特徴量重要度を用いた決定木機械学習モデルの新潟県沖波高予測への適用

長岡技術科学大学大学院 環境社会基盤工学専攻　学生会員　堀内 颯人
長岡技術科学大学 環境社会基盤工学　正会員　細山田 得三

１．はじめに

我が国の沿岸域は経済的に重要であり，洋上風力発電，貿易，レジャー等に大きな役割を果たしている．水災害の増加により，沿岸域の防災・減災がより重要視され，波浪予測技術の向上が急務とされている．現在，気象庁の全球波浪数値予報モデルGPV（GWM）を使用して，沿岸域の様々な活動における判断が行われている．近年，機械学習技術が波浪予測において注目を浴び，物理モデルに匹敵する予測精度を持つ例も増えている．

Tomらの事例では，GWM等の全球モデルによる予測の空間解像度を改善するために機械学習モデルGradient Boosting Decision Tree(GBDT)の実装の一つであるXGBoostを用いた結果, GWM以上の精度を持つ波浪予測システムとなった. 具体的には, GWMの予測値の生データを特徴量，那覇港と中城港のNOWPHASの有義波・波高観測値を教師データとしてモデルを構築しており，Tomらは機械学習を用いてある点のデータを任意の点におけるデータへの変換を行なったと述べている．こうした機械学習を用いた波浪予測事例の共通した課題として機械学習手法の独自化が挙げられる．他分野の機械学習においては, モデルに有効･有意とされる説明変数（特徴量）を生データから作成・選択する特徴量エンジニアリングが不可欠である．また，XGBoostをはじめとして機械学習は予測の根拠を明示することためのの手法があるが，それを用いてモデルを説明した例はない．従って，機械学習を用いた波浪予測は多分に改善の余地がある．

以上から本研究では，他分野で一般的とされる機械学習の手法を用いた1週間先までの波高予測システムを構築するとともに，使用データの違いによる予測精度の評価・検討をする．このシステム構築に当たり, 留意すべき事項として, 低波浪時かつ1週間先までの波高を回帰分析・多変量時系列解析的に予測すること，入出力をテーブルデータとして扱うこと，波浪予測における特徴量エンジニアリングの最適手法を模索することを示す．特に本研究のシステムは海上工事の作業可否判断の元となる0.5mから1mの低波浪を対象とする．

２． 機械学習モデル

本研究で使用した機械学習モデルはGradient Boosting Decision Tree（GBDT）とTabNetである．

２．１ Xgboost

GBDT（勾配ブースティング木）の学習過程は，まず初期モデルを作成し，それから指定された本数の決定木を順次追加していく手法である．各決定木は，前の決定木による予測と実際の目的変数との誤差を最小化するように構築され、徐々にモデルの精度が向上する．決定木の重要性は，モデルの予測が目的変数に近づくにつれて減少し、最終的には各決定木の寄与が調整され，高度な予測性能を発揮する．最終的に，予測対象のデータに対して，複数の決定木の葉のウェイトの和を計算して最終予測値が得られる．

本研究では，GBDTの実装ライブラリの一つであるXGBoostを用いる． XGBoostには，モデルが各特徴量を重視する程度を示すFeature Importance算出機能が備わっており，決定木の分岐の条件として採択された回数を大量に作成した特徴量をGBDTに入力し，Feature Importanceを算出することで特徴量の要不要を判断する．

２．２ 特徴量重要度（Feature Importance）

XGBoostのFeature Importanceは，機械学習モデルの解釈と特徴量の重要性を理解するための重要な概念である．Feature Importanceは，モデルの各決定木で特徴量が分割に使用された回数と，それに伴う分割のゲイン（目的関数の減少）を考慮して計算される．特徴量が多くの分割で重要な役割を果たし，目的関数を大幅に減少させた場合，その特徴量は高い重要度の値を持つ．この値は，特徴量がモデルの予測にどれだけ寄与しているかを示し，高い値を持つ特徴量は予測に重要である．

