기상청

[기상위성 자료를 활용한 여름철 자외선 산출기술 개발]

참 가 번 호	220136	팀 명 ※ 반드시 참가신청 시 작성한 팀명	멀티플레이어스

<목차>

1. 서론

1.1. 과제 이해

본 과제의 목적은 기상위성 데이터를 활용하여 여름철 자외선 값을 예측하는 기술을 개발하는 것이며, 모델의 성능은 RMSE로 평가합니다.

본 과제의 주제인 자외선은 자외선A, 자외선B, 자외선C로 구성되는데 이 중 오존층에 흡수되어 지표까지 도달하지 못하는 자외선C를 제외하고, 지표까지 도달하는 자외선A와 자외선B가 우리 피부에 영향을 줍니다. 자외선은 유리창과 구름을 통과할 뿐더러 피부 깊숙히 침투해서 피부 노화를 일으키는 주범이어서, 일 년 중 자외선 지수가 특히 높은 여름철에 급성 피부 환자 수가 급증합니다. 이러한 자외선으로 인한 피해를 막기 위해 기상청에서 자외선 지수 예보 서비스를 제공하고 있는데, 현재 우리나라의 자외선 지수 관측지점이 15개밖에 되지 않아 전국적인 해상도가 다소 낮은 현황입니다. 서비스의 해상도를 높이기 위해서는 더 많은 지역에 대한 자외선 데이터가 필요한데, 자외선 측정 지점의 개수를 늘리는 대신에 분 단위로 계속해서 들어오는 기상위성 데이터를 활용하여 자외선을 산출할 수 있는 기술을 개발한다면 시간과 비

용 측면에서 아주 경제적이고 유의미한 기술을 개발하는 것이라고 할 수 있습니다. 현재 15개 관측 지점에서 생산되고 있는 자외선 값을 기상위성 데이터만을 활용하여 근사해내는 것이 본 과제의 궁극적인 목적입니다.

1.2. 데이터 이해 및 EDA

본 과제 수행을 위해 기상청 날씨마루에서 자외선 데이터와 기상위성 데이터를 다운로드하였습니다. 천리안위성 2A호가 수집하는 기상위성 데이터를 독립변수(feature)로, 전국 15개 지점에서 생산하는 자외선 데이터를 종속변수(label)로 하여 훈련 데이터로 모델에 넣어 모델이 그 상관관계를 학습하여, 그 결과 기상위성 데이터만 주어졌을 때 자외선 값을 잘 예측할 수 있는 모델을 만드는 것이 본 과제의 목적입니다.

기상위성이 수집하는 데이터로는 구름, 안개, 수증기, 황사, 오존량 등등 다양한 대기 정보를 수집하는 16개 채널이 있으며, 태양천정각, 위성천정각, 대기외일사량, 지면타입, 날짜, 시간 정보와 관측지점번호, 위도, 경도와 같은 데이터를 포함합니다. 이 중 지면타입, 관측지점번호와 같은 범주형 변수들은 이후 자외선 지수 예측에 영향을 미치지 않는 것으로 판단하여 학습에서 제외하였고, 수치형 변수 중 16개 채널 데이터들은 채널별로 범위의 차이가 크기 때문에 이후 정규화 처리를 진행하였으며, 시간 관련 데이터는 연속성을 위해 벡터화 처리를 진행하였습니다. 자세한 내용은 2절 데이터 전처리에서 상세히 기술하였습니다.

학습 데이터 기간은 2020~2021년(24개월), 평가 데이터 기간은 2022년 6월(1개월)입니다. 국내 15개 관측지점에서 10분 간격으로 수집된 데이터이므로, 학습 데이터의 수는 15(관측지점수) * 6(10분 간격) * 24(시) * 731일(366일 + 365일)의 곱으로 약 160만 개입니다.

학습 데이터 중에서 이상치가 포함된 데이터는 52,623개로 전체의 3% 가량이었는데, 대부분의 이상치들이 하나의 데이터 안에서 중복되어 관찰되는 양상을 보였습니다. 또 관측지점별로 데이터를 나누어 시간의 흐름에 따라 시각화해 보았을 때 연속적인 이상치의 군집이 있었는데, 해당 지점의 관측 기구 고장과 같은 이유로 추측됩니다. 146 지점 같은 경우 2020년 12월부터 2021년 2월까지 3개월 동안의 자외선 데이터가 모두 이상치였습니다. 모델을 훈련시키기 전에 이상치를 제거하거나 대체하는 등의 적절한 처리가 필요한데, 이상치 비율이 전체의 3% 가량 밖에 되지 않기 때문에 단순 제거하는 것도 충분히 가능하지만 본 팀은 시계열 데이터의 속성을 보존하기 위해 이상치 대체 기법을 선택하였습니다.

