV06	11 시 (명목형)	IV24	1 일 전 강수량 (연속형)
IV07	12 시 (명목형)	IV25	1 일 전 하늘상태 (순서형)
IV08	13 시 (명목형)	IV26	1 일 전 기온 (연속형)
IV09	14 시 (명목형)	IV27	1 일 전 풍향 (연속형)
IV10	15 시 (명목형)	IV28	1 일 전 풍속 (연속형)
IV11	16 시 (명목형)	IV29	1 일 전 일사량 (연속형)
IV12	17 시 (명목형)	IV30	강수형태 (명목형)
IV13	18 시 (명목형)	IV31	습도 (연속형)
IV14	19 시 (명목형)	IV32	강수량 (연속형)
IV15	20 시 (명목형)	IV33	하늘상태 (순서형)
IV16	2 일 전 습도 (연속형)	IV34	기온 (연속형)
IV17	2 일 전 강수량 (연속형)	IV35	풍향 (연속형)
IV18	2 일 전 하늘상태 (순서형)	IV36	풍속 (연속형)

2.2 예측 모델 구성

본 연구는 랜덤 포레스트와 유사하지만 다른 접근 방식을 갖는 조건부 랜덤 포레스트를 이용하여 단기 일사량 예측 모델을 구성한다. 이러한 이유로 조건부 랜덤 포레스트의 나무 구조는 랜덤 포레스트의 나무 구조보다 데이터를 편향적으로 학습하지 않으므로, 평가 집합(Test Set)에서 출력 변수를 예측할 때, 훈련 집합(Training Set)에서 모델 학습의 과적합(Overfitting) 문제 해결에 더 적합하기 때문이다[6]. 이뿐만 아니라 예측값을 평균화하는 랜덤 포레스트의 나무 구조와 달리, 조건부 랜덤 포레스트는 입력 변수의 가중치를 평균화하여 예측값을 도출하기 때문에 변수 중요도를 나타내었을 때, 더욱 효과적으로 모델 구조에 관해 설명이 가능하다. 본 논문에서는 적은 데이터 셋을

다루므로, 많은 데이터 셋을 요구하는 심층 신경망(DNN: Deep Neural Network)이나 Boosting 계열의 알고리즘들(예: XGBoost, LightGBM)을 고려하지 않았다[7].

또한, 본 연구는 점차 많은 데이터 셋을 학습하기 위해 시계열 교차검증(TSCV: Time Series Cross-Validation)을 적용한다. 시계열 교차검증은 그림 1 과 같이 각 예측 시점에서 예측 모델의 훈련 집합은 첫 시점부터 이전의 관측 시점까지 구성된다. 그리하여 각 예측 시점에 구성된 예측 모델은 1 시점 뒤부터 15 시점 뒤까지 다중 시점의 일사량을 예측하며, 각 시점의 예측 정확도를 계산하고 이를 평균값을 계산하여 예측 모델의 성능을 평가한다.



(그림 1) 다중 시점 예측을 위한 시계열 교차검증

3. 실험 및 평가

본 논문에서 제안한 예측 모델의 성능을 평가하기 위해, 전체 데이터 셋을 일자별로 분리하였으며, 6 월 1 일부터 2 일은 입력 변수를 위해 데이터를 이용하며, 3 일부터 23 일 총 3 주의 기간은 훈련 집합으로 24 일부터 30 일 총 1 주는 평가 집합으로 선정하여 실험을 진행하였다. 예측 기법으로 조건부 랜덤 포레스트와 예측 성능을 비교하기 위해, 다중선형회귀, 의사결정나무(DT: Decision Tree), GBM(Gradient Boosting Machine), 랜덤 포레스트(RF: Random Forest)로 총 4 가지의 기법들을 이용하였다. 실험 환경은 R 3.5.1 버전의 RStudio 1.1453 버전에서 진행하였다. Grid Search 를 통해 각 예측 기법에 관한 최적의 초매개변수(Hyperparameter) 값을 선정하였으며, 이는 표 3 과 같다.

<표 3> 각 예측 기법에서 선정된 초매개변수의 값

예측 기법	패키지	초매개변수의 값
GBM	*gbm*	• *distribution*: gaussian • *shrinkage*: 0.001 • *interaction.depth*: 5 • *bag.fraction*: 0.5 • *n.trees*: 3000 • *cv.folds*: 5
RF	*randomForest*	• *mtry*: 12 • *ntree*: 128
CRF	*party*	• *mtry*: 6 • *ntree*: 500

예측 모델의 예측 정확도를 평가하기 위해, 제곱근평균제곱오차(RMSE: Root Mean Square Error) 및 평균절대오차(MAE: Mean Absolute Error)를 사용하였으며, 이는 식 1, 2 와 같다. A^t 와 F^t 는 실제 관측된 일사량과 일사량 예측값을 나타내며, n 은 관측치의 수이다.

$$RMSE = \sqrt{\sum (F^t - A^t)^2 / n} \quad (1)$$

$$MAE = 1/n \times \sum |F^t - A^t| \quad (2)$$

표 4 와 5 는 각 예측 시점에 관한 예측 모델들의 RMSE 와 MAE 의 결과이다. 표에서 붉게(Red) 표기된 것은 낮은 예측 성능을 나타내며 푸르게(Blue) 표기된 것은 우수한 예측 성능을 나타낸다. 아래의 표에서 나타낸 것과 같이, 현재 시점과 예측 시점의 간격이 멀어질수록 예측 성능이 저하된다는 것을 확인할 수 있다. 또한, 조건부 랜덤 포레스트는 다른 예측 기법보다 더욱 우수한 예측 성능을 보인다는 것을 확인할 수 있었다.