２．３ TabNet

Neural Networkは代表的な機械学習アルゴリズムの一つである．線形式と活性化式，パラメータの自動的な最適化で，データ間の依存を表現する．

TabNetはNeural NetworkをベースにTreeモデルを模した深層学習モデルである．TabNetは深層学習モデルでありながらGBDTと同様にテーブルデータに有効であり，Attention（図-2中のAttentive Transformerに相当）によるFeature Importanceの算出機能を持つ．Attentionは学習により入力値が重要であるか否かを適切に決定しているNeural Networkであり，これにより，TabNetは入力データ内の特徴の重要性を適切に評価し，前ステップでの重要な特徴に焦点を当てつつ，段階的にデータの構造とパターンを理解する．

３．使用データ

本研究では低波浪予測システムに波浪観測値と気象観測値を使用する。波浪観測値はNOWPHAS観測値を用いており、NOWPHASの波高観測点は，対象地点である新潟港を含む日本海側の観測点（青森西，伏木富山，輪島，柴山，藍島）とし，期間は2014年から2018年とする．新潟港含む日本海側の複数観測点は複数の波高データの影響を知るために用いている．気象観測値はJPA-55を用いており、風速データはより古いデータからも有効な傾向を見出し，ノイズとしないという点からJPA-55を選出した．JRA-55より取得した風速データは，緯度35-50，経度85-140の範囲における地表から10m上空のものを用いている．これらは，緯度経度1.25度の格子間隔で配置されている．

４．実験

４．１ 予測システム

本研究の波高予測システムでは2時間ごとあるいは6時間ごとに更新される波浪観測値及び気象解析値から選択済み特徴量を作成する．この特徴量は，観測更新時点から過去任意ステップ分のデータから作成される．その特徴量を複数のあらかじめ学習済みのモデルに入力し，各モデルから予測値が得られる．その後，最適な特徴量の組み合わせを選定する特徴量選択，特徴量の更新に伴う再学習を行い，各々の特徴量の組み合わせごとに予測値精度を評価する．

４．２ 風速データの選定と比較

XGBoostのFeature Importanceを活用し、予測期間が1日先から一週間先それぞれの波高予測において，最適な風速地点をデータ取得範囲内の全地点から20地点選定する．全地点からランダムに20地点を選び，予測モデルを構築し，その地点を用いて予測を行う．この過程を300回繰り返すことで，各地点における特徴量重要度の値が得られる．各地点における特徴量重要度の合計値を選択された回数で割ることにより，各地点の重要度が算出される．重要度の値が高い上位20地点を最適な風速地点とし，それらの地点を用いた予測モデルと，ランダムに選択された地点を使用したモデルの予測精度を比較する．

５． 結果と考察

５．１ 風速データの選定

本研究では予測期間ごとに風速データの影響をXGBoostのFeature Importanceを活用することで，予測に最適なデータの選定を行った．

図-3より，予測期間が長くなるにつれ，予測対象地点の新潟港より西側の地点に位置する風速データが予測において重要視されることが明らかとなった．これは，気象パターンが一般的に西から東に向かって移動するため，西側のデータがより重要であると考えられている．温帯低気圧が偏西風に乗って移動し，これが悪天候をもたらす要因であるため，

