자원·환경경제연구 제 28 권 제 4 호
Environmental and Resource Economics Review
Volume 28, Number 4, December 2019 : pp. 467~495
DOI: https://doi.org/10.15266/KEREA.2019.28.4.467

미세먼지의 영향을 고려한 머신러닝 기반 태양광 발전량 예측

성상경* · 조영상**

요 약 : 태양광 발전과 같은 신재생에너지의 불확실성은 전력계통의 유연성을 저해하며, 이를 방지하기 위해서는 정확한 발전량의 사전 예측이 중요하다. 본 연구는 미세먼지 농도를 포함한 기상자료를 이용하여 태양광 발전량을 예측하는 것을 목적으로 한다. 본 연구에서는 2016년 1월 1일부터 2018년 9월 30일까지의 발전량, 기상자료, 미세먼지 농도 자료를 이용하고 머신러닝 기반의 RBF 커널 함수를 사용한 서포트 벡터 머신을 적용하여 태양광 발전량을 예측하였다. 예측변수에 미세먼지 농도 반영 유무에 따른 태양광 발전량 예측 모델의 성능을 비교한 결과 미세먼지 농도를 반영한 발전량 예측 모델의 성능이 더 우수한 것으로 나타났다. 미세먼지를 고려한 예측 모형은 미세먼지를 고려하지 않은 예측 모형 대비 6~20시 예측 모형에서는 1.43%, 12~14시 예측 모형에서는 3.60%, 13시 예측 모형에서는 3.88%만큼 오차가 감소하였다. 특히 발전량이 많은 주간 시간대에 미세먼지 농도를 반영하는 모형의 예측 정확도가 더 뛰어난 것으로 나타났다.

주제어 : 머신러닝, 태양광 발전량 예측, 미세먼지, 서포트 벡터 머신

JEL 분류 : C35, Q20, Q47

접수일(2019년 8월 8일), 수정일(2019년 9월 6일), 게재확정일(2019년 9월 9일)

* 연세대학교 일반대학원 공과대학 산업공학과 석사과정, 제1저자(e-mail: ssk1205@hanmail.net)
** 연세대학교 공과대학 산업공학과 부교수, 교신저자(e-mail: y.cho@yonsei.ac.kr)

Prediction of Photovoltaic Power Generation Based on Machine Learning Considering the Influence of Particulate Matter

Sangkyung Sung* and Youngsang Cho**

ABSTRACT : Uncertainty of renewable energy such as photovoltaic(PV) power is detrimental to the flexibility of the power system. Therefore, precise prediction of PV power generation is important to make the power system stable. The purpose of this study is to forecast PV power generation using meteorological data including particulate matter(PM). In this study, PV power generation is predicted by support vector machine using RBF kernel function based on machine learning. Comparing the forecasting performances by including or excluding PM variable in predictor variables, we find that the forecasting model considering PM is better. Forecasting models considering PM variable show error reduction of 1.43%, 3.60%, and 3.88% in forecasting power generation between 6am~8pm, between 12pm~2pm, and at 1pm, respectively. Especially, the accuracy of the forecasting model including PM variable is increased in daytime when PV power generation is high.

Keywords : Machine learning, Photovoltaic power forecasting, Particulate matter, Support vector machine

Received: August 8, 2019. Revised: September 6, 2019. Accepted: September 9, 2019.

* Master student, Department of Industrial Engineering, College of Engineering, Graduate School, Yonsei University, First author(e-mail: ssk1205@hanmail.net)

** Associate Professor, Department of Industrial Engineering, College of Engineering, Yonsei University, Corresponding author(e-mail: y.cho@yonsei.ac.kr)

I. 서 론

전 세계적으로『신 기후체제』에 대응하기 위하여 온실가스 배출을 줄이기 위한 노력을 하고 있으며, 우리나라 역시 에너지 분야에서 온실가스 배출을 감소시키기 위한 정책을 추진 중이다. 2018년 7월 정부는『2030년 국가 온실가스 감축목표 달성을 위한 기본 로드맵 수정안』을 발표하였으며, 여기에는 전력 수요관리 강화, 석탄화력 발전 비중 축소 및 재생에너지 발전 비중 확대 등의 정책 변화를 통해 온실가스 발생량을 감소시키겠다는 내용이 포함되어 있다.

『재생에너지 3020 이행계획』 및『제8차 전력수급기본계획』에 따르면 정부는 원전과 석탄발전을 단계적으로 감축하고 재생에너지 등의 분산형전원을 대폭 확대하는 정책을 추진하려고 한다. 이에 따른 전원믹스 변화가 향후 사회적 비용을 증가시킬 수 있으나(김광인 외, 2019) 정부는 친환경 발전원 구성을 달성하기 위해 2030년 총 발전량 중 재생에너지 비중을 20%, 63.8GW까지 확대하고, 재생에너지 신규설비의 95% 이상을 태양광, 풍력 등 청정에너지로 공급할 계획이다. 특히 2030년 약 36.5GW를 태양광으로 발전하는 것을 목표로, 도시형 자가용 태양광 확대, 소규모(100kW 이하) 사업지원 및 협동조합을 통한 참여 활성화, 농촌지역 태양광 활성화 정책 등을 시행하고 있다.

하지만 태양광 발전은 일사량에 의존하여 발전하는 특성으로 변동적 신재생에너지로 분류되며, 일반적으로 대규모의 변동적 신재생에너지가 전력계통에 병입될 때 전력계통의 유연성을 저해하는 요소로 작용한다. 변동적 신재생에너지는 출력을 예측하기 어려운 불확실성과 출력의 변화폭이 큰 변동성을 지니고 있어 전력계통 운영에 어려움을 준다. 전력계통의 안전도 저하를 방지하고 재생에너지의 이용을 극대화하기 위해서는 변동적 신재생에너지 발전 예측 시스템 구축이 중요하지만, 태양광 발전의 경우 발전량이 지역별 기상조건에 크게 영향을 받아 정확한 발전량 예측에 어려움이 존재한다. 따라서 태양광 발전량 사전 예측의 정확도 제고를 통해 태양광 발전소의 안정적 전력계통 연계를 높이고, 이를 통해 효율적인 에너지 관리 및 국가 재생에너지 확대 정책의 실현 가능성을 높이는 것이 필요하다.

본 연구는 태양광 발전량이 기상상태에 크게 영향을 받는다는 사실에 초점을 두고, 기상자료를 활용하여 머신러닝 기반 태양광 발전량 사전 예측 모델을 수립하고 그 성능을

평가하는 것을 목적으로 한다. 특히 태양광 발전량의 예측 정확도를 제고하기 위해 일반적인 기상자료(기온, 강수량, 풍향, 풍속, 습도, 운량)와 더불어 최근 사회적 이슈가 되고 있는 미세먼지 농도 자료(PM_{10}, $PM_{2.5}$)를 함께 고려하였다.

최근 우리나라의 초미세먼지 농도는 OECD 회원국 가운데 가장 높은 것으로 나타났다. OECD의 세계 각국 연평균 초미세먼지 농도 통계 자료에 따르면 2017년 우리나라 연평균 초미세먼지 농도는 25.1μg/m^3으로 회원국 가운데 가장 높은 것으로 나타나고 있으며[1], 이로 인해 사회적 불만과 개인 건강에 대한 우려가 증가하고 있는 상황이다. 실제 서미숙 외(2017)는 미세먼지 농도의 상승이 사람들의 주관적인 삶의 만족도를 하락시킨다는 연구 결과를 제시하였다. 미세먼지는 태양광 발전량에도 영향을 미치는데, 최근 언론 보도 등에 따르면 미세먼지로 인해 태양광 발전량이 평균 19% 감소되는 것으로 보고되고 있다.[2][3][4]

이에 본 연구에서는 한국환경공단에서 제공하고 있는 미세먼지 농도 자료(PM_{10}, $PM_{2.5}$)를 기상자료와 함께 머신러닝 기반 태양광 발전량 예측 모델에 반영하여 미세먼지 농도의 반영 여부에 따라 예측 모델의 정확도가 어떻게 변화하는지를 분석하였다. 본 연구 결과는 태양광 발전량 사전 예측 정확도 제고를 통해 신재생에너지의 불확실성을 완화할 수 있어 향후 재생에너지 발전 예측 시스템 구축에 활용이 가능하며, 동시에 전력계통의 안정적이고 효율적인 운영에도 일정 부분 기여할 것으로 기대된다.

본 논문의 구성은 다음과 같다. 제2장에서는 태양광 발전에 영향을 미치는 요인들을 살펴보고, 태양광 발전량 예측 관련 선행연구를 고찰한다. 제3장에서는 연구방법과 태양광 발전량 예측 모형을 설정하고, 제4장에서는 머신러닝 알고리즘을 활용하여 미세먼지 농도의 반영 유무에 따른 태양광 발전량 예측 모델의 성능을 비교 분석하였다. 마지막 제5장에서는 본 연구의 주요 결과를 살펴보고, 의의와 한계점을 논의한다.

---

1) OECD.Stat[Website], (2019.9.1), https://stats.oecd.org/index.aspx?queryid=72722)
2) 서유진,“미세먼지 심한 날, 태양광 발전량 19% 줄어”, 중앙일보, 2019.3.13
3) 국회 산업통상자원중소 벤처기업위원회 소속 김삼화 의원실,“미세먼지 높았던 3.1-6일과 직전 6일 발전량 비교”
4) 이는 태양광 발전소 12곳에서 미세먼지 비상저감 조치가 시행된 6일간 태양광 발전량 데이터이므로 표본이 작긴 하지만, 공기 중 미세먼지가 태양광 발전 효율에 상당 부분 영향을 미친다고는 판단된다.

Ⅱ. 선행연구

1. 태양광 발전에 영향을 미치는 요인

차왕철 외(2014)에 따르면 태양광 발전소의 발전량에 영향을 미치는 요소는 설비요소, 지리 · 지형 요소, 기상요소로 크게 구분할 수 있다. 여기서 설비요소는 태양광 발전 시스템의 설치방향, 각도, 고정식/추적식 등과 같은 발전설비 형식과 모듈, 인버터 등의 효율을 의미하며, 지리 · 지형 요소는 발전소 설치 지역의 위도, 경도, 고도 등을 의미한다.

