기계 학습을 이용한 천리안 위성 데이터 기반 지면 및 지상 온도 예측

김자현*·김재훈**·나상우***·백대환****

*고려대학교(세종) 공공정책대학 빅데이터전공

**고려대학교(세종) 공공정책대학 빅데이터전공

***고려대학교(세종) 공공정책대학 빅데이터전공

***고려대학교(세종) 공공정책대학 빅데이터전공

Ground and Surface Temperature Prediction Using Machine Learning with Cheollian Satellite Data

요약

본 연구에서는 인공위성 자료를 활용한 지면 온도 측정 방법에 대해 서술하며, 이를 위해 다중 밴드의 자료를 알고리즘을 활용하여 산출한다. 본 연구의 목적은 데이터 기반 예측 모델을 통해 지면 온도를 예측하는 것이다. 그러나, 기존의 인공위성 자료에는 결측치가 많아, 이를 보안하기 위해 데이터 전처리가 필요하였다. 이 후에는 SVR(Support Vector Regressor), kNN(k-Nearest Neighbor), XGBoot Regressor, LGBM Regressor, LSTM(Long Short-Term Memory) 방법을 이용하여 지면 온도를 산출하였다. 이들 각 목델을 최적화한 후, 그들의 예측 성능을 비교하였다. 결과적으로, LSTM 모델이 R2 값이 0.722, RMSE ㄱ 값이 2.95 로 가장 우수하였으므로, 이 모델을 최적의 모델로 선택하였다.

Key words : Meteorological Satellite Data, SVR, kNN, XGBoost, LGBM, LSTM

1. 서론

최근 지구 온난화 현상이 증가하면서 폭염이나 한파와 같은 전 지구적 이상기후 현상이 빈번히 발생하고 있다. 이로 인해 농작물의 생산에 큰 어려움을 겪고 있는 상황이다. 특히, 토양의 온도는 식물의 생육, 미생물의 활동, 토양의 생성 과정 등에 중요한 영향을 미친다. 토양 온도가 낮아지면 유기물의 분해 속도가 늦어져서 다량의 유기물이 쌓이게 되고, 반대로 토양 온도가 높아지면 유기물의 분해가 빨라져서 무기화 작용이 촉진된다. 따라서, 식물의 성장에 있어서 지면 온도를 정확히 파악하고 적절한 대응을 하는 것이 매우 중요하다. 기존의 지면 온도 측정 방법은 온도계의 구부를 지면에 노출되지 않게 묻고 모세관의 윗부분을 약간 치켜 세우는 방식이었다. 그러나 이 방식은 날씨가 좋지 않을 때는 바람과 비로 인해 온도계의 수감 부분이 쉽게 노출되며, 일사량이나 지면 상태에 따라 크게 변동하므로 일정한 상태를 유지하며 측정하는 것이 어렵다. 이에 따라, 인공위성 데이터를 이용하여 지면 온도를 예측하면 기상 상태에 영향을 받지 않고 정확한 측정이 가능하다. 이렇게 얻어진 지면 온도

정보는 농업 기상학 뿐만 아니라 토목, 건축 등 각종 산업 분야에서 널리 활용될 것으로 기대된다.

2. 관련연구

2.1 지면 온도

지면온도는 일반적으로 맨땅 또는 짧은 잔디 밑의 온도를 의미한다. 실질적으로는 온도계의 수감부가 노출되지 않을 정도로 지면에 얇게 묻어서 측정한다. 적설이 있는 경우, 적설의 보온 효과로 인해 지면온도는 0℃ 내외로 유지되는 특성이 있다[1].

2.2 SVR(Support Vector Regressor)

SVR 은 데이터에 노이즈가 존재한다는 가정하에 동작하는 모델로, 이를 고려하여 노이즈가 포함된 실제 값을 완벽히 추정하는 것을 목표로 하지 않는다. 이는 적정 범위 내에서 실제값과 예측값의 차이를 허용하는 것을 의미한다. 본 연구에서 활용한 커널 SVR 모델은 원공간(Input space)에서의 데이터를 매핑함수 $\phi(x)$를 통해 고차원 공간(Feature space)로 변환하고, 이 고차원 공간에서 데이터를 잘 설명하는 선형회귀선을 찾는 방식으로 동작한다.

