데이터 개방 활용을 위한 태양광 발전 데이터 비식별화 기법

김민호, 홍석재, 임정택, 함경선, 김태형*
한국전자기술연구원
minhokim@keti.re.kr, sjhong@keti.re.kr, jtlim@keti.re.kr
ksham@keti.re.kr, thkim@keti.re.kr*

Anonymization Strategy for Solar Energy Data in Open Sharing Environments

Minho Kim, Seokjae Hong, Jeongtaek Lim, Kyung Sun Ham, Taehyoung Kim*
Korea Electronics Technology Institute

요 약

본 논문은 태양광 발전소의 시계열 데이터를 다수의 외부 연구자와 공유할 때 발생할 수 있는 지역 및 특성 식별 가능 문제를 해결하기 위해, 유효전력, 무효전력, 전류 항목을 중심으로 한 비식별화 기법을 제안한다. 고해상도의 전력 데이터는 인공지능 기반 예측과 제어 기술 개발에 핵심적이지만, 발전소의 위치나 규모가 간접적으로 노출될 수 있어 공공 데이터 개방 및 학술 공유에 제약이 따른다. 이에 따라 본 연구는 각 항목의 전기적 특성과 데이터 분포에 따라 맞춤형 정규화 기법을 적용하였으며, 전압은 재식별 위험도가 낮아 비식별화 대상에서 제외하였다. 또한, 비식별화 전후 데이터를 이용해 머신러닝 기반 예측 실험을 수행한 결과, 데이터 활용성의 손실 없이 정보 보존이 가능함을 확인하였다.

Ⅰ. 서 론

최근 재생에너지의 보급 확대와 더불어, 태양광 발전소의 운영 데이터를 활용한 다양한 인공지능 기반 분석, 예측, 최적 제어 기술의 개발이 활발히 이루어지고 있다. 이러한 기술은 실시간 모니터링, 고장 진단, 예측 제어 등 고도화된 서비스를 가능하게 하며, 이를 위해 1분 또는 1초 단위의 고해상도 발전량, 전력 품질, 기상 정보 등을 포함하는 시계열 데이터가 핵심 자산으로 강조되고 있다. 이러한 고정밀 데이터는 데이터 생산자의 보안이 있어야 하는 경우가 대부분이므로, 기술 발전을 위해 개방된 환경에서의 공유가 제한된다. 따라서 본 연구에서는 태양광 데이터 허브 구축을 전제로, 데이터를 외부와 공유하되 개별 발전소나 지역의 식별이 불가능하게 하기 위한 비식별화 기법을 적용하였다. 특히, 유효전력(Active Power), 무효전력(Reactive Power), 전류(Current) 등의 전력 관련 주요 항목에 대해 특성에 맞는 스케일링 및 구조적 변환을 수행하였으며, 전압(Voltage)은 전기적 특성과 영향력을 고려하여 비식별화 대상에서 제외하였다. 본 논문은 이러한 전력 데이터 비식별화 방안의 설계 근거와 적용 결과를 제시하고, 데이터 활용성과 보호 수준 간의 균형을 조율한 방안을 제시한다.

Ⅱ. 본론

1. 비식별화의 필요성과 목적

태양광 발전소 데이터는 특정 시간대의 발전 출력, 전압, 전류, 무효전력 등 다양한 전력 품질 요소를 포함한다. 이러한 데이터를 시간 단위로 정밀하게 수집하는 경우, 누적된 시계열 데이터를 통해 기상 패턴, 일조 시간, 출력 분포 특성 등을 기반으로 발전소의 지리적 위치 추론이 가능하다[1][2]. 특히 국내처럼 지역별 일사량 패턴이 뚜렷한 경우, 발전량 곡선만으로도 데이터 출처 지역 및 발전소를 식별할 가능성이 있다.

본 연구에서는 국가 또는 지자체 차원에서 구축된 데이터 허브에 태양광 발전 데이터가 외부 기관이나 산업체와의 협력을 목적으로 공유되는 시나리오를 가정하였다. 이 과정에서 데이터를 활용하는 예비 연구자의 연구 효용성을 보장하되, 데이터의 익명성을 보장하기 위하여 출력 특성을 정규화하여 설비 추정을 어렵게 하기 위한 방식을 도입하였다. 이는 단순한 개인정보 보호의 차원을 넘어, 지역 기반 식별 가능성 자체를 제거하고, 공정한 인공지능 학습 환경을 제공하기 위한 비식별화 목적이다.

