DOI:10.20040/j.cnki.1000-7709.2023.20220677

基于机器学习算法的万家寨水库排沙预测研究

颜小飞[1,2],郭秀吉[1,2],孙龙飞[1,2]

(1. 黄河水利委员会黄河水利科学研究院,河南 郑州 450003;
2. 水利部黄河下游河道与河口治理重点实验室,河南 郑州 450003)

摘要:为克服水库排沙多因素、非线性复杂关系建立难题,实现水库排沙准确预测,利用万家寨水库2002~2020年水沙系列数据,基于XGBoost、KNN、SVR、RF四种机器学习算法分别预测分析水库出库含沙量。结果表明,利用机器学习算法可有效预测综合考虑不同影响因素的水库排沙;不同机器学习算法在水库排沙预测的适用性有所不同,对比之下,基于RF算法建立的水库排沙预测模型的决定系数R^2最高为0.934 9,平均绝对误差及均方根误差均最小,分别为2.974、4.886,其预测效果更优于其他三种算法。研究成果可为水库排沙精确预测及调度方案优化提供参考。

关键词:水库排沙;含沙量;机器学习算法;预测模型

中图分类号:TV145 **文献标志码**:A **文章编号**:1000-7709(2023)03-0079-04

1 概况

黄河万家寨水利枢纽位于黄河中游上段托克托至龙口峡谷河段内(图1)。万家寨水库于1998年10月下闸蓄水,水库总库容为$8.96\times10^8\ m^3$,调节库容为$4.45\times10^8\ m^3$,死库容为$4.51\times10^8\ m^3$,设计调洪库容为$3.02\times10^8\ m^3$;最高蓄水位为980.00 m,正常蓄水位为977.00 m,校核洪水位为979.10 m,汛期限制水位为966.00 m,最低发电水位为952.00 m。枢纽主要任务是供水结合发电调峰,同时兼有防洪、防凌作用,干流入库站头道拐水文站归属黄河水利委员会,出库站万家寨水文站为水库专用站。为准确预测万家寨水库排沙量,考虑到水库排沙的影响因素较多,非线性复杂关系较难建立,拟采用机器学习方法预测水库排沙,并对比分析不同机器学习算法的预测效果,以期为水库调度及安全运行提供参考。

图1 万家寨水库区域位置示意图

Fig. 1 The location of Wanjiazhai Reservoir

2 研究方法

研究方法为:①给出不同机器学习算法的基本原理;②确定水库排沙过程的主要影响因素,并构建综合考虑不同影响因素的水库排沙预测模型;③提出基于不同机器学习算法的水库排沙预测流程;④通过万家寨水库实例分析,对比不同算法模型的预测准确性,并最终确定优选的机器学习算法水库排沙预测模型。

2.1 机器学习算法基本原理

2.1.1 XGBoost算法

XGBoost算法是在梯度提升决策树GBDT算法基础上,在目标函数中引入正则项以约束损失函数的下降和模型整体的复杂度,防止模型过

收稿日期:2022-04-07,**修回日期**:2022-06-13

基金项目:国家重点研发计划(2021YFC3200400);黄河水利科学研究院科技发展基金专项项目(黄科发202102);黄河水利科学研究院基本科研业务费专项(HKY-JBYW-2018-18)

作者简介:颜小飞(1985-),男,硕士、工程师,研究方向为测绘物探仪器应用,E-mail:yxf513@126.com

通讯作者:孙龙飞(1993-),男,博士、工程师,研究方向为水工结构数值模拟与施工质量实时控制,E-mail:1287215017@qq.com

拟合，同时对损失函数使用二阶泰勒展开，直接利用了损失函数的一、二阶导数值，进一步优化了模型效率与精度[1]。XGBoost 算法的目标函数为：

$$Q=\sum_{i=1}^{n} l(y_i,\hat{y}_i)+\sum_{j=1}^{m}\Omega(f_j) \tag{1}$$

式中，y_i 为样本真实值；$\hat{y}_i$ 为样本的预测值；l 为反映 y_i 与 $\hat{y}_i$ 差异的损失函数；n 为样本数；$\Omega(f_j)$为正则项，用于控制模型复杂度，避免过拟合；f_j 为第 j 个树的模型；j 为分类回归树个数。