2.3 LGBM Regressor

Light GBM 은 Gradient Boosting 프레임워크를 기반으로 하는 Tree 기반 학습 알고리즘이다. 다른 알고리즘들과 차별화된 점은, Light GBM 은 Tree 가 수직적으로 확장된다는 점이다. 즉, Light GBM 은 leaf-wise 방식을 취하며, 다른 알고리즘들은 주로 level-wise 방식을 사용한다. Light GBM 은 확장을 위해 max delta loss 를 가진 leaf 를 선택하게 된다. 동일한 leaf 를 확장할 때, leaf-wise 알고리즘은 level-wise 알고리즘에 비해 더 큰 손실을 줄일 수 있다.

2.4 LSTM(Long Short-Term Memory)

LSTM 은 기존의 RNN(Recurrent Neural Network)모델의 한계인 기울기 소실(Vanishing Gradient) 문제를 해결하기 위해 개발된 모델로, 입력 게이트, 출력 게이트, 망각 게이트로 구성된 셀을 추가하여 개선되었다. LSTM 모델은 과거 학습 정보를 잘 기억하고 새로운 학습 결과에 반영할 수 있어 시계열 데이터나 예측 문제에 대한 성능이 우수하다. 본 연구에서는 기상 위성 데이터와 같은 시계열 데이터에 LSTM 모델을 적용하는 방법을 제안한다. 이를 위해 LSTM 모델의 성능 향상을 위한 활성화 함수 및 하이퍼파라미터 조정에 대한 실증적인 실험을 수행하였다. 실험 결과, 활성화 함수로는 relu, 최적화 알고리즘으로는 adam 을 사용했을 때 가장 좋은 성능을 보였다. relu 함수는 전체적인 출력이 발산하는 문제를 해결하기 위해 nomalization 과정을 통해 분석하였다[2].

3. 지면/지상 온도 예측

3.1 전처리

모델을 구축하기에 위해 변수를 선택해야 한다. 천리안 2A 호의 밴드자료를 설명변수로 두고, 지면 온도, 지상온도를 목적변수로 두었다. 설명변수로는

파랑가시밴드, 초록 가시밴드, 빨강 가시밴드, 식생 가시밴드, 권운 밴드, 눈/얼음 채널, 야간안개/하층운 밴드, 상층 수증기 밴드, 중층 수증기 밴드, 하층 수증기 밴드, 구름상 밴드, 오존 밴드, 대기창 밴드, 깨끗한 대기창 밴드, 오염된 대기창 밴드, 이산화탄소(CO2) 밴드, 30 일 청천 빨강 가시밴드, 30 일 청천 대기창밴드, 태양천정각이 사용되었다. 목적변수로 지면 온도[3]가 사용되었다. 제공된 기상위성 자료 중 목적변수로 중요한 지면 온도의 결측치가 많다. 따라서 제공된 자료 중 기상청과 겹치는 지점만 목적변수로 사용하였다.

기존 데이터는 10 분 측정단위로, 기상청 자료와 맞지 않기에, 1 시간 단위로 다운 샘플링을 진행하였다. 데이터 수가 많다고 판단되어 결측치 행은 제거하였다. 데이터는 20, 21 년 자료를 합친 것으로 총 1609323 개가 된다.

3.2 지면온도 예측 성능비교

112 번 기상 관측 지점(인천 광역시 중구 동인천동)의 데이터를 사용해서 머신러닝 모델 svr, knn, xgboost, lgbm boost 과 신경망 모델 lstm 중 21 년 겨울 test 결과 도출 후 최적 모델 선정한다. 선정한 최적 모델을 활용하여 21 년 사계절 모두 예측모델 생성한다.

