랜덤 포레스트를 이용한 태양광 발전량 예측

이웅희 [1], 김영훈 [2]
한양대학교 컴퓨터공학과
{woongheelee[1], nongaussian[2]}@hanyang.ac.kr

Predicting Photovoltaic Power Generation with Random Forests

Woonghee Lee[1], Younghoon Kim[2]
Dept. of Computer Science and Engineering, Hayang University

요 약

태양광 발전 방식은 기존 고갈 가능성이 있는 에너지를 대체하기 위해 많은 개발이 이루어져왔다. 태양광 발전 모듈의 인버터에는 발전량에 영향을 주는 다양한 속성들이 계측되어 저장된다. 본 연구에서는 이런 데이터에, 발전량에 영향을 주는 외부 요인인 기상 데이터를 추가하고, 랜덤 포레스트를 써서 과거 몇일까지의 데이터를 고려했을 때 가장 예측 성능이 높은지 실험을 통해 검증하였다. 2 일 전부터 최대 365 일 전까지의 데이터를 고려한 결과 5 일 정도의 과거 데이터를 고려했을 때 예측 성능이 가장 높고, 고려하는 기간이 길어질수록 예측 성능이 떨어지는 경향을 보였다.

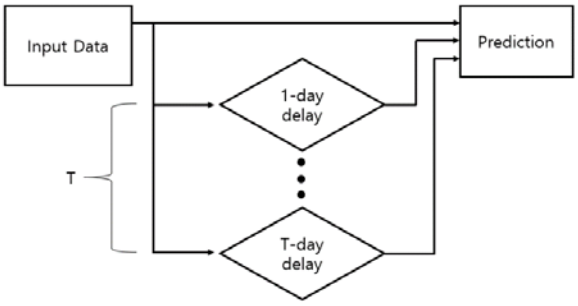
1. 서론

태양광 발전 에너지는 기존의 고갈 가능성이 있는 석유, 석탄과 같은 에너지 발전 방식과 사고 위험 가능성이 있는 원자력 에너지 발전 방식을 대체 하기 위해 많은 연구가 이루어져 왔다. 태양광 발전은 발전기의 도움 없이 태양전지를 이용하여 태양빛을 직접 전기 에너지로 변환시키는 발전 방식으로 광전 변환기를 써서 태양빛을 전기 에너지로 변환 시킨다.

태양광 발전에 쓰이는 인버터에는 국내의 법적 제도에 의해 여러 계측 정보를 한곳으로 모아 놓도록 되어있다. 수집 되는 요소는 인버터의 전력, 전압뿐 아니라, 일사량, 부품 온도, 고장 내용 등 다양한 데이터가 수집되고 있다. 태양광 발전은 기상 상태, 지리학적 위치와 같은 여러 변수에 따라 발전량 변동이 발생하는 데[1], 인버터에 수집 된 데이터 분석을 통해 발전량을 예측하는 연구가 다수 진행되어오고 있다[2, 3].

발전량 예측에는 발전량 예측 기간을 기준으로 크게 세 가지 예측 방식의 연구가 있다. 첫째, 현재 예측으로 현재 시각에서부터 앞으로 약 6 시간 후까지의 발전량을 예측하는 방법이다. 둘째, 단기 예측으로 7 일 이후의 발전량을 예측하는 방법이다. 셋째, 장기 예측으로 월 단위, 연 단위 발전량을 예측에 관한 방법이다. 본 연구에서는 10 분에서 15 분 단위의 초 단기 데이터를 일 단위로 전처리 하여 날짜를 늘려가며, 랜덤 포레스트(Random Forest)를 이용하여 발전량 예측에 적절한 일별 기간 데이터의 크기를 실험을 통해 검증하고자 한다.

2. 랜덤 포레스트



(그림 1) 트리 예측과정의 데이터 흐름도

L. Breiman[4]에 의해 랜덤 포레스트의 구체적 아이디어가 제안된 이후 여러 가지 분야에서 응용되어왔다. 랜덤 포레스트는 다수의 결정트리(Decision Tree)로부터 예측된 값의 평균 또는 가중 평균을 출력하는 앙상블 학습의 한 방법이다. 랜덤 포레스트는 결정트리의 계층적 특성에서 발생 가능한 에러의 전파 가능성을 줄여주고, 다수의 결정 트리를 사용하여 과학습(overfitting) 을 방지하여 예측의 정확도를 높인다.

랜덤 포레스트에 주요한 파라미터는 랜덤 포레스트를 구성하는 결정 트리의 개수와 트리의 최대 깊이가 있다. 결정 트리의 개수를 늘리면 연산량이 늘어나서 속도가 느려지는 반면, 주어진 데이터에 대한 과학습을 피할 수 있다. 한편, 트리의 최대 깊이가 작으면 데이터에 대한 과소학습(underfitting)이 발생하고, 허용 깊이가 너무 깊으면, 과학습이 발생한다. 따라서, 실험에서는 깊이의 튜닝을 통해 가장 적절한 값을 찾아내는 것이 중요하다.

