Modeling and Simulation 建模与仿真, 2023, 12(4), 4183-4194
Published Online July 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.124381

Hans汉斯

基于特征选择的SSA-XGBoost水质量预测研究

赵　桐，刘媛华*

上海理工大学管理学院，上海

收稿日期：2023年6月5日；录用日期：2023年7月21日；发布日期：2023年7月28日

摘　要

为了能够更好的实现水资源的利用，针对目前对水质预测研究中存在的特征参数复杂、单一模型预测模型精度和适应度欠佳等问题，提出了一种基于XGBoost的水质预测模型。首先利用主成分分析方法对特征进行选择，降低问题复杂度和计算成本，并对数据中的缺失值进行填充，其次采用麻雀搜索算法(SSA)对XGBoost模型中的参数进行优化，采用优化后的参数对水质进行预测。最后在不同实验条件下对水质进行预测，实验结果证明，本文提出的SSA-XGBoost方法与现有方法相比，具有更优秀的性能。

关键词

水质预测，XGBoost，麻雀搜索算法，特征选择

Research on SSA-XGBoost Water Quality Prediction Based on Feature Selection

Tong Zhao, Yuanhua Liu*

Business School, University of Shanghai for Science and Technology, Shanghai

Received: Jun. 5th, 2023; accepted: Jul. 21st, 2023; published: Jul. 28th, 2023

Abstract

In order to better realize the utilization of water resources, a water quality prediction model based on XGBoost is proposed in view of the problems existing in the current research on water quality prediction, such as complex characteristic parameters, poor precision and fitness of a single model prediction model, etc. Firstly, the principal component analysis method is used to select features, reduce problem complexity and computational costs, and fill in missing values in the data. Secondly, the sparrow search algorithm (SSA) is used to optimize the parameters in the XGBoost

*通讯作者。

文章引用: 赵桐, 刘媛华. 基于特征选择的 SSA-XGBoost 水质量预测研究[J]. 建模与仿真, 2023, 12(4): 4183-4194.
DOI: 10.12677/mos.2023.124381

model, and the optimized parameters are used to predict water quality. Finally, water quality was predicted under different experimental conditions, and the experimental results showed that the SSA-XGBoost method proposed in this paper has better performance compared to existing methods.

Keywords

Water Quality Prediction, XGBoost, Sparrow Search Algorithm, Feature Selection

1. 引言

水是一种巨大的自然资源，水资源在饮用水、农业、娱乐和工业用水等各种方面都至关重要，但这些水资源很大程度上会受到工业、人类行为或其他自然过程的污染，对环境和人类健康都产生了直接影响，导致疾病和死亡率持续增加，准确、灵敏的水质预测模型，能够有效的服务于水污染的治理和水资源的利用，因此对水质进行预测是非常必要的一项研究。

水质预测研究主要集中在机器学习模型研究方面，由于机器学习模型在处理非线性等复杂数据时具有更高精度、鲁棒性、有效性以及可靠性，因此在处理水质相关数据方面有显著的优势。Heddam 等人[1]使用了具有乙状激活功能、放射状、在线顺序和最佳修剪特性的 ELM 神经网络模型，并与 MLP 和 MLR 进行了比较，对溶解氧指标进行预测，实验证明此 ELM 神经网络模型预测溶解氧的准确性更高。Mitrović 等人[2]采用了 18 个水质特征指标作为蒙特卡洛模拟的 ANN 模型的输入量，采用 WQ 单变量输出的方式，对水质进行预测，模型预测效果优秀，适用于多目标场景，具有高精度、效率高等特点。Tiwari 等人[3]采用多输入变量对水质指数(WQI)进行预测，此研究采用了两种聚类技术，即模糊 C-均值(FCM)和基于 ANFIS 的减法聚类(SC1-ANFIS)，通过实验证明，SC1-ANFIS 对 WQI 的预测性优于 FCM。Rankinen 等人[4]提出了可管理非正态误差分布的广义线性模型(GLM)和可处理非线性和缺失数据的增强回归树(BRT)模型，考虑到气候变化、农业措施和环境政策等间接因素，对未来各种情景下的水质情况进行预测。Ahmed [5]等使用两个 ANN 模型(即 FFNN 和 RBFNN)预测 Surma 河的溶解氧(DO)，实验发现两个 ANN 模型都具有较好的预测能力，相对而言 FFNN 比 RBFNN 预测精度更高一些，此水质预测模型可以应用于水管理和处理系统。查文舒等[6]通过全连接神经网络、卷积神经网络、循环神经网络等多种网络结构进行微分方程的求解，大幅提高泛化能力与应用价值。张皓等[7]提出一种多重 T-S 型模糊神经网络 PID 温度控制算法，利用 T-S 型模糊神经网络的单输出特性，建立能分别输出 PID3 个参数的 3 重网络模型，模型稳定性高，抗干扰能力强。李晶晶等[8]以长短期记忆(LSTM)网络为基础提出了一种新的数据驱动空间负荷预测方法，分析神经网络内部的时序，避免数据消沉现象，确定训练数据空间的相关性，提高了预测速度。陆继翔等[9]提出了一种基于卷积神经网络(CNN)和 LSTM 网络的混合模型短期负荷预测方法，将海量的历史负荷数据、气象数据、日期信息以及峰谷电价数据按时间滑动窗口构造连续特征图作为输入，先采用 CNN 提取特征向量，将特征向量以时序序列方式构造并作为 LSTM 网络输入数据，再采用 LSTM 网络进行短期负荷预测，预测精度得到明显提升。

