Journal of Water Resources Research 水资源研究, 2021, 10(1), 44-52
Published Online February 2021 in Hans. http://www.hanspub.org/journal/jwrr
https://doi.org/10.12677/jwrr.2021.101005

Hans汉斯

基于机器集成学习的中长期径流预报研究

吕盼成，王丽萍，刘　源

华北电力大学，北京
Email: 18811591811@163.com

收稿日期：2020年12月18日；录用日期：2021年1月29日；发布日期：2021年2月18日

摘　要

中长期径流预报对水库优化调度及水资源优化开发利用都有着重要的意义。首先采用基于Boosting算法的梯度提升回归树(Gradient Boosting Decision Tree, GBRT)和极端梯度提升树(Extreme Gradient Boosting, XGBoost)、基于Bagging算法的随机森林(Random Forest, RF)和极端随机树(Extreme Random Tree, ET)四种算法作为预报模型对锦屏一级水库月平均入库流量序列进行预报，并对预测结果进行对比分析。结果显示，RF预测效果最差，XGBoost预测效果最好。进一步选用其中预测效果较好的三个方法ET、XGBoost、GBRT作为初级学习器，以Logistic回归作为次学习器，进行Stacking集成学习预测。结果表明，Stacking集成学习的预测效果要优于单一模型中预测效果最好的XGBoost，其预测值的结果和实测值更为接近，为中长期径流预报提供了新思路。

关键词

径流预报，集成学习，机器学习，锦屏一级水库

Research on Medium and Long-Term Runoff Forecasting Based on Machine Integrated Learning

Pancheng Lv, Liping Wang, Yuan Liu

North China Electric Power University, Beijing
Email: 18811591811@163.com

Received: Dec. 18th, 2020; accepted: Jan. 29th, 2021; published: Feb. 18th, 2021

Abstract

Medium and long-term runoff forecast is of great significance to the optimal operation of reservoirs, de-

作者简介：吕盼成，出生于 1995 年 1 月，安徽省马鞍山市人，硕士研究生，研究方向为径流预报。

文章引用: 吕盼成, 王丽萍, 刘源. 基于机器集成学习的中长期径流预报研究[J]. 水资源研究, 2021, 10(1): 44-52.
DOI: 10.12677/jwrr.2021.101005

velopment and utilization of water resources. Firstly, the gradient boosting decision tree (GBRT) and extreme gradient boosting (XGBoost) based on boosting algorithm are selected. There is also random forest (RF) and extreme random tree (ET) based on bagging algorithm. These four algorithms are used as forecasting models to forecast the average monthly inflow of the Jinping-I Reservoir, and then the prediction results are analyzed and compared. The results showed that the RF prediction was the worst, and XGBoost was the best. Then, the three methods with better prediction effect are ET, XGBoost and GBRT as primary learners, logistic regression as secondary learners, and stacking ensemble learning to predict. The first mock exam results show that the prediction result of Stacking ensemble learning is better than that of XGBoost with the best prediction result in a single model. The predicted value is closer to the measured value, which provides a new idea for medium and long-term runoff forecast.

Keywords

Runoff Forecast, Ensemble Learning, Machine Learning, The Jinping-1 Reservoir

1. 引言

中长期径流预报是指根据前期的水文气象等要素，应用数理统计法或者物理成因法等对未来的一段时间(通常指 3 天以上 1 年以内)的径流进行预测[1]。由于径流过程是一个高度复杂的非线性过程，在人类活动和气候变换等因素的作用下，流域的径流形式发生了新的变化，导致传统中长期径流预报的精度较低。因此以机器学习为基础的数据挖掘技术在径流预报中逐渐受到了关注。

机器学习是一种人工智能，可以利用算法或者逻辑从数据中抽取模型，可以深度挖掘大数据的深度价值和内在联系[2]。将机器学习应用于水文领域，对提升径流预报的有效性有着重要作用。如：李伶杰等[3]利用随机森林选取预报因子，并建立随机森林和支持向量机模型对龙江水库开展径流预报研究，总体精度较高，但是支持向量机泛化能力更强。左岗岗[4]分别采用 SVM、GBDT、DNN 对渭河流域的月径流和年径流进行预测，在年径流预测中 SVM 表现最好，而在月径流上，GBDT 综合表现水平最好。许斌等[5]引入 RF 和 GBDT 两类机器学习算法，对丹江口水库未来一段时间的径流序列进行预报，得到两类模型精度相似，可用于丹江口中长期的径流预报。

