

基于输入优化与 XGBoost 的污水处理厂运行电耗调控研究

王子建¹, 宗轩宇¹, 余海洋¹, 周旭雯¹, 彭赵旭^{1*}

(1. 郑州大学土木与水利学院, 河南郑州, 450001)

摘要: 为了探索降低污水厂处理吨水能耗的调试方法, 以进水流量、进水氨氮、出水 COD、MLSS 等水厂日常运维表现的十四个参数为对象, 采用 OLS、Cp 准则、Lasso 等九种方法对其与吨水电耗的相关性进行筛选, 并以 8 个相关性最强的参数为自变量、吨水电耗为因变量, 利用 XGBoost 进行建模。结果表明, 所建 XGBoost 中对吨水电耗的影响强弱依次是进水流量、出水 TN、进水氨氮、好氧池末端 DO、进水 TP、氨氮负荷、厌氧池末端 ORP、出水 TP (F score 依次是 118.0、105.0、82.0、79.0、78.0、66.0、64.0、56.0)。发现处理水量的规模效应、TN 的处理程度是影响水厂能耗的关键因素。本研究结果可以为水厂的智慧节能降耗提出一定的理论借鉴与技术支持。

关键词: 节能; 吨水电耗; 变量优选; 水量; TN

Research on Operation and Consumption Control of Wastewater Treatment Plants based on Input Optimization and XGBOOST

Zijian Wang¹, Xuanyu Zong¹, Haiyang Yu¹, Xuwen Zhou¹, Zhaoxu Peng^{1*}

(1. School of Civil and Hydraulic Engineering, Zhengzhou University, Henan, 450001)

Abstract: In order to explore the decrease in the treatment of tonne water energy consumption, the 14 parameters of the daily operation and maintenance performance of water supply, water ammonia nitrogen, effluent COD, MLSS, using OLS, CP guidelines, lasso, etc. Nine methods were screened to the correlation between ton water power consumption, and the strongest parameters of 8 relevance were independent variables, and the tons of water consumption were modified by XGBoost. The results show that the impact of the tons of water power consumption in the XGBoost is strong in turn is an inlet flow, TN, ammonia nitrogen, aerobic tank end DO, water TP, ammonia nitrogen load, anaerobic pool end ORP, and water TP (F score is 118.0, 105.0, 82.0, 79.0, 78.0, 66.0, 64.0, 56.0) in turn. It is found that the scale effect of treatment water is found, the degree of processing of TN is a key factor affecting the energy consumption of water plant. The results of this study can provide a certain theoretical reference and technical support for the wisdom energy saving reduction of the water plant.

Keywords: Energy saving; ton water power consumption; variable is preferred; water volume; TN

基金项目: 河南省高等学校重点科研项目(20B560018); 国家水体污染控制与治理科技重大专项(2015ZX07204-002); 郑州大学 2021 年度大学生创新创业训练计划项目。

作者简介: 王子建(2001-), 男, 山东潍坊人, 郑州大学本科生, accamonster@163.com

* 彭赵旭(1983-), 男, 郑州大学副教授, 主要从事污水脱氮除磷研究, pzx@zzu.edu.cn

1. 引言

当前我国多数污水处理厂普遍存在运行能耗高的情况（年耗电相当于 8 万吨标准煤），2020 年全国约产生 11 亿 m^3 城镇污水，处理总费用近 14 亿元，其中曝气电费约 7 亿元（主要能耗在曝气环节）。随着我国碳达峰碳中和战略的推行，提高污水处理效率、降低污水处理厂能耗成为亟待解决的行业问题。

