Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Random Forest para Modelagem Hidrológica na Cidade de São Paulo

Fernando L. S. Filho [1], Marcos G. Quiles[2]
UNIFESP, São José dos Campos, SP
Leonardo B. L. Santos[3]
CEMADEN, São José dos Campos, SP

Em Aprendizagem de Máquina, modelos do tipo *ensemble* são aqueles que combinam diversos outros modelos para fazer uma predição. Um exemplo muito utilizado na literatura é o *Random Forest* [2]. Esse algoritmo é construído sobre o conceito de *bagging* [1], termo usado pela primeira vez como um acrônimo para *bootstrap aggregating*. O processo de *bootstrapping* consiste na formação de novos conjuntos de dados replicados do conjunto de dados de treinamento original, cada um com o mesmo tamanho, escolhido aleatoriamente, mas com reposição. Para cada novo replicar, uma árvore de decisão é criada e a previsão final, no caso de regressão, é obtida pela média do resultado de cada preditor individual, sendo o processo chamado *aggregating*.

Em [3], os autores fazem uma revisão das aplicações de *Random Forest* nas ciências da água, destacando o grande potencial desta técnica para fins de previsão e inferência para diversos tipos de modelagem como vazão, temperatura, evapotranspiração. Neste trabalho buscamos responder a seguinte pergunta: sob quais circunstâncias, é possível atingir um bom desempenho na previsão do nível de um rio utilizando dados de estações pluviométricas e fluviométricas em uma bacia urbanizada utilizando *Random Forest*?

Neste trabalho, temos como foco do estudo a bacia do rio Tamanduateí, altamente urbanizada, especialmente na porção que fica dentro da cidade de São Paulo. Os dados das estações foram obtidos do site do Departamento de Águas e Energia Elétrica (DAEE) e eles têm amostragem 10-minutal. O objetivo da modelagem é prever o nível do rio na estação do Mercado Municipal, que é a mais próxima do exutório, com algumas horas de antecedência. Para realizar os experimentos, baixamos os dados das estações de interesse para o primeiro trimestre de todos os anos de 2018 a 2022. A escolha do primeiro trimestre se deve ao fato de que é um dos períodos com maior ocorrência de precipitação e alagamentos nessa região.

Para a realização dos experimentos, usamos os dados de 2018 e 2019 para primeiro treino, os de 2020 para validação (e posteriormente também para compor o conjunto de treino), e os dados de 2021 e 2022 para teste. Dispondo de 6 estações fluviométricas e 10 estações pluviométricas, decidimos testar diferentes combinações de entradas a fim de prever como saída o nível da estação no exutório num horizonte de 2 horas, inicialmente usando apenas informações do instante atual para previsão do futuro. Por exemplo, em um primeiro experimento utilizamos como entrada apenas a estação pluviométrica apenas no exutório. No segundo, usamos todas as estações pluviométricas disponíveis. No sétimo experimento, usamos todas as estações pluviométricas e fluviométricas disponíveis. Além disso, em uma segunda rodada de experimentos, utilizamos não apenas informações atuais, mas também *lag features*, isto é, valores passados das séries temporais como entradas no modelo para predição de valores futuros. Vale ressaltar ainda que utilizamos

[1] fernando.saraiva@unifesp.br
[2] quiles@unifesp.br
[3] santoslbl@gmail.com

uma técnica bayesiana para otimização de hiperparâmetros, por meio da biblioteca *Optuna*. Para avaliação do desempenho, utilizamos duas métricas: o erro quadrático médio (RMSE) e a eficiência de Nash–Sutcliffe (NSE). Esta última é uma métrica bastante usada no contexto hidrológico, variando de menos infinito a 1, sendo 1 uma medida de predição perfeita.

Ao realizar os experimentos, observamos que aquele que levou a melhores métricas no conjunto de treino foi o experimento 7, aquele que tinha como entradas todas as estações pluviométricas e todas as fluviométricas. Entretanto, o que apresentou melhores métricas no conjunto de teste foi o experimento 5, aquele que tinha como entradas todas as estações pluviométricas e apenas a estação fluviométrica do exutório. Com o experimento 5, chegamos a um NSE de 0.713 usando apenas valores atuais na previsão. Usando ainda *lag features*, chegamos a um NSE de 0.757, um valor considerado bom. Aplicamos ainda técnicas de seleção de atributos e mostramos que é possível manter o desempenho do modelo praticamente inalterado reduzindo significativamente a quantidade de entradas do modelo.

Testamos ainda o desempenho do modelo para diferentes horizontes de predição. Para o horizonte mais curto de 1 hora, chegamos a um desempenho extraordinário, com NSE de 0.918. Para o horizonte de 4 horas obtivemos o último desempenho tolerável, com NSE de 0.389. A partir de 5 horas, tivemos valores de NSE menores do que 0.36, que indicam predições não qualificadas.

O fato de que o uso das estações a montante para predição do nível do exutório leva a excelentes métricas no treino mas não no teste é um indício de que, de um ano para o outro, as relações entre os níveis de diferentes estações pode mudar. O problema da mudança de distribuições ao longo do tempo não é incomum no trato com séries temporais e essa questão deverá ser melhor explorada em trabalhos futuros. Porém, o trabalho foi bem-sucedido em mostrar que é possível obter ótimas previsões de curto prazo no exutório utilizando como entradas de um modelo de *Random Forest* apenas as estações pluviométricas e a memória da própria estação fluviométrica do exutório. Em trabalhos futuros, pretendemos ainda comparar o desempenho de diferentes modelos mais modernos também baseados em *ensemble* de árvores de decisão, como XGBoost, LightGBM e Catboost. Explorando algoritmos mais modernos, podemos aumentar o horizonte de predição com desempenhos aceitáveis, tornando, assim, nosso modelo ainda mais útil para fins práticos.

Agradecimentos

Os autores agradecem ao CNPq, processo 446053/2023-6.

Referências

[1] L. Breiman. “Bagging predictors”. Em: **Machine learning** 24 (1996), pp. 123–140.

[2] L. Breiman. “Random forests”. Em: **Machine learning** 45 (2001), pp. 5–32.

[3] H. Tyralis, G. Papacharalampous e A. Langousis. “A brief review of random forests for water scientists and practitioners and their recent history in water resources”. Em: **Water** 11.5 (2019), p. 910.