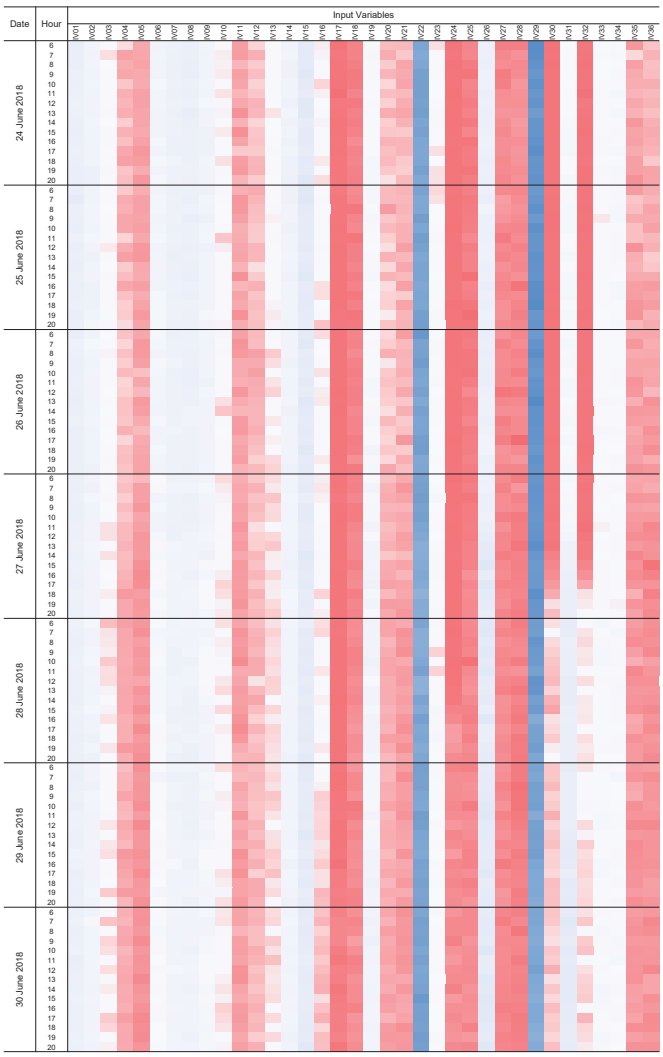
<표 4> 각 예측 시점에서 예측 모델들의 RMSE

예측 시점	MLR	DT	GBM	RF	CRF
1	199.56	189.08	147.68	169.40	138.77
2	227.03	195.91	160.56	174.92	153.06
3	250.22	207.50	168.36	178.94	163.00
4	265.22	214.04	173.87	182.27	165.88
5	300.24	221.01	177.73	183.52	166.68
6	306.06	230.33	179.54	183.64	170.06
7	318.77	232.26	181.25	183.21	171.35
8	321.69	229.44	182.22	184.24	171.60
9	323.54	231.31	183.02	185.12	169.67
10	327.85	231.13	182.91	185.64	171.43
11	336.13	226.83	182.85	186.79	171.05
12	335.62	221.47	182.48	186.10	171.01
13	336.86	216.80	182.86	185.64	169.37
14	338.19	210.27	182.79	185.73	169.13
15	338.87	201.23	183.03	186.10	169.21

<표 5> 각 예측 시점에서 예측 모델들의 MAE

예측 시점	MLR	DT	GBM	RF	CRF
1	140.54	127.58	106.27	125.20	100.52
2	154.92	134.67	114.79	129.16	110.15
3	167.05	141.21	119.83	132.11	118.38
4	175.09	145.18	123.58	133.02	119.93
5	188.80	148.93	125.83	133.94	120.51
6	196.21	154.29	126.05	134.08	122.53
7	205.31	156.23	126.66	133.10	122.15
8	209.63	153.40	126.96	133.89	121.71
9	210.26	155.30	127.28	134.59	120.23
10	212.71	155.57	127.40	134.74	122.07
11	217.35	151.59	127.19	135.25	121.61
12	218.11	146.59	127.15	135.36	120.44
13	219.39	142.56	127.35	134.71	119.67
14	219.95	136.87	126.84	134.96	120.35
15	220.54	132.34	127.02	134.99	118.54

그림 2 는 단기 일사량 예측 모델을 구성하기 위해, 조건부 랜덤 포레스트를 시계열 교차검증을 이용해 학습 시켜 구성된 예측 모델의 변수 중요도를 히트맵 그래프를 통해 시간대별로 나타낸 것이다. 그림에서 붉게(Red) 표기된 것은 낮은 변수 중요도를 나타내며, 푸르게(Blue) 표기된 것은 높은 변수 중요도를 나타낼 뿐만 아니라 주요 변수라고 판단할 수 있다.