在水质预测的相关研究中，存在着影响因子众多、数据指标复杂以及单一模型预测精度低等问题，

因此本文采用主成分分析方法(PCA)作为特征选择的方法，采用 XGBoost 作为预测模型，并利用麻雀搜索算法(SSA)对 XGBoost 模型的参数进行优化。

2. 关键技术

2.1. PCA 主成分分析

主成分分析法作为多元统计中的重要部分，是一种较为常见的无监督的数据降维方法，通过某种线性投影，将高维的数据映射到低维的空间中，并使得投影中维度上的数据方差最大。

假设有 n 个样本，且每个样本有 p 个变量，则可以构成一个 $n \times p$ 的原始数据矩阵，将原始数据进行标准化处理，计算方法如(1)所示：

$$\mathbb{Z}_{ij} = \frac{X_{ij} - X_i}{S_i} \tag{1}$$

式中，$\mathbb{Z}_{ij}$ 为标准化后的数据，X_{ij} 为原始数据，X_i 是第 i 个指标的样本均值；S_i 为第 i 个指标的标准差。

基于标准化的矩阵，计算相关系数 R。根据相关系数矩阵 R 的特征方程，求解 R 的特征值和特征向量，R 的特征值为 $\lambda_i\left(i=1,2,\cdots,p\right)$ 且 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$，$\lambda_i$ 是主成分特征向量所对应的特征值，即各主成分的方差值，其大小代表了原始样本在主成分中所占的比重，每个特征值对应的特征向量为 $l_{gi}\left(i=1,2,\cdots,p\right)$，通过这些特征向量把标准化的指标转化为主成分[10]，计算方法如(2)所示：

$$F_g = Z \times L_g\left(g=1,2,\cdots,p\right) \tag{2}$$

计算贡献率 λ_i 和累计贡献率 η_i，计算方法如(3) (4)所示：

$$\tau_i = \frac{\lambda_i}{\sum_{k=1}^{p}\lambda_k}\left(i=1,2,\cdots,p\right) \tag{3}$$

$$\eta_i = \frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{p}\lambda_k}\left(i=1,2,\cdots,p\right) \tag{4}$$

确定主成分并计算各主成分综合得分：首先要确定主成分的个数，主要方法有两种[11] [12]，一是主成分方差累计贡献率大于 80%、二是各主成分特征值大于 1.0，然后由主成分的方差贡献率通过加权求和法得出主成分的综合得分。

2.2. XGBoost 算法

XGBoost 是基于 CART 树的一种 boosting 算法，它是通过多个学习器的学习，来不断降低模型值和实际值的差。其基本思想是不断生成新的树，每棵树都是基于上一颗树和目标值的差值来进行学习。模型输出表达式为 $y_i = \sum_{k=1}^{t} f_k\left(x_i\right)$，其中：$K$ 为树的总个数，f_k 表示第 k 颗树，y_i 表示样本 x_i 的预测结果。

模型的目标函数由两部分组成，一是模型误差，即样本真实值和预测值之间的差值，二是模型的结构误差，即正则项，用于限制模型的复杂度。目标函数的计算方法如(5)所示

$$Obj\left(\theta\right) = \sum_{i=1}^{n} l\left(y_i, \hat{y}\right) + \sum_{k=1}^{k}\Omega\left(f_k\right) \tag{5}$$

其中：$l\left(y_i, \hat{y}\right)$ 为样本 x_i 的损失函数，$\Omega\left(f_k\right)$ 表示第 k 颗树的正则项。

XGBoost 通过不断地分裂添加树，每次添加树的过程即为学习一个新函数 $f\left(x\right)$，去拟合前一次预测

的残差。当训练完成得到 k 棵树，对样本的分数进行预测，每个叶子节点对应一个分数，将每颗树的分数相加即可得到该样本的预测值。计算方法如(6)所示：

$$\hat{y}_i^{(t)}=\sum_{k=1}^{t}f_k\left(x_i\right)=\hat{y}_i^{(t-1)}+f_t\left(x_i\right) \tag{6}$$

其中：f_k 表示第 k 棵树，$\hat{y}_i^{(t)}$ 表示组合 t 棵树模型对样本 x_i 的预测结果。

优化目标函数。损失函数采用均方误差，目标函数为：

$$\begin{aligned}Obj\left(\theta\right)&=\sum_{i=1}^{n}\left(y_i-\left(\hat{y}_i^{(t-1)}+f_t\left(x_i\right)\right)\right)^2+\Omega\left(f_t\right)+C1\\&=\sum_{i=1}^{n}\left[2\left(y_i-\hat{y}_i^{(t-1)}\right)f_t\left(x_i\right)+f_t\left(x_i\right)^2\right]+\Omega\left(f_t\right)+C1\end{aligned} \tag{7}$$