然而传统机器学习有时候只能得到几个有偏好的模型，得到结果存在一定的误差，因此应用于实际情况往往不是很理想。Stacking 集成学习则可以通过引入次学习器，提高单一学习算法的预测效果，使预测结果更接近实际结果。鉴于此，本文将 Stacking 集成思想引入到现有的基于机器学习的径流预报模型中，以雅砻江流域锦屏一级水库为研究对象，预测锦屏一级水库月平均入库流量。首先，采用 GBRT、XGBoost、RF、ET 进行预测，并对预测结果进行统计分析，评价各单一算法可靠性。在此基础上，结合 Stacking 集成学习理论，进一步提升预测效果，并对预测结果进行分析。

2. 研究方法

2.1. 基于 Boosting 的单一算法

Boosting 算法，是一种可以用来减小监督式学习中偏差的机器学习算法，其中各个预测函数必须按照顺序迭代生成。Boosting 算法工作机制为：先从初始训练集训练出一个基学习器，再根据基学习器的表现对样本分

布进行调整，然后基于调整后的样本分布来训练下一个基学习器；如此重复进行，直至整个集成结果达到退出条件，然后将这些学习器进行加权结合[6]。其具体流程描述如图 1 所示。

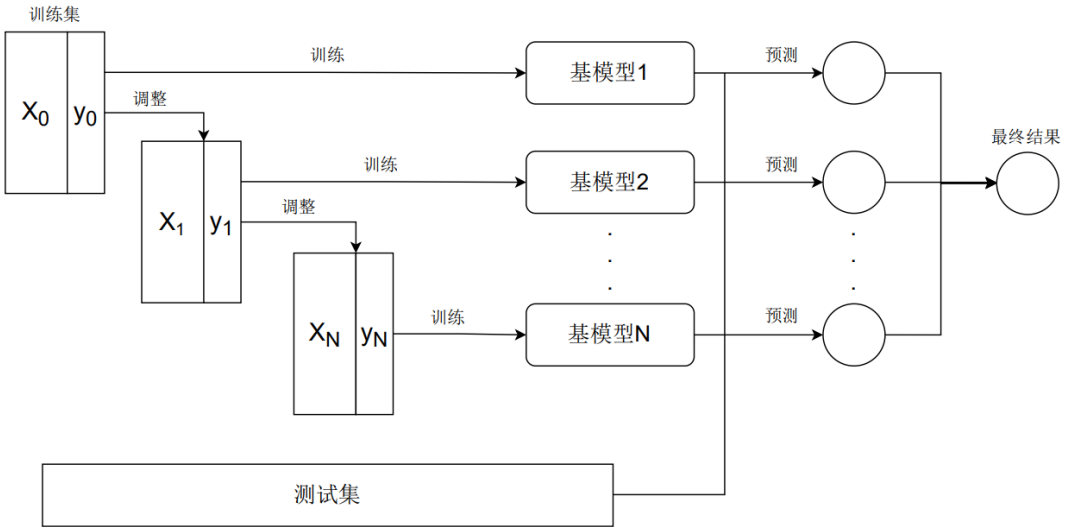

Figure 1. Boosting algorithm description diagram
图 1. Boosting 算法描述示意图

GBRT 是一种以回归树为基本分类器用 Boosting 策略训练出来的模型。其原理是由多棵决策树构成，输出为每棵决策树输出结果的累加，利用梯度提升和回归决策树的组合方式，每次建立新的决策树模型都是在之前模型损失函数的梯度下降方向，使得决策树模型能够不断的改进[7]。

XGBoost 则是在梯度提升决策树的基础上进行了改进，其优势表现在数据处理效率高、效果好、泛化能力强。作为 GBRT 的高效实现，XGBoost 主要从以下三个方面做了优化：

1) 算法本身优化：在损失函数上，加上了正则化部分，并对误差部分做了二阶泰勒展开，使得结果更加准确。

2) 运行效率优化：决策树的建立过程中采用了并行化的选择，提高了算法的运行速度。

3) 健壮性优化：算法加入了 L1 和 L2 正则化项，可以有效的防止过拟合，其泛化能力更强。

2.2. 基于 Bagging 的单一算法

Bagging 算法，又称装袋算法，是一种可降低方差的机器学习算法，与 Boosting 最主要区别是 Bagging 的各个预测函数是并行生成的，可提高运行效率，当与其他回归算法结合时，则可以提高准确率和稳定性。其原理是给定包含 N 个样本的数据集，先随机取出一个样本放入采样中，再把该样本放回初始数据集，使得下次采样时该样本仍有可能被选中[8]。其算法描述如图 2 所示。