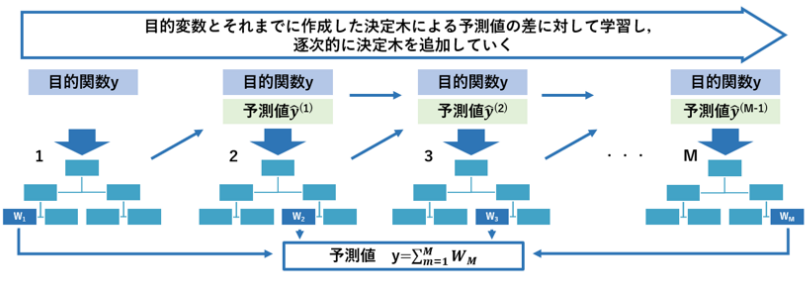

図-1 GBDTの学習，予測

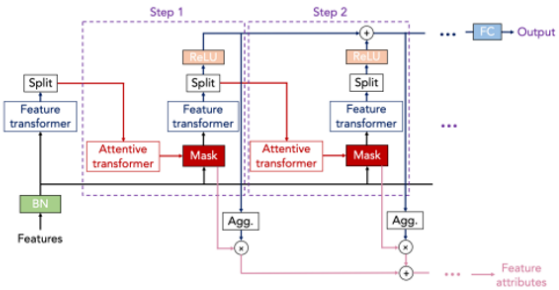

図-2 TabNetの概略

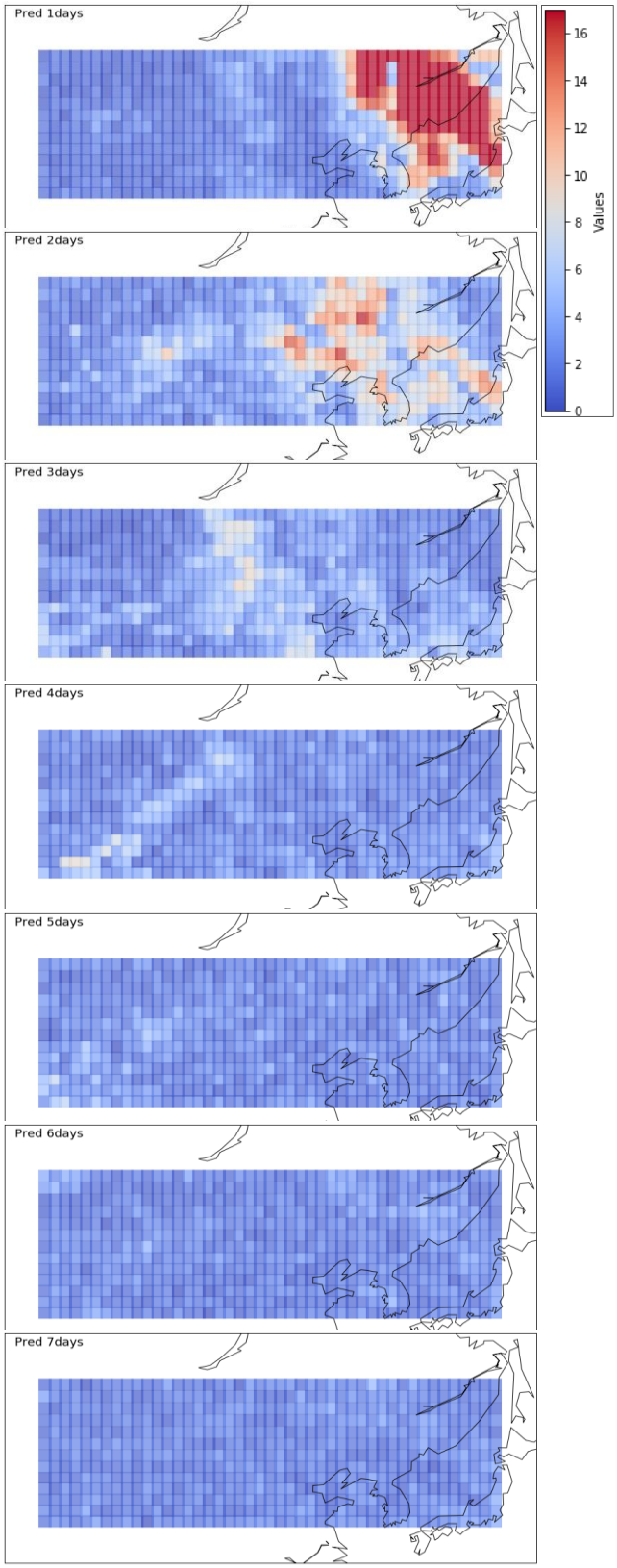

図-3 各予測期間における新潟港波高予測のための特徴量重要度の空間分布

天気に左右される波高の予測において，予測対象地点が西側の地点からの影響を受けやすいといえる．

予測期間が長くなるにつれ，特徴量重要度の値が減少し，重要視する地点の範囲が広がる傾向がある．予測期間が長くなることで，気象パターンや気象条件が変動しやすくなる．長期的な予測において，季節や気象イベントの変動を考慮する必要が生じ，広い地域の情報がより重要になるため，特定の地点に依存しなくなると考えられる．つまり，予測期間が長くなることで特定の地点に限定されない広域なデータ情報がより重要となり，特定地点の風速データの影響が相対的に低減する傾向があるといえる．