태양광 발전량에 가장 큰 영향을 미치는 요소는 일사량, 기온 등의 기상요소이다(손정훈 외, 2019). 태양광 발전은 태양의 빛 에너지를 변환시켜 전력을 생산하므로 태양에서 오는 빛의 복사를 의미하는 일사량이 발전량에 가장 큰 영향을 미치는 요인이다. 또한 태양전지는 반도체로 구성되어 있으며 이는 고온에서 효율이 떨어지는 특성을 가지고 있어 온도 역시 발전량에 영향을 미친다(이강혁 외, 2016). 상대습도는 대기 중의 수증기량을 나타내는데, 강수 또는 구름이 많은 날은 상대습도가 높아 일사가 직접적으로 차폐된다. 맑은 날에도 대기 중의 수증기는 일사를 산란시키는 작용을 한다. 따라서 상대습도가 증가할수록 태양광 발전량이 감소하는 경향이 있어 상대습도 또한 태양광 발전량에 대한 변수로 작용할 수 있다. 풍속의 경우 바람이 강할수록 태양광 패널의 표면 온도를 하강시켜 발전량에 직간접적으로 영향을 줄 수 있다(이순환 외, 2014). 그 외 운량, 풍향, 강수량 등도 태양광 발전량에 대해 영향을 미치는 것으로 보고되고 있다(이강혁 외, 2016).

최근 기상 요인들 중 태양광 발전에 영향을 미치는 요인으로 미세먼지가 언급되고 있다. 미세먼지는 직경에 따라 크게 PM_{10}과 $PM_{2.5}$ 등으로 구분되는데, PM_{10}은 1,000분의 10mm보다 작은 먼지이며, $PM_{2.5}$는 1,000분의 2.5mm보다 작은 먼지로 머리카락 직경의 1/20~1/30 크기보다 작은 입자이다.[5] 미세먼지 발생원은 흙먼지, 바닷물의 소금 및 꽃가루 등의 자연적 발생원과 화석연료에서 나오는 매연, 자동차 배기가스, 건설현장의 날림 먼지 및 소각장 연기 등의 인위적 발생원이 있다. 미세먼지는 발생원으로부터 고체상태의 미세먼지로 나오는 경우와 발생원에서 가스 상태로 배출된 물질이 공기 중의 다

5) 에어코리아[웹사이트], (2019.9.1), https://www.airkorea.or.kr/web/airMatter?pMENU_NO=130

른 물질과 화학반응을 일으켜 미세먼지가 되는 경우로 나누어질 수 있다(환경부, 2016). 한진목 외(2018)에 따르면 대기 중의 누적 미세먼지 농도 증가에 비례하여 태양전지 셀의 오염면적이 증가한다. 이는 미세먼지가 태양전지 셀의 오염원으로서 태양광 발전 시설의 발전 출력에 적지 않은 영향을 준다는 것을 의미한다. 또한 공기 중의 미세먼지 자체가 태양 복사에 직접적인 영향을 줄 수 있는데, 이는 미세먼지에 의한 반사 또는 산란이 발생하기 때문이다. 빛의 반사는 입사하는 빛이 물체에 부딪칠 때 진행 방향이 바뀌어 나가는 현상이고, 빛의 산란은 태양 빛이 공기 중의 질소, 산소, 먼지 등과 같은 작은 입자들과 부딪칠 때 빛이 사방으로 재방출되는 현상이다. 이 과정에서 원래 지표면에 도달해야 할 복사가 줄어들기 때문에 일사량은 감소하게 된다(조덕기 외, 1996).

미세먼지가 태양광 발전에 미치는 간접적인 영향으로 구름에 미치는 영향이 있다. 구름이 생기기 위해서는 응결핵이 필요하며, 이 응결핵은 구름의 주성분인 물방울이 되기 위해서 수증기가 달라붙을 수 있는 작은 입자를 말한다. 보통의 경우 대기 중의 소금이나 흙 알갱이가 응결핵의 역할을 하지만, 만약 미세먼지가 이 역할을 하게 되면 일반적인 경우보다 작은 입자가 응결핵이 된다. 이로 인해 구름 입자의 크기가 작아지게 되고, 구름의 표면적이 늘어 더 밝은 구름이 형성되며, 이는 태양복사를 더 많이 반사시키게 된다(Twomey, 1977). 즉, 대기 중의 미세먼지 농도 역시 태양광 발전량 예측에 있어서 중요한 변수로써 고려해야 함을 알 수 있다.

2. 태양광 발전량 예측 관련 선행연구

최근 전 세계적으로 태양광 발전량이 크게 증가함에 따라 태양광 발전량 예측 관련 연구가 많이 진행되고 있다. Sharma et al.(2011)은 스마트 그리드 망에서 재생에너지의 변동성 및 불확실성으로 인한 재생에너지 발전량 예측의 어려움을 해결하고자 하였다. 이에 머신러닝 기술을 이용하여 미국 National Weather Service(NWS)의 기상자료를 기반으로 한 태양광 발전량 예측 모델을 제시하였다. 하루 중 정오시간의 데이터를 기준으로 10개월 동안의 기상 및 발전량 데이터를 분석에 사용하였는데, 1~8월 데이터를 훈련데이터로, 9~10월 데이터를 테스트데이터로 사용하였다. 또한 기존 운량자료 기반 예측 모델과 다중커널함수를 이용한 Support Vector Machine(SVM) 예측 모델의 성능을 Root Mean Squared Error(RMSE) 값을 이용하여 비교하였는데, 연구 결과 SVM 예측

모델의 성능(RMSE 128W/m^2)이 기존 운량자료 기반 예측 모델의 성능(RMSE 178W/m^2)보다 27% 더 정확한 것으로 나타났다.

da Silva Fonseca Jr. et al.(2012)은 일본 기타큐슈 지역 1MW 태양광 발전소의 발전량 예측을 위해 운량을 포함한 기상자료 데이터를 SVM 모델에 적용하였다. 2008년부터 2010년까지의 데이터 중 마지막 1년간의 데이터가 예측 용도로 사용되었다. 운량 정보의 사용 유무에 따라 RMSE, Mean Absolute Error(MAE) 값을 이용하여 예측 성능을 비교한 결과, 운량 정보를 제외했을 경우가 운량 정보를 사용했을 경우 대비 RMSE가 32%, MAE가 42% 이상 증가하여 운량 정보가 예측에 중요한 변수임을 발견하였다.

Shi et al.(2012)은 중국지역에서의 기상상태 자료를 기반으로 SVM을 활용하여 태양광 발전량 예측 알고리즘을 제안하였다. 10개월간 15분 데이터를 활용하여 이후 24시간 태양광 발전량을 예측하였으며, 기상상태를 맑음, 흐림, 안개, 비 4단계로 구분하였다. 기상상태로 구분된 4가지 예측 모델을 생성하여 각 모델의 오차를 비교한 결과 맑음, 안개 모델이 타 모델보다 성능이 우수하였다. 대상지역인 중국 남부 지역에는 맑고, 안개가 낀 날의 비중이 많으므로 연구 결과의 적용이 가능하다는 결론을 제시하였다.

Mei and Ma(2013)는 과거 기상자료 및 발전량 데이터를 활용하여 태양광 발전량 예측을 위한 Radial Basis Function(RBF) 커널 기반 SVM 모델을 생성하였다. 여기서 SVM 모델의 학습속도를 올리기 위해 러프세트이론(rough set theory)을 활용하여 불필요한 변수를 제거하였다. 또한 계절 속성, 일조시간, 일사량, 기온 등과 같은 날씨 특성이 유사한 날들을 기준으로 과거 기상 데이터를 여러 형태로 분류하고, 전체 데이터를 사용하는 대신 유사한 날들 별로 데이터를 일부 사용하여 모델을 간략화하였다.

Ramli et al.(2015)은 사우디아라비아 지역 태양광 발전소의 발전량을 예측하기 위해 SVM 및 Artificial Neural Networks(ANN) 모델을 사용하여 예측 정확도와 예측 처리속도 측면에서 두 모델의 성능을 비교하였다. 연구 결과 SVM 모델의 예측 정확도(RMSE 18.3~31.1W/m^2)가 ANN 모델의 예측 정확도(RMSE 61.7~95.5W/m^2) 대비 더 우수했으며, SVM 모델의 예측속도(2.15초) 또한 ANN 모델의 예측속도(4.56초) 대비 더 빠른 것으로 나타나 SVM 모델의 예측 성능이 ANN 모델보다 전반적으로 우수하다는 것을 확인하였다.

Das et al.(2018)은 태양광 발전량 예측 모델과 모델 최적화 연구에 대한 리뷰를 진행하면서 SVM 모델은 최근 태양광 발전량 예측에 널리 사용되고 있으며, 매우 유연한 비

선형 모델임을 언급하였다. 저자들에 따르면 SVM 모델의 강점은 ANN 모델과 달리 학습 데이터의 양에 크게 의존하지 않고 학습이 가능하며, Neural Networks(NN) 모델처럼 국지 최솟값에 빠지는 문제가 발생하지 않는다는 것이다. 따라서 SVM 모델을 사용하면 태양광 예측과 관련된 복잡한 수학적 문제를 단순화할 수 있다는 점을 강조하였다.

Vakili et al.(2017)은 이란 테헤란 지역의 1일(1day) 태양광 발전량을 예측하였다. 예측변수로는 기존 연구에서 사용하던 기상자료(상대습도, 풍속, 최고 · 최저 기온)뿐만 아니라 미세먼지 데이터를 함께 사용하였다. 태양광 발전량 학습을 위해 ANN 모델을 적용하였으며 예측변수에 미세먼지 자료의 포함 유무에 따른 결과를 비교하였다. 연구결과 미세먼지 자료를 포함한 경우가 미포함의 경우 대비 평균절대백분율오차(Mean Average Percentage Error, MAPE)가 2.99% 감소하였다.