2. 항목별 비식별화 처리 전략

본 연구에서는 태양광 발전 데이터 중에서도 식별 가능성이 높고 물리적 해석이 중요한 항목들을 선별하여 비식별화를 수행하였다. 그 대상은 유효전력, 일(Day) 누적 발전량, 무효전력, 전류 항목이며, 각각의 전기적 특성과 데이터 분포 특성을 고려해 정규화 방식과 범위를 다르게 설정하였다. 반면, 전압은 재식별 위험도가 낮고, 전력계통의 안정성 유지를 위해 실질적인 수치 보존이 필요하므로 비식별화 대상에서 제외하였다.

유효전력은 태양광 인버터를 통해 계통에 실질적으로 공급되는 순간 전력이다. 일반적으로 발전기의 최대 출력에 가까운 값을 가지며, 장비 용량 및 시스템 크기를 직접적으로 반영한다. 정규화 전의 유효전력 데이터를 그대로 공개할 경우, 발전소의 최대 출력 용량 유추가 가능하다. 유효전력은 전체 데이터 구간에 대하여 최소 및 최대 범위를 기준으로 Min-Max 정규화를 적용하였다. 정규화 범위는 0에서 100 사이의 비율 값으로 설정하여, 발전량의 상대적인 추세는 유지하되 실제 절댓값은 유추할 수 없도록 하였다.

누적 발전량은 하루 동안 발생한 전력량의 누적값이며, 일사량 총합 또는 지역 기상 조건을 기반으로 하루 최대 발전량을 유추할 수 있다. 하루 최종 누적값이 크면 발전량이 높은 지역임을 추론 가능하며, 이는 지역 식별의 단서가 될 수 있다. 하루 단위로 데이터를 분할 후, 각 날짜의 최대 누적 발전량을 기준으로 해당 하루치 데이터를 정규화하였다. 이를 통해 증가 패턴은 유지하되, 발전 총량 수치는 숨길 수 있도록 설계하였다.

무효전력은 전압 유지와 위상 보정을 위한 전력이다. 실제 에너지로 소비되지는 않으나, 전력계통의 운전 안정성을 결정짓는 주요 요소이다. 특히 부호가 존재하므로, 부호 정보는 그대로 유지해야 한다. 무효전력은 Min-Max 정규화 대신 절댓값 기준의 MaxAbs 정규화를 적용하였다. 정규화 범위는 -100에서 100까지의 비율 값 범위로 설정하여 부호 정보는 그대로 유지된다. 이렇게 하면 공급 및 흡수 특성을 학습에 사용할 수 있으면서도, 수치적 크기는 비식별화할 수 있다.

전류는 태양광 인버터의 직류 출력 또는 3상 교류 전류로, 출력 전력 및 인버터 용량과 비례한다. 특히 AC 전류는 발전소의 최대출력과 전압과의 관계를 통해 유효전력을 추정할 수 있는 간접 정보가 된다. 각 전류 항목에 대해 개별적으로 Min-Max 정규화를 수행하였으며, 정규화 범위는 모두 0에서 100까지의 비율 값으로 설정하였다. 전류는 값의 절대적 크기보다는 변화율 혹은 부하 흐름과 같은 패턴이 중요하므로, 정규화를 통해 설비 유추 위험을 줄이면서 데이터의 분석 가치는 유지하였다.

항목	정규화 수식	정규화 범위
유효전력	$P_{norm} = \frac{P - P_{min}}{P_{max} - P_{min}} \times 100$	0 ~ 100
누적 발전량	$E_{norm}(t) = \frac{E(t)}{E_{max}} \times 100$	0 ~ 100
무효전력	$Q_{norm} = \frac{Q}{\lvert Q_{max} \rvert} \times 100$	-100 ~ +100
전류	$I_{norm} = \frac{I - I_{min}}{I_{max} - I_{min}} \times 100$	0 ~ 100

표 1 항목별 비식별화 방법

3. 비식별화에 따른 데이터 활용성 비교

본 연구에서는 정규화 기반 비식별화 기법이 태양광 발전 데이터의 활용 가능성에 미치는 영향을 정량적으로 평가하기 위해, 비식별화 처리 전후의 데이터를 활용하여 동일한 조건으로 발전량 예측 모델을 학습하고 그 성능을 비교하였다.