2.1.2 KNN 算法

KNN 算法的基本思想是将当前新数据的每个特征与具有相似特征的样本数据值相匹配，然后输出样本数据中最相似的 K 个数据的属性值[2]。其定义为：

$$D(X,Y)=\left(\sum_{i=1}^{n}|x_i-y_i|^p\right)^{1/p} \tag{2}$$

式中，$D(X,Y)$为 X、Y 之间距离，用以衡量样本间的相似程度；x_i 为样本 X 的第 i 个特征；y_i 为样本 Y 的第 i 个特征；p 为距离计算方式，当 $p=1$ 时为曼哈顿距离，当 $p=2$ 时为欧拉距离。

2.1.3 SVR 算法

SVR 算法是在支持向量机 SVM 分类的基础上，引入核函数和损失函数，通过非线性映射将数据映射至高维特征空间，找到最优拟合超平面，使所有训练样本与该面的总偏差最小，以解决非线性回归问题的方法。给定训练样本集 $D=\{(x_i,y_i)|i=1,2,\cdots,n\}$，SVR 的目标是找到一个回归函数 $f(x)$，使其与实际输出 y 尽可能接近[3]：

$$f(x)=\boldsymbol{\omega}^{\mathrm{T}}\varphi(x)+\boldsymbol{b} \tag{3}$$

式中，$\boldsymbol{\omega}$ 为法向量；$\varphi(\cdot)$ 为非线性映射函数；$\boldsymbol{b}$ 为偏移量。

引入松弛变量 ξ_i、ξ_i^*，则最优化问题转化为：

$$\begin{cases}\min\left(\dfrac{1}{2}\|\boldsymbol{\omega}\|^2+C\displaystyle\sum_{i=1}^{n}(\xi_i,\xi_i^*)\right)\\ \text{s.t.}\begin{cases}f(x_i)-y_i\leqslant\varepsilon+\xi_i \quad i=1,2,\cdots,n\\ y_i-f(x_i)\leqslant\varepsilon+\xi_i^*\\ \xi_i\geqslant 0\\ \xi_i^*\geqslant 0\end{cases}\end{cases} \tag{4}$$

式中，C 为惩罚因子；ε 为不敏感损失函数。

2.1.4 RF 算法

RF 算法是一种包含多个决策树的算法，其利用随机方式生成每棵决策树的每个节点，再由不同节点分叉形成多个决策树，每棵决策树之间无关联，进而构成一个“随机森林”，并最终通过汇总决策树的结果，进行因变量的回归和分类[4]。其基本步骤如下。

步骤 1 从原始样本集中随机抽取 N 个训练样本，有放回的进行 k 轮抽取，得到 k 个相互独立的训练集，用选择好的训练集训练一个决策树。

步骤 2 对于不同的训练集，可建立不同的训练模型，且决策树的每个节点根据具体问题进行分裂。每棵树遵循分枝优度准则，一直到不能再分裂为止。

步骤 3 取不同模型预测结果的平均值作为最后的预测结果。

2.1.5 不同机器学习算法使用范围及优缺点对比

对比 XGBoost、KNN、SVR、RF 四种机器学习算法的使用范围及优缺点，结果见表 1。

表 1 不同机器学习算法使用范围及优缺点对比

Tab. 1 Comparison of serviceable range and advantages and disadvantages of different machine learning algorithms

算法类型	使用范围	优点	缺点
XGBoost	分类和回归	灵活性强、正则化防止过拟合	需遍历数据集、预排序过程复杂、消耗内存
KNN	分类和回归	思想简单、训练时间复杂度低、准确度较高	计算量大、预测速度相比逻辑回归算法较慢
SVR	回归	计算复杂度较低、可解决高维问题及非线性问题	对参数和核函数的选择较敏感
RF	分类和回归	对大样本训练速度快、模型方差小、泛化能力强	模型效果会受划分较多的特征影响