국내 연구의 경우, 먼저 송재주 외(2014)는 국내 실증단지 내 발전단지의 실시간 기상자료 예측값을 이용하여 태양광 발전량을 예측하기 위한 모델을 제시하였다. 저자들은 일사량 단기 예측을 위해 신경망 모델을 사용하였으며, 일사량 예측을 위한 예측변수는 날짜, 시간, 기온, 습도, 풍향, 풍속, 기압 등을 사용하였다. 2012년 1년간의 데이터를 이용하여 월별 일사량 예측을 수행한 결과 단순히 신경망 모델만 적용했을 경우(MAPE 27%, RMSE 107W/m^2)보다 신경망 모델에 정오시간대 오차 보정을 적용할 때(MAPE 25%, RMSE 98W/m^2) 일사량 예측값 평균은 실측값에 더 근접하고 정확도가 향상되었다. 일사량 단기 예측값과 실측값의 상관도를 분석한 결과 결정계수(R^2)는 0.85 이상으로 높은 상관도를 보였다.

이강혁과 김우제(2016)는 2013~2014년간의 발전량, 기상실측, 기상예보 데이터를 활용하여 24시간 사전 태양광 발전량 예측을 수행하였다. 연구에서는 먼저 Support Vector Regression(SVR)을 이용하여 일사량 예측을 실시하고, 이 예측된 일사량으로부터 최종 발전량을 예측하는 과정을 거쳤다. 그리고 발전량 예측 성능을 저해하는 요인으로 구름 두께의 불확실성을 고려하였다. 분석을 통해 하늘상태가 맑은 날에는 태양광 발전량 예측 성능이 좋고, 구름이 많은 날에는 구름 두께의 불확실성으로 예측 성능이 저하된다는 결론이 제시되었다.

배국열 외(2017)는 대표적인 머신러닝 기법인 ANN과 RBF 커널 기반의 SVM을 이용한 태양광 출력 예측 기술을 제안하고 해당 기법의 예측오차 확률 분포함수를 도출하였다. 또한 기상 입력변수와 관련하여 날씨 예측오차가 태양광 출력 예측 정확도에 미치는

영향을 분석하였다. 연구에서는 태양광 출력 예측 성능을 비교하기 위해 RMSE, Mean Relative Error(MRE), R^2를 평가지표로 사용하였으며, 비교 결과 SVM 기법의 성능(RMSE 58.72W/m^2, MRE 3.82%, R^2 0.9562)이 ANN 기법의 성능(RMSE 71.41W/m^2, MRE 5.47%, R^2 0.9234)보다 우수하였으며, 기후 예측 오차를 반영하여도 SVM 기법이 더 우수하다는 결론을 제시하였다.

이건주 외(2017)는 태양광 발전소 최적 입지 선정을 위한 연구에서 태양광 발전량에 영향을 미치는 기상요인들을 분석하였다. 2014년 영암발전소의 발전량 데이터와 기상자료를 활용하였는데, 연구 결과 일조량과 일사량은 태양광 발전량과 양의 상관관계, 습도와 미세먼지는 음의 상관관계를 가지는 것으로 나타났다. 특히 미세먼지의 경우 약 100μg/m^3 이상의 미세먼지 농도에서 뚜렷한 음의 상관관계를 보여 미세먼지가 태양광의 발전량에 일정 부분 영향을 미치는 것으로 나타났다.

본 연구에서는 선행연구에서 머신러닝 기법으로 성능이 우수하다고 밝혀진 RBF 커널을 활용한 SVM 모델 기법을 적용하여 태양광 발전량 예측을 수행하였다. 기존 국내외의 연구들은 기상자료를 활용하여 태양광 발전량을 예측하는 연구는 많은 반면, 미세먼지의 효과를 살펴보는 연구는 많지 않은 실정이다. 이건주 외(2017)는 상관관계 분석을 통해 미세먼지가 태양광 발전량에 미치는 것을 밝혀냈지만, 머신러닝 기반의 태양광 발전량 예측 모델의 예측 성능에 대한 미세먼지의 영향을 실증 분석하는 본 연구와는 연구 방향에 있어 차이가 있다. 또한 Vakili et al.(2017)은 기상자료와 미세먼지 데이터를 함께 이용하여 ANN 모델 기반 1일 태양광 발전량 예측을 수행하였기 때문에 SVM 모델 기반 1시간 단위의 태양광 발전량을 예측하는 본 연구와는 차이가 있다.

III. 연구방법

1. 머신러닝 개요

머신러닝은 인공지능의 한 분야로서, 데이터라는 형태로 얻어지는 경험(experience)으로부터 특정한 목표 작업(task)에 대한 성능(performance)을 향상시키는 과정이라고 정의할 수 있다(조성준 외, 2016). 머신러닝은 통계학, 인공지능, 컴퓨터 과학이 종합적으로 관련된 연구 분야이며, 예측 분석이나 통계적 머신러닝으로도 불린다. 최근 영화

추천, 음식 주문, 쇼핑, 맞춤형 온라인 라디오 방송과 사진에서 친구 얼굴을 찾아주는 일까지 일상생활에서 경험하는 많은 웹사이트와 기기들이 머신러닝 알고리즘을 핵심기술로 채택하고 있다.

머신러닝의 종류는 일반적으로 지도학습, 비지도학습, 강화학습으로 구분된다. 먼저 지도학습은 머신러닝에서 가장 기본이 되고 구현하기 쉬운 알고리즘으로 입력과 출력 데이터가 있고, 주어진 입력으로부터 출력을 예측하고자 할 때 사용한다. 지도학습은 학습결과를 바탕으로 무엇을 예측하느냐에 따라 회귀(regression)와 분류(classification)로 구분할 수 있다. 다음으로 비지도학습은 출력값이나 정보가 주어지지 않는 상태에서 컴퓨터를 학습시킨다. 비지도학습은 레이블이 없는 데이터를 사용하는 데, 그 목적은 데이터 내 어떠한 관계(relationships)를 찾아내는 것이다. 이는 결과 정보가 없는 데이터들에 대해 특정 패턴을 찾아 특성이 비슷한 데이터를 합쳐서 군집화(clustering)하는 학습방법이다(문성은 외, 2016). 마지막으로 강화학습은 최근 알파고의 학습으로 유명해진 기법이다. 강화학습의 핵심은 보상(reward)으로, 컴퓨터는 보상을 받는 행위를 위해서 스스로 문제를 찾아나가며 초반에는 어느 정도 인간의 개입이 들어갈 수 있다. 강화학습의 목표는 보상의 최대치가 되는 행동을 하는 것으로 이 최대치의 보상을 얻기 위해 컴퓨터는 끊임없이 학습을 한다.

본 연구에서는 기상자료 및 미세먼지 농도자료라는 입력 데이터와 태양광 발전량이라는 출력 데이터를 활용하여 주어진 학습 데이터로부터 학습된 알고리즘이 태양광 발전량을 얼마나 정확히 예측하는지를 분석하고자 한다. 따라서 본 연구에서는 머신러닝 기법 중에서 입력과 출력 데이터를 바탕으로 한 지도학습 기법을 활용하였으며, 그 중에서도 비선형에서도 우수한 성능을 나타내는 SVM 알고리즘을 적용하였다.

2. 서포트 벡터 머신(Support Vector Machine)

서포트 벡터 머신(Support Vector Machine, SVM)은 구조적 위험 최소화(Structural Risk Minimization, SRM) 원리에 바탕을 둔 머신러닝 지도학습 방법이다. SRM은 기대되는 위험의 상한 경계를 최소화하므로 훈련 데이터의 에러를 최소화하는 특징이 있다(Nasien et al., 2010).

SVM의 기본 아이디어는 학습 데이터를 2개의 클래스로 구분하는 선형 평면을 구하는 것으로, 2차원 선형 분류 문제는 <그림 1>과 같다. 여기서 회색으로 표시된 데이터는 서포트 벡터(support vector)이며, 두 클래스 사이의 가장 큰 마진(margin)을 결정한다.

x를 예측변수, y를 응답변수라고 할 때, 학습데이터는 $(x_1,y_1),\ldots.,(x_l,y_l),\ x\in R^n$, $y\in +1,-1$로 표현될 수 있으며 학습 데이터를 분류하는 최적의 평면을 구하는 문제는 식 (1)이 최소가 되는 w와 b를 구하는 문제가 되며 제약조건은 식 (2)와 같다. 여기서 w는 이 평면과 수직인 법선벡터이며, b는 w가 0일 때의 절편을 의미한다.

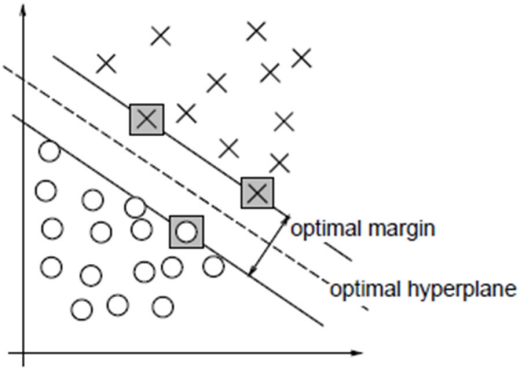
$$\textit{minimize}\quad \Phi(w)=\frac{1}{2}\parallel w\parallel^2=\frac{1}{2}w^T\cdot w \tag{1}$$

$$\textit{subject to}\quad y_i(w^T\cdot x_i+b)\geq 1\quad i=1,2,3,\ldots..,l \tag{2}$$

이 최적화 문제를 라그랑지안 문제로 변환하면 식 (3)으로 표현되고, 이를 쌍대(dual) 문제로 다시 변환하면 식 (4)와 같으며 제약조건은 식 (5)가 된다. 여기서 α_i는 라그랑지안 승수이다. 따라서 최적의 평면을 구하기 위해서는 식 (4)가 최대가 되는 α를 구하면 된다(Vapnik, 1995).