对于目标函数中的正则项，从每一棵回归树考虑，其模型可表示为：

$$f_t\left(x\right)=\omega_{q(x)},\omega\in R^T,q:R^d\to\left\{1,2,\cdots,T\right\} \tag{8}$$

其中：ω 为叶子节点 q 的分数，$q\left(x\right)$ 表示样本 x 对应的叶子节点，T 为该树的叶子节点个数。ω_j^2 为其中一棵回归树。

为了避免过拟合，对树上叶子节点的分数 ω 进行正则化，XGBoost 的目标函数可写为：

$$Obj\left(\theta\right)=\sum_{i=1}^{n}l\left(y_i,\hat{y}^{(t-1)}+f_t\left(x_i\right)\right)+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2+C \tag{9}$$

其中：γ 为叶子个数，ω_j^2 表示 ω 的 L2 模平方。

利用泰勒展开式去将目标函数进行进一步的变形，且令 $g_i=\partial_{\hat{y}^{(t-1)}}l\left(y_i,\hat{y}^{(t-1)}\right)$，$h_i=\partial^2{}_{\hat{y}}l\left(y_i,\hat{y}^{(t-1)}\right)$，由于在第 t 棵树，y_i 是真实值，即已知，第 t 颗回归树是根据前面的 $t-1$ 颗回归树的残差得来的，相当于 $t-1$ 颗树的值 $\hat{y}_i^{(t-1)}$ 是已知的，因此 $l\left(y_i,\hat{y}^{(t-1)}\right)$ 是常数。去除所有常数项，并将 $\sum_{i=1}^{n}\left[g_if_t\left(x_i\right)+\frac{1}{2}h_if_t\left(x_i\right)^2\right]$ 看作是每个样本在第 t 棵树的叶子节点的分数相关函数的结果之和，则目标函数可表示为：

$$\begin{aligned}Obj^t\left(\theta\right)&\approx\sum_{i=1}^{n}\left[g_if_t\left(x_i\right)+\frac{1}{2}h_if_t\left(x_i\right)^2\right]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2\\&=\sum_{i=1}^{T}\left[\left(\sum_{i\in I_j}g_i\right)\omega_j+\frac{1}{2}\left(\sum_{i\in I_j}h_i+\lambda\right)\omega_j^2\right]+\gamma T\end{aligned} \tag{10}$$

式中：T 为第 t 棵树中总叶子节点的个数；$I_j=\left\{i\middle|q\left(x_i\right)=j\right\}$ 表示在第 j 个叶子节点上的样本；ω_j 为第 j 个叶子节点的分数值。定义 $G_j=\sum_{i\in I}g_i$，$H_j=\sum_{i\in I}h_i$，通过对 ω_j 求导等于 0，可以得到 $\omega_J^*=-\frac{G_j}{H_j+\lambda}$，则目标函数表示为：

$$Obj^*=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T \tag{11}$$

2.3. 麻雀搜索算法

麻雀作为一种群居类动物，种类繁多，对环境的适应性较强，有较高的灵敏度，飞行能力强。在麻雀觅食过程中，具有不同的分工，具体可以分为发现者和加入者。发现者和加入者的身份是动态切换的，只要能够寻找到更丰富的食物来源，每只麻雀都可以成为发现者，但发现者和加入者所占整个种群数量

的比重是不变的。

假设麻雀种群的初始规模数是 n，用 $X=\left\{X_{1,1},X_{1,2},\cdots,X_{2,1},\cdots,X_{n,d}\right\}$ 表示。d 表示麻雀个体所附带的维度。算法中，发现者有较强搜索能力即具备较好适应度值，因此更容易搜寻到食物。在整个空间中，其位置更接近最优解的位置。在每轮迭代搜索的过程中，发现者会进行位置更新，计算方式为：

$$X_{i,j}\left(t+1\right)=\begin{cases}X_{i,j}\left(t\right)\cdot\exp\left(-\dfrac{i}{\alpha\cdot t_{\max}}\right), & R<ST\\ X_{i,j}\left(t\right)+Q\cdot L, & R\geq ST\end{cases}\tag{12}$$

其中，X_{ij} 表示种群中第 i 只麻雀在第 j 维的位置；t 是算法当前的迭代次数，$t_{\max}$ 是最大迭代次数；α 是(0, 1]之间的随机值；R 的取值范围是[0, 1]，表示算法中麻雀个体遇到危险时的预警值；ST 的取值范围是[0.5, 1]，表示安全值；Q 是服从正态分布的随机数；L 表示大下为 $l\times d$，元素都是 1 的矩阵。

当 $R\geq ST$ 时，表示部分麻雀已发现危险，发现者按正态分布随机移动到当前位置附近。当 $R<ST$ 时，表示此时麻雀群体搜索的环境周围不存在危险，发现者可以进行大范围的搜索操作，往外搜索食物。随着种群迭代次数的增加，$\exp\left(-\dfrac{i}{\alpha\cdot t_{\max}}\right)$ 项的取值范围将随之减少，即对应到麻雀个体的每一维上的值都将减少。

