# 최대 파고 예측에서의 Permutation Importance 분석: 제주도 지귀도 데이터 사례 연구

홍건우, 최길한, 유광운, 이나경, 김용강*
국립공주대학교

redgil77@smail.kongju.ac.kr, choikilhan12@smail.kongju.ac.kr, yugwangun@smail.kongju.ac.kr, dlskrud468@smail.kongju.ac.kr, *ygkim@smail.kongju.ac.kr

# Permutation Importance Analysis in Maximum Wave Height Prediction: A Case Study of Jiguido, Jeju Island

Geonwoo Hong, Gilhan Choi, Gwangun Yu, Nakyeong Lee, Yonggang Kim*
Kongju National University

요 약

본 연구는 지귀도 등표기상관측 데이터를 활용하여 Long Short-Term Memory(LSTM) 모델로 최대 파고를 예측하고, Permutation Importance 를 통해 특성 중요도를 분석하였다. 결과적으로, 최대 파고와 유의 파고가 예측에 중요한 요소임을 확인하였으며, 이는 해양 예측 모델의 정확도 향상에 기여할 것으로 기대된다.

## Ⅰ. 서론

기후 변화와 자연 재해의 빈도 증가로 인해 정확한 해양 예측이 점점 더 중요해지고 있다. 정확한 예측을 위해서는 예측에 영향이 많아가는 feature 를 선정하는 것이 중요한데 등표기상관측으로 측정된 데이터를 활용해서 어떤 feature 가 예측에 도움이 되고 어떤 feature 는 도움이 되지 않는지를 연구하려고 한다. 더 나아가서는 최대 파고 예측은 해양 및 항해 활동의 안전을 보장하는 데 중요한 역할을 한다. 본 연구는 지귀도에서 등표기상관측으로 측정된 데이터를 활용하여 최대 파고를 예측하고, Long Short-Term Memory(LSTM) 모델을 이용해 각 특성의 중요도를 평가함으로써 파고 예측에 가장 영향력 있는 변수를 식별하는 것을 목표로 한다.

## Ⅱ. 모델 구축

본 연구에서 사용된 데이터는 제주도 지귀도에 위치한 등표기상관측소에서 2004 년 1 월 1 일부터 2015 년 12 월 31 일까지 수집된 대략 11 만 개의 데이터로, 13 개의 기상 및 해양 관련 특성으로 구성된다. 자세한 13 개의 feature 들은 Table 1 를 보면 알 수 있다. 이렇게 수집된 데이터는 결측치 제거 후 표준화 과정을 거쳐 모델 학습에 적합한 형태로 변환되었다. LSTM 신경망 모델을 구축하여 시계열 데이터의 패턴을 학습하고, 이를 바탕으로 최대 파고를 예측하였다. Adam 최적화 알고리즘을 사용하였으며[1], Adam 최적화 알고리즘은 각 파라미터의 기울기의 대한 첫번째 모멘트와 두번째 모멘트를 추정하여 업데이트 하는 방식이다. $\alpha$는 학습률, $\beta_1$ 은 모멘텀 지수 $\beta_2$ 는 RMSProp 지수, $\epsilon$ 은 수치적 안정성을 위한 변수, $m_t$ 는 모멘텀 모멘트, $v_t$ 는 RMSProp 모멘트를, $t$ 시간을 나타낸다. Adam 최적화가 이루어지는 과정은 다음과 같다.

| Index | Input Feature |
|---|---|
| 0 | Wind Speed |
| 1 | Wind Direction |
| 2 | Maximum Gust Direction |
| 3 | Maximum Gust Speed |
| 4 | Sea Level Pressure |
| 5 | Temperature |
| 6 | Minimum Daily Temperature |
| 7 | Maximum Daily Temperature |
| 8 | Water Temperature |
| 9 | Maximum Wave Height |
| 10 | Significant Wave Height |
| 11 | Wave Period |
| 12 | Water Level |

Table 1. Feature Index

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \quad (1)$$
$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \quad (2)$$
$$\check{m}_t = \frac{m_t}{1-\beta_1^t}, \check{v}_t = \frac{v_t}{1-\beta_2^t} \quad (3)$$
$$\theta_{t+1} = \theta_t - \frac{\alpha \check{m}_t}{\sqrt{\check{v}_t + \epsilon}} \quad (4)$$