〈그림 1〉 2차원 선형 분류 문제

(출처: Cortes and Vapnik, 1995)



$$\textit{minimize} \quad L(w,b,\alpha) = \frac{1}{2} w^T \cdot w - \sum_i \alpha_i (y_i (w^T \cdot x_i + b) - 1), \ \alpha_i \geq 0 \tag{3}$$

$$\textit{maximize} \quad L(\alpha) = \sum_k \alpha_k - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{4}$$

$$\textit{subject to} \quad \sum_i \alpha_i y_i = 0, \ \alpha_i \geq 0 \ \text{for} \ \ all \ \ \alpha_i \tag{5}$$

선형 분류 문제 중에는 에러 없이는 학습 데이터를 구분할 수 없는 경우가 있는데, <그림 2>는 이러한 상황을 나타내고 있다.

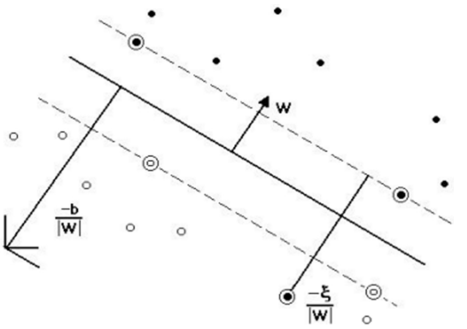
이때 slack 변수ξ를 적용하여 문제를 해결할 수 있는데, slack 변수는 오차를 허용하면서 선형적으로 분류를 할 때 사용하는 변수이다. slack 변수를 고려하여 최적의 평면을 구하기 위한 문제와 제약조건은 식 (6)과 식 (7)로 표현된다(Vapnik, 1995). 여기서 C는 일반화 파라미터(regularization parameter)이며, 오버피팅(overfitting)을 조절하기 위한 페널티(penalty) 항이다(Cortes et al., 1995).

$$\textit{maximize} \quad L(\alpha) = \sum_k \alpha_k - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{6}$$

$$\textit{subject to} \quad \sum_i \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C \quad \text{for} \ \ all \, \alpha_i \tag{7}$$

〈그림 2〉 소프트 마진 분류 문제

(출처: 구경민, 2002)



SVM 알고리즘은 비선형 분류 문제에도 사용할 수 있다. 비선형 분류를 위해서는 주어진 데이터를 고차원 특징 공간으로 매핑(mapping)하는 작업이 필요하다. 이를 효율적으로 실행하기 위해 커널 함수를 사용한다(Nasien et al., 2010). <그림 3>은 데이터를 고차원 공간으로 매핑하여 비선형 분류 문제를 선형 분류 문제로 변환하는 과정을 나타냈다.

비선형 입력 데이터는 아래 식 (8)에서처럼 $\phi(x)$ 함수에 의해 고차원 공간으로 매핑된다. 보통 커널함수 $k(x_i,x_j)$를 사용하는 데, $k(x_i,x_j) = \phi(x_i)\phi(x_j)$라고 하면, 식 (8)은 식 (9)로 표시할 수 있으며 제약조건은 식 (10)과 같다. 따라서 식 (9)가 최대가 되는 α를 찾으면 최적 평면을 구할 수 있다.

$$maximize \quad L(\alpha) = \sum_k \alpha_k - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \Phi(x_i)\Phi(x_j) \tag{8}$$

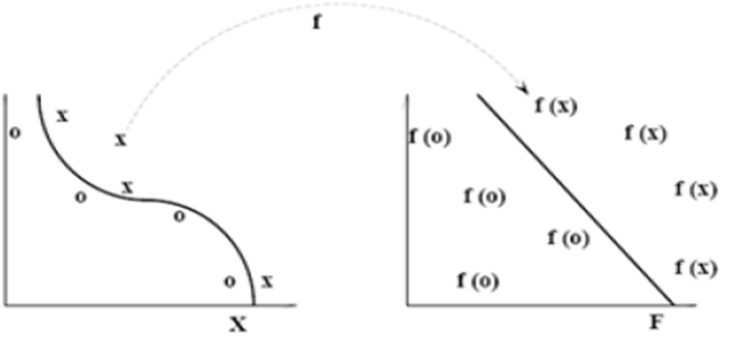
$$maximize \quad L(\alpha) = \sum_k \alpha_k - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j k(x_i, x_j) \tag{9}$$

$$subject\ to \quad \sum_i \alpha_i y_i = 0,\ 0 \le \alpha_i \le C \text{ for } all\ \alpha_i \tag{10}$$

SVM은 분류 문제를 해결하기 위해 개발되었지만 회귀문제 영역까지 적용되고 있다.

〈그림 3〉 고차원공간으로 매핑

(출처: Nasien et al., 2010)



회귀문제를 위한 SVM 기법을 SVR이라고 한다. <그림 4>는 SVR의 비선형 회귀문제 및 파라미터를 그림으로 나타낸 것이다.

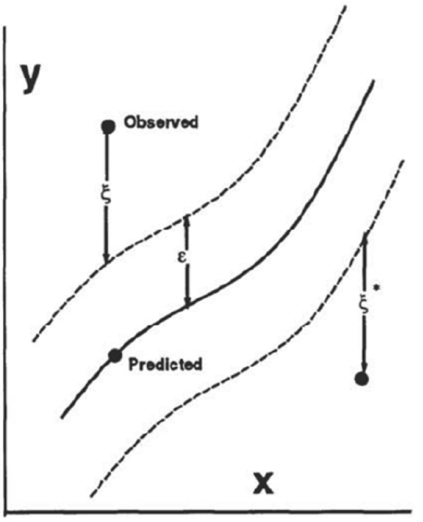
<그림 4>에서 ε은 한계점(threshold)을 결정하는 파라미터이고, ξ는 오차를 허용하기 위한 slack 변수이다. SVR은 모든 예측값들이 실제값으로부터 ε범위 안에 놓이도록 한다. 이를 만족하는 최적의 평면을 구하기 위한 문제는 식 (11)이 최소가 되는 w와 b를 구하면 되며 이때 제약조건의 식 (12)와 같다(Drucker et al., 1997). 이후 과정은 앞서 분류 문제와 동일하게 쌍대문제로 변환하고 커널함수를 적용하여 학습 데이터들을 고차원공간으로 매핑시킨 후 문제를 해결할 수 있다.

$$\textit{minimize} \quad \Phi(w) = \frac{1}{2} w^T \cdot w + C \sum_i (\xi_i + \xi_i^*) \tag{11}$$

$$\textit{subject to} \ \ y_i - (w^T \cdot x_i + b) \le \epsilon + \xi_i, (w^T \cdot x_i + b) - y_i \le \epsilon + \xi_i^*, \ \xi_i, \xi_i^* \ge 0 \tag{12}$$

〈그림 4〉 SVR의 비선형 회귀문제 및 파라미터

(출처: Drucker et al., 1997)



본 연구에서는 태양광 발전량 예측에 있어서 가장 널리 사용되고 있는 RBF 커널 함수를 적용한 SVR기법을 사용하였다. RBF 커널 함수는 식 (13)으로 나타낼 수 있고, 이론적으로 데이터들을 무한대의 차원으로 매핑이 가능하다(Das et al., 2018).

$$k(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2),\ \gamma > 0 \tag{13}$$

RBF 커널 함수에서 x_i, x_j는 각각의 데이터 포인트이며, $\parallel x_i - x_j \parallel$ 는 유클리디안 거리, γ는 가우시안 커널의 폭을 제어하는 매개변수이다. RBF 커널 SVM은 매개변수 C 및 γ가 사용자에 의해 세팅되어야 한다. C는 데이터 샘플들이 다른 클래스에 놓이는 것을 허용하는 정도를 결정하고, γ는 결정경계의 곡률을 결정한다. 따라서 좋은 성능을 얻기 위해서는 최적의 매개변수 C 및 γ를 찾는 것이 중요하다.

IV. 연구결과

1. 데이터

본 연구는 기상자료(일사량, 기온, 강수량, 습도, 풍속, 풍향, 운량)와 미세먼지 농도자료(PM_{10}, $PM_{2.5}$)를 반영하여 머신러닝 기반 태양광 발전량 예측 모델을 수립하고, 변수로 미세먼지 농도자료를 포함하는지 여부에 따라 태양광 발전량 예측 모델의 정확도가 어떻게 변화하는지를 분석하고자 한다.

태양광 발전량 예측을 위해서는 먼저 분석 대상인 태양광 발전소의 선정이 필요하다. 공공데이터포털에서 한국남부발전(주)의 발전소 통합 발전정보를 데이터로 제공하고 있으며, 본 연구에서는 한국남부발전(주)의 태양광 발전소 중 준공시기가 다소 오래된 부산 태양광 발전소를 연구대상으로 선정하였다. 현재 기준으로 준공시기가 오래되었기 때문에 타 발전소보다 더 오랜 기간의 발전량 데이터를 얻을 수 있기 때문이다. 한국남부발전(주)의 부산 복합태양광은 전체 용량이 390kW이며, 한국남부발전(주)의 부산발전본부 내 설치되어 운영 중이다. 부산 복합태양광 발전소의 기본 정보는 <표 1>과 같다.

〈표 1〉 부산 복합태양광 발전소 기본 정보

(출처: 한국남부발전(주))

구분	내용	비고
발전소명	부산 복합태양광	
주소	부산광역시 사하구 감천항로 7	부산발전본부 내
준공시기	2008년 7월	
설치용량	389.96kW	
모듈	170W × 1,428개, 200W × 736개	단결정 실리콘 태양전지
인버터	100kW × 3대, 50kW × 3대	

태양광 발전량 데이터는 2016년 1월 1일부터 2018년 9월 30일까지 2년 9개월 동안의 시간당 발전량 데이터를 사용하였다. 태양광 발전량에 영향을 미치는 요소 중 설비요소와 지리 · 지형 요소는 고정적인 요인이므로 고려 대상에서 제외하였고, 일사량, 기온 및 미세먼지 농도 등의 기상요소만 고려하였다.