1.3. 사용언어 및 개발환경

본 팀은 과제 수행을 위하여 프로그래밍 언어 Python을 사용하였고, 모델 개발은 구글 클라우드 기반 개발환경인 Google Colaboratory에서 진행하였습니다. reproduction 가능하도록 seed 값은 42로 고정하였습니다.

2. 데이터 전처리

2.1. 데이터 병합

먼저 기상청 날씨마루에서 학습 데이터 2020~2021년(24개월)과 평가 데이터 2022년 6월(1개월)를 다운로드 하였습니다. 총 25개월의 데이터(uv_2001.csv, uv_2002.csv, uv_2003.csv, … , uv_2111.csv, uv_2112.csv, uv_2206.csv)를 Pandas Dataframe으로 변환하여 하나의 Dataframe으로 병합하였습니다.

2.2. 이상치 처리

학습 데이터에는 –999.0이라는 값이 전체의 3% 가량을 차지하고 있었습니다. 이 이상치는 자외선 데이터와 band1-16 데이터에서 나타났는데, 이상치 이외 정상값들의 분포와 비교해 보았을 때 정상범위로 보기 힘든 값이기 때문에 제거하거나 대체하는 작업이 필요했습니다. 본 팀은 기존 데이터의 시계열 속성을 보존하기 위해 이상치가 포함된 데이터를 제거하지 않고 적절한 값으로 대체하는 전처리를 수행하였습니다.

먼저 전체 데이터에서 –999.0 값들을 찾아 결측치 NaN으로 변경한 다음, 결측치 처리 함수를 사용하였습니다. Sklearn의 knn imputer 모듈을 사용하였는데 이는 KNN(k-Nearest Neighbors) 알고리즘을 사용하는 함수로, 결측치가 없는 가까운 k개의 데이터의 평균을 계산하여 결측치를 대체합니다. k=4로, 즉 가까운 4개 데이터를 탐색하는 것으로 설정하였고 전체 데이터를 관측지점별로 나누어 각각 knn inputation을 진행하였습니다.

2.3. 변수 변환

date_time 변수에는 년, 월, 일, 시, 분 데이터가 한꺼번에 들어 있는데 이 중 월, 시 데이터를 sin, cos 공식으로 변환하여 month_sin, month_cos, hour_sin, hour_cos 네 개의 변수를 생성하였습니다. 시간 변수를 변환한 이유는, 예를 들어 시 데이터의 경우 0시부터 23시까지 24개의 값을 갖는데, 이 수치형 데이터를 이러한 전처리 없이 그대로 사용하게 되면 23시와 0시의 차이를 컴퓨터가 23으로 받아들이게 됩니다. sin, cos 변환을 거친 두 개 변수를 함께 사용함으로써 컴퓨터가 23시와 0시의 차이를 23 차이가 아닌 1 차이로 학습할 수 있도록 시간 변수를 벡터화 하였습니다.

2.4. 파생변수 생성

기존 변수 solarza를 사용하여 새로운 파생변수 uv_calculated를 만들었습니다.

```
x = cos(solarza*(2*pi/360)) - 0.01
uv_calculated = 11 * (((abs(x)+x)/2)**2)
```

위 수식은 지면에서 수직방향으로 태양과 이루는 각도인 태양천정각 변수 solarza를 활용해서 오존이나 구름 등 다른 외부 요인들의 영향이 배제된, 맑은 날의 태양 에너지만으로 계산된 순수 자외선값을 근사하는 수식입니다. solarza 값을 라디안 단위로 변환하여 cos 공식을 사용해서 수식화하였고, 조정값 –0.01, 진폭 11 등의 상수는 실제 자외선 데이터의 분포와 유사하도록 조정하여 실험적으로 얻은 상수값입니다. 이렇게 생성한 파생변수 uv_calculated를 독립변수로 함께 사용해서 모델을 학습시켰을 때 모델의 성능이 향상되는 것을 확인하였습니다.

2.5. 정규화

band1-16, 경도, 위도, solarza(태양천정각), esr(대기외일사량), uv_calculated와 같은 수치형(numeric) 데이터들을 살펴보았을 때 변수별로 수치의 범위의 차이가 컸습니다. band1과 band7 변수를 예로 들면, 범위(최댓값과 최솟값의 차)가 각각 1.16, 168.46으로 그 범위의 차이가 변수별로 너무 크기 때문에 Sklearn의 Minmax scaler를 사용해서 모든 값을 0~1 사이로 정규화하는 전처리를 진행하였습니다.