当一些加入者找不到食物补充能量时，会监控发现者在捕食过程中的行为。当发现者搜索到丰富的食物后，加入者会离开自己所在的位置去抢夺发现者的食物，如果能够抢到食物就会进行补充能量，否则会被迫去其他区域觅食。加入者的位置更新描述如下：

$$X_{i,j}\left(t+1\right)=\begin{cases}Q\cdot\exp\left(\dfrac{X_{\omega}-X_{i,j}\left(t\right)}{i^2}\right), & i>\dfrac{1}{2}\\ X_{i,j}\left(t\right)+\left|X_{i,j}\left(t\right)-X_{p}\left(t\right)\right|\cdot A^{*}\cdot L, & i\leq\dfrac{1}{2}\end{cases}\tag{13}$$

其中，X_{p} 表示发现者适应度值最优的位置；X_{ω} 表示当前空间中适应度值最差的位置；A 是维度 $l\times d$，元素都是 1 或者−1 的矩阵；A^{*} 满足关系式 $A^{*}=A^{\mathrm{T}}\left(AA^{\mathrm{T}}\right)^{-1}$。当 $i>\dfrac{n}{2}$ 时，表明该加入者处于十分饥饿的状态，利用一个标准正态分布随机数与以自然对数为底指数函数的积，控制其取值符合正态分布，即获取更多的能量。当 $i\leq\dfrac{n}{2}$ 时，其过程可解释为在当前最优位置附近随机找到一处位置，且每一维据最优位置方差较小，值较为稳定。

觅食过程中麻雀个体遇到危险时，会往内部或者其他同伴靠拢。该过程的麻雀个体更新位置的方法如下：

$$X_{i,j}\left(t+1\right)=\begin{cases}X_{b}\left(t\right)+\beta\cdot\left|X_{i,j}\left(t\right)-X_{b}\left(t\right)\right|, & f_i\neq f_b\\ X_{i,j}\left(t\right)+K\cdot\dfrac{\left|X_{i,j}\left(t\right)-X_{\omega}\left(t\right)\right|}{\left(f_i-f_{\omega}\right)+\varepsilon}, & f_i=f_b\end{cases}\tag{14}$$

其中，X_{b} 是当前的全局最优位置；β 表示步长控制参数，满足均值为 0，方差为 1 的正态分布的随机值；K 是[−1, 1]之间的随机值，表示麻雀的移动方向；ε 是接近零的常数，防止分母为 0 的情况出现；f_t 表示第 i 只麻雀的适应度值；f_b 和 f_{ω} 表示当前麻雀种群的最优和最差适应度值。当 $f_t\neq f_b$ 时，表示第 i 只麻雀的在觅食圈的外围，较容易受到外来者的攻击；当 $f_t=f_b$ 时，表示一些麻雀意识到了危险，需要向

周围的同伴靠拢来保障自己的安全[13]。

3. SSA-XGBoost 模型预测模型建立

准确、灵敏的水质预测模型对水资源的有效利用和管控具有重要意义，由于溶解氧与水质指标参数具有复杂的非线性关系，且单一模型对水质预测精度欠佳，因此本文提出了基于 XGBoost 的水质预测模型，通过麻雀算法中个体位置的更新，实现对 XGBoost 中参数的优化。本文选取溶解氧作为模型输出，以此来准确高效的判断水质情况。溶解氧是指溶到水体中的分子氧，其来源主要包括水体和大气平衡状态下溶解到水体中的氧以及水体中进行化学、生物反应形成的氧。水中的溶解氧含量如果较高将会有利于水中污染物的降解，可以加快水的净化速度，如果溶解氧的含量较低则水中污染物降解的速度较慢。溶解氧不仅是衡量水质的重要指标，也是水体净化的重要因素。因此采用溶解氧作为衡量水质的标准，通过预测溶解氧实现对水质的预测。

通过缺失值填充、特征选择和参数优化三个方面结合，提出水质预测模型 SSA-XGBoost。溶解氧的影响因素包括 pH、电导率、浊度、高锰酸钾指数、氨氮、总磷、总氮。针对溶解氧影响因素众多且关系复杂的问题，本文通过 PCA 方法对水质参数进行相关性分析以选择模型的输入特征，减少冗余信息导致的误差，降低问题复杂度。而针对采集数据中存在缺失值的问题，通过皮尔逊系数对不同缺失值填充方法进行分析比较，以此寻找最优的缺失值填充方法。