현재 기상 예보자료 중 기온, 풍향, 풍속, 습도, 운량 데이터는 3시간 단위로 제공하고 있으며, 강수량 데이터는 6시간 단위로 제공하고 있다. 운량 데이터는 관측 자료의 경우 0~10까지의 수치로 표현되어 있지만, 예보 자료의 경우 0~2까지를 "맑음", 3~5를 "구름 조금", 6~8을 "구름 많음", 9~10을 "흐림"으로 범주화하여 등급으로 제공된다. 미세먼지 농도 예보자료도 매일 4회(5시, 11시, 17시, 23시) 수치가 아닌 등급으로 제공되며, PM_{10}, $PM_{2.5}$ 농도 수치 기준으로"매우 나쁨", "나쁨", "보통", "좋음" 4단계로 구분하여 표현된다.

본 연구에서는 태양광 발전량 예측을 위해 2016년 1월 1일부터 2018년 9월 30일까지 매 시간별 기상 자료, 미세먼지 농도 관측 데이터를 이용하였다.[6] 기상자료는 기상자료개방포털의 과거 시간별 종관기상관측 자료를 활용하였고, 미세먼지 농도 자료는 한국환경공단 에어코리아(Air Korea)의 전국 실시간 대기오염도 및 측정소별 관측자료를 사용하였으며, 부산광역시 측정소 중에서 부산 복합태양광 발전소에서 가장 가까운 부산 대신동 측정소의 자료를 사용하였다.[7]

6) 머신러닝 알고리즘에 적용할 변수(월, 시간, 기온, 강수량, 풍향, 풍속, 습도, 운량, PM_{10}, $PM_{2.5}$) 데이터 중 누락이 발생한 시간의 데이터는 제외하였음.

7) 대기오염 관측소에서 발전소까지는 지도 상에서 직선거리 약 3.2km 정도 떨어져 있음.

머신러닝 알고리즘을 적용하기 위해서는 앞에서 살펴본 바와 같이 학습 데이터(training data)와 평가 데이터(test data)가 필요하다. 본 연구에서는 전체 데이터 중 학습 데이터의 비중을 75%, 평가 데이터의 비중 25%로 하고, 랜덤 샘플링을 통해 학습 데이터 세트(training data set)와 평가 데이터 세트(test data set)를 구성하였다.

또한 각 예측변수(기온, 강수량, 풍향, 풍속, 습도, 운량, PM_{10}, $PM_{2.5}$) 데이터의 단위가 다르므로 머신러닝 알고리즘 적용 시 예측변수에 대한 데이터 전처리 과정을 진행하였다. 여러 데이터 전처리 방법 중 태양광 발전량 예측에서는 정규화가 널리 사용되고 있다(Das et al., 2018). 각 데이터의 변환 공식은 식 (14)와 같다.

$$Data_{Normal} = \frac{Data_{actual} - Data_{\min}}{Data_{\max} - Data_{\min}} \tag{14}$$

본 연구의 목적은 일사량 및 태양광 발전량 예측 모델을 수립함에 있어서 미세먼지 농도의 영향을 분석하는 것이며, 이를 위해 예측변수에 미세먼지 농도를 포함한 예측 모델과 미세먼지 농도를 제외한 예측 모델의 성능을 서로 비교하였다. 모델의 성능 평가 지표는 MAE를 사용하였으며, 이는 식 (15)와 같다. 여기서 $W_{forecasted}$는 예측값, W_{true}는 실제값을 나타낸다. 모델별로 MAE 값을 구하고 비교하여 최소의 MAE 값을 가지는 모델을 최적의 예측 모델로 간주하였다.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \left| W_{forecasted} - W_{true} \right| \tag{15}$$

또한 발전 시간대를 고려하여 예측 모델을 시간대별로 3가지 유형으로 나누어 미세먼지 농도 포함 유무에 따른 영향을 분석하였다. 먼저 하루 중 태양광 발전이 시작되는 6시부터 발전이 종료되는 20시까지의 데이터를 사용하여 예측 모델을 생성하고, 다음으로 하루 중 태양광 발전량이 비교적 많은 시간대인 12~14시 데이터를 사용하여 예측 모델을 수립하였다. 마지막으로 태양광 발전량이 하루 중 최대가 되는 13시 데이터만을 사용하여 예측 모델을 만들었다. 이때 12,733개의 데이터 중 분석에 사용된 데이터는 각각 8,902개, 2,031개, 657개이다. 이를 통해 각 시간대별 예측 모델을 수립하는 과정에서 시간 변수가 예측에 어떠한 영향을 미치는지를 확인하고자 하였다.

2. 일사량 예측 모델

현재 우리나라 기상자료개방포털에서 과거 일사량 관측치는 제공하지만, 기상청 동네예보 시스템에서 일사량 예보 데이터를 제공하고 있지 않다. 따라서 태양광 발전량 예

〈표 2〉 미세먼지를 포함한 예측변수 조합별 일사량 예측 성능

예측변수	MAE(MJ/m)		
	6~20시*	12~14시	13시**
월, (시간), PM_{10}, $PM_{2.5}$, 기온, 강수량, 풍향, 풍속, 습도, 운량	0.1584	0.2309	0.2409
월, (시간), PM_{10}, 기온, 강수량, 풍향, 풍속, 습도, 운량	0.1592	0.2315	0.2403
월, (시간), $PM_{2.5}$, 기온, 강수량, 풍향, 풍속, 습도, 운량	0.1606	0.2329	0.2425
(시간), PM_{10}, $PM_{2.5}$, 기온, 강수량, 풍향, 풍속, 습도, 운량	0.2311	0.3367	0.3160
월, PM_{10}, $PM_{2.5}$, 기온, 강수량, 풍향, 풍속, 습도, 운량	0.4206	0.2271	-
월, (시간), PM_{10}, $PM_{2.5}$, 강수량, 풍향, 풍속, 습도, 운량	0.1624	0.2401	0.2424
월, (시간), PM_{10}, $PM_{2.5}$, 기온, 풍향, 풍속, 습도, 운량	0.1586	0.2314	0.2423
월, (시간), PM_{10}, $PM_{2.5}$, 기온, 강수량, 풍속, 습도, 운량	0.1578	0.2301	0.2381
월, (시간), PM_{10}, $PM_{2.5}$, 기온, 강수량, 풍향, 습도, 운량	0.1587	0.2292	0.2367
월, (시간), PM_{10}, $PM_{2.5}$, 기온, 강수량, 풍향, 풍속, 운량	0.1649	0.2566	0.2421
월, (시간), PM_{10}, $PM_{2.5}$, 기온, 강수량, 풍향, 풍속, 습도	0.2253	0.3310	0.3459
월, (시간), PM_{10}, $PM_{2.5}$, 기온, 강수량, 습도, 운량	**0.1576**	0.2280	0.2325
월, (시간), PM_{10}, 기온, 강수량, 습도, 운량	0.1583	0.2281	**0.2240**
월, (시간), $PM_{2.5}$, 기온, 강수량, 습도, 운량	0.1587	0.2293	0.2354
월, (시간), PM_{10}, $PM_{2.5}$, 기온, 습도, 운량	0.1578	0.2264	0.2400
월, (시간), PM_{10}, 기온, 습도, 운량	0.1585	0.2264	0.2329
월, (시간), $PM_{2.5}$, 기온, 습도, 운량	0.1589	0.2285	0.2399
월, PM_{10}, $PM_{2.5}$, 기온, 강수량, 습도, 운량	-	0.2241	-
월, PM_{10}, 기온, 강수량, 습도, 운량	-	**0.2227**	-
월, $PM_{2.5}$, 기온, 강수량, 습도, 운량	-	0.2254	-
월, PM_{10}, $PM_{2.5}$, 기온, 습도, 운량	-	0.2243	-
월, PM_{10}, 기온, 습도, 운량	-	0.2229	-
월, $PM_{2.5}$, 기온, 습도, 운량	-	0.2253	-

* 6~20시 모형의 경우 시간 변수를 제외했을 경우 오차가 커지므로 시간 변수 및 다른 변수를 제외한 경우는 고려하지 않음.
** 13시 모형의 경우 모든 데이터가 13시 기준으로 동일하므로 시간 변수는 고려하지 않음.

측을 위해서는 일사량을 먼저 예측하여야 한다. 일사량 예측 모델을 수립함에 있어서 고려해야 할 예측변수는 월, 시간, 미세먼지 농도(PM_{10}, $PM_{2.5}$), 기온, 강수량, 풍향, 풍속, 습도, 운량이다. 이 중 예측에 도움이 되는 변수와 그렇지 않은 변수를 찾아내기 위해 예측변수들의 조합을 여러 가지 경우로 나누어 분석을 수행하였으며, 각각의 경우마다 예측 알고리즘을 생성하고, 생성된 예측 모델의 일사량 예측치와 과거 일사량 관측치를 비교하여 최적 예측 모델을 선정하였다.

분석은 위에서 언급한 3가지 유형의 시간대별로 진행되었다. 먼저 미세먼지 농도를 포함한 예측변수 조합 중 일사량 예측 모델의 MAE 값이 최소인 모형을 찾는 작업을 수행하였다. 예측에 적용된 머신러닝 알고리즘 SVM의 매개변수 C 및 γ 값은 그리드서치를 통해 도출된 최적인 값을 사용하였다. 예측변수 조합별 MAE 값 결과는 <표 2>와 같다.

미세먼지 농도가 제외된 일사량 예측 모델은 앞서 구한 미세먼지 농도가 포함된 일사량 예측 모델의 최적 예측변수 중 미세먼지 농도 변수를 제외한 예측변수 조합을 적용하였다. 미세먼지 농도 포함 유무에 따른 최적의 일사량 예측 모델은 <표 3>과 같다.

3가지 시간대 유형별로 최적의 예측변수 조합이 다르긴 하지만 공통적으로 풍향, 풍속 변수가 예측변수에서 제외될 경우 일사량 예측 모델의 MAE 값이 작아지고, 미세먼지 농도 자료가 포함될 때 일사량 예측 모델의 정확도가 증가하는 것으로 나타났다. 선정된 일사량 예측 모델을 통해 예측된 일사량 데이터는 다음 절의 태양광 발전량 예측 모델의 예측변수로 사용된다.

