

アンサンブル学習スタッキングを用いた河川流量推定におけるXGBoostのハイパーパラメータの影響分析

熊本大学水文研究室 学生会員 坂口大珠
 熊本大学くまもと水循環・減災研究教育センター 正会員 石田桂
 熊本大学水文研究室 学生会員 横尾和樹
 熊本大学水文研究室 学生会員 永里起義

1. はじめに

近年、地球温暖化の影響により洪水や渇水が甚大化・頻発化することが予想されている。そのため高精度な降雨流出モデリング技術開発が求められている。昨今、機械学習の中でも特に深層学習と呼ばれる手法が注目されており、水文分野でも多くの適用がなされている。降雨流出モデリングにおいてもLong Short-Term Memory (LSTM) NetworkやConvolutional Neural Network (CNN)を用いて、高い流量推定結果を示している。また、これらの深層学習を弱学習器とし、アンサンブル学習を導入することで単体の学習器より高い結果を得られることが期待されている。降雨流出モデリングにおいて、深層学習にXGBoostを用いてアンサンブル学習させた研究はある。しかし、XGBoostにおいて適切な内部変数の設定がなされていない。本研究では、学習方法の違う深層学習手法にXGBoostを用いてアンサンブル学習をおこす。それぞれの結果を比較し河川流量推定における精度向上を試みた。

2. 手法

(1) アンサンブル学習

アンサンブル学習には、いくつかの手法がある。本研究では、そのなかでもスタッキングを用いた。スタッキングとは、複数の学習器を階層的に積み重ねる手法である。本研究では、学習の仕方が異なる3つの深層学習手法 (LSTM, CNN, Multi Layer Perceptron (MLP)) を1段階目の学習とした。それぞれの深層学習手法で、入力データには日平均降水量および気温を用い、対象データは日平均流量とした。2段階目の学習にはXGBoostを用いた。2段階目の学習には、1段階目で出力した推定河川流量を入力データとし、同様に対象データを日平均流量として学習した。

(2) XGBoost

XGBoostとはGradient BoostingとRandom Forestsを組み合わせた手法である。XGBoostでは、まず1つの回帰木で学習を行う。その後、回帰木を追加して学習を繰り返し、予測値と観測値の誤差を補正する。モデルの予測の数式は以下で表される。

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^k f_{k(x_i), f_k} \in F$$

$$where F = f(x) = w_q(x)$$

$$q: R^m \rightarrow T, w \in R^T$$

k は k 番目の回帰木、 F は回帰木の集合、 $f_{k(x_i)}$ はデータ x_i の重み、 $\hat{y}_i$ は予測値、 w は回帰木における葉の重み、 T は葉の数である。

3. 対象地・データセット

本研究では手取川を対象地とした。流域面積は809km²、幹線流路長は72kmの石川県最大の第一級河川である。入力データである降水量と気温データはそれぞれAsian Precipitation - Highly Resolved Observational Data Integration Towards Evaluation of Water Resources (APHRODITE)とEuropean Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5)から取得した。APHRODITEは観測データを元にした水平解像度が約4 kmのグリッドデータである。ERA5は水平解像度が0.25°×0.25°であり、時間解像度は1時間のものが公開されている。ERA5から手取川流域における流域平均の気温を算出し、一時間間隔のデータから日データに換算した。対象データである流量データは、国土交通省の水文水質データベース (WIS) から取得した。手取川の下流側にある観測所のうち最も流量データが揃っていた鶴来観測所を対象とした。流量データも一時間毎の値から日平均流量を求めた。対象期間は2002～2015年とした。本研究ではデータセットを2002年～2009年のモデルの学習を行う訓練期間、2010年～2012年の過学習が生じていないか確認を行う検証期間、2013～2015年のモデルの評価を行うテスト期間の3つに分けた。

各深層学習法の設定は以下に行った。各深層学習手法に共通するハイパーパラメータの設定を表-1に示す。また各深層学習手法ごとにもハイパ

表-1 深層学習手法共通のパラメータ

共通するハイパーパラメータ	
入力データの長さ	180,365
バッチサイズ	256
Patience	30
最適化関数	Adam
バッチ選択	Shuffle

一パラメータを設定した。LSTMにおいては、Hidden Satate Length (HSL) は10, 25, 50, 75, 100 のものを用いることとした。つまり、Input Data Length (IDL) が2種類、HSLが5種類の計10組の組み合わせを学習した。CNNにおいては畳み込み層トプリーグ層の組み合わせをNumber of Layers (NLY) とみなし、NLYが3, 4, 5のときを試した。つまり、IDLが2種類、NLYが3種類の計6組の組み合わせで学習を行う。MLPにおいてはHidden Node Size (HNS) を8, 16, 32とし、それぞれで学習を行った。つまり、入力データの長さIDLが2種類、HNSが3種類の計6組の組み合わせで学習した。それぞれの組み合わせで100回の学習する。合計で2200回の学習を行う。