其具体实现步骤如下：

步骤 1. 对水质相关数据进行采集。

步骤 2. 根据主成分分析从候选参数中选择输入特征，降低问题复杂度。

步骤 3. 通过皮尔逊系数对不同缺失值填充方法进行分析，选择最优的缺失值填充方法。

步骤 4. 初始化设置水质预测模型的种群数量 pop 为 30，对个体的位置、种群边界和最大迭代次数进行初始化，计算适应度值。

步骤 5. 根据适应度函数更新个体的位置。

步骤 6. 判断是否满足终止条件，终止条件即达到最大迭代次数或适应度值达到设定阈值，满足终止条件则输出 XGBoost 最优参数，否则返回步骤 4。

步骤 7. 以获取到的最优参数代入到 XGBoost 中，得到水质预测模型。

步骤 8. 在线运行阶段，根据采集的参数计算输入特征，并利用 XGBoost 模型进行水质预测。

模型的处理流程如图 1 所示。

4. 实验分析

4.1. 特征选择与数据预处理

本文使用的数据取自 2023 年 2 月 1 日至 5 日的上海市太湖流域以及长江流域的明星路桥、临江、吴淞口、前卫村桥、七效港西桥等 19 个断面的 551 个水质样本数据，监测站点每 4 小时发布一次实时数据。在模型建立前要对数据进行降维处理，确定影响水质溶解氧的变量数目，使得样本数据更为直观方便。首先利用主成分分析方法计算出水质指标的累计方差贡献率，将方差贡献率累加大于 80%的指标作为选取的特征变量，各特征的方差贡献率如图 2 所示，其中电导率、浊度、高锰酸盐指数和总磷四个参数的方差贡献率累加超过 80%，因此选用电导率、浊度、高锰酸盐指数和总磷作为水质预测模型的输入变量。

对于样本数据由于温度、传感器故障、检验操作步骤等情况存在数据缺失的问题，为了提高预测准确度，需要对数据进行预处理。根据本文数据特点采用零值填充、平均值填充、最小值填充三种缺失值填充方法对样本数据进行填充，并通过皮尔逊相关系数(Pearson Correlation Coefficient)对不同缺失值填充

方法效果进行评估，其中皮尔逊相关系数计算公式如式(15)所示：

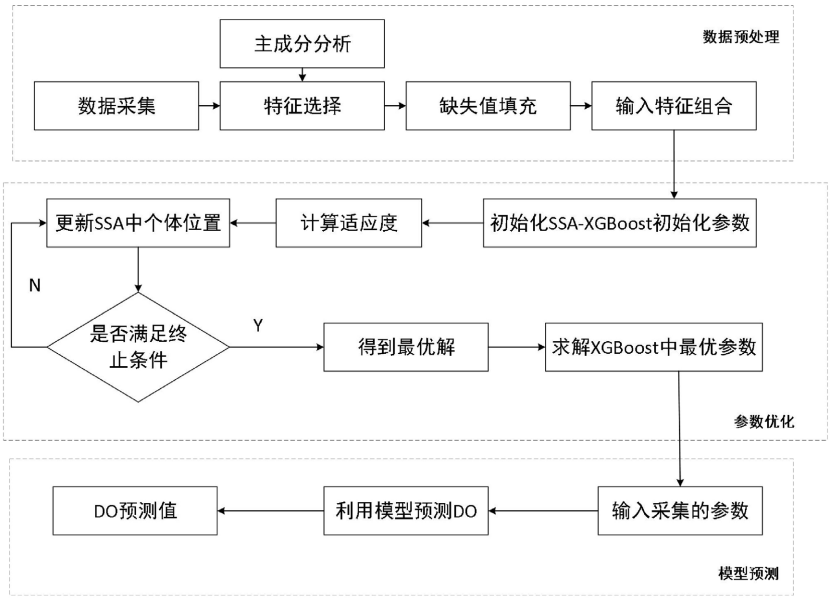

Figure 1. Model flowchart
图 1. 模型流程图

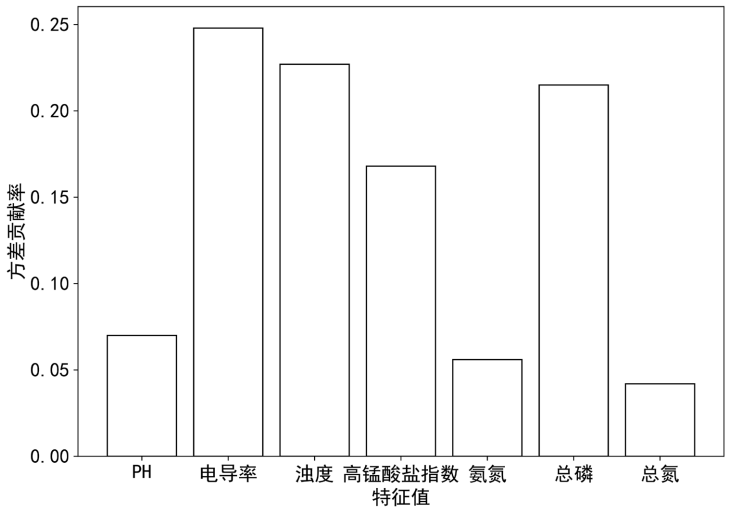

Figure 2. Variance contribution rate of each feature
图 2. 各特征方差贡献率

$$\rho_{XY}=\frac{\operatorname{cov}\left(Para,DO\right)}{\sigma_{Para}\sigma_{DO}}=\frac{n\sum_{i=1}^{n}para_i do-\sum_{i=1}^{n}para_i\sum_{i=1}^{n}do_i}{\sqrt{n\sum_{i=1}^{n}para_i^2-\left(\sum_{i=1}^{n}para_i\right)^2}-\sqrt{n\sum_{i=1}^{n}do_i^2-\left(\sum_{i=1}^{n}do_i\right)^2}} \tag{15}$$

