

Insight into machine-learning model and prediction result: Case of random forest predictor for ground-motion intensit

*Hisahiko Kubo¹, Takashi Kunugi¹, Wataru Suzuki¹, Takeshi Kimura¹, Shin Aoi¹

1. National Research Institute for Earth Science and Disaster Resilience

1. はじめに

機械学習を用いたアルゴリズムは我々の日常の中に浸透しつつあり、地球物理学の分野でも様々な場面で活躍の場を広げている (e.g., Kong et al. 2018; Bergen et al. 2019)。他方で、機械学習、特にニューラルネットワークに関しては、「ブラックボックス」問題が往々にして指摘されている。機械学習を用いたシステムを社会実装していく際には、ブラックボックス問題に関連する様々な社会的要請を答えていく必要があり、透明性や説明責任が今後の機械学習関連の研究開発ではますます重要になっていくと想定される (e.g., 福島・他 2017)。本研究ではこれまで著者らが行ってきた地震動指標のランダムフォレスト予測器 (e.g., 久保・他 2018a 人工知能学会; 久保・他 2018b 地震学会; 久保・他 2019 人工知能学会) を題材として、その機械学習モデルと予測結果の解釈を行う。

久保・他 (2018a) では、地震動指標の一つである地表最大加速度 (PGA) を予測対象とし、震央距離・モーメントマグニチュード (Mw)・震源深さ・予測地点でのS波速度1400m/s層の深さ (D1400)・予測地点の地下深さ30mまでの平均S波速度 (Vs30) の5つを説明変数とする予測器をランダムフォレスト (RF) によって構築している。RFは、木構造を模したモデルによって回帰・分類を行う決定木を複数用いて、決定木の森を構成し予測を行うアンサンブル学習アルゴリズムである (e.g., Breiman 2001; Zhou 2017)。決定木自体はif文の集合体であり、ホワイトボックス型の手法である。しかしながら複数の決定木を組み合わせるRFの場合、一つ一つのif文が最終的な予測結果にどのように効いているのかわかりづらく、ニューラルネットワークほどではないにしてもブラックボックス的な要素を多く含む手法とみなすことができる。特に木構造が深い決定木を多数用いる場合には、その傾向が顕著となる。

2. ランダムフォレスト予測器

久保・他 (2018a) と同様のやり方で、RF予測器を用意した。本研究で用いたRFは、最大深さ100の決定木計100で構成される。学習データは、久保・他 (2019) で用いたデータセットを一部リバイスしたものを利用した。

3. 説明変数の重要度

作成した予測器において「どの説明変数が重要なのか」を知ることは、その機械学習モデルと予測結果を理解する第一歩である。RFおよび決定木では木構造における各説明変数へのアクセス頻度から、各説明変数の重要度を算出できる。以下ではこれをアクセス重要度と呼ぶ。2. のRF予測器におけるアクセス重要度は [震央距離, Mw, 震源深さ, D1400, Vs30] = [0.40, 0.26, 0.10, 0.12, 0.12] となる。震源距離とMwの値が他に比べて大きく、二つを合わせると全体の三分の二を占めていることがわかる。これは震源からの距離と地震の規模でPGAの大きさの大体が決まることを示しており、これまでの地震学の知見と整合する。

ただしアクセス頻度に基づく重要度の計算はconsistencyを満たさないことが先行研究によって指摘されている (e.g., Lundberg et al. 2018)。そこでPermutation Importance (PI; Breiman 2001; Fisher et al. 2018) を用いた重要度の評価を行った。PIは、ある一つの説明変数に関してデータ配列の順序をシャッフルしたデータを用意し、シャッフルしたことがモデルの精度にどの程度影響を与えるかを評価することで、重要度を求めるものである。ELI5を用いて計算したところ、[震央距離, Mw, 震源深さ, D1400, Vs30] = [0.9038±0.0083, 0.4836±0.0074, 0.1035±0.0015, 0.1498±0.0020, 0.1372±0.0023] が得られ、PIでもアクセス重要度と同じ傾向が見られた。ただし、震央距離・Mw・その他三つの間の重要度の差がより強調されている。なおPermutation Importanceが負の値だと、その説明変数の順序を変えた場合に予測器のパフォーマンスが向上したことを示すが、今回はそのような変数はなかったことがわかる。

加えてShap (SHapleyAdditive exPlanations; Lundberg and Lee 2017; Lundberg et al. 2018)を用いた評価も行った。Shapはゲーム理論のShapley valueに基づき、各特徴量が個々の予測にどの程度寄与しているかを測る指標であるが、アクセス重要度・PIと同様の結果が得られた。