RF 作为一种监督式集成学习模型，采用 Bagging 思想利用多棵决策树对样本进行训练的预测的一种分类器。对于一个输入样本，m 棵树会有 m 个分类结果，而 RF 集成了所有的分类投票结果，将投票最多的类别指定为最终的输出，处理回归问题时，以每棵决策树输出的均值为最终结果。ET 与 RF 十分相似，都是有许多随机树构成，主要区别有两点：

1) 对于每个决策树的训练集，RF 是对采样集进行随机采样，以其结果作为每个决策树的训练集。而 ET 使用的是所有样本，只有特征是随机选择的；

2) RF 在特征点的划分上是和传统决策树一样，会基于信息增益、信息增益率、基尼系数、均方差等原则来选择最优特征值；而 ET 在划分决策树上是随机的选择一个特征值进行划分的。

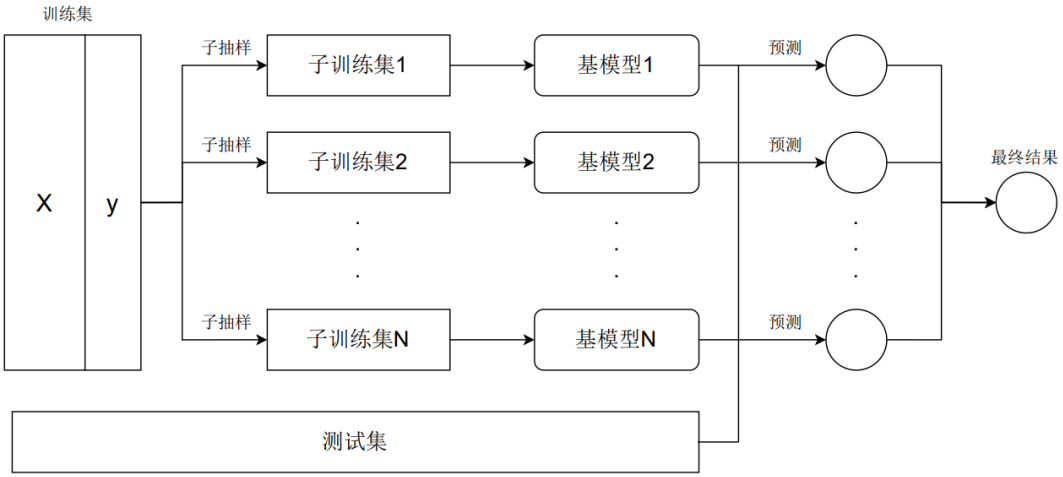

Figure 2. Bagging algorithm description diagram
图 2. Bagging 算法描述示意图

2.3. 基于 Stacking 的集成算法

在 Bagging 和 Boosting 中，弱学习器一般为同一模型，如 RF、ET、GBRT 和 XGBoost 的弱学习器都是决策树。因此想要将不同的模型结合在一起，综合考虑不同模型的优势，则可以采用 Stacking 集成方法。Stacking 集成的思路是先将原始数据的特征作为输入，选取一系列弱学习器作为初级学习器，初级学习器的输出作为次级学习器的输入，最后的得到的输出则为 Stacking 预测结果。其算法描述如图 3 所示。

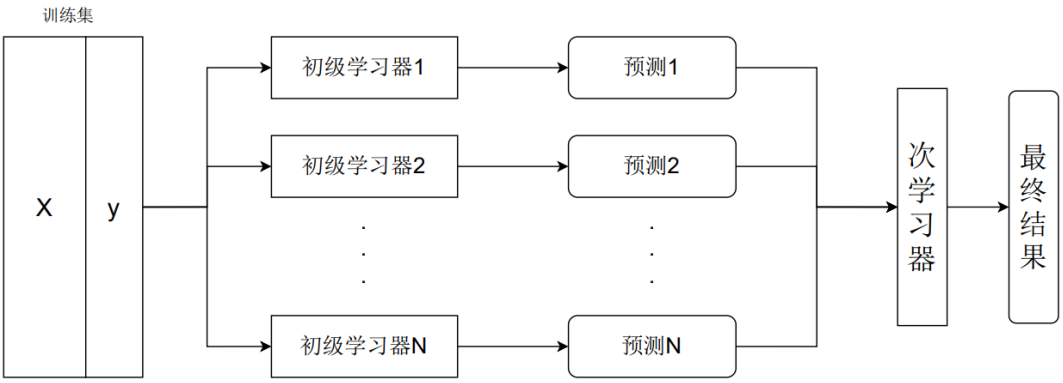

Figure 3. Stacking algorithm description diagram
图 3. Stacking 算法描述示意图

在初级学习器的选取上，学习器个数过多可能会导致过拟合，一般选用 2~3 个模型作为初级学习器，效果会最好。因此，本文在 RF、ET、GBRT 和 XGBoost 中选取结果较好的 3 个模型作为初级学习器，以第一层预测的结果作为预报特征，并采用 Logistic 回归作为次级学习器对最终的结果进行预测。