Algorithm of Adam

(1) 에서는 기울기의 지수의 이동평균을 구하고 (2) 에서는 기울기의 제곱에 대한 지수 이동 평균을 구한다. 그후 (3)에서 바이어스 보정을 한 뒤 (4) 최종적으로 파라미터를 업데이트을 한다. 그리고 과적합을 방지하기

위한 Early Stopping 기법이 적용되었다[2]. Early Stopping 기법은 매 epoch 마다 검증 데이터에 대한 loss 를 측정하고 훈련 데이터에 대한 loss 는 감소하나 검증 데이터의 대한 loss 가 증가하는 시점에서 학습을 멈추도록 조정하는 방식이다.

## Ⅲ. 특성 중요도 분석

훈련된 모델을 바탕으로 Permutation Importance 방법을 사용하여 각 특성의 중요도를 계산하였다.[3] 이 방법은 모델의 예측 성능에 각 특성이 얼마나 기여하는지를 정량적으로 평가하여, 중요도가 높은 특성을 식별할 수 있게 한다. Permutation Importance 방법을 적용시키기 위해서는 훈련된 모델 $f$ , feature matrix $x$ , 타겟 변수 $y$ 그리고 error 측정 기준 $L(y, f(x))$ 가 필요하다. 여기서 측정기준은 MSE 방식을 이용했다. 알고리즘은 아래의 순서로 진행된다.

$$e_{orig} = L(y, f(x)) \quad (1)$$

$$e_{perm} = L\left(y, f(x_{perm})\right) \quad (2)$$

$$FI_j = \frac{e_{perm}}{e_{orig}} \quad (3\_1)$$

$$FI_j = e_{perm} - e_{orig} \quad (3\_2)$$

### Algorithm of Permutation Importance

$e_{orig}$ 는 원래 모델의 예측 오차를, $e_{perm}$ 은 변형된 데이터의 예측 오차를, $FI_j$ 는 특성의 importance 를 나타낸다 이때 $j$ 는 평가하고자 하는 특성의 인덱스를 나타낸다. (3_1)는 비율로서 나타내는 방식이고 (3_2)는 차이로 나타내는 방식이다.

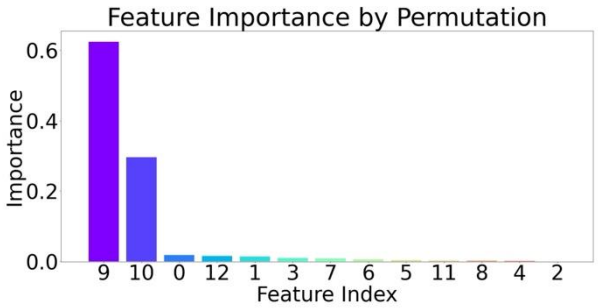


그림 1 특성 인덱스별 중요도

그림 1 을 보면 중요도가 가장 높은 특성은 최대 파고(feature index 9)로, 전체 중요도의 약 64.13%를 차지하며, 이는 최대 파고 예측에 가장 큰 영향을 미치는 요소이다. 뒤를 이어 유의 파고(feature index 10)는 약 30.77%의 중요도로 두 번째로 큰 영향을 미치는 것으로 나타났다. 이외의 특성들은 상대적으로 낮은 중요도를 보이며, 특히 수온은 중요도가 0 으로 나타나 최대 파고 예측에 있어 영향력이 없음을 보여준다.