3.2.1 머신러닝

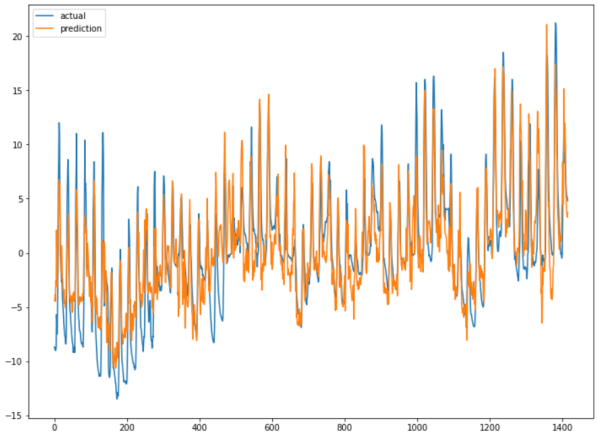
모든 모델에 pipeline 을 활용하여 standard scaler 적용했다.

(1) SVR

초모수 값을 C=50000, gamma=0.1 로 설정. 겨울 test 결과 <표-1>와 같이 도출했다. <그림-1>은 예측 값(주황)과 실제 값(파랑)의 차이를 도식화한 그래프이다.

mse	rmse	mae	r2
8.76	2.96	2.32	0.72

<표- 1> svr mse, rmse, mae, r2 결과



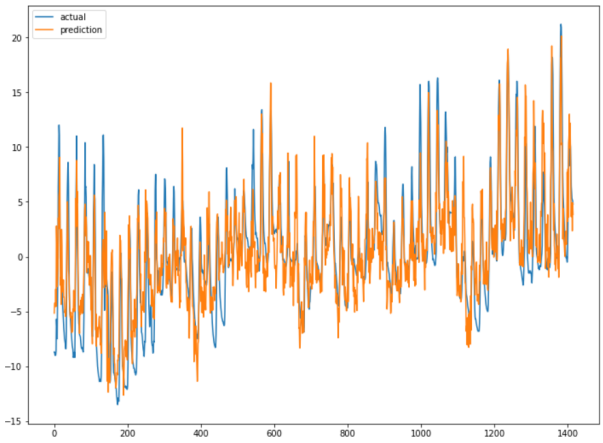
<그림- 1> svr 예측과 실제의 차이

(2) LGBM

초모수 값을 booster='gbtree', colsample_bytree=0.75, learning_rate=0.1, max_depth=5, alpha=1, n_estimators=10000 로 설정했다. 겨울 test 결과 <표-4>와 같이 도출했다. <그림-4>은 예측 값(주황)과 실제 값(파랑)의 차이를 도식화한 그래프이다.

mse	rmse	mae	r2
9.03	3.00	2.34	0.71

<표- 2> lgbm mse, rmse, mae, r2 결과



<그림- 2> lgbm 예측과 실제의 차이

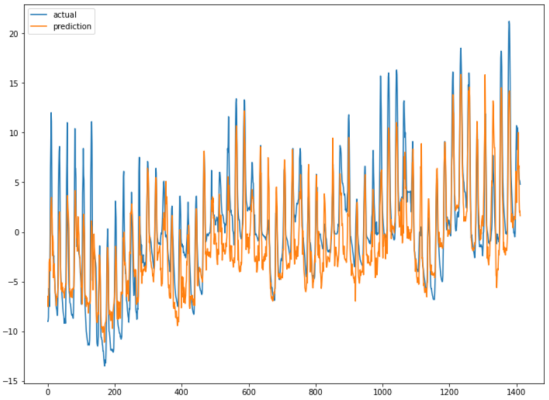
3.2.2 신경망 학습

(1) LSTM

초모수 값을 epochs=100, batch_size=24, loss='mean_squared_error'로 설정했다. 겨울 test 결과 <표-n>와 같이 도출했다. <그림-n>은 예측 값(주황)과 실제 값(파랑)의 차이를 도식화한 그래프이다.

