2024 날씨
빅데이터 콘테스트

날씨 빅데이터 콘테스트

기상관련 강수분야 최종 공모안

수치모델 앙상블을 활용한 강수량 예측

기상청

접수번호	240210
팀명	벤치클라이머
팀원	정정룡, 황영우, 정성훈

1. 분석 배경 및 목표

강수량은 의류, 토목, 관광사업 등 대부분의 분야에 영향을 주며, 농업, 수자원 관리, 재난 관리 등의 분야에는 직접적이며 치명적인 영향을 끼친다. 따라서 정확한 강수량 예측은 각 분야에서의 피해를 최소화하고 실질적인 이득을 얻는 데에 큰 기여를 할 수 있다. 현재 기상청에서는 상세한 기상 및 예상 강수량을 제공하고 있으나, 예보의 불확실성이 소수 존재하기에 이를 개선할 필요성이 있다. 이에 본 공모안에서는 주어진 수치모델 앙상블 강수 확률 자료를 활용하여 예측 모델을 생성하는 것을 목표로 한다.

2. 데이터 및 변수 정의

활용한 데이터는 기상청 날씨마루에서 제공한 예측 및 관측자료를 활용하였다. 해당 자료는 A년부터 C년까지 연속된 3년중에 5월 ~ 9월의 데이터가 존재하며, 총 20개의 지점에서 관측되었다. 분석을 진행하면서 모델 생성에 필요하다고 판단한 파생 변수를 생성하였다. 파생변수 생성 과정을 거쳐 사용한 변수는 다음과 같다.

변수명	정의	변수명	정의	변수명	정의
basis_index	예측 시각 및 관측소	v07	10.0mm이상 누적 확률	v05_ind	2.0mm ~ 5.0mm 확률
dh	예측시각 - 발표시각	v08	20.0mm이상 누적 확률	v06_ind	5.0mm ~ 10.0mm 확률
v01	0.1mm이상 누적 확률	v09	30.0mm이상 누적 확률	v07_ind	10.0mm ~ 20.0mm 확률
v02	0.2mm이상 누적 확률	v00_ind	0.1mm미만 확률	v08_ind	20.0mm ~ 30.0mm 확률
v03	0.5mm이상 누적 확률	v01_ind	0.1mm ~ 0.2mm 확률	v09_ind	30.0mm이상 확률
v04	1.0mm이상 누적 확률	v02_ind	0.2mm ~ 0.5mm 확률	vv	실강수량
v05	2.0mm이상 누적 확률	v03_ind	0.5mm ~ 1.0mm 확률	v_expect	예측강수량
v06	5.0mm이상 누적 확률	v04_ind	1.0mm ~ 2.0mm 확률	class_interval	강수계급

<표 1> 사용 변수 정의 : 파생변수

3. EDA 및 전처리

3.1 결측값 처리

학습데이터에서 vv의 값이 -999인 경우는 학습에 노이즈 발생할 수 있기 때문에 삭제하고 모델에 활용했다.

대상 변수	처리	요건
vv	제거	값이 -999인 경우

<표 2> 결측값 처리

3.2 발표 / 예측날짜 누락

3.2.1 발표날짜 누락

모든 관측소에서 B년도 05-13 09:00 ~ 05-15 21:00까지 강수확률을 발표하지 않았다.

3.2.2 예측날짜 누락

각 관측소에 따라 하나의 발표 시점에 80행의 예측을 모두 진행하지 않은 날짜가 존재한다. 이 중 3.3.1의 경우를 제외한 196개의 시점들을 각각 확인하였다.

관측소	발표날짜	누락 예측날짜
STN002	C/05/01 09:00	C/05/11 00:00 ~ C/05/11 09:00

<표 3> 예측날짜 누락 예시

3.3 파생변수 생성

3.3.1. basis_index

제공된 데이터는 시계열 데이터임을 확인하였다. 기존에는 발표 시각을 기준으로 정렬되어 있으나, 활용한 모델(Bidiredction LSTM)에서 정렬 순서가 매우 중요하므로 예측 시각을 기준으로 정렬하고자 하였다. 또한 관측소별로 강수확률을 제공하기 때문에 이를 고려하여 생성하고자 하는 정렬 기준 변수에 추가하였다. 예측 시각의 정보를 담고 있는 ef_year, ef_month, ef_day, ef_hour 네 변수를 통합하고, stn4contest를 추가하여 변수를 생성하였다.

3.3.2. v0X_ind

제공된 예측 자료는 기상청에서 생성한 앙상블 모델의 구간별 강수량 누적 확률값이다. 모델 생성에 각 구간 별 확률값이 필요하다고 판단하여 이를 단순 계산하여 v00_ind ~ v09_ind를 생성하였다. v00_ind는 강수량이 0.1mm 미만일 확률을 나타내는 변수이다.