3. 입력 데이터 구성 및 전처리

데이터는 군산에 위치한 공장 내에 설치된 태양광 발전 모듈의 인버터 네 대에서 기록된 태양광 발전량에 관한 것으로 2012 년 3 월부터 2015 년 12 월까지 기록된 데이터가 쓰였다. 원본 데이터는 10 분에서 15 분 단위로 인버터에 기록 된 데이터이다. 데이터의 인스턴스 개수는 196,406 개로 전처리 과정을 통해 일별 데이터로 변환하였다. 또한 전처리 과정에서 발전량 예측에 불필요한 속성은 제거하였다. 전처리 된 데이터에 기상청에서 얻은 일별 날씨 데이터를 조합하였다.

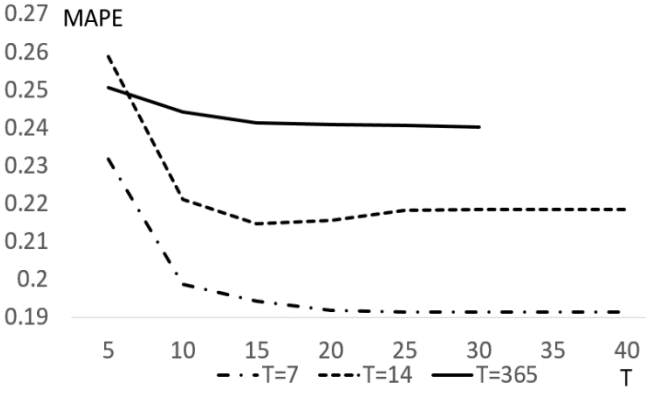
<표 1> 인버터와 기상청으로부터 얻은 속성

인버터 데이터	발전량	경사면 일사량	수평면 일사량	대기 온도	모듈 온도
기상청 데이터	평균기온	최대기온	최소기온	강수량	

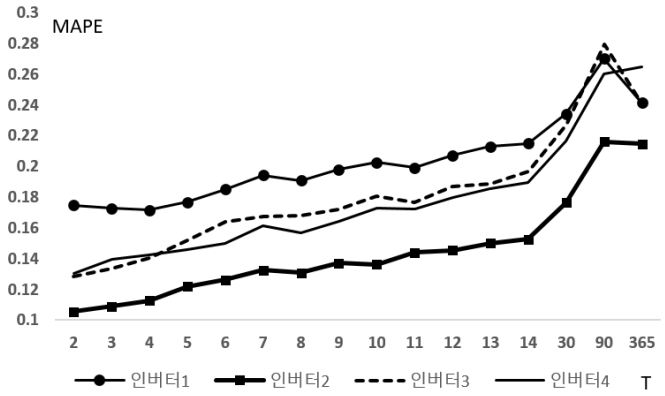
태양광 발전용 인버터에서 얻은 데이터의 구체적인 속성은 날짜, 저장 시점, 인버터의 전기적 특성들, 10 – 15 분 간격의 구간 발전량, 인버터 설치 지점의 경사면 일사량, 수평면 일사량, 대기온도, 모듈 온도이다. 이 중 전기적 특성은 외부 특성이 아닌 구간 발전량에 종속되어 있는 속성이므로 해당 속성은 삭제하였다. 따라서 구간 발전량을 발전량 예측 문제의 종속변수로 설정하고, 일사량과 대기 온도, 모듈 온도를 독립 변수로 설정 하였다. 본 연구에서는 일일 단위의 평균 발전량 예측이 목표이므로, 10 – 15 분 간격으로 측정된 발전 데이터를 일일 평균 발전량 데이터로 전처리 하였다. 한편, 태양광 발전은 기상 상황에 따라 발전량이 변하므로 주어진 데이터에 기상청에서 얻은 데이터를 추가하였다. 기상청[1]에서 얻은 데이터는 일별 데이터로, 평균 대기온도, 최대 대기온도, 최저 대기온도, 강수량이다. 이 데이터를 일별로 정리된 인버터의 발전량 데이터에 조합하여 기본 데이터를 구성하였다. 표 1 은 발전량예측을 위해 사용된 일별 속성값을 요약하여 보여준다.