2.2 水库排沙预测模型构建

水库排沙过程的主要影响因素包括入库流量 Q_1、入库含沙量 S_1、出库流量 Q_2、坝前水位 Z_{W}、坝前水位差 ΔZ_{W}、累计淤积量 G 共 6 个输入变量。其中需要说明的是“坝前水位差 ΔZ_{W}”为考虑坝前水位变化对排沙影响的时效性，即计算前一天的水位与当天水位的差值，对于当天排沙的影响所引入的变量。

此外，以出库含沙量 S 作为唯一输出变量，建立各影响因素与水库出库含沙量的综合预测模型，所建立的模型为：

$$S=f(Q_1,S_1,Q_2,Z_{\mathrm{W}},\Delta Z_{\mathrm{W}},G) \tag{5}$$

式中，S 为现有数据中出库含沙量；$f(\ )$为回归函数。

2.3 基于机器学习算法的水库排沙预测流程

利用机器学习算法综合考虑各影响因素的水库排沙预测步骤如下。

步骤 1 选择合适的样本数据，并对数据进行归一化预处理。

步骤 2 进行数据分割，确定训练样本和测试样本，其中输入、输出变量见式(5)。

步骤 3 将训练样本代入不同机器学习算法程序中进行训练，同时调整优化算法参数组合，最

终得到综合考虑各影响因素的水库排沙预测模型。

步骤 4　将测试数据的输入变量代入模型进行计算,得到预测出库含沙量,并与实际出库含沙量作比较,以评估不同模型预测精度,以平均绝对误差 M_{MAE}、均方根误差 R_{RMSE}、决定系数 R^2 作为评估指标。

步骤 5　确定优选的机器学习算法的水库排沙预测模型。

3　万家寨水库排沙预测

3.1　原始数据统计处理

以万家寨水库 2002～2020 年水沙系列数据为基础数据,统计数据的输入、输出变量。为排除闸门开闭这一未知情况的影响,剔除出库含沙量在 1 kg/m^3 以下的数据(认为是由闸门关闭引起)。部分原始数据输入输出变量统计见表 2。

表 2　部分原始数据输入输出变量统计

Tab. 2　Part of input and output variable statistics of raw data

入库流量	入库含沙量	出库流量	坝前水位/m	坝前水位差/m	累计淤积量/10^8 m^3	出库含沙量
2 830	2.05	2 870	953.24	0.190	3.172	1.96
1 620	5.58	1 600	955.08	−1.453	3.183	2.90
⋮	⋮	⋮	⋮	⋮	⋮	⋮
257	2.79	138	959.56	−0.997	4.616	16.70
596	2.57	620	953.47	1.375	4.548	4.13
⋮	⋮	⋮	⋮	⋮	⋮	⋮

注:入库流量、出库流量单位为 m^3/s;入库含沙量、出库含沙量单位为 kg/m^3。

将所有原始数据中 454 个(80%)样本数据用于训练,剩余 113 个(20%)样本数据用于预测。此外,考虑各变量之间的量纲差异,为消除不同变量之间量纲差异带来的影响,对所有数据进行归一化无量纲预处理,方法为:

$$\omega' = (\omega - \bar{\omega})/\sigma \tag{6}$$

式中,ω'为归一化后数据;ω 为原始数据;$\bar{\omega}$ 为原始数据平均值;σ 为原始数据标准差。

3.2　不同机器学习算法模型预测结果对比分析

将所有数据代入 XGBoost、KNN、SVR、RF 四种算法中建立预测模型,得到不同机器学习算法模型预测得到的出库含沙量与实际出库含沙量对比结果见图 2。由图 2 可知,整体上不同机器学习算法所建立模型得到的预测出库含沙量与实际出库含沙量的分布情况基本一致,且除个别点外,绝大部分数据点结果均接近,表明了机器学习算法应用于综合考虑各影响因素的水库排沙预测的有效性。同时,相比较之下,RF、XGBoost、KNN 三种算法特征点(出库含沙量大于 30 kg/m^3 的样本点)预测结果的准确性优于 SVR