mse	rmse	mae	r2
8.70	2.95	2.37	0.72

<표- 3> lstm mse, rmse, mae, r2 결과



<그림- 3> lstm 예측과 실제의 차이

3.2.3 최적 모델 선정

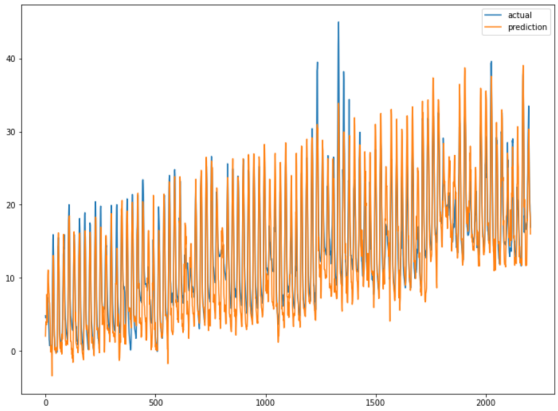
3 가지 모델 비교 결과 lstm 신경망 학습이 가장 좋은 성능을 가진 모델이라고 판단했다. 따라서 이어지는 목차는 lstm 모델로 21 년 봄, 여름, 가을 데이터로 예측을 실행해 보았다.

3.3 지면온도 예측 LSTM 모델링

(1) 봄 (21 년 3 월, 4 월, 5 월)

mse	rmse	mae	r2
9.45	3.07	2.38	0.83

<표- 4>lstm 봄 mse, rmse, mae, r2 결과

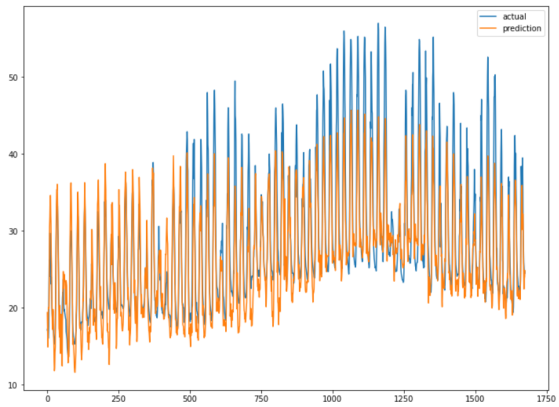


<그림- 4> lstm 봄 예측과 실제의 차이

(2) 여름 (21 년 6 월, 7 월, 8 월)

mse	rmse	mae	r2
18.45	4.29	3.14	0.76

<표- 5> lstm 여름 mse, rmse, mae, r2 결과

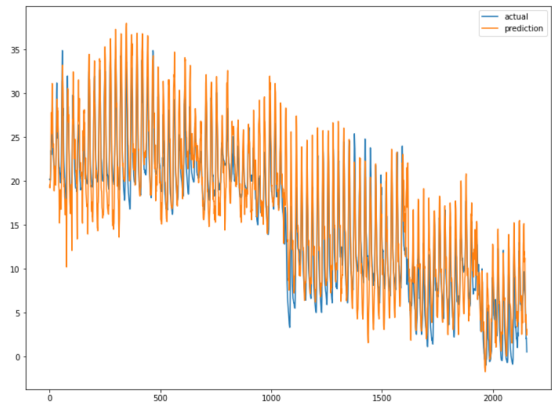


<그림- 5> lstm 여름 예측과 실제의 차이

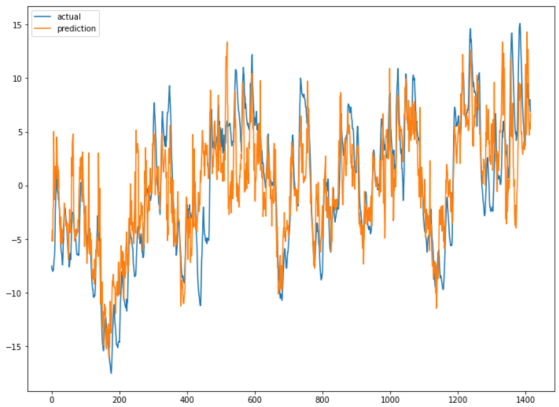
(3) 가을 (21 년 9 월, 10 월, 11 월)