2.4. 评价指标的选取

回归评价指标主要包括平均绝对误差、均方误差、均方根误差、平均绝对百分误差、拟合优度等[9]。本文选取均方误差(MSE)和拟合优度(R^2)作为模型模拟精度的评价指标。MSE 是各数据偏离真实值的距离平方和的平均数，即误差平方和的平均数，在相同预测长度中，其值越小说明预测结果越好；R^2 是回归平方和在总平方和中所占比率，其值越接近 1，表明其拟合预测性能越好。计算公式分别如下所示：

$$MSE = \frac{1}{N}\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2 \tag{1}$$

$$R^2 = 1 - \frac{\sum_i \left(y_i - \widehat{y_i}\right)^2}{\sum_i \left(y_i - \overline{y_i}\right)^2} \tag{2}$$

同时，根据《水文情报规范》中的中长期预报的规范要求，本文选取预测值和实测值之间的相对误差范围在20%以内的预测为合格预测，误差范围在10%以内的预测为优秀预测。

3. 实例分析

锦屏一级水电站水库位于四川省雅砻江流域，是雅砻江干流下游河段的龙头水库，库容约 77.6×10^8 m^3，水库主要用于发电，兼蓄水、拦沙和防洪。同时，由于该水库是下游河段的控制性水库，对下游其他水库具有显著的补偿效益。因此，准确预报该水电站水库月平均径流量，有利于指导该地区水资源综合开发利用、科学管理和优化调度。

3.1. 单一算法结果对比分析

本文选取锦屏一级1990~2012年共计23年的月径流数据，选取的预报因子为降水量、月平均气温、月平均水汽压、月平均相对湿度、前一年月径流值以及多年平均月径流值，其中月尺度气象数据可从中国气象数据网站上直接获取。以1990年1月~2011年12月共264个月的资料作为训练集进行建模，并以2012年1月~2012年12月共计12个月的资料作为验证集进行验证。分别用RF、ET、GBRT、XGBoost四种算法作为回归模型进行预测，并通过网格搜索对各个模型进行调参得到最佳预测效果。训练集各单一模型的拟合值与实测值的对比结果如图4所示。

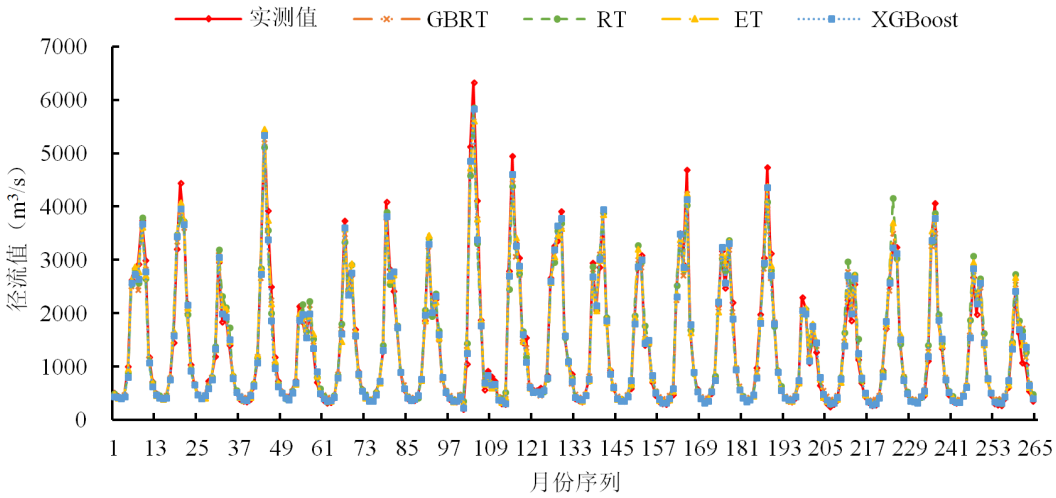

Figure 4. Comparison of the fitted and measured values of the training set RF, ET, GBRT, XGBoost
图 4. 训练集 RF、ET、GBRT、XGBoost 拟合值与实测值对比结果图