〈표 3〉 일사량 예측 모델 성능 비교

구분	예측변수		MAE(MJ/m^2)
	공통 예측변수	미세먼지 반영	
6~20시	월, 시간, 기온, 강수량, 습도, 운량	×	0.1594
		PM_{10}, $PM_{2.5}$	0.1576
12~14시	월, 기온, 강수량, 습도, 운량	×	0.2286
		PM_{10}	0.2227
13시	월, 기온, 강수량, 습도, 운량	×	0.2279
		PM_{10}	0.2240

3. 태양광 발전량 예측 모델

본 연구에서 태양광 발전량 예측을 위해 사용한 예측변수는 월, 시간, 미세먼지 농도(PM_{10}, $PM_{2.5}$), 기온, 강수량, 풍향, 풍속, 습도, 운량, 일사량이다. 앞 절에서의 일사량 예측 모델 수립과 마찬가지로 예측에 도움이 되는 변수와 그렇지 않은 변수를 찾아내기 위해 예측변수들의 조합을 여러 가지 경우로 나누어 분석을 수행하였다. 일사량 예측과 동일하게 각각의 경우마다 예측 알고리즘을 생성하고, 생성된 태양광 발전량 예측 모델의 MAE 값을 비교를 통해 예측 성능을 비교 평가하였다.

3가지 유형의 시간대별로 태양광 발전량 예측 모델을 수립하였으며, 이때 예측변수 중 일사량 변수는 관측값을 사용하였다. 먼저 미세먼지 농도를 포함한 예측변수 조합 중 태양광 발전량 예측 성능을 MAE 값을 중심으로 비교하였다. 예측에 적용된 머신러닝 알고리즘 SVM 의 매개변수 C 및 γ 값은 그리드서치를 통해 도출된 최적인 값을 사용하였다. 예측변수 조합별 MAE 값 결과는 <표 4>와 같다.

일사량 예측 모델과 마찬가지로 풍향, 풍속 변수가 예측변수에서 제외될 때 MAE 값이 작아지는 것으로 나타났으며, 미세먼지 농도 자료는 $PM_{2.5}$를 제외하고 PM_{10}만 사용했을 때의 정확도가 더 높은 것으로 나타났다. 이는 PM_{10} 변수와 $PM_{2.5}$ 변수간 다중공선성이 높기 때문인 것으로 판단된다. 실제 전체 12,733개의 시간별 데이터에서 PM_{10} 변수와 $PM_{2.5}$ 변수 간 상관계수는 0.81로 매우 높은 수준이다.

또한 태양광 발전량 예측에 있어서 월, 시간 데이터의 변수 사용 유무에 따라 오차가 크게 변화하므로 태양광 발전량이 계절 및 시간에 크게 영향을 받음을 알 수 있고, 기상자료 중에서는 기온, 습도, 운량 데이터가 태양광 발전량 예측에 중요한 요인임을 알 수 있다.

최적의 태양광 발전량 예측 모델은 모든 예측변수 조합 중 예측성능을 MAE로 비교하여 가장 작은 MAE 값을 가지는 모델을 최적의 예측 모델로 선정하였다. 또한 미세먼지 농도가 제외된 태양광 발전량 예측 모델은 앞서 구한 미세먼지 농도가 포함된 태양광 발전량 예측 모델의 최적 예측변수 중 미세먼지 농도 변수를 제외한 예측변수 조합을 적용하였다. 이를 통해 도출된 미세먼지 농도 포함 유무에 따른 최적의 태양광 발전량 예측 모델은 <표 5>와 같으며, 미세먼지 농도 자료가 포함될 때 태양광 발전량 예측 모델의 성능이 보다 뛰어남을 알 수 있다.

〈표 4〉 미세먼지 포함한 예측변수 조합별 태양광 발전량 예측 성능

예측변수	MAE(kWh)		
	6~20시*	12~14시	13시**
월, (시간), PM_{10}, $PM_{2.5}$, 기온, 강수량, 풍향, 풍속, 습도, 운량, 일사량	4.3623	9.5626	9.3797
월, (시간), PM_{10}, 기온, 강수량, 풍향, 풍속, 습도, 운량, 일사량	4.3532	9.5294	9.1519
월, (시간), $PM_{2.5}$, 기온, 강수량, 풍향, 풍속, 습도, 운량, 일사량	4.3733	9.6383	9.3013
(시간), PM_{10}, $PM_{2.5}$, 기온, 강수량, 풍향, 풍속, 습도, 운량, 일사량	4.7969	9.7702	9.3614
월, PM_{10}, $PM_{2.5}$, 기온, 강수량, 풍향, 풍속, 습도, 운량, 일사량	6.4219	9.7192	-
월, (시간), PM_{10}, $PM_{2.5}$, 강수량, 풍향, 풍속, 습도, 운량, 일사량	4.4544	9.6979	9.4738
월, (시간), PM_{10}, $PM_{2.5}$, 기온, 풍향, 풍속, 습도, 운량, 일사량	4.3658	9.5725	9.3662
월, (시간), PM_{10}, $PM_{2.5}$, 기온, 강수량, 풍속, 습도, 운량, 일사량	4.2833	9.5835	9.2215
월, (시간), PM_{10}, $PM_{2.5}$, 기온, 강수량, 풍향, 습도, 운량, 일사량	4.3663	9.5747	9.3731
월, (시간), PM_{10}, $PM_{2.5}$, 기온, 강수량, 풍향, 풍속, 운량, 일사량	4.4187	9.6384	9.4466
월, (시간), PM_{10}, $PM_{2.5}$, 기온, 강수량, 풍향, 풍속, 습도, 일사량	4.7904	11.3515	10.6465
월, (시간), PM_{10}, $PM_{2.5}$, 기온, 강수량, 습도, 운량, 일사량	4.2759	9.5476	9.1397
월, (시간), PM_{10}, 기온, 강수량, 습도, 운량, 일사량	4.2463	**9.4162**	8.9178
월, (시간), $PM_{2.5}$, 기온, 강수량, 습도, 운량, 일사량	4.2934	9.6167	9.0570
월, (시간), PM_{10}, $PM_{2.5}$, 기온, 습도, 운량, 일사량	4.2689	9.5708	9.1394
월, (시간), PM_{10}, 기온, 습도, 운량, 일사량	**4.2362**	9.4350	**8.8602**
월, (시간), $PM_{2.5}$, 기온, 습도, 운량, 일사량	4.2815	9.6252	9.0531
월, PM_{10}, $PM_{2.5}$, 기온, 강수량, 습도, 운량, 일사량	-	9.6872	-
월, PM_{10}, 기온, 강수량, 습도, 운량, 일사량	-	9.7073	-
월, $PM_{2.5}$, 기온, 강수량, 습도, 운량, 일사량	-	9.7754	-
월, PM_{10}, $PM_{2.5}$, 기온, 습도, 운량, 일사량	-	9.6988	-
월, PM_{10}, 기온, 습도, 운량, 일사량	-	9.6261	-
월, $PM_{2.5}$, 기온, 습도, 운량, 일사량	-	9.7312	-

* 6~20시 모형의 경우 시간 변수를 제외했을 경우 오차가 커지므로 시간 변수 및 다른 변수를 제외한 경우는 고려하지 않음.
** 13시 모형의 경우 모든 데이터가 13시 기준으로 동일하므로 시간 변수는 고려하지 않음.

〈표 5〉 태양광 발전량 예측 모델 성능 비교

구분	예측변수		MAE(kWh)
	공통 예측변수	미세먼지 반영	
6~20시	월, 시간, 기온, 습도, 운량, 일사량	×	4.2678
		PM_{10}	4.2362
12~14시	월, 시간, 기온, 강수량, 습도, 운량, 일사량	×	9.4776
		PM_{10}	9.4162
13시	월, 기온, 습도, 운량, 일사량	×	9.0417
		PM_{10}	8.8602

4. 태양광 발전량 예측 모델 성능 비교

지금까지 3가지 유형의 시간대별로 일사량 예측 모델 및 태양광 발전량 예측 모델을 살펴보았고, 본 절에서는 미세먼지 농도를 예측변수에 함께 고려했을 경우와 그렇지 않을 경우의 태양광 발전량 예측 모델의 성능을 비교 분석한다. 현재 기상청에서는 일사량 예보자료를 제공하지 않으므로, 본 연구에서는 태양광 발전량 예측을 위해서 먼저 주어진 기상정보를 이용하여 일사량을 예측하고, 이 예측된 일사량을 태양광 발전량 예측모델의 예측변수 데이터로 사용하였다. 이는 SVR을 이용하여 24시간 앞의 태양광 발전량을 예측한 이강혁과 김우제(2016)의 연구와 동일한 접근법이라고 할 수 있다.