アンサンブル学習に用いるXGBoostは以下の通りに設定した。学習率を0.1, 0.3, 0.5の3種類、決定木の葉の重みの下限を1, 5, 10の3種類変動させた。木の深さの最大値は6, 木の数は100, 誤差関数には二乗平均平方根誤差 (RMSE), Early StoppingのPatienceを20とした。また入力データに関しても単一の学習器だけでなく、異なる手法をまたいで入力データとした場合も学習する。

4. 結果

(1) 深層学習の結果

深層学習の結果について図—1 (LSTM の結果のみ) に示す。LSTM を用いた場合、訓練期間においてNSEの最大値は0.807~0.847, 中央値は0.746~0.791。検証期間において最大値は0.742~0.786である, 中央値は0.692~0.716である。また、テスト期間において最大値は0.680~0.713, 中央値は0.601~0.646である。CNNを用いた場合、訓練期間においてNSEの最大値は0.863~0.925, 中央値は0.808~0.857。検証期間において最大値は0.635~0.677, 中央値は0.512~0.547である。また、テスト期間において最大値は0.577~0.631, 中央値は0.431~0.492である。MLPを用いた場合、訓練期間においてNSEの最大値は0.916~0.936, 中央値は0.864~0.904。検証期間において最大値は0.575~0.652, 中央値は0.487~0.623である。また、テスト期間において最大値は0.474~0.579, 中央値は0.385~0.498である。

(2) アンサンブル学習の結果

アンサンブル学習を行った結果を表—1 に示す。入力データを異なる深層学習手法またいで入力とした場合、NSEは訓練期間では0.996~0.998, 検証期間では0.650~0.689, テスト期間では0.522~0.610である。入力データをLSTMのみ用いた場合、NSEは訓練期間では0.930~0.956, 検証期間では0.751~0.777, テスト期間では0.595~0.705である。入力データをCNNのみ用いた場合、NSEは訓練期間では0.999~0.990, 検証期間では0.586~0.670, テスト

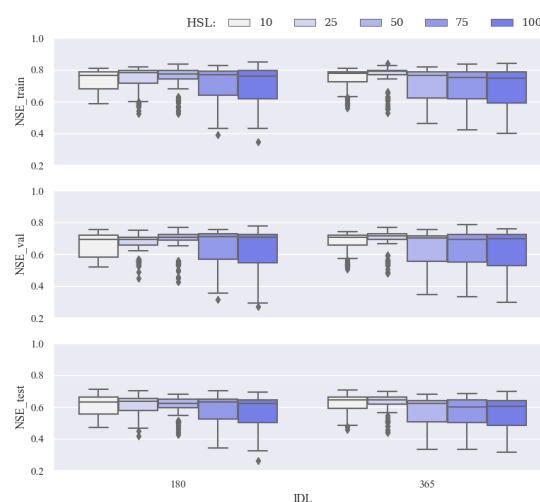


図-1 LSTMのパラメーター別精度

表-2 アンサンブル学習後の結果 (LSTM)

学習率	木の葉の重み	訓練期間	検証期間	テスト期間
0.1	1	0.951	0.777	0.666
0.1	5	0.952	0.766	0.689
0.1	10	0.956	0.760	0.705
0.3	1	0.945	0.764	0.635
0.3	5	0.942	0.773	0.642
0.3	10	0.930	0.767	0.685
0.5	1	0.950	0.768	0.595
0.5	5	0.931	0.762	0.605
0.5	10	0.955	0.751	0.663

期間では0.510~0.615である。入力データをMLPのみ用いた場合、NSEは訓練期間では0.998~0.979, 検証期間では0.603~0.639, テスト期間では0.506~0.538である。

5. まとめ

本研究では、深層学習にアンサンブル学習を行い、第2段階においてハイパーパラメータを影響分析を行った。以上のことからハイパーパラメータの設定の重要性が示された。また学習器の組み合わせを変更することにより精度が変動することも見込めるため今後も検討を行っていく必要がある。

参考文献

- 1) 坂口 大珠, 石田 桂, 永里 赳義, 横尾 和樹, 木山 真人, 尼崎 太樹: アンサンブル学習による河川流量推定における弱学習器に対する考察, 第2回 AI・データサイエンスシンポジウム, Nor.2021.