其中，Para 为水质指标，cov(Para, DO)为水质指标和溶解氧(DO)之间的协方差，σ_{Para} 和 σ_{DO} 为水质指标

和溶解氧的标准差。ρ_{XY} 取值范围为[−1, 1]，其中皮尔逊系数越接近 1，代表水质指标与 DO 的相关性越高。

不同缺失值填充方法的皮尔逊系数对比图如图 3 所示。为节约计算成本，采用不同的缺失值填充方法，选取对 DO 影响程度最大的 4 个参数进行相关性分析，其中图 2 为参数与 DO 的皮尔逊相关系数对比情况，采用平均值填充方法使参数与 DO 的相关性有显著提升。

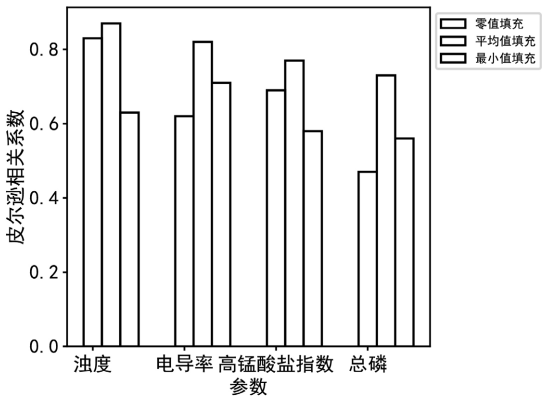

Figure 3. Pearson coefficient analysis of different missing value filling methods
图 3. 不同缺失值填充方法皮尔逊系数分析

4.2. 仿真环境与评价指标

基于 SSA-XGBoost 的溶解氧预测模型是在 Intel(R) Core(TM) i7-10510U (8 核)，内存 16 GB，Win10 64 位操作系统，编程语言为 python 的开发环境中进行仿真实验。采用 SSA 对学习目标参数进行优化。n_estimator 为学习器的数量，learning_rate 为学习率，max_depth = 365 为叶最大深度，gamma 为损失减小阈值。优化后的 XGBoost 在训练过程中的参数取值为 n_estimator = 417，learning_rate = 0.51，max_depth = 365，gamma = 0.83。

为了更准确的验证模型的预测效果，本文采用均方根误差(RMSE)、平均绝对误差(MAE)、决定系数(R^2)的两个评价指。如式：

均方根误差(RMSE)：

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i-y_i\right)^2} \tag{16}$$

平均绝对误差(MAE)：

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i-y_i\right| \tag{17}$$

决定系数(R2)：

$$R^2=1-\frac{\sum_i\left(\hat{y}^{(i)}-y^{(i)}\right)^2}{\sum_i\left(\bar{y}-y^{(i)}\right)^2} \tag{18}$$

式中，y_i 为第 t 天的溶解氧含量；$\hat{y}_i$ 为第 t 天的溶解氧含量的预测值；n 为预测样本数。均方根误差是来

衡量观测值同真值之间的偏差，RMSE 指标越小，说明模型的预测精度越高；决定系数是用来评价模型系数拟合优度，R2 越大越好。当预测值与真实值完全一致时，R2 达到最大值 1。

4.3. 不同数据量的模型性能评价

为测试模型在不同规模数据集中的性能，本文取 500 条样本数据分为 10 组，每组五十条，在不同数据量下进行实验，以证明模型在不同数据量下的鲁棒性，采用 SSA-XGBoost 模型进行仿真实验并对预测结果进行了统计分析，如图 4~图 6 所示。

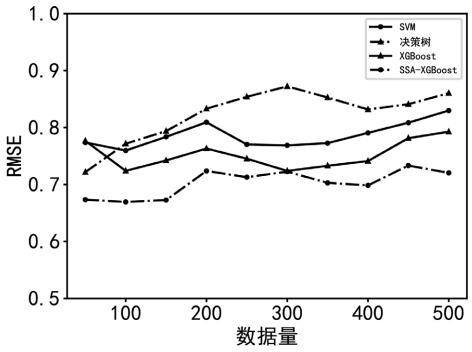

Figure 4. RMSE under different data volumes
图 4. 不同数据量下的 RMSE

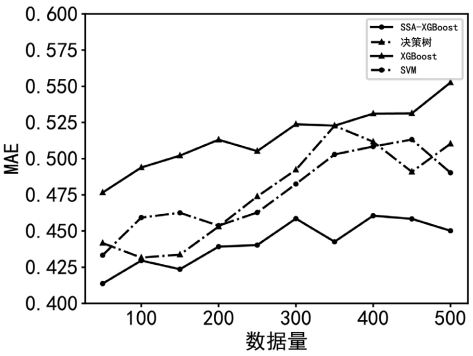

Figure 5. MAE under different data volumes
图 5. 不同数据量下的 MAE

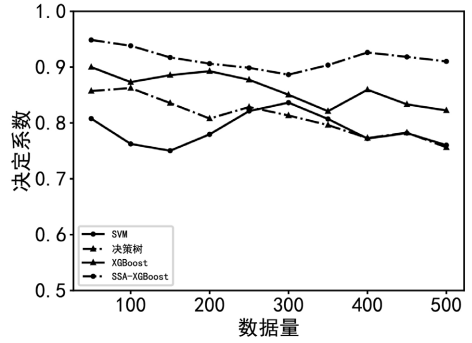

Figure 6. Coefficient of determination under different data volumes
图 6. 不同数据量下的决定系数