2.6. 변수 선택

관측지점번호, 지면타입, 관측고도, 위성천정각 변수는 예측에 영향을 미치는 것으로 보이지 않아 학습에 사용하지 않았습니다.

2.7. 데이터 선택 및 추출

우리 과제는 일 년 내내의 자외선 값을 예측하는 모델이 아니라 여름철 자외선 값을 예측하는 모델을 개발하는 것입니다. 시간의 흐름에 따라 2년을 시각화해서 펼쳐 보았을 때 분명한 데이터의 주기성이 있고, 여름철 3개월과 겨울철 3개월을 추출하여 자외선 분포를 비교해 보았을 때도 분포에서 확연한 차이를 보였습니다. 또 기상위성 데이터를 사용하는 데 있어서도 눈/얼음과 같이 겨울철의 자외선 지수에 영향을 주는 요인이 여름철과는 다를 것이기 때문에 데이터의 계절성을 고려하는 것이 필수적이라고 판단하였습니다.

본 팀은 여름철과 데이터 분포의 차이가 큰 다른 계절 데이터를 학습에 사용하는 것이 오히려 모델의 적절한 학습을 방해할 것이라는 가설을 세웠고, 실제로 RNN 모델의 한 종류인 GRU 모델로 이를 실험해 보았을 때 2년 전체 데이터를 모두 학습에 사용했을 때 RMSE 0.7331에서 여름철 데이터(전체 데이터의 25%)만 사용했을 때 0.6791로 RMSE가 줄어 모델의 예측 성능이 향상됨을 확인하였습니다. 따라서 평가 데이터 기간이 2022년 6월이므로 시간적으로 먼 계절은 제외하고 인접한 5, 6, 7월만 추출하여 모델 학습에 사용하였습니다.

3. 분석기법

3.1. 모델 선택

데이터 전처리를 끝내고 다양한 모델링을 시도해 보았습니다.

빠르게 모델 성능 수준을 파악하기 위해 가장 먼저 Python의 오토 머신러닝 라이브러리 Pycaret을 사용하였습니다. 결과는 Light Gradient Boosting Machine, Cat Boost Regressor, Extra Trees Regressor, Gradient Boosting Regressor, Random Forest Regressor 순으로 성능이 좋았고, RMSE 0.7 내외 수준의 성능을 보였습니다.

이어서 Tensorflow의 Keras 라이브러리를 사용해서 Dense layer만으로 기본적인 딥러닝 모델을 구축해 보았을 때 RMSE 0.72 수준의 성능을 보였습니다.

보유한 데이터가 시계열 데이터이므로 시계열 데이터 처리에 적합한 RNN(순환신경망)에서 좋은 성능을 낼 것으로 기대하였고, 따라서 RNN의 여러 종류인 Simple RNN, LSTM, GRU 모델들을 학습시켜 보았습니다. Simple RNN은 시계열 데이터 학습에 적합한 RNN의 가장 기본적이고 복잡한 모델이고, 그것에 게이트를 추가하여 기존 simple RNN의 기울기 소실 문제를 해결하여 장기기억이 가능하도록 한 것이 LSTM, 이 아이디어를 가져가면서 매개변수를 줄인 것이 GRU입니다. GRU는 단순하고 빠르다는 장점 때문에 데이터셋이 작거나 모델 설계 시 반복 시도를 많이 해야 할 경우 적합한 것으로 알려져 있는데, LSTM과 GRU는 어느 것이 성능 면에서 낫다고 할 수 없으며 주어진 문제와 하이퍼파라미터의 설정 등에 따라 결과가 달라집니다. 본 팀이 세 가지 RNN 모델을 구축하여 성능을 비교해 보았을 때 본 과제에서는 LSTM이 가장 성능이 좋은 것으로 확인되어 최종적으로 LSTM 모델을 선택하였습니다.