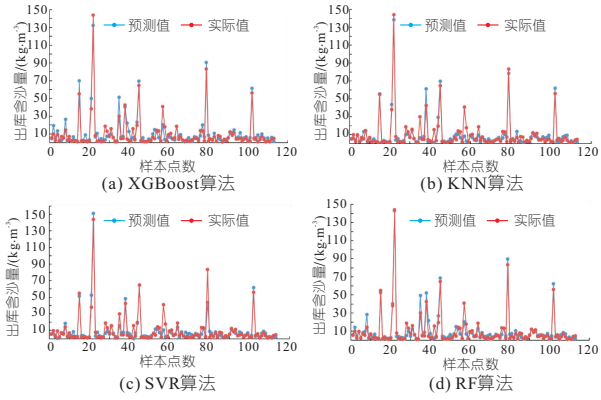

图 2　不同机器学习算法模型预测值与实际值对比结果

Fig. 2　Comparison of predicted and actual values of different machine learning algorithm models

算法预测结果的准确性。

统计预测出库含沙量与实际出库含沙量之间相关性见图 3。由图 3 可见,不同机器学习算法模型所得预测出库含沙量与实际出库含沙量之间均满足线性关系,其关系表达式的斜率均在 0.93 以上,接近 1;同时各模型决定系数 R^2 均在 0.88 以上(表 3),表明两者间相关性良好。相比较之下,XGBoost 算法与 KNN 算法模型 R^2 分别为 0.911 8、0.910 4,两者预测结果相近;而 RF 算法模型 R^2 最高为 0.934 9,SVR 算法模型 R^2 最低为 0.885 0。进一步统计不同模型预测值与实际值两者间的平均绝对误差 M_{MAE} 及均方根误差 R_{RMSE},结果见表 3。

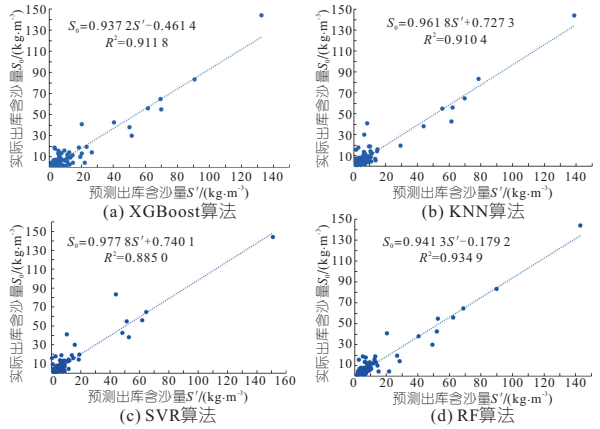

图 3　不同机器学习算法模型预测值与实际值相关性分析

Fig. 3　Correlation analysis of predicted and actual values of different machine learning algorithm models

表 3　不同机器学习算法模型预测结果统计

Tab. 3　Prediction results statistics of different machine learning algorithm models

算法类型	决定系数 R^2	平均绝对误差 M_{MAE}	均方根误差 R_{RMSE}
XGBoost	0.911 8	3.672	5.701
KNN	0.910 4	3.054	5.559
SVM	0.885 0	3.327	6.270
RF	0.934 9	2.974	4.886

由表 3 可知，各模型下预测值与实际值之间误差均相对较小，表明机器学习算法应用于水库排沙预测有效，在一定程度上可实现综合考虑不同影响因素的水库排沙预测。其中，RF 算法模型的平均绝对误差 M_{MAE} 为 2.974，均方根误差 R_{RMSE} 为 4.886 均最小，同时结合其决定系数 R^2 最高，表明针对万家寨水库排沙预测过程，在现有数据条件下，RF 算法模型预测精度优于其他三种算法模型。