앞에서 도출한 3가지 시간대별 최적의 예측변수 조합을 이용한 태양광 발전량 예측 모델을 적용하여 예측값과 실제값을 비교하여 MAE를 구하였다. 미세먼지 농도를 반영했을 경우와 반영하지 않았을 경우의 성능 차이를 비교하였으며, 그 결과는 <표 6>과 같다. <표 6>에서 6~20시 태양광 발전량 예측 모델의 MAE가 12~14시, 13시 태양광 발전량

〈표 6〉 시간대별 태양광 발전량 예측 모델 성능 비교 결과

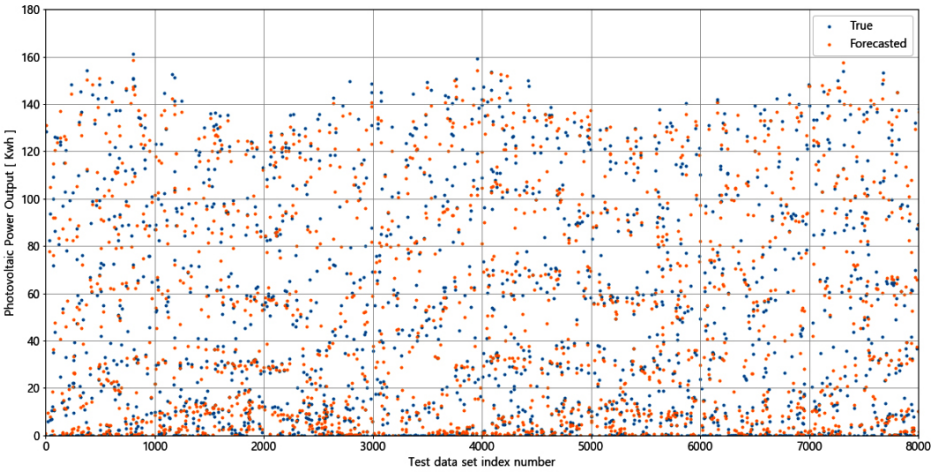
구분	발전량 MAE(kWh)		미세먼지 미반영 대비 오차 감소율
	미세먼지 반영	미세먼지 미반영	
6~20시	6.1867	6.2765	1.43%
12~14시	12.2203	12.6764	3.60%
13시	12.4539	12.9565	3.88%

예측 모델의 MAE 대비 상대적으로 낮아 성능이 더 우수함을 알 수 있다. 이는 6~20시 태양광 발전량 예측 모델의 예측 변수 중 시간 변수가 포함됨에 따라 다양한 시간대의 대량의 학습 데이터가 분석에 사용되면서 예측 성능이 상대적으로 향상된 결과라 판단된다.

태양광 발전이 시작되는 6시부터 종료되는 20시까지의 발전량 예측 모델에서 미세먼지 농도를 반영했을 경우가 미반영했을 경우보다 오차가 1.43% 감소하였다. 태양광 발전량이 많은 주간 12~14시 시간대의 발전량 예측 모델에서는 미세먼지 농도를 반영했을 경우 오차가 3.6% 감소하였으며, 태양광 발전량이 최대인 13시 시간대의 발전량 예측 모델에서는 미세먼지 농도를 반영했을 경우 오차가 3.88% 감소하였다. 즉, 미세먼지 농도를 고려한 태양광 발전량 예측 모델이 미세먼지를 고려하지 않은 모델에 비해 예측 성능이 우수하다는 결론을 얻을 수 있다. 또한 시간대별로 태양광 발전량 예측 성능을 비교할 때 태양광 발전이 일어나는 전체 시간대(6~20시)보다 태양광 발전량이 비교적 많은 주간 시간대(12~14시, 13시)의 발전량을 예측할 경우 미세먼지 농도를 반영하는 모델이 예측 성능이 더 뛰어남을 알 수 있다.

미세먼지 농도 데이터와 일사량 예측값을 반영한 태양광 발전량 예측 모델의 평가 데이터에 대한 실제값과 예측값을 그래프로 나타내면 <그림 5>, <그림 6>, <그림 7>과 같다. 가로축은 랜덤으로 샘플링된 평가 데이터 세트의 인덱스 번호를 의미하고 세로축은

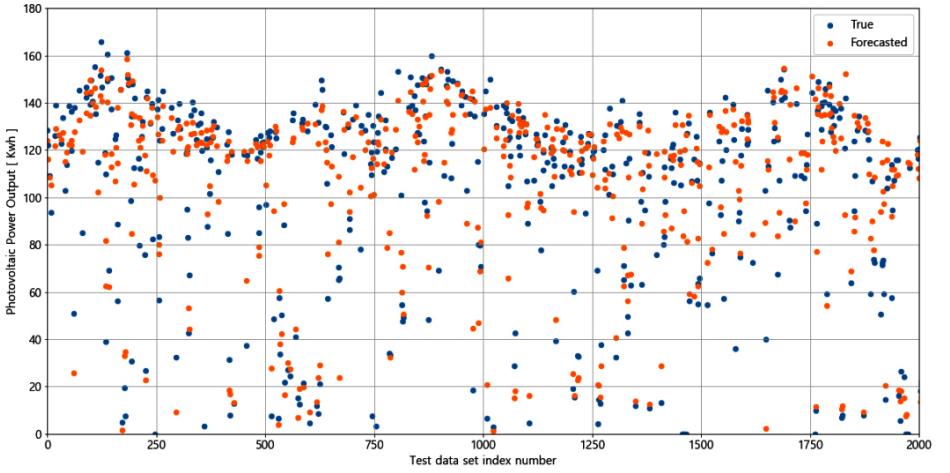
〈그림 5〉 6~20시 태양광 발전량 예측 모델 그래프(일사량 예측값, 미세먼지 농도 반영)



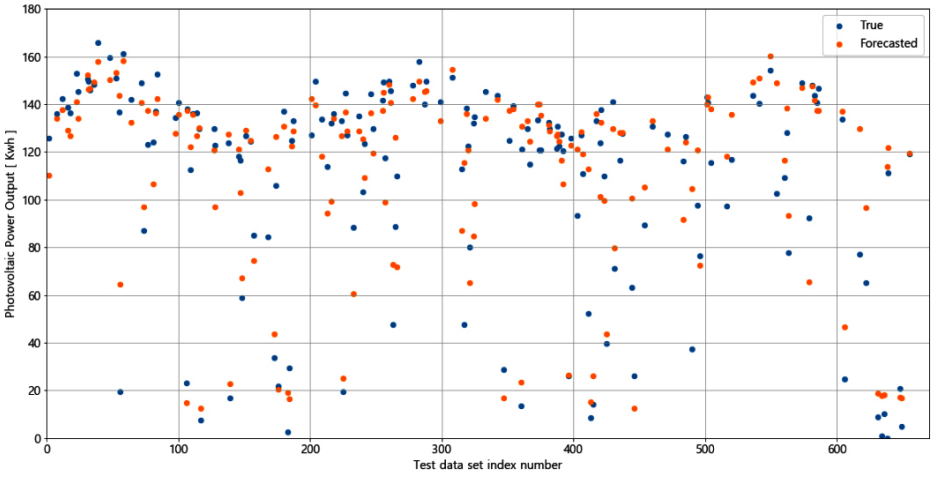
태양광 발전량을 나타낸다. 여기서 청색 포인트는 실제값, 주황색 포인트는 예측값을 나타내며, 동일한 세로축에 대해서 청색 포인트와 주황색 포인트의 간격은 태양광 발전량의 오차를 의미한다.

샘플 수가 적은 12~14시, 13시 예측 모델을 살펴보면 발전량이 큰 경우 대비 작은 경

〈그림 6〉 12~14시 태양광 발전량 예측 모델 그래프(일사량 예측값, 미세먼지 농도 반영)



〈그림 7〉 13시 태양광 발전량 예측 모델 그래프(일사량 예측값, 미세먼지 농도 반영)



우가 예측값과 실제값의 오차가 비교적 크다는 것을 알 수 있다. 이는 태양광 발전량 예측 모델을 생성하기 위해 사용되는 학습 데이터의 불균형성 때문인 것으로 판단된다. 현재 본 연구에 사용한 데이터는 태양광 발전 성능이 준수한 경우의 기상조건과 태양광 발전 성능이 저조한 경우의 기상조건 데이터를 포함하고 있으며, 기상조건이 양호한 날이 그렇지 않은 날 대비 상대적으로 많은 비중을 차지하고 있다. 즉, 기상상태가 나쁜 날의 학습 데이터 수가 상대적으로 적어 해당조건에 대해 태양광 발전량 예측 모델이 충분히 학습하지 못한 결과라고 판단된다.

V. 결 론

전 세계적으로『신 기후체제』에 대응하기 위하여 온실가스 배출을 줄이기 위한 노력을 하고 있다. 우리나라도 에너지 분야의 온실가스 배출을 감소시키기 위한 다양한 정책을 추진하고 있으며, 특히 국내에서는 태양광 발전량의 비중이 크게 확대될 것으로 전망되고 있다. 하지만 태양광 발전과 같은 신재생 에너지는 출력을 예측하기 어려운 불확실성과 출력의 변화폭이 큰 변동성을 지니고 있어 전력계통에 병입될 때 전력계통의 유연성을 저해할 수 있다. 따라서 신재생 에너지로 인한 전력계통의 안정도 저하를 방지하기 위해서는 신재생 에너지 발전량의 사전 예측이 매우 중요하다고 할 수 있다.

이러한 상황에서 본 연구는 기상자료를 활용하여 머신러닝 기반의 태양광 발전량 예측 모델을 수립하였으며, 특히 최근 사회 환경 분야에서 큰 이슈가 되고 있는 미세먼지의 영향을 함께 고려하였다. 2016년 1월 1일부터 2018년 9월 30일까지 기간 동안의 발전량, 기상자료, 미세먼지 농도자료를 머신러닝 알고리즘에 사용하였으며, 예측 모형은 여러 선행연구에서 머신러닝 지도학습 알고리즘 중 비선형 문제에 비교적 성능이 우수하다고 알려진 RBF 커널 함수 기반의 SVM 모델을 적용하였다. 보다 정확한 태양광 발전량 예측을 위해서는 향후 다양한 예측 모형 및 기법들과 본 연구 결과를 비교 분석하는 추가적인 연구가 필요하다고 판단된다.

예측변수에 미세먼지 농도를 포함한 태양광 발전량 예측 모델과 미세먼지 농도를 제외한 예측 모델의 성능을 평가지표 MAE를 사용하여 서로 비교한 결과 최종적으로 미세먼지 농도를 반영한 발전량 예측 모델의 성능이 더 우수하다는 것을 확인하였다. 미세먼

지를 고려한 예측 모형은 미세먼지를 고려하지 않은 예측 모형 대비 6~20시 예측 모형에서는 1.43%, 12~14시 예측 모형에서는 3.60%, 13시 예측 모형에서는 3.88%만큼 오차가 감소하였다. 특히 태양광 발전량이 많은 주간 시간대에 발전량을 예측할 경우 성능이 더 좋다는 사실을 확인하였다.