５．２ 風速データの違いによる精度比較

最適な20地点の風速データを用いたモデルとランダムに選択された20地点の風速データを用いたモデルの精度を比較した．

図-4より，XGBoostでの波高予測において，1日先予測を除く他の予測期間において，特定の地点を選定して使用した場合，予測精度はランダムに選択した場合と殆ど変化しないもしくは悪化する結果が得られた．この結果は，風速データを用いた波高予測において，特定の地点の選定が予測期間に依存することを示しており，1日先予測では特定の地点の風速データが有益である可能性が高い一方，長期予測においては特定の地点の選定は困難であると考えられる．よって，風速データを活用する波高予測において，予測期間に合わせた柔軟な地点選定の手法の採用が必要である．また，本研究では地点選定プロセスにおいて特徴量重要度の値を抽出する際に繰り返し回数を300回としたが，回数が制限されていたことから，選定地点の最適性が十分に評価されなかった可能性がある．

また，TabNetにおいては，1日，2日，3日，4日，7日先予測では，選定された地点を使用した場合，予測精度はランダムに選択した場合より精度が向上した．しかし，5日，6日先予測では悪化する結果が

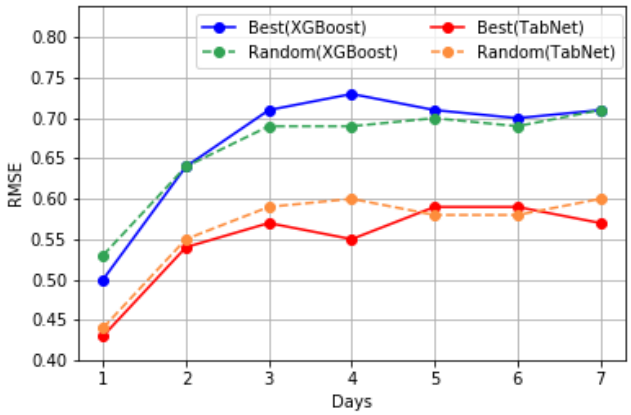

図-4 XGBoost と TabNet を用いた各予測期間における風速データの違いによる精度比較

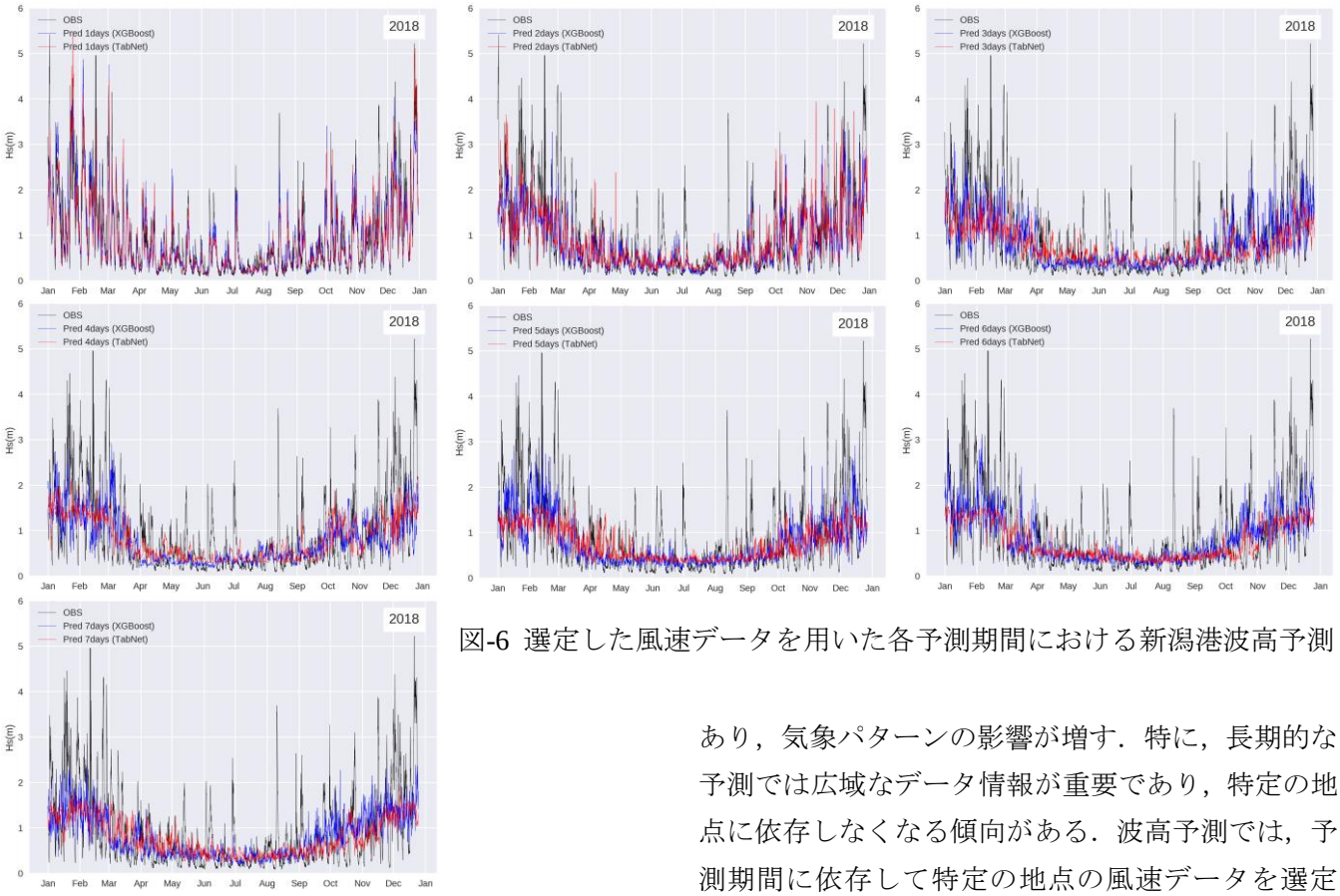

図-6 選定した風速データを用いた各予測期間における新潟港波高予測

得られ，これはXGBoostと同様予測期間が長くなるにつれ風速データの地点選定が困難なためである．

XGBoostとTabNet両方の予測モデルにおいて，2日，3日，4日，7日先予測では，精度の向上と悪化に差異が見られる．これは，XGBoostが不要データに対する頑強性を有しているため，ランダムに選択された地点の中に不要なデータが含まれても予測精度への影響が限定的であった可能性が考えられる．