训练集各模型的 R^2 分别为：XGBoost (0.9874)、ET (0.9851)、GBRT (0.9812)、RF (0.9726)。同时，由上图可以看出在训练集中，各模型拟合结果与实测值总体比较接近，且各模型之间相差不大，说明在训练集中各个模型中都能很好的拟合出实际径流值。各个模型的预测结果和实际值的对比结果如表1所示。

由表1可得，合格次数从高到低的排序分别是XGBoost (9次)、GBRT (8次)、ET (8次)、RF (6次)；优秀次数从高到底的排序分别是XGBoost (3次)、GBRT (3次)、ET (2次)、RF (2次)。各单模型预测结果与实测值对比如图5所示。

Table 1. Comparison results of RF, ET, GBRT, XGBoost and measured values
表 1. RF、ET、GBRT、XGBoost 与实测值对比结果

月份	实测值(m^3/s)	RF 月径流预测模型		ET 月径流模型		GBRT 月径流模型		XGBoost 月径流模型	
		预测值(m^3/s)	相对误差	预测值(m^3/s)	相对误差	预测值(m^3/s)	相对误差	预测值(m^3/s)	相对误差
1	261	357	36.8	313	19.9	351	34.5	292	11.9
2	233	221	−5.2	273	17.2	269	15.5	258	10.7
3	250	334	33.6	397	58.8	297	18.8	287	14.8
4	401	460	14.7	479	19.5	457	14.0	447	11.5
5	664	654	−1.5	599	−9.8	626	−5.7	542	−18.4
6	1822	2296	26.0	2071	13.7	2085	14.4	1931	6.0
7	4558	3221	−29.3	3428	−24.8	3378	−25.9	3521	−22.8
8	2284	2708	18.6	1858	−18.7	1901	−16.8	1771	−22.5
9	2692	3080	14.4	3421	27.1	2763	2.6	2501	−7.1
10	1838	1130	−38.5	1492	−18.8	1267	−31.1	1357	−26.2
11	800	654	−18.3	728	−9.0	829	3.6	727	−9.1
12	390	531	36.2	523	34.1	490	25.6	458	17.4

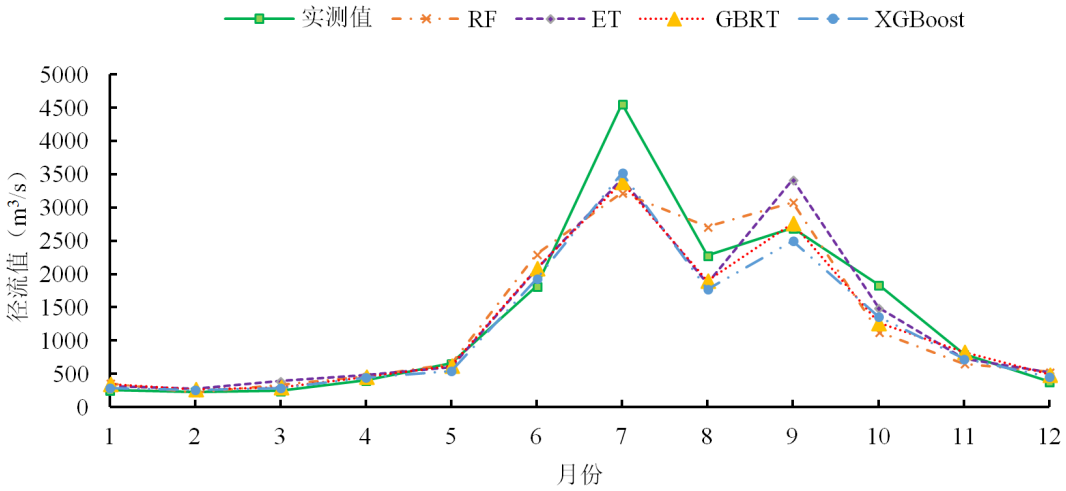

Figure 5. Comparison of the predicted and measured values of RF, ET, GBRT, and XGBoost
图 5. RF、ET、GBRT、XGBoost 预测值与实测值对比结果图

各个模型的预测结果回归指标、优秀率以及合格率如表 2 所示。

Table 2. RF, ET, GBRT, XGBoost evaluation index results
表 2. RF、ET、GBRT、XGBoost 评价指标结果