3.3.3. v_expect

기상청에서 누적확률을 예측하기 위한 앙상블 모델로 계산되는 예측 강수량을 확인하고 활용하고자 하였다. 이에 3.3.2에서 생성한 각 구간 별 확률에 구간의 중간값을 곱해 합하였다. 수식은 다음과 같다.

$$\begin{aligned}\text{v\_expect} = \quad & 0.05 \times v00\_ind + 0.15 \times v01\_ind + 0.35 \times v02\_ind + 0.75 \times v03\_ind + \\ & 1.5 \times v04\_ind + 3.5 \times v05\_ind + 7.5 \times v06\_ind + 15 \times v07\_ind + \\ & 25 \times v08\_ind + 30 \times v09\_ind\end{aligned}$$

4. 모델링

강수량 및 강수구간 예측을 위해 본 공모안에서는 시계열 및 딥러닝 모형을 주로 활용하여 진행하였다.

4.1 모델 소개

- LSTM : 기울기 소실 문제로 인해 긴 시퀀스에 대한 학습에 어려움을 겪는 기존 순환 신경망(RNN)의 한계를 극복한 유형. 장기간 정보를 유지할 수 있는 메모리 셀이 존재하며, input, forget, ouput 3개의 게이트를 사용하여 정보의 흐름을 제어. 장기 종속성 처리에 강하며 시계열 예측에 주로 사용.
- BiLSTM : 두 개의 LSTM으로 구성되는 모델로서, 하나는 처음부터 끝까지 처리하고 다른 하나는 끝부터 처음까지 데이터 처리. 과거와 미래의 정보를 모두 고려함으로써 더 높은 성능을 기대할 수 있으나, 그 만큼 계산 및 메모리 비용이 높음.
- GRU : LSTM과 유사하나 더 간단한 구조의 RNN 유형. LSTM과 달리 reset, update 두 개의 게이트를 사용하며, 별도의 메모리 셀이 없는 것이 특징. 계산 복잡성을 줄이면서 비슷한 성능을 달성하는 것이 목표인 모델. LSTM과 동일한 도메인에서 사용됨.

3가지의 모델을 각각 단일로 사용했을 때 Bidrection LSTM이 가장 학습이 안정적이며 성능도 우수했다. 이외에도 머신러닝 모델을 활용했으나 성능이 저조하여 최종적으로 Bidirection LSTM 모델을 활용하여 세부적인 튜닝을 진행했다.

4.2 Custom Loss function & 사용자 정의 학습 루프

강수량 예측모델의 성능은 본 공모안에서 제시된 CSI가 최적인 방향으로 학습되어야한다. CSI를 단일로 손실함수로 만들 경우 성능이 저조하여 RMSE와 CSI 지표를 혼합하여 만든 커스텀 Loss function을 기준으로 학습을 진행하였다. 또한 행을 기준으로 계산할 경우 CSI는 항상 0 또는 1의 값을 가지는 문제가 있다. 이를 해결하기 위해 배치 단위로 계산되어 학습을 진행하도록 사용자 정의 학습 루프를 활용했다. 최종 Loss는 아래와 같다. Beta는 40으로 두어 CSI에 더 많은 가중을 두도록 했다.

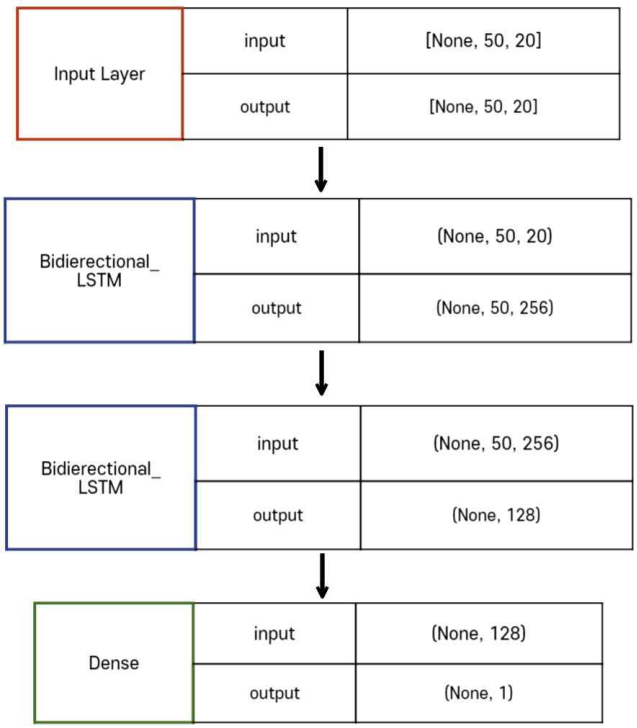
$$Loss = RMSE + \beta(1 - CSI)$$

4.3 데이터 변환

모델에 활용하는 변수들의 분포를 살펴보았을 때 대부분 skewed 되어 성능향상을 위해 이를 정규화 및 표준화를 진행했다. 이에 Yeo-Johnson 변환 후 표준화를 진행했다. Yeo=Johnson 변환은 비정규 분포의 데이터를 정규 분포에 가깝게 변환하기 위해 사용되는 기법으로 0의 값과 음수값이 포함되어도 사용할 수 있다는 장점이 있다.