4 结论

a. 利用万家寨水沙系列数据，基于不同机器学习算法，建立综合考虑入库流量、入库含沙量、出库流量、坝前水位、坝前水位差、累计淤积量影响的水库排沙预测模型。

b. 不同机器学习算法模型得到的预测出库含沙量与实际出库含沙量的分布情况基本一致，预测出库含沙量与实际出库含沙量之间相关性良好，各模型决定系数 R^2 均在 0.88 以上。

c. 基于 RF 算法所建立模型决定系数 R^2 最高为 0.934 9，且平均绝对误差 M_{MAE} 为 2.974，均方根误差 R_{RMSE} 为 4.886 均最小，表明相对于其他模型，RF 算法模型在水库排沙预测方面具有更高的准确性和精度。

参考文献：

[1] 王梦雅，刘丽冰，熊桂龙，等. 面向袋式除尘器的大数据挖掘 XGBoost 优化算法研究[J]. 电子测量与仪器学报，2020，34(7)：159-167.

[2] 周鑫，谢晖，付山，等. 基于 KNN 算法的中心带孔圆板拉深—翻孔变形方式的研究[J]. 锻压技术，2021，46(7)：53-59.

[3] 刘泉声，王栋，朱元广，等. 支持向量回归算法在地应力场反演中的应用[J]. 岩土力学，2020，41(增刊 1)：319-328.

[4] 吴芳，李映雪，张缘园，等. 基于机器学习算法的冬小麦不同生育时期生物量高光谱估算[J]. 麦类作物学报，2019，39(2)：217-224.

Research on Sand Discharge Prediction of Wanjiazhai Reservoir Based on Machine Learning Algorithms

YAN Xiao-fei[1,2], GUO Xiu-ji[1,2], SUN Long-fei [1,2]

(1. Yellow River Institute of Hydraulic Research, Yellow River Conservancy Commission, Zhengzhou 450003, China;
2. Key Laboratory of Lower Yellow River Channel and Estuary Regulation, MWR, Zhengzhou 450003, China)

Abstract: In order to overcome the difficult problem of establishing multi-factor and non-linear complex relationship of reservoir sand discharge and achieve its accurate prediction, four machine learning algorithms including XGBoost, KNN, SVR and RF were used to predict and analyze the sand content of reservoir outflow based on the series data of Wanjiazhai reservoir from 2002 to 2020, respectively. The results show that the use of machine learning algorithms can effectively realize the reservoir discharge prediction considering different influencing factors. The applicability of different machine learning algorithms in reservoir discharge prediction varies. In comparison, the highest coefficient of determination R2 of the reservoir discharge prediction model based on RF algorithm is 0.9349, and the corresponding average absolute error and root mean square error are the smallest, which are 2.974 and 4.886, respectively. The prediction effect of the RF algorithm is better than the other three algorithms. The proposed method can provide a theoretical basis for accurate prediction of reservoir sand discharge and optimization of scheduling scheme.

Key words: reservoir sand discharge; sand content; machine learning algorithm; prediction model

(上接第 107 页)

Numerical Simulation of Hydraulic Characteristics of Different Side Vertical Seam Fishway

LI Yang[1], HAN Lei[1], TIAN Zhen-hua[1], DI Gao-jian[1], LI Shu-hang[1], YE Kun-he[2]

(1. Heilongjiang Province Hydraulic Research Institute, Harbin 100050, China;
2. School of Water Conservancy and Electric Power, Heilongjiang University, Harbin 150080, China)

Abstract: Vertical seam fishway has gradually attracted the attention of hydraulic engineering field because it can adapt to large amplitude water level, obvious energy dissipation effect and stable flow pattern. In this paper, the influence of the length width ratio of the pond on the hydraulic characteristics of the opposite side vertical slit fishway was studied by numerical simulation. The results show that the change of the length width ratio of the pond has little impact on the dissipation rate per unit volume; The attenuation along the main flow area first increases with the increase of the length width ratio of the pond, and then remains unchanged. The larger the length width ratio of the chamber is, the closer the maximum velocity in the chamber is to the vertical joint; The ratio of the main flow velocity to the maximum flow velocity (the maximum variation of the main flow velocity along the way) is within a certain range. The length width ratio of the pond is between 1.00-1.13, which can obtain better water flow pattern, larger reflux area and better water quality in the mainstream area of the fishway.

Key words: aspect ratio of pond chamber; opposite side type; vertical seam fishway; hydraulic characteristics