模型	RF	ET	GBRT	XGBoost
MSE	242085	185886	163864	137347
R^2	0.8064	0.8507	0.8435	0.8696
合格率	0.5	0.67	0.67	0.75
优秀率	0.17	0.17	0.25	0.25

从上图可以看出各个模型与实测结果的总体拟合结果都比较好，但是不同的模型结果还是有些差距的。在非汛期各个模型与的拟合结果与实测值都很接近，但是在汛期误差则相对比较大。通过上表可以得到 RF 的均方误差最大，并且 R^2、合格率以及优秀率都最小，综合四个模型，其效果最差。GBRT 和 ET 总体比较接近，

却各有优劣，GBRT 的 MSE 和优秀率均优于 ET，但是其 R^2 差于 ET。XGBoost 各项指标在四个模型中均为最优，因此，在单一算法的结果中 XGBoost 的预测结果最好。

3.2. 集成算法结果对比分析

由于单一模型的计算结果中，RF 的计算结果相较于剩下三个模型来说，精度偏低，所以在 Stacking 结合策略集成学习模型中，选用 GBRT、XGBoost 和 ET 作为初级学习器，Logistic 回归作为次学习器，采用 5 折交叉验证对初级学习器的结果进行 Stacking 集成。

以 GBRT 为例，首先把整个数据集分为训练集和测试集两部分，然后把训练集分成 5 份，先拿出其中 4 份作为训练集另外 1 份作为测试集，在第一次交叉验证后会得到一个关于这一份测试集的预测值 TrainP1，然后对原来整个数据集的预测值 TestP1。5 折交叉验证，即将上述过程进行 5 次，得到针对 Training Data 数据预测 5 列数据 TrainP1，TrainP2，TrainP3，TrainP4，TrainP5 以及对 Testing Data 数据预测的 5 列数据 TestP1，TestP2，TestP3，TestP4，TestP5。此时得到的新特征 1 就是由 TrainP1，TrainP2，TrainP3，TrainP4，TrainP5 拼凑而成，而新特征 1 对应的 TestP 即为 estP1，TestP2，TestP3，TestP4，TestP5 的平均值，这就是 Stacking 的一个完整的流程。接着再对 ET 和 XGBoost 这两个模型重复以上步骤即可得到新特征 2 和新特征 3 以及对应的 TestP。最后将得到的新特征以及对应的预测值进入到下一层模型采用 Logistic 进行进一步训练得到最终的结果。其过程如图 6 所示。