(그림 2) 각 예측 시점에서 예측 모델의 변수 중요도

그림 2 에서 확인할 수 있듯이, 과거 2 일의 일사량 관측치는 예측 모델에서 매우 중요한 입력 변수이며, 실제 단기 일사량을 예측할 때 주요 요인이라는 것을 알 수 있다. 또한, 2018 년 6 월 27 일 이후로 강수형태 및 강수량의 변수 중요도가 높아졌다는 것을 확인할 수 있다. 이는 2018 년 6 월 1 일부터 27 일 오후까지 비가 오지 않아 값이 0 으로 측정되어 예측 모델을 학습할 때 중요한 변수라는 것을 판단하지 못했으나, 비가 온 시점부터는 비로 인해 일사량이 없다는 것을 예측 모델 학습에 인지하고 이에 관한 가중치를 높임으로 변수 중요도가 높아졌다는 것을 확인하였다. 그 외에도 하늘상태와 기온은 일사량 예측 모델의 주요 변수라는 것을 확인하였다.

4. 결 론

본 논문은 불충분한 데이터 셋에서 정확한 일사량 예측을 수행하기 위해, 조건부 랜덤 포레스트 기반의 다중 단기 일사량 예측 기법을 제안하였다. 태양광 발전 시스템에 적용 가능성을 높이기 위해, 기상청의 동네예보에서 제공하는 정보를 이용하여 예측 모델의 입력 변수를 구성하였다. 다음으로 적은 데이터 셋에서도 효과적으로 모델을 학습할 수 있는 조건부 랜덤 포레스트를 이용하여 예측 모델을 학습하고, 더욱더 많은 데이터의 학습과 최근 일사량 패턴 및 추세를 반영하기 위해 시계열 교차검증을 적용하였다. 예측 모델은 점 예측 방식이 아닌 다중 예측 방식으로 1 시점 뒤 시점부터 15 시점 뒤 시점까지 총 15 시점을 예측하여 예측 불확실성을 대비하는 데 도움을 줄 수 있었다. 제안한 예측 기법은 다양한 예측 기법들과 비교하여 더욱 우수한 예측 성능을 보였으며, 변수 중요도를 통해 과거 일사량과 기온, 하늘상태 등이 향후 일사량을 예측할 때 주요 변수라는 것을 확인할 수 있었다.

본 논문에서는 실제 수집된 태양광 발전 시스템의 일사량 데이터가 한 달 치만 수집되어 다양한 환경의 일사량 데이터를 통해 실험을 진행하기가 어려웠다. 향후, 일사량 데이터를 수집하여 다양한 기간, 지역 등을 고려하여 범용성을 가질 수 있는 일사량 예측 모델을 개발할 계획이다.

사사문구

이 연구는 2019 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단-에너지클라우드기술개발사업 (No. 2019M3F2A1073184) 및 한국전력공사의 2018 년 착수 에너지 거점대학 클러스터 사업(No. R18XA05)의 지원을 받아 수행된 연구임.

참고문헌

[1] X. Huang, J. Shi, B. Gao, Y. Tai, Z. Chen, and J. Zhang, “Forecasting Hourly Solar Irradiance Using Hybrid Wavelet Transformation and Elman Model in Smart Grid,” *IEEE Access,* Vol. 7, pp. 139909-139923, 2019.

[2] J. Kim, J. Moon, E. Hwang, and P. Kang, “Recurrent inception convolution neural network for multi short-term load forecasting,” *Energy and Buildings,* Vol. 194, pp. 328-341, 2019.

[3] M. Paulescu and E. Paulescu, “Short-term forecasting of solar irradiance,” *Renewable Energy,* Vol. 143, pp. 985-994, 2019.

[4] Y. Kwon, A. Kwasinski, and A. Kwasinski, “Solar Irradiance Forecast Using Naïve Bayes Classifier Based on Publicly Available Weather Forecasting Variables,” *Energies,* Vol. 12, p. 1529, 2019.

[5] S. Jung, J. Moon, S. Park, and E. Hwang, “A Probabilistic Short-Term Solar Radiation Prediction Scheme Based on Attention Mechanism for Smart Island,” *KIISE Transactions on Computing Practices,* Vol. 25, No. 12, pp. 602-609, 2019.

[6] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, “Bias in random forest variable importance measures: Illustrations, sources and a solution,” *BMC Bioinformatics,* Vol. 8, p. 25, 2007.

[7] J. Moon, J. Kim, P. Kang, and E. Hwang, “Solving the Cold-Start Problem in Short-Term Load Forecasting Using Tree-Based Methods,” *Energies,* Vol. 13, p. 886, 2020.