mse	rmse	mae	r2
9.51	3.08	2.46	0.85

<표- 6> lstm 가을 mse, rmse, mae, r2 결과



<그림- 6> lstm 가을 예측과 실제의 차이



<그림- 7> svr 지상온도 예측과 실제의 차이

3.4 실험결과

머신러닝 모델을 적용한 결과인 <표-1>, <표-2>, <표-3>, <표-4>의 내용과 신경망 모델을 적용한 결과인 <표-5>의 내용을 보았을 때 신경망 모델(LSTM)이 가장 결과가 좋았다. 따라서 2020 년도 데이터를 학습 시킨 LSTM 모델에 2021 년 봄, 여름, 가을, 겨울 데이터를 시험 데이터로 넣어 결과를 도출했다. 시험 결과 결정 계수 r2 가 평균 0.8 로 꽤 좋은 성능을 보였다고 판단할 수 있다.

3.5 지상온도 예측 성능비교

지면온도 예측과 마찬가지로 2021 년 겨울 (1,2 월) 데이터를 시험데이터로 활용하여 모델들의 성능을 비교했다.

3.5.1 머신러닝 모델

(1) SVR

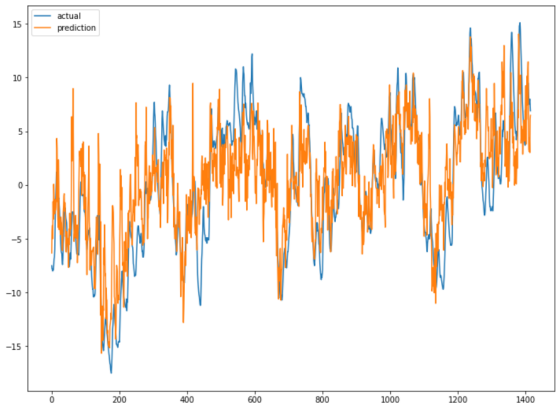
mse	rmse	mae	r2
13.22	3.63	2.79	0.68

<표- 7> svr 지상온도 예측 mse, rmse, mae, r2 결과

(2) LGBM

mse	rmse	mae	r2
14.48	3.80	2.96	0.65

<표- 8> lgbm 지상온도 예측 mse, rmse, mae, r2 결과



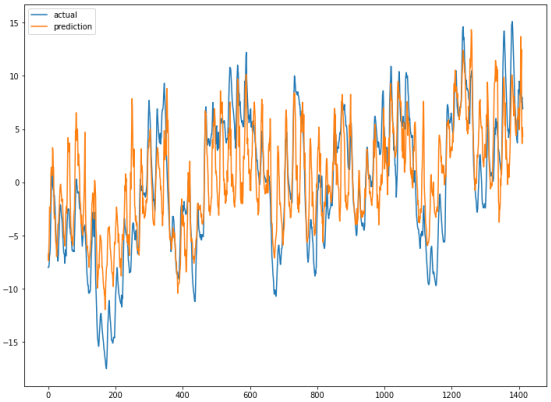
<그림- 8> lgbm 지상온도 예측과 실제의 차이

3.5.2 신경망 학습 모델

(1) LSTM

mse	rmse	mae	r2
15.62	3.95	3.18	0.62

<표- 9> lstm 지상온도 예측 mse, rmse, mae, r2 결과



<그림- 9> lstm 지상온도 예측과 실제의 차이

3.5.3 최적모델 선정

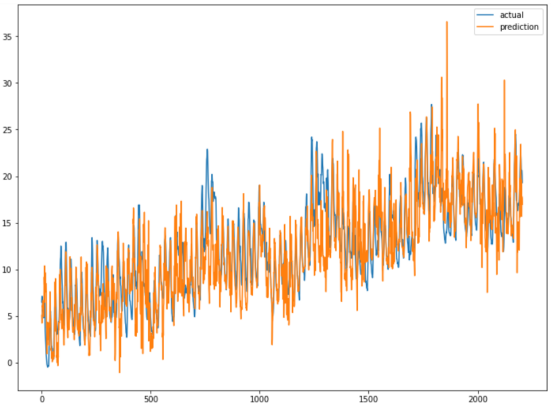
3 가지 모델 비교 결과 SVR 이 가장 좋은 성능을 가진 모델이라고 판단했다. 따라서 이어지는 목차는 SVR 모델로 21 년 봄, 여름, 가을 데이터로 예측을 실행해 보았다.