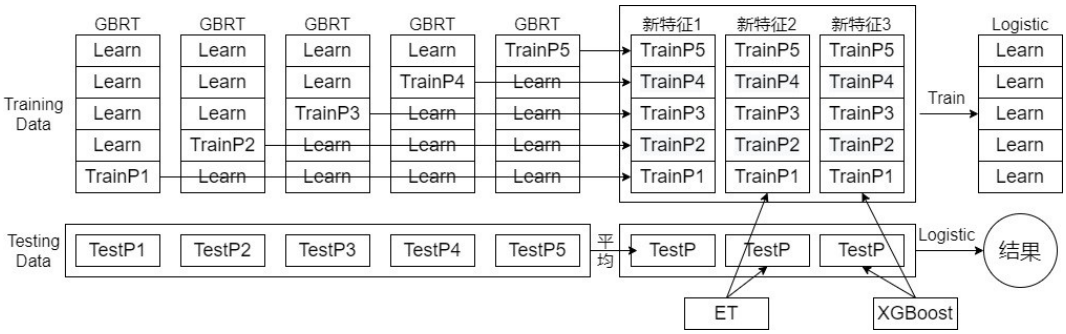

Figure 6. 5-Fold cross-validation diagram
图 6. 5 折交叉验证示意图

训练集中 XGBoost 和 Stacking 集成的拟合结果与实测值对比如图 7 所示。

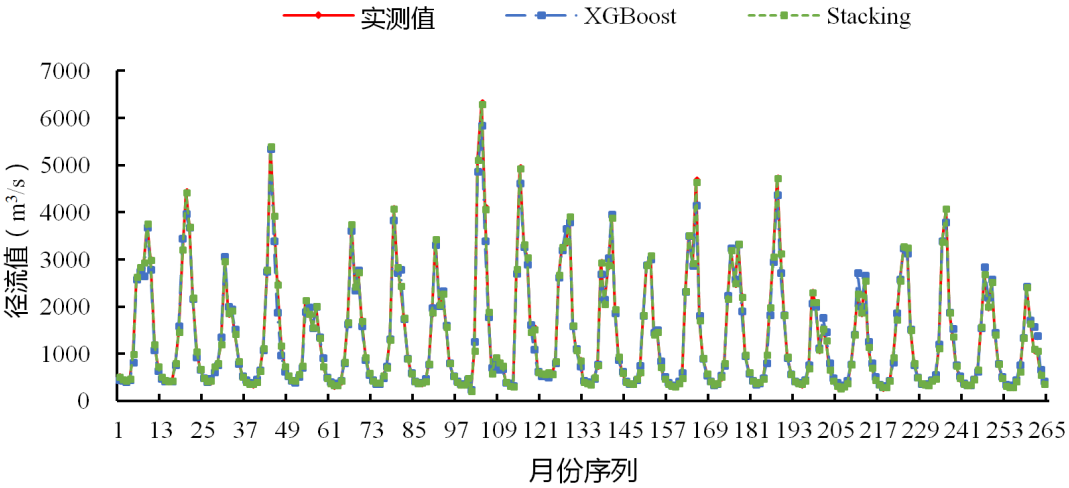

Figure 7. Comparison of the fitted and measured values of the training set XGBoost and Stacking integration
图 7. 训练集 XGBoost、Stacking 集成拟合值与实测值对比结果图

训练集 Stacking 集成的 R^2 为 0.9941，相较于 XGBoost 有所提升，说明 Stacking 集成在训练集中与实测值更为接近，其拟合效果最好。将 Stacking 集成算法和单一模型结果最好的 XGBoost 与实测值的拟合结果如图 8 所。

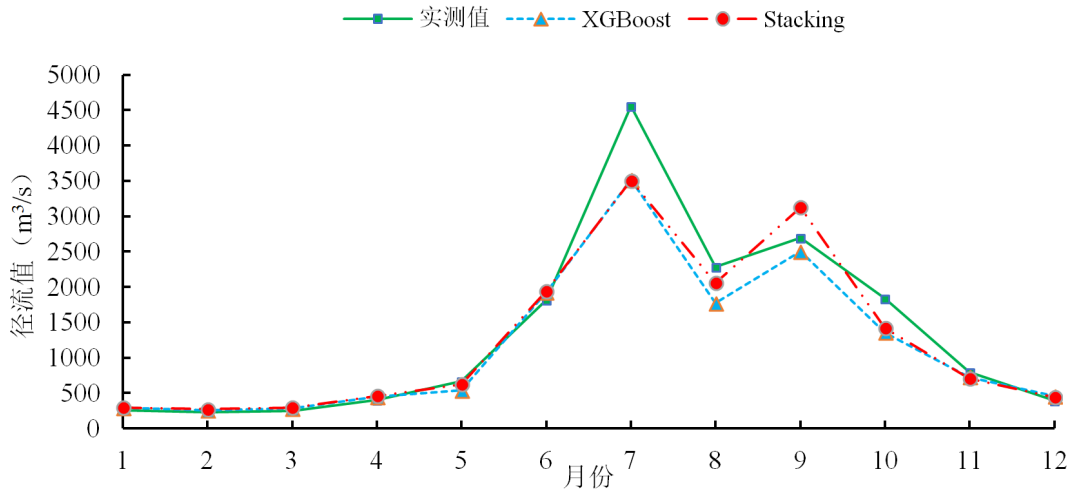

Figure 8. Comparison of the predicted and measured values of XGBoostand Stacking integration
图 8. XGBoost、Stacking 集成预测值与实测值对比结果图

Stacking 集成 MSE、R^2、合格率以及优秀率与 XGBoost 的对比结果如表 3 所示。

Table 3. Stacking integration and XGBoost evaluation index results
表 3. Stacking 集成与 XGBoost 评价指标结果

	MSE	R^2	合格率	优秀率
Stacking 集成	129751	0.8934	0.83	0.25
XGBoost	137347	0.8696	0.75	0.25

从上图可以看出，Stacking 集成和 XGBoost 总体上与实测值更为接近。从上表的对比结果中可得 Stacking 集成除了优秀率效果没有发生变化外，其他三个指标都有正向提升，其中：MSE 下降了 5.854%、R^2 提升了 2.737%、合格率提升了 10.667%。从 Stacking 集成和 XGBoost 的对比结果中，进一步得到：Stacking 集成相对于各单一机器学习模型而言，根据相关评价指标，可认为其预报效果最好。由此可见，相对于单一算法，Stacking 集成对中长期径流预报结果更符合实际。