본 연구에서는 당일의 발전량을 예측하기 위해 사용한 일일 데이터를 T=2 일전부터 하루씩 늘려가며 T=14 일 전까지 날짜 그리고 T=30 일 전 날짜, T=90 일 전 날짜, T=365 일 전 날짜 별로 속성을 확장하여 다시 전처리 하는 시스템을 구성했다. (그림 1)과 같이 T 일 전까지의 데이터는 다음과 같이 만들어졌다. 어떤 날짜 d의 데이터 속성들의 집합을 M_{d}라고 하면, 각각의 데이터는 한 줄에 $\{M_{\mathrm{d-T+1}}, \ldots, M_d\}$로 이루어져 있다. 이를 통해 날짜 d의 발전량을 예측한다. 본 연구에서는 몇일 전까지의 데이터를 고려해야 예측률이 높아지는지는 실험으로 검증하고자 하였다. 따라서, 일차적인 전처리 과정을 통해 얻은 데이터를 T 값을 변화시키며 생성하였다. 따라서 원본 데이터가 총 N 개의 인스턴스를 가지고 있고 속성의 개수가 K 라고

[1] http://www.kma.go.kr/



(그림 3) 트리 깊이에 따른 랜덤 포레스트 성능



(그림 2) 발전량 예측의 오차율 결과

할 때, T 일까지의 기간으로 전처리한 데이터는 $(N - T + 1) \times TK$ 크기의 행렬 형태가 된다.

4. 입력 데이터 구성 및 전처리

본 연구는 Intel i5-5600 CPU, 8GB memory, Windows 운영체제에서 Weka 를 이용해 수행되었다. 랜덤 포레스트의 예측 성능 향상을 위해 트리의 최대 깊이를 5 에서 5 씩 40 까지 늘려가며 T=7, 14, 365 인 경우에 대해 10-겹 교차 검증(10-fold cross validation)으로 예측 성능을 확인하였다. 이 때, 트리의 개수는 100 개로 고정하였다. (그림 2)는 트리의 깊이 변화에 따른 예측 오차를 나타낸다. 예측 성능을 평가하는 오차율 계산 방법은 평균 절대 백분율 오차(MAPE)를 사용하였다. 평균 절대 백분율 오차 M은 N 개의 인스턴스에 관해 관측값(또는 실제값)을 A_i, 예측값을 F_i라고 할 때, 아래와 같은 식으로 나타낸다.

$$M = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{A_i - F_i}{A_i} \right|$$

(그림 2)에서 평균 절대값 오차는 T 값에 따라 트리의깊이가 15 또는 20 일 때 예측 오차가 가장 적고, 그 이상 깊이를 더해도 성능의 변화가 적거나 오히려 성능이 감소하는 모습을 보였다. 한편, T=365 일 때 트리의 최대 깊이가 25 이상일 경우, 실험 환경에서 메모리 한계상 랜덤 포레스트가 계산되지 않는다. 따라서 각각의 인버터에 대한 발전량 예측에 관한 실험에서는 트리의 최대 깊이를 15 로 고정하여 진행하였다.

실험 결과는 (그림 3)과 같다. 인버터에 따라 예측 오차가 T=2 또는 T=4 일 때 최소가 되고, T 가 커질수록 점점 평균 절대 백분율 오차가 증가하는 경향을 보였다. 이는 T 가 늘어남에 따라 데이터의 크기가 증가하는 데, 랜덤 포레스트가 모델을 만드는 과정에서 데이터 크기만큼 과학습이 일어나기 때문인 것으로 분석된다.

5. 결론

본 연구에서는 태양광 발전 인버터로부터 10-15 분 단위의 데이터가 주어졌을 때, 일별 전처리와 기상 데이터 추가를 통해 당일의 태양광 발전량을 예측하였다. 이 때, 과거 몇일 전 까지의 데이터를 고려했을 때 예측 율 성능이 높은지 실험을 통해 검증하였다. 실험에서 인버터에 따라 과거 2 일 전까지 또는 4 일 전까지의 데이터를 고려했을 때 예측 오차가 가장 적었고, 날짜를 늘릴수록 점점 예측율이 떨어짐을 확인하였다. 이는 랜덤 포레스트를 통한 예측 모델 생성에서 고려할 날짜가 늘어날 때, 과학습이 발생하는 것으로 분석된다. 추후 과학습 발생이 적은 알고리즘을 적용하여, 발전량 예측 성능을 더욱 높이고자 한다.

6. 사사의 글

본 연구는 산업통산자원부 에너지기술개발사업(20153010011980, 과제명: 태양 광발전 운영효율 향상을 위한 통합관리 시스템 개발)의 연구비 지원으로 수행함

참고문헌

[1] Sophie Pelland, J. Remund, J. Kleissl, T. Oozeki and K. De Brabandere, "Photovoltaic and Solar Forecasting: State of the art" *IEA PVPS*, Task 14, pp. 1 – 36, 2013.

[2] J. Junior, T. Oozeki, H. Ohtake, K. Shimose, T. Takashima and K. Ogimoto, "Forecasting Regional Photovoltaic Power Generation-A Comparison of Strategies to Obtain One-Day-Ahead Data" *Energy Procedia,* vol. 57, pp. 1337-1345, 2014.

[3] M. Almeida, O. Perpinan, and L. Narvarte, "PV Power Forecast Using a Nonparametric PV Model" Solar Energy, vol. 115, pp. 354-368, 2015.

[4] Leo Breiman, "Random Forests" *Machine Learning,* vol. 45, no. 1, pp. 5 -32, 2001.