3.2. 데이터셋 구성

LSTM 모델의 input으로는 예측하고자 하는 자외선 값의 시간대로부터 이전 5개 데이터까지 묶어서 6개의 데이터를 한꺼번에 input으로 주었습니다. 예측하고자 하는 종속변수 y가 6번째 행의 자외선 데이터 하나일 때, 이때의 독립변수 X는 6행(1번째 행부터 6번째 행까지) * 23열(band1-16, solarza(태양천정각), esr(대기외일사량), month_sin, month_cos, hour_sin, hour_cos, uv_calculated)의 데이터셋입니다. 이렇게 데이터셋을 구성하여 학습을 진행하였기 때문에 처음

5행의 자외선 데이터는 예측할 수 없었는데, 예측이 불가능한 시간대가 밤 12~1시 사이기 때문에 예측값을 0으로 채워 넣었습니다.

3.3. LSTM 모델 구축

Tensorflow Keras 라이브러리를 사용하여 LSTM layer(hidden unit = 16)와 Dense layer(hidden unit = 1)를 연결하여 Sequential 모델을 구축하고, 손실함수(loss function)는 RMSE를, 최적화 알고리즘은 Momentum과 RMSprop의 장점을 결합한 Adam 알고리즘을 사용하였습니다. 하이퍼파라미터 학습률(learning rate)은 0.001로 초기설정하고 이후에 학습을 진행하면서 모델의 개선이 더딜 때 감소 조정되도록 코딩하였는데 해당 내용은 3.4절 모델 학습에서 상세히 기술하였습니다.

3.4. 모델 학습

학습은 배치 사이즈 64로 에폭 200회 진행하였습니다. 검증 데이터에 대한 validation loss(RMSE)가 최솟값일 때에만 모델을 저장하도록 call back 함수를 코딩했고, ReduceLROnPlateau 함수를 사용하여 3회 에폭 동안 validation loss(RMSE)가 감소하지 않으면 학습률에 0.8을 곱하여 모델의 개선을 유도하였는데 학습률의 최솟값은 1e-7까지로 제한하였습니다.

또한 데이터가 많지 않아 K-fold 교차검증 기법을 사용하여 전체 학습 데이터를 5세트로 나누어 학습을 진행하였습니다. K-fold 교차검증 기법은 시간 소요가 크다는 단점이 있으나 총 데이터 수가 적을 때 모델의 정확도를 향상시킬 수 있는 기법입니다. 학습이 완료된 후 validation loss를 기준으로 상위 3개 모델만 선택하여 0.4, 0.3, 0.3의 가중치(weight)를 주는 방법으로 앙상블하여 최종 예측값을 도출하였습니다.

4. 분석결과

지금까지 데이터 병합, 이상치 처리, 변수 변환, 파생변수 생성, 정규화, 변수 선택, 데이터 선택 및 추출 등의 데이터 전처리를 거쳐 다양한 머신러닝 및 딥러닝 모델을 구축하여 성능을 비교해 보고 최적의 모델 LSTM을 선택하여 최적의 데이터 셋을 구성하고 하이퍼파라미터 설정 및 조정을 거쳐 모델 학습을 완료하였습니다.

K-fold 교차검증 기법을 사용하여 전체 데이터를 5개 세트로 나누어 학습한 결과, 총 5개 모델의 RMSE를 평균낸 값은 0.6492785였으며, 상위 3개 모델만 선택하여 0.4, 0.3, 0.3의 가중치를 주어 앙상블한 최종 결과는 RMSE 0.648252로 전체 평균보다 조금 더 높은 성능을 얻을 수 있었습니다.

5. 활용방안 및 기대효과

본 과제를 통하여 기상위성 데이터를 활용한 자외선 산출기술의 정확도가 현재 관측지점에서 생산하고 있는 값에 근사한 수준으로 높아진다면, 지상에 있는 자외선 관측지점을 대체하여 경제적인 비용을 줄이는 것뿐 아니라 현재의 자외선 지수 예보 서비스의 품질을 지역과 시간면에서 모두 비약적으로 높일 수 있습니다.

현재 대한민국 기상청에서는 국내 15개 관측지점에서 생산하는 값으로 국내에 한하여 자외선 지수 예보 서비스를 제공하고 있는데, 천리안위성 2A호가 수집하는 기상위성 데이터를 사

용하여 자외선 지수를 산출할 수 있게 된다면 한반도 전역을 포함하여 아시아-태평양 지역에 대해서는 2분 간격으로, 전 지구에 대해서는 10분 간격으로 자외선 지수 예보가 가능하게 됩니다. 자외선 수치가 높아 그로 인한 피해가 크지만 관련 시설 및 서비스가 미비한 국가들을 주 타겟으로 하여 자외선 지수 예보 서비스를 제공한다면 글로벌 경쟁력을 확보하고 외교적으로 유리한 위치를 점할 수 있을 것입니다.