4. 总结

机器学习作为近几年兴起的一种热门算法，在各种竞赛中都取得了很好的成绩。本文选取机器学习算法中的集成算法对锦屏一级的月径流进行了模拟和预测。在单一算法中，除了 RF 月径流预测结果较差外，GBRT、XGBoost 和 ET 模型结果表现都比较好，其中 XGBoost 月径流模型行综合效果最好，能够满足一般中长期径流预报精度要求。

在此基础上，择优选取其中三个模型作为初级学习器，以其结果作为下一层的输入，选取 Logistic 回归作为次学习器，采用 5 折交叉验证，构建基于 Stacking 集成策略的预测模型进行预测。其最终结果相对于单一模型的结果都有所提升，体现了 Stacking 集成算法在径流预报中的优势，为中长期径流预报提供了新的可行思路，为分析和研究中长期径流的变化规律提供了新的研究方法。

基金项目

国家自然科学基金(51709105)；中央高校基本科研业务费专项资金资助(2020MS026; 2019MS031)。

参考文献

[1] 朱双. 流域中长期水文预报与水资源承载力评价方法研究[D]: [博士学位论文]. 武汉: 华中科技大学, 2017.
ZHU Shuang. Studies on watershed long-term hydrological forecast and evaluation method of water resources carrying capacity. Ph.D. Thesis, Wuhan: Huazhong University of Science & Technology. 2017. (in Chinese)

[2] KIRK M. Python 机器学习实践: 测试驱动的开发方法[M]. 北京: 机械工业出版社, 2018.
KIRK M. Thoughtful machine learning with Python: A test-driven approach. Beijing: China Machine Press, 2018. (in Chinese)

[3] 李伶杰, 王银堂, 胡庆芳, 刘定忠, 张安富, 巴亚荃. 基于随机森林与支持向量机的水库长期径流预报[J]. 水利水运工程学报, 2020(4): 33-40.
LI Lingjie, WANG Yintang, HU Qingfang, LIU Dingzhong, ZHANG Anfu and BAYAQUAN. Long-term reservoir runoff forecast based on random forest and support vector machine. Journal of Water Resources and Water Transport Engineering, 2020(4): 33-40. (in Chinese)

[4] 左岗岗. 基于机器学习的渭河流域径流预测系统研究[D]: [硕士学位论文]. 西安: 西安理工大学, 2017.
ZUO Ganggang. The research of WEI River runoff prediction system based on machine learning. Master's Thesis, Xi'an: Xi'an University of Technology, 2017. (in Chinese)

[5] 许斌, 杨凤根, 郦于杰. 两类集成学习算法在中长期径流预报中的应用[J]. 水力发电, 2020, 46(4): 21-24+34.
XU Bin, YANG Fenggen and LI Yujie. Application of two types of integrated learning algorithms in mid- and long-term runoff forecasting. Hydropower, 2020, 46(4): 21-24+34. (in Chinese)

[6] 胡丹. 面向视觉跟踪的深度学习模型设计与优化研究[D]: [博士学位论文]. 西安: 西北工业大学, 2017.
HU Dan. Model design and optimization of deep learning for visual tracking. Ph.D. Thesis, Xi'an: Northwestern Polytechnical University, 2017. (in Chinese)

[7] 陈宏, 邓芳明, 吴翔, 付智辉. 基于梯度提升决策树的电力电子电路故障诊断[J]. 测控技术, 2017, 36(5): 9-12+20.
CHEN Hong, DENG Fangming, WU Xiang and FU Zhihui. Power electronic circuit fault diagnosis based on gradient boosting decision tree. Measurement and Control Technology, 2017, 36(5): 9-12+20. (in Chinese)

[8] 侯舒凯. 基于集成学习方法的 MINIST 手写数字识别[J]. 通讯世界, 2018(8): 236-237.
HOU Shukai. MINIST handwritten digit recognition based on integrated learning method. Communication World, 2018(8): 236-237. (in Chinese)

[9] 宋俊杰. 三峡流域中长期径流预报模型精度评定综合分析及优化方法研究[D]: [硕士学位论文]. 武汉: 华中科技大学, 2013.
SONG Junjie. Comprehensive analysis and optimization method of the accuracy assessment of the medium and long-term runoff forecast model in the Three Gorges Basin. Master's Thesis, Wuhan: Huazhong University of Science and Technology, 2013. (in Chinese)