3.6 지상온도 예측 SVR 모델링

(1) 봄

mse	rmse	mae	r2
10.33	3.21	2.38	0.61

<표-10> svr 지상온도 예측 mse, rmse, mae, r2 결과

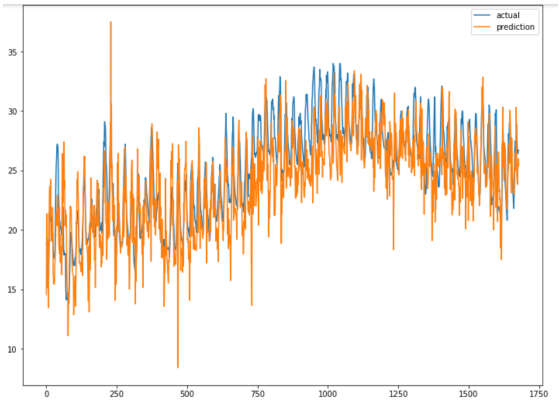


<그림- 10> svr 지상온도 예측과 실제의 차이

(2) 여름

mse	rmse	mae	r2
7.77	2.78	2.15	0.53

<표-11> svr 지상온도 예측 mse, rmse, mae, r2 결과

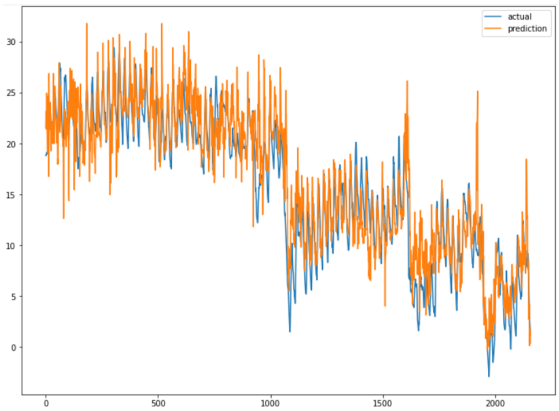


<그림- 11> svr 지상온도 예측과 실제의 차이

(3) 가을

mse	rmse	mae	r2
9.43	3.07	2.29	0.81

<표-12> svr 지상온도 예측 mse, rmse, mae, r2 결과



<그림- 12> svr 지상온도 예측과 실제의 차이

4. 결론

본 연구에서는 인공위성을 활용한 지면온도 예측을 위하여 인공위성 자료 데이터만을 이용하였다. 다양한 기법중 LSTM 기법의 정확도가 가장 높기에 이 방법을 제안한다. 기존 측정 방법에서 벗어나 인공 위성만을 활용한 지면온도, 지상온도 측정법을 시사한다. 이를 통해

지면온도, 지상온도를 이용하는 다양한 작업의 능률을 향상 시킬 것으로 기대된다.

참고문헌

[1] 종합 기후감시변화정보. 한국기상청.

[2] 정종진, 김지연. (2020). LSTM 을 이용한 주가예측 모델의 학습방법에 따른 성능분석. 디지털융복합연구, 18(11), 259-266.

[3] 기상자료개방포털-종관기상관측(ASOS) - 자료. 한국기상청. https://data.kma.go.kr/

[4] 이지윤. (2018). 서포트 벡터 머신(Support Vector Machine)의 개념과 사용법 https://leejiyoon52.github.io/Support-Vecter-Regression/

[5] Nu Ri Lee. (2020). LightGBM 모델의 정의와 파라미터 튜닝 방법. https://nurilee.com/2020/04/03/lightgbm-definition-parameter-tuning/