4.4 모델 구조

최종적으로 활용한 모델 구조는 아래의 그림과 같다.



<그림 1> 모델 구조

첫 번째 레이어는 Bidirection LSTM의 유닛을 128로 두고 dropout을 0.1의 비율로 설정했다. 시간적 연속성을 유지하기 위해 별도의 Dropout 레이어를 두지 않았다. 두 번째 레어이는 Bidirection LSTM의 유닛을 64로 두었다. 이후 Dense 레이어를 통해 최종 예측값을 출력하도록 설정했다.

모델링을 진행할 때 과적합보다 과소적합의 경향을 보여 유닛수를 너무 크게 두지않았고 Dropout 또한 첫 번째 레이어에서만 작은 비율로 설정했다. 이외에 잔차 연결, Attention 구조 등을 활용해보았으나 과소적합이 심하게 일어나 이를 활용하기 위해서는 매우 긴 학습시간이 소요될 것으로 예상되어 위와 같은 간결한 모델 구조로 진행했다.

4.5 모델 학습 및 평가 방법

학습데이터를 살펴보면 A, B, C년도로 구분된다. 이때 5~9월로 제공되기 때문에 연도를 결합하기에는 시간적 연속성이 깨지기 때문에 부적합하다. 이에 검증데이터와 가장 가까운 연도인 C년도로 학습하고 평가지표는 B,C년도의 loss의 평균으로 설정했다.

4.6 모델 파라미터 설정

파라미터	비고
epoch	60
Sequence Length	50
Batch Size	64
Beta	40
Learning Rate	0.0007
Optimizer	Adam
Active Function	sigmoid, tanh(Dense 레이어에서는 linear)

<표 4> 파라미터 설정

epoch는 60으로 두고 25번의 epoch에서 Loss가 개선되지 않으면 Ealry Stop하도록 설정했다. 과소적합의 위험성이 크기 때문에 Batch Size를 64로 두어 안정적으로 학습하도록 했으며, 학습율도 마찬가의 이유로 0.0007로 설정했다. Active Function의 경우 Bidirection LSTM 레이어의 입력게이트, 망각 게이트, 출력 게이트에서 전부 sigmoid 함수를 사용했고, 셀 상태 업데이트에서는 tanh 함수로 설정했다. Relu나 leaky Relu 등의 함수를 활용할 경우 성능이 저조했음을 확인했다. Beta는 상기 설명했듯 CSI의 가중을 더 두기 위해서 40으로 설정했다.

각 파라미터는 grid search와 같은 방법으로 최적의 값을 찾았다. 하지만 Batch Size와 학습율의 경우는 시간과 컴퓨팅 파워가 충분하다면 더 작은 값으로 설정하여 더욱 안정적으로 학습할 수 있음을 파악했다.

5. 모델 예측결과

최종 평가지표의 결과는 아래와 같다

A년도 CSI: 0.08917, RMSE: 4.760, Loss: 41.1929
B년도 CSI: 0.08073, RMSE: 3.494, Loss: 40.2648

Average Loss: 40.7288

검증 결과의 CSI는 0.075이다.

6. 활용방안 및 기대효과

- 기상청의 앙상블 모델은 강수 누적 확률 예측에서 낮은 성능을 보이며, 실제로 확률이 가장 높은 강수 등급과 class_interval이 대부분 일치하지 않다. 이를 보완하기 위해 Bidirection LSTM 모델을 사용하면 보다 정확하고 먼 시일의 강수 예측이 가능함을 확인했다. 발빠른 강수 예측은 각 분야에서의 이익을 증대시키고, 장마철의 인명 피해를 줄이는 데 크게 기여할 수 있다.
- 시간과 컴퓨팅 파워가 충분하다면 batch size와 학습율을 더 낮추어 학습할 시 더욱 안정적으로 수렴하며 성능도 높아질 것으로 파악된다.
- 이외에 각 행의 dh값이 작을수록 vv와 상관계수가 높음을 확인하여 각 행에 대한 가중치를 dh값에 따라 계산 후 적용하면 더 좋은 결과가 있을 것으로 예상한다. 수식은 아래와 같다. 적절한 alpha 값을 설정하지 못해 본 모델에서는 적용하지 않았다.

$$w = e^{-\alpha * dh / dh_{\max}}$$

7. 참고문헌

Zaremba, Wojciech, Ilya Sutskever, and Oriol Vinyals. "Recurrent neural network regularization." arXiv preprint arXiv:1409.2329 (2014).

Shi, Jimeng, Mahek Jain, and Giri Narasimhan. "Time series forecasting (tsf) using various deep learning models." arXiv preprint arXiv:2204.11115 (2022).

홍성재, et al. "기계학습의 LSTM 을 적용한 지상 기상변수 예측모델 개발." 대기 31.1 (2021): 73-83.