ただ，全予測期間にわたってTabNetの予測精度はXGBoostを上回る傾向が見られ，機械学習モデルを用いた波高予測では，TabNetモデルがより適しているといえる．図-6より，XGBoostとTabNet両方の予測モデルは，波高1m以下の低波浪時は，比較的適切なフィッティングを示すが，高波浪時には，変動の傾向を十分に捉えきれていないことがわかる．

６．結論

本研究は波高予測における風速データの選定と風速データの違いによる精度比較に焦点を当てた．本研究では，予測期間ごとに風速データの影響を評価し，最適なデータを選定する方法を提案した．予測期間が長くなるほど西側の風速データが重要であり，気象パターンの影響が増す．特に，長期的な予測では広域なデータ情報が重要であり，特定の地点に依存しなくなる傾向がある．波高予測では，予測期間に依存して特定の地点の風速データを選定する必要があり，特に短期予測においては特定地点のデータが有益であると示した．一方，長期予測においては特定地点の選定が難しく，予測期間ごとの柔軟な地点選定が必要であると考えられる．TabNetは全予測期間においてXGBoostを上回る予測精度を示した．ただ，高波浪時の予測には改善の余地がある．本研究は，波高予測のためにモデルと風速データの選定を検討する重要性を強調し，予測期間に合わせた柔軟な手法が必要であることを示した．

参考文献

1) Tracey H. A. Tom, 間瀬 肇, 池本 藍, 川中 龍児, 武田 将英, 原 知聡, 金 洙列: GWM と XGBoost を用いた 1 週間波浪予測: 土木学会論文集 B3（海洋開発） vol.77, ppI_7-I_12:2021.
2) 国土交通省港湾局 全国港湾海洋波浪情報網: リアルタイムナウファス: https://www.mlit.go.jp/kowan/nowphas/:最終アクセス 2023/10/15
3) JRA-55 project: https://jra.kishou.go.jp/JRA-55/index_ja.html:最終アクセス 2023/10/15
4) 門脇 大輔, 阪田 隆司, 保坂 桂佑, 平松 雄司: Kaggle で勝つデータ分析の技術: 技術評論社: 2019.