由图可知，随着训练数据集中样本数量从 100~500 变化，SSA-XGBoost 的 RMSE 和决定系数虽有波动，但总体保持平稳，在不同数据量情况下，SSA-XGBoost 都具有最小的 RMSE、MAE 和最大的决定系数，而其他模型的性能则随数据量增加出现明显的下降。

4.4. 与现有方法性能的比较

在进行水质指标溶解氧的预测问题时，将经过主成分分析特征选择的溶解氧数据作为 SSA-XGBoost 预测模型的输入，取 80%的数据为训练集，20%数据为测试集，并与支持向量机(SVM)、XGBoost、决策树、SSA-XGBoost 预测模型进行对比分析，由图 7~图 9 可知，SSA-XGBoost 的均方根误差和决定系数都具有最好的性能且波动较小。

4.5. 预测结果

采用 SSA-XGBoost 模型对溶解氧进行预测，取数据集中的 70%作为模型的训练集，取数据集中的 30%作为模型的测试集，在不同实验条件下，SSA-XGBoost 模型都具有最好的预测性能，在测试集中对溶解氧真实值和 SSA-XGBoost 预测值进行对比，如图 10 所示。由图可知，采用 SSA-XGBoost 模型的预测值和真实值拟合程度较高，具有良好的预测能力。