실험 대상은 1분 단위 유효전력 데이터와 수치예보모델(Numerical Weather Prediction, NWP) 기반의 기상 데이터를 결합한 시계열 데이터 셋으로, 종속 변수는 1시간 단위로 집계된 태양광 발전량이다. 이를 위해 원본 및 비식별화 데이터의 kW 단위의 값을 kWh 단위로 계산하고, 60분 단위로 시간 해상도를 변환하였다. 독립 변수는 수치예보모델에서 추출한 일사량 및 구름양 관련 특성을 사용하였으며, XGBoost 회귀 모델을 기반으로 학습을 수행하였다[3]. 모든 실험은 동일한 파라미터 조건에서 이루어졌으며, 학습 및 테스트 데이터는 일관된 방식으로 분할되었다. 예측 성능 평가는 평균 절대 오차(Mean Absolute Error, MAE)의 정규화 지표인 nMAE(normalized MAE)와 결정계수(R^2, Coefficient of Determination)를 사용하였다. nMAE는 데이터의 크기 단위 또는 값의 분포 범위에 따른 차이를 정규화한다. 따라서, 본 연구에서의 비식별화 전후 데이터의 예측 오차를 상대적인 관점에서 비교하기에 적합하다. 결정계수는 회귀 모델이 종속 변수의 분산을 독립 변수로부터 얼마나 잘 설명하는지 나타내는 지표이다. 예측 수행 결과와 성능 지표는 다음과 같다.

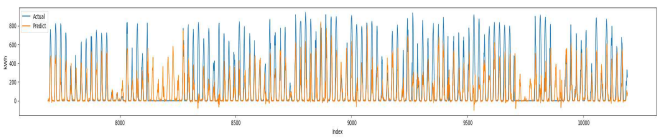

그림 1 원본 데이터의 태양광 발전량 예측 그래프

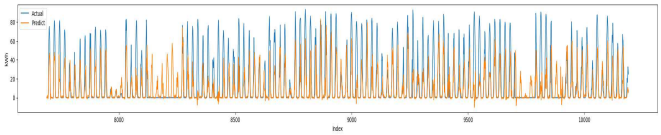

그림 2 비식별화 데이터의 태양광 발전량 예측 그래프

구분	nMAE	R^2
원본 데이터	0.65527	0.40603
비식별화 데이터	0.65530	0.40604

표 2 원본 데이터와 비식별화 데이터의 성능 지표

실험 결과, 각 모델의 성능 지표에서 소수점 셋째 자리 이하에서 극히 미세한 차이만을 보였으며, 사실상 비식별화 처리에 따른 성능 저하가 없다고 판단할 수 있다. 특히 결정계수 값이 거의 유사하다는 점에서, 정규화 처리에도 불구하고 기상 변수와 시간당 발전량 간의 상관 구조가 손상되지 않았음을 확인할 수 있다. 이는 비식별화된 데이터가 예측 모델 학습에 필요한 정보를 충분히 보존하고 있음을 의미한다.

이러한 결과는 정규화 기반의 비식별화 기법이 데이터의 재식별 위험을 효과적으로 낮추면서도, 예측 분석 및 모델링에 필요한 정보의 유용성을 유지할 수 있음을 시사한다. 특히 공공 또는 산업적 목적의 태양광 발전 데이터 공유 시, 비식별화 데이터를 통해 충분한 모델 학습과 예측이 가능함을 보여주는 경험적 사례로서 의미를 가진다.

Ⅲ. 결론

본 논문에서는 태양광 발전 데이터 중 재식별 위험도가 높은 항목에 대한 정규화 기반 비식별화 기법을 제안하였다. 특히 무효전력은 부호 보존 방식으로 정규화하고, 전류는 항목별로 값의 범위를 조정함으로써 설비 용량 추정을 어렵게 하였다. 또한, 실험을 통해 비식별화 이후에도 예측 정확도가 일정 수준 유지됨을 확인하였다. 이로써 본 연구는 정규화 기반 비식별화 기법이 정보 보호와 데이터 활용성 간의 균형을 실현할 수 있는 실효성 있는 방안임을 실험적으로 확인하였으며, 이는 향후 태양광 발전 데이터를 포함한 다양한 에너지 데이터의 공공 활용 기반 마련에 기여할 수 있을 것이다.

ACKNOWLEDGMENT

이 논문은 2023년도 정부(산업통상자원부)의 재원으로 한국에너지기술평가원의 지원을 받아 수행된 연구임(RS-2023-0023170, 분산형 재생에너지 시스템 개방형 통합 플랫폼 개발)

참 고 문 헌

[1] Heo, B. N., & Lee, J. H. (2023). An analysis methodology for the power generation of a solar power plant considering weather, location, and installation conditions. Journal of Korea Society of Industrial Information Systems, 28(6), 91-98.

[2] 김완수 and 조하현. (2019). 지역별 기상자료를 고려한 태양광발전출력 모형 연구. 전기학회논문지, 68(9), 1109-1117.

[3] Tianqi Chen and Carlos Guestrin, “XGBoost: A Scalable Tree Boosting System,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785 - 794, 2016.