

機械学習を用いた風水害による罹災率の予測モデルの作成

九州大学 学生会員 ○森 康平 正会員 笠間 清伸 古川 全太郎

1. はじめに

近年、日本では集中豪雨や地震、それによる土砂災害の被害が頻発している。気象庁によると¹⁾、将来の気温上昇に伴って21世紀末の日本の日降水量の最大値は20世紀末と比較して約12%増加、日降水量200mm以上の年間日数は約1.5倍に増加すると予測されている。この激甚化していく気候変動に対して本文では、機械学習アルゴリズムを活用して風水害による罹災率を高精度で予測するモデルを作成することを目的とした^{2),3)}。

2. モデル作成プロセス

2.1 目的変数と説明変数

モデル作成に用いたデータは全て都道府県単位であり、主に自然災害関連、防災関連、地形関連の3つの観点から、総数73項目、2008から2017年にかけてのデータを収集した。自然災害関連として、人口や県内総生産のような経済指標および土砂災害や降雨のデータを26項目、防災関連として、消防機関の整備状況や避難箇所等の防災におけるハードおよびソフト対策のデータを26項目、地形データとして、標高、傾斜、森林率、道路率等のデータを21項目収集した。このデータのうち、目的変数として、風水害による被害者数を各都道府県の人口で除した値である罹災率を採用し、それ以外のデータを説明変数として設定した。

2.2 機械学習アルゴリズム

本文では機械学習アルゴリズムはランダムフォレスト(RF)、XGBoost、LightGBMを用いた。これらを用いて、収集したデータをトレーニングデータ：テストデータ = 7.5 : 2.5に分割し、トレーニングデータから作成したモデルによるテストデータの予測精度の比較を決定係数 (R^2) と平方平均二乗誤差 (RMSE) を基準に行った。これらはそれぞれ式(1)、(2)で表される。

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (2)$$

この評価の結果から、考察を行うのに最適なアルゴリズムを決定した。

2.3 ランダムフォレスト(RF)

ランダムフォレストは、ブートストラップサンプリングによって少しずつ異なる決定木から構築される。この複数の決定木すべてに対して予測を行い、結果の平均を最終的な予測として用いる。本文では回帰分析を行った。

2.4 勾配ブースティング

モデル作成に用いた XGBoost と LightGBM は勾配ブースティングの一種であり、複数の決定木を組み合わせより強力なモデルを構築する手法である。その予測値 $\hat{y}_i$ は式(3)で算出される。

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^n K = f_k(x_i), f_k \in \mathcal{F} \quad (3)$$

ここで、 x_i は入力、 $\hat{y}_i$ は出力、 q はツリーの構造、 T はツリー内のリーフの数である。また、最適化される目的関数はトレーニング損失と正則化項で構成される。

3. 結果および考察

RF、XGBoost および LightGBM の予測精度を比較した結果を表1に示した。予測精度の結果から、XGBoost が R^2 は最も高く、RMSE は最も低かったため、XGBoost が最も高精度なアルゴリズムであると判断した。

これより、XGBoost を用いたより詳細な分析結果を示す。罹災率の予測値と正解値を比較した結果を図1に示す。

図1中央の赤い直線は予測値=正解値であり、この直線にプロットが近いほど正解値と予測値が近いことを示す。図1からも、XGBoost のプロットの多くが他と比較して正解値に近いことが読み取ることができる。(a) に示すような外れ値を予測することはできていないが、概ね正解値に近い値を予測できている。また、罹災率の正解値が1%を超えるケースに関して、7個中6個のデータの予測値が正解値の50%を下回ったことから、罹災率が約1%を超えると予測精度が低下することが考えられた。

次に、罹災率に関する説明変数の重要度を算出し、上位10変数とその重要度を図2に示す。この重要度の計算結果から、主に土砂災害の発生回数や短時間における最大雨量の重要度が非常に高いことがわかった。特に土砂災害に関連する項目として、河川の被害箇所、がけ崩れの被害箇所、土石流発生回数および砂防施設の被害箇所が含まれており、特に河川の被害箇所の重要度は最も高かった。このことから、洪水等の水害が罹災率に大きな影響を与ると考えられる。また、降水量に関する指標として、最大3時間、6時間、12時間雨量が含まれていた。この要因として、近年ゲリラ豪雨や線状降水帯による局地的な大雨の被害が増加していることが挙げられる。このような短時間に集中的に発生する降雨も罹災率に大きな影響を与えると考えられる。