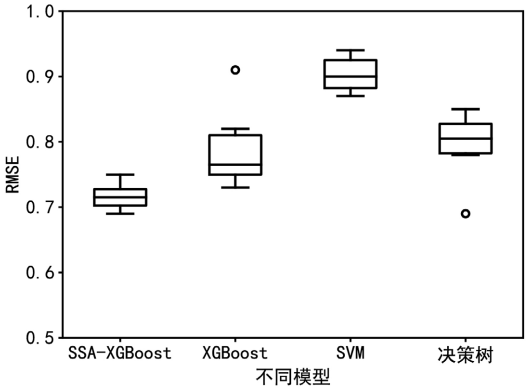

Figure 7. RMSE under different models
图 7. 不同模型下的 RMSE

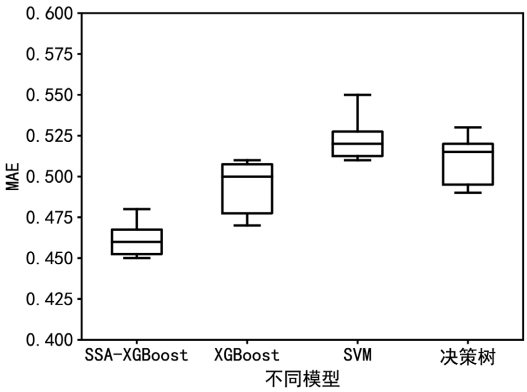

Figure 8. MAE under different models
图 8. 不同模型下的 MAE

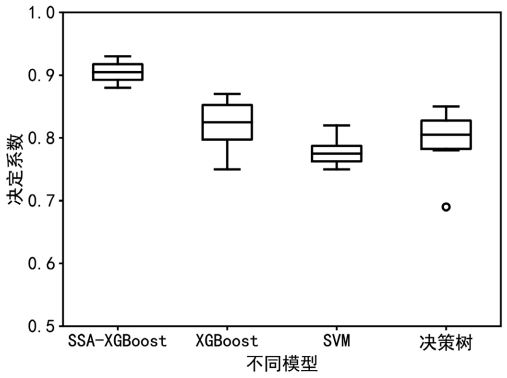

Figure 9. Coefficient of determination under different models
图 9. 不同模型下的决定系数

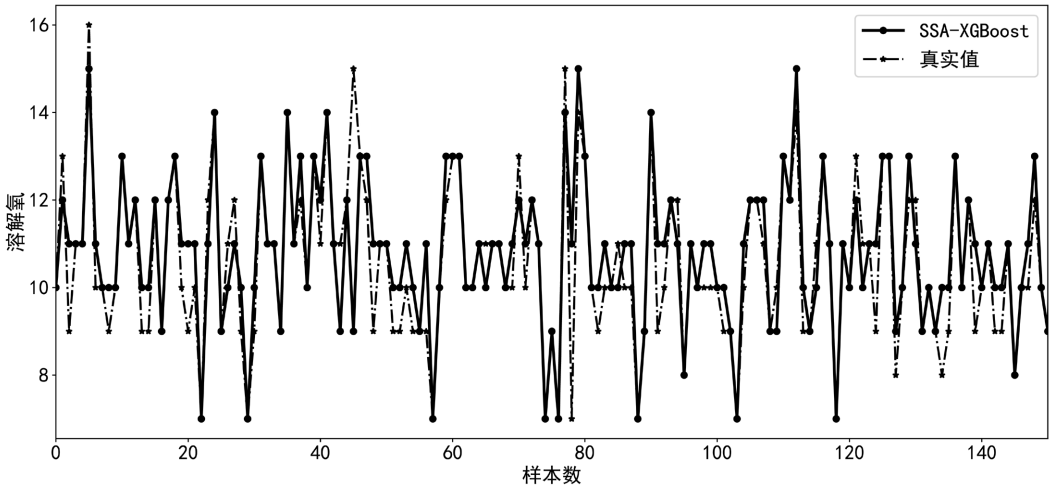

Figure 10. Prediction Results
图 10. 预测结果图

5. 结论与展望

本文采用主成分分析(PCA)进行特征选择，结合麻雀搜索算法和 XGBoost 算法，提出了 SSA-XGBoost 预测模型，以最优超参数实现水质预测。研究采用 PCA 分析水质指标与溶解氧之间的相关性，确定了预测模型的输入特征，降低了变量之间的耦合性，消除了信息冗余对预测精度的影响，通过皮尔逊系数分析方法确定了最优缺失值填充方法为平均值填充。实验通过 SSA-XGBoost、SVM、XGBoost、决策树四种算法对上海市的水质指标溶解氧进行预测，测试结果表明，本文提出的 SSA-XGBoost 方法预测误差更小，且该方法预测结果的 RMSE、R^2 波动均优于其他现有模型。表明 SSA-XGBoost 模型可以更好地预测上海地区未来的水质变化。

参考文献

[1] Heddam, S. and Kisi, O. (2017) Extreme Learning Machines: A New Approach for Modeling Dissolved Oxygen (DO) Concentration with and without Water Quality Variables as Predictors. *Environmental Science and Pollution Research*, **24**, 16702-16724. https://doi.org/10.1007/s11356-017-9283-z

[2] Mitrović, T., Antanasijević, D., Lazović, S., Perić-Grujić, A. and Ristić, M. (2019) Virtual Water Quality Monitoring

at Inactive Monitoring Sites Using Monte Carlo Optimized Artificial Neural Networks: A Case Study of Danube River (Serbia). *Science of the Total Environment*, **654**, 1000-1009. https://doi.org/10.1016/j.scitotenv.2018.11.189

[3] Tiwari, S., Babbar, R. and Kaur, G. (2018) Performance Evaluation of Two Anfis Models for Predicting Water Quality Index of River Satluj (India). *Advances in Civil Engineering*, **2018**, 1-10. https://doi.org/10.1155/2018/8971079

[4] Rankinen, K., Cano Bernal, J.E., Holmberg, M., Vuorio, K. and Granlund, K. (2019) Identifying Multiple Stressors That Influence Eutrophication in a Finnish Agricultural River. *Science of the Total Environment*, **658**, 1278-1292. https://doi.org/10.1016/j.scitotenv.2018.12.294

[5] Ahmed, A.A.M. (2017) Prediction of Dissolved Oxygen in Surma River by Biochemical Oxygen Demand and Chemical Oxygen Demand Using the Artificial Neural Networks (ANNs). *Journal of King Saud University—Engineering Sciences*, **29**, 151-158. https://doi.org/10.1016/j.jksues.2014.05.001

[6] 查文舒, 李道伦, 沈路航, 张雯, 刘旭亮. 基于神经网络的偏微分方程求解方法研究综述[J]. 力学学报, 2022, 54(3): 543-556.

[7] 张皓, 涂雅培, 高瑜翔, 唐军, 黄天赐. 基于多重模糊神经网络的PID温度控制算法[J]. 西华大学学报(自然科学版), 2023, 42(4): 58-65+81.

[8] 李晶晶, 张永敏, 田桂林, 崔胜胜, 严洁. 基于LSTM神经网络的数据驱动空间负荷预测方法[J]. 电子设计工程, 2022, 30(22): 154-157.

[9] 陆继翔, 张琪培, 杨志宏, 涂孟夫, 陆进军, 彭晖. 基于 CNN-LSTM 混合神经网络模型的短期负荷预测方法[J]. 电力系统自动化, 2019, 43(8): 131-137.

[10] 韩伟, 李钢. 主成分分析在地区科技竞争力评测中的应用[J]. 数理统计与管理, 2006(5): 512-517.

[11] 方红卫, 孙世群, 朱雨龙, 等. 主成分分析法在水质评价中的应用及分析[J]. 环境科学与管理, 2009, 34(12): 152-154.

[12] 刘臣辉, 吕信红, 范海燕. 主成分分析法用于环境质量评价的探讨[J]. 环境科学与管理, 2011, 36(3): 183-186.

[13] Xue, J. and Shen, B. (2020) A Novel Swarm Intelligence Optimization Approach: Sparrow Search Algorithm. *Systems Science & Control Engineering*, **8**, 22-34. https://doi.org/10.1080/21642583.2019.1708830