## Ⅳ. 결론

이 연구는 제주도 지귀도에서 수집된 데이터를 사용하여 LSTM 모델로 최대 파고를 예측하고, 각 특성의 중요도를 분석하였다. 특성 중요도 결과에 따르면, 최대 파고와 유의 파고가 예측에 크게 기여하는 주요 요소로 확인되었다. 이러한 발견은 해양 예측 모델의 정확성을 과 불필요한 학습을 줄일 수 있어서 예측 모델을 보다 빠르고 정확하게 학습시킬 수 있다. 이 연구는 기후 변화에 따른 해양 환경 예측의 효율성을 높이는 데 기여할 수 있을 것이다. 지금까지의 전체 흐름을 의사코드와 그림으로 나타내면 다음과 같다

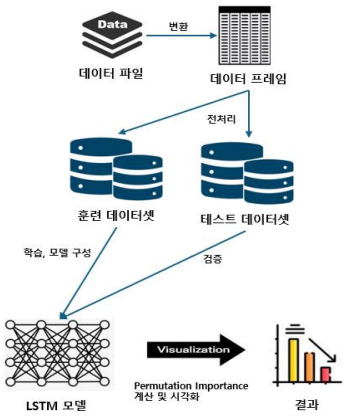


그림 2 전체 흐름도

**Algorithm 1** LSTM 모델을 이용한 데이터 분석 및 Permutation Importance 계산

**입력:** 데이터 파일 (2005-2018.csv)
**출력:** 모델 평가 및 특징 중요도
**1. 데이터 불러오기 및 전처리**
- 데이터 파일을 불러와 데이터프레임으로 변환 후 비어있는 값 제거
- 훈련 및 테스트 데이터 분할 및 데이터 표준화
**2. 데이터셋 생성**
**function** CREATE_DATASET(data, seq_len, pred_days, target_index)
  - 빈 X, Y 리스트 생성
  **for** 각 시퀀스에 대해 **do**
    - X에 시퀀스 추가 및 Y에 타깃 값 추가
  **end for**
  **return** X, Y
**end function**
- 훈련 및 테스트 데이터셋 생성
**3. LSTM 모델 구성 및 훈련**
- 모델 구성 (LSTM 레이어 + Dense 레이어)
- Adam 옵티마이저:
$m_0 \leftarrow 0$
$v_0 \leftarrow 0$
$t \leftarrow 0$
**반복 시작**
**while** 수렴하지 않음 **do**
  $t \leftarrow t + 1$
  $g_t \leftarrow$ gradient 계산
  $m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$ ▷ 모멘텀 업데이트
  $v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t^2$ ▷ 속도 업데이트
  $\hat{m}_t \leftarrow \frac{m_t}{1-\beta_1^t}$ ▷ 편향 보정 모멘텀
  $\hat{v}_t \leftarrow \frac{v_t}{1-\beta_2^t}$ ▷ 편향 보정 속도
  $\theta_{t+1} \leftarrow \theta_t - \frac{\alpha \hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon}$ ▷ 파라미터 업데이트
**end while**
- 조기 중단 콜백 설정 후 모델 훈련 (훈련 데이터로)
**4. 모델 예측 및 평가**
- 테스트 데이터로 모델 예측
- RMSE, MAE, R2 계산
**5. Permutation Importance 계산**
**function** CALCULATE_PERMUTATION_IMPORTANCE(f, x, y, metric=MSE, repeats=5)
  **(5.1)** 기본 성능 측정: $e_{\text{orig}} = L(y, f(x))$
  **for** 각 특성에 대해 **do**
    **(5.2)** 중요도 초기화
    **for** 반복 횟수에 대해 **do**
      **(5.3)** X 데이터 특성 섞은 후 성능 측정
      **(5.4)** 계산된 성능: $e_{\text{perm}} = L(y, f(x_{\text{perm}}))$
      **(5.5)** 중요도 계산: $\text{FI}_j = \frac{e_{\text{perm}}}{e_{\text{orig}}}$ 또는 $\text{FI}_j = e_{\text{perm}} - e_{\text{orig}}$
    **end for**
    **(5.6)** 중요도 저장
  **end for**
  **return** 중요도
**end function**


## ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. RS-2022-00166739).


## 참 고 문 헌


[1] NEWEY, Whitney K. Adaptive estimation of regression models via moment restrictions. Journal of Econometrics, 1988, 38.3: 301-339.

[2] YING, Xue. An overview of overfitting and its solutions. In: Journal of physics: Conference series. IOP Publishing, 2019. p. 022022.

[3] BREIMAN, Leo. Random forests. Machine learning, 2001, 45: 5-32.