気象庁によると¹⁾将来的な降水量の変化や土砂災害の増加はおおよそ予測されている。そのため、土砂災害の発生回数や短時間における最大雨量といった変数の重要度が上位にきているのは、将来予測に大きな影響を与えることを示唆している。

しかしながら、罹災率の予測に関して、図1 (a) に示した外れ値は、2011年に紀伊半島に上陸した台風12号による水害の影響を大きく受けている。機械学習による予測の欠点として、このような突発的な大災害の予測が難しいということが考えられる。 R^2 やRMSE の値はこのような大きな外れ値の影響を強く受けるため、 R^2 やRMSEなどの目に見える形ではあまり高い精度は見込めない可能性が高い。

4. まとめ

本文では、風水害による罹災率予測モデルの作成を行った。モデルの作成にはXGBoostを用いて予測

表1 各アルゴリズムの予測精度

アルゴリズム	R^2	RMSE
RF	0.119	1.071
XGBoost	0.165	1.043
LightGBM	0.129	1.065

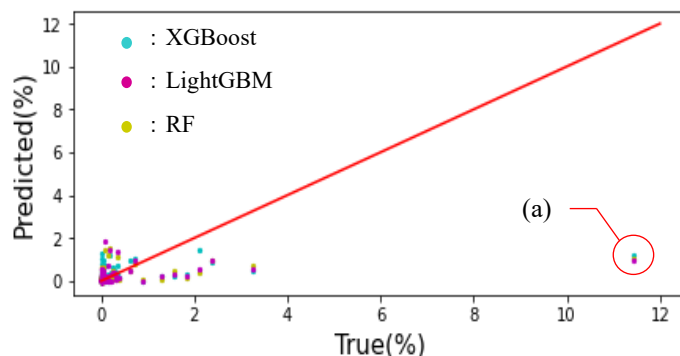


図1 罹災率の予測値と正解値

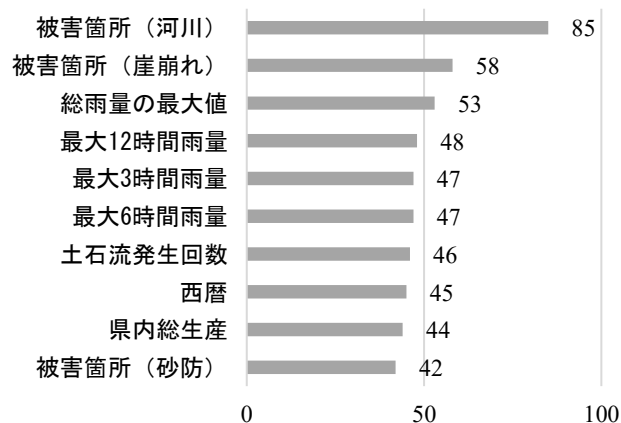


図2 罹災率に関する重要度 (XGBoost) (上位10位)

精度を評価した。データは、外れ値の影響は大きいものの、概ね正解値に近い値を予測できた。また、罹災率に関する重要度を評価した結果、土砂災害発生回数や雨量など、予測可能な項目の影響が大きいということがわかった。今後は作成したモデルを用いて、気候変動による罹災率の将来予測を実施する予定である。

謝辞：本文は JSPS 科学研究費 JP19H00812 と JP20H00266の助成を受けたものです。

参考文献

- 1) 気象庁：日本の気候変動2020 https://www.data.jma.go.jp/cpdinfo/ccj/2020/pdf/cc2020_honpen.pdf
- 2) Scalable and Flexible Gradient Boosting. Introduction to Boosted Trees. <https://xgboost.readthedocs.io/en/latest/tutorials/model.html>
- 3) Andreas C.Muller, Sarah Guido：Python ではじめる機械学習 O'REILLY (2017)