본 연구는 기상자료를 활용한 머신러닝 기반의 태양광 발전량 예측 모델에서 미세먼지의 농도를 함께 고려하였다는 점에서 큰 의의가 있다. 특히 태양광 발전량 예측에서 미세먼지 농도자료를 반영할 경우 PM_{10}, $PM_{2.5}$ 데이터 모두 사용하는 것보다 PM_{10} 데이터만을 사용하는 것이 오히려 예측 성능이 더 좋은 것으로 나타났다. 또한 본 연구에서는 최적의 일사량 및 발전량 예측 모델을 얻기 위해서 기상자료로 구성된 다양한 예측변수 조합을 검토하였으며, 기상자료 변수 중 풍향, 풍속 정보는 예측에 있어서 오히려 성능을 떨어뜨리는 요소로 확인되었다. 반면에 기온, 습도, 운량 정보는 태양광 발전량 예측에 있어 반드시 필요한 요소로 나타났다. 이러한 결과는 향후 태양광 발전량 예측을 수행하는 다른 알고리즘의 개발 및 적용 시 유용하게 활용될 수 있을 것으로 기대된다. 본 연구에서 도출된 결론을 바탕으로 미세먼지 농도 및 기상자료를 활용하여 태양광 사전 예측 모델을 수립한다면 태양광 발전량 예측 정확도를 제고할 수 있을 것이며, 국내 전력계통의 안정화에 어느 정도 기여할 수 있을 것이라 판단된다.

하지만 본 연구는 태양광 발전량 예측 모델을 수립하는 과정에서 일부 한계점을 가지고 있다. 본 연구에서 일사량 및 태양광 발전량을 예측하는 모델을 수립하는 과정에서 과거 1시간 단위로 구성된 기상자료 및 미세먼지 관측 데이터를 활용하였다. 하지만 현재 기상정보 및 미세먼지 농도 예보자료는 3시간 또는 6시간 단위로 발표되고 1시간 단위의 수치로 제공되지 않아 1시간 단위의 태양광 발전량 단기 예측에 일정 정도 어려움이 있는 실정이다. 물론 현재의 예보자료를 머신러닝 예측 알고리즘에 적용하면 태양광 발전량 예측이 가능하지만, 어느 정도의 오차가 발생할 수 있다. 이를 해결하기 위해 3시간 또는 6시간 단위의 데이터에 보간법을 적용하여 각 1시간 단위의 추정된 값을 사용하면 오차를 줄일 수 있지만, 궁극적인 대안으로 향후 기상청 및 관련 기관들이 기상정보 및 미세먼지 농도자료를 1시간 단위의 수치화 데이터 형태로 제공할 수 있다면 본 연구의 한계점을 어느 정도 극복할 수 있을 것이다.

[References]

관계부처합동, “2030년 국가 온실가스 감축목표 달성을 위한 기본 로드맵 수정안”, 2018.

관계부처합동, “미세먼지 관리 종합대책”, 2017.

구경민, “Support Vector Machine 을 이용한 micro array gene expression data 의 분류”, Doctoral dissertation, 연세대학교 대학원, 2002.

국회 산업통상자원중소 벤처기업위원회 소속 김삼화 의원실, “미세먼지 높았던 3.1-6일과 직전 6일 발전량 비교”, 2019.

김광인 · 김현수 · 조인구, “에너지 전환정책과 발전의 사회적 비용 -제7차와 제8차 전력수급기본계획 비교- ”, 「자원 · 환경경제연구」, 제28권 제1호, 2019, pp. 147~176.

문성은 · 장수범 · 이정혁 · 이종석, “기계학습 및 딥러닝 기술동향”, 「한국통신학회지 (정보와통신)」, 제33권 제10호, 2016, pp. 49~56.

배국열 · 장한승 · 성단근, “기계학습 기반의 태양광 출력예측 및 예측 오차 분석”, 「한국통신학회 학술대회논문집」, 2017, pp. 13~14.

산업통상자원부, “재생에너지 3020 이행계획”, 2017.

산업통상자원부, “제8차 전력수급기본계획”, 2017.

서미숙 · 조홍종, “미세먼지가 삶의 만족도에 미치는 영향:WTP 추정을 중심으로”, 「자원 · 환경경제연구」, 제26권 제3호, 2017, pp. 417~449.

서유진, “미세먼지 심한 날, 태양광 발전량 19% 줄어”, 「중앙일보」, 2019.3.13.

손정훈 · 정수종, “태양광 발전량에 영향을 미치는 미세먼지에 관한 연구”, 「한국환경정책학회 학술대회논문집」, 2019, pp. 52~54.

송재주 · 정윤수 · 이상호, “태양광 발전을 위한 발전량 예측 모델 분석”, 「디지털융복합연구」, 제12권 제3호, 2014, pp. 243~248.

안드레아스 뮐러 · 세라 가이도, 『파이썬 라이브러리를 활용한 머신러닝』, 한빛미디어, 2017.

에너지경제연구원, “신재생에너지 보급 확산을 대비한 전력계통 유연성 강화방안 연구”, 2017.

에어코리아[웹사이트], (2019.9.1), https://www.airkorea.or.kr/web/airMatter?pMENU_NO=130.

이강혁 · 김우제, “서포트 벡터 회귀를 이용한 24시간 앞의 태양광 발전량 예측”, 「한국정보기술학회 논문지」, 제14권 제3호, 2016, pp. 175~183.

이건주 · 이기현 · 강성우, “미세먼지와 기상정보 기반의 AHP 분석을 통하여 태양광 발전소 최

적입지선정에 대한 사례연구", 「대한안전경영과학회지」, 제19권 제4호, 2017, pp. 157~167.

이순환 · 김해동 · 조창범, "국지 기상 요소에 의한 태양광 발전량 변동특성에 관한 연구", 「한국환경과학회지」, 제23권 제11호, 2014, pp. 1943~1951.

조덕기 · 이태규 · 전일수 · 전홍석, 오정무, "고산지대의 일사량 특성분석", 「한국태양에너지학회 논문집」, 제16권, 제2호, 1996, pp. 49~63.

조성준 · 강석호, "머신러닝(인공지능)의 산업 응용", 「ie 매거진」, 제23권 제2호, 2016, pp. 34~38.

차왕철 · 박정호 · 조욱래 · 김재철, "지리 · 지형 · 기상자료를 활용한 태양광발전량 예측에 관한 연구", 「한국조명 · 전기설비학회 2014 춘계학술대회 논문집」, 2014, pp. 211~212.

한진목 · 최수광 · 김세웅 · 정영관, "공기 중의 미세먼지에 의한 태양전지의 오염에 관한 연구", 「한국수소 및 신에너지학회 논문집」, 제29권 제3호, 2018, pp. 292~298.

환경부, "교토의정서 이후 신 기후체제 파리협정 길라잡이", 2016.

환경부, "바로 알면 보인다. 미세먼지, 도대체 뭘까?", 2016.

Cortes, C., and V. Vapnik, "Support-vector networks," *Machine Learning*, Vol. 20, No. 3, 1995, pp. 273~297.

da Silva Fonseca Jr, J. G., T. Oozeki, T. Takashima, G. Koshimizu, Y. Uchida, and K. Ogimoto, "Use of support vector regression and numerically predicted cloudiness to forecast power output of a photovoltaic power plant in Kitakyushu, Japan," *Progress in photovoltaics: Research and applications*, Vol. 20, No. 7, 2012, pp. 874~882.

Das, U. K., K. S. Tey, M. Seyedmahmoudian, S. Mekhilef, M. Y. I. Idris, W. Van Deventer, B. Horan, and A. Stojcevski, "Forecasting of photovoltaic power generation and model optimization: A review," *Renewable and Sustainable Energy Reviews*, Vol. 81, 2018, pp. 912~928.

Drucker, H., C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," *In Advances in neural information processing systems*, 1997, pp. 155~161.

Mei, H. W., and J. J. Ma, "Photovoltaic Power Generation Forecasting Model with Improved Support Vector Machine Regression Based on Rough Set and Similar Day," *In Advanced Materials Research*, Vol. 805, 2013, pp. 114~120.

Müller, K. R., S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE transactions on neural networks*, Vol. 12, No. 2, 2001, pp.

181~201.

Müller, K. R., A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, "Predicting time series with support vector machines," *In International Conference on Artificial Neural Networks*, Springer, Berlin, Heidelberg, 1997, pp. 999~1004.

Nasien, D., S. S. Yuhaniz, and H. Haron, "Statistical learning theory and support vector machines," *In 2010 Second International Conference on Computer Research and Development*, 2010, pp. 760~764.

OECD.Stat[Website], (2019.9.1), https://stats.oecd.org/index.aspx?queryid=72722.

Ramli, M. A., S. Twaha, and Y. A. Al-Turki, "Investigating the performance of support vector machine and artificial neural networks in predicting solar radiation on a tilted surface: Saudi Arabia case study," *Energy conversion and management*, 105, 2015, pp. 442~452.

Sharma, N., P. Sharma, D. Irwin, and P. Shenoy, "Predicting solar generation from weather forecasts using machine learning," *In 2011 IEEE international conference on smart grid communications (SmartGridComm)*, 2011, pp. 528~533.

Shi, J., W. J. Lee, Y. Liu, Y. Yang, and P. Wang, "Forecasting power output of photovoltaic systems based on weather classification and support vector machines," *IEEE Transactions on Industry Applications*, Vol. 48, NO. 3, 2012, pp. 1064~1069.

Smola, A. J., and B. Schölkopf, "A tutorial on support vector regression. Statistics and computing," *Statistics and Computing*, Vol. 14, No. 3, 2004, pp. 199~222.

Twomey, S., "The influence of pollution on the shortwave albedo of clouds," *Journal of the atmospheric sciences*, Vol. 34, No. 7, 1977, pp. 1149~1152.

Vakili, M., S. R. Sabbagh-Yazdi, S. Khosrojerdi, and K. Kalhor, "Evaluating the effect of particulate matter pollution on estimation of daily global solar radiation using artificial neural network modeling based on meteorological data," *Journal of cleaner production*, Vol. 141, 2017, pp. 1275~1285.

Vapnik, V., The nature of statistical learning theory, Springer science & business media, 1995.