目前为了降低曝气能耗，常见的方法有自动化控制曝气流程、优化运行方式、改进设备等^[1-4]。青岛张村河水质净化厂引进曝气精确分配与控制系统，以空气流量信号作为控制信号，溶解氧和出水氨氮信号作为辅助控制信号，可使污水厂吨水鼓风机耗电由 $0.133 \text{ kW}\cdot\text{h}/\text{m}^3$ 降至 $0.100 \text{ kW}\cdot\text{h}/\text{m}^3$ ，降低 24.8%^[5]。魏全源等提出更换空气悬浮离心鼓风机、曝气头，及加强运行管理等措施。采用空气悬浮离心鼓风机较单级高速离心鼓风机节省约 12%，较罗茨风机节省 31%^[6]。张晓霞提出合理使用变频器、内部优化、曝气设备设计优化等方法。应用变频泵与原工艺相比可以节约 12% 的能源^[7]。由此可见，目前的节能方式对水厂日常运维综合表现方面考虑得不够全面，且存在设备成本高、探头布置难、更新难度大等问题。若能以水厂日常监测指标为基础，建立吨水电耗与处理目标之间的响应关系，是降低运维能耗的可行之路，但是目前该方面研究还鲜有报道。

鉴于此，本文通过在水厂日常监测指标中解析各参数与水厂电耗响应关系，通过基于统计学的方法与 R 语言、python 等工具进行参数筛选，并将其对吨水电耗的影响进行权重排序，进而将优选变量作为输入进行数学模型构建，并考虑了模拟结果的精准度。本研究的结果意在为污水厂节能降耗提供针对性的意见，对降低其实际运维能耗有较强的参考价值。

2. 材料与方法

2.1 水厂简介

河南省某城镇污水处理厂设计处理规模为 3 万 t/d ，总变化系数为 1.42。污水处理工艺采用 AAO 氧化沟+深度处理工艺，出水水质标准按照《城镇污水处理厂污染物排放标准》（GB18918—2002）的一级 A 标准达标排放。污泥处理采用带式浓缩脱水一体机脱水干化后进行卫生填埋。污水处理工艺流程如图。

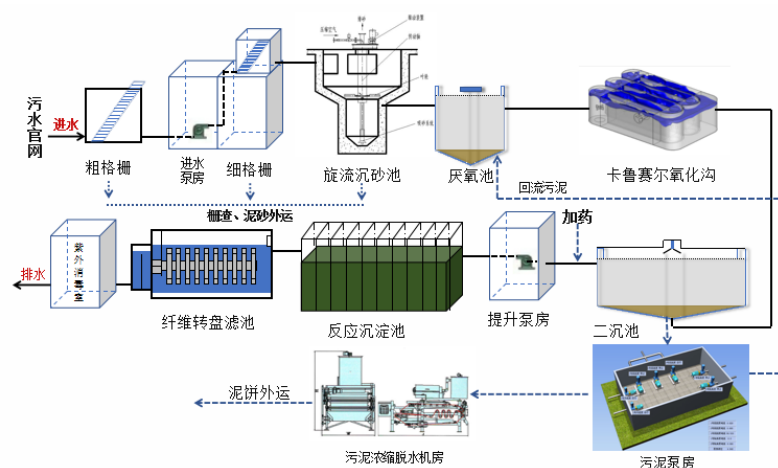


图 1 某污水处理厂工艺流程图

选择该厂 2020 年 12 个月的运行数据为研究对象，包括进水流量 (x_1)、进水氨氮浓度 (x_2)、进水 COD 浓度 (x_3)、进水 TN 浓度 (x_4)、进水 TP 浓度 (x_5)、出水氨氮浓度 (x_6)、出水 COD 浓度 (x_7)、出水 TN 浓度 (x_8)、出水 TP 浓度 (x_9)、MLSS (x_{10})、好氧池末端 DO (x_{11})、厌氧池末端 ORP (x_{12})、有机负荷 (x_{13})、氨氮负荷 (x_{14})。

2.2 数据筛选

2.2.1 数据预处理

为了消除不同量纲与单位的影响,采用 min-max 标准化方式 (Min-Max Normalization) 对数据进行数据归一化处理,如公式 (1) 所示^[8]:

$$x = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

其中 x_{min} 数据最小值, x_{max} 为样本数据最大值。

2.2.2 筛选方法

为了消除个别方法对筛选结果的影响,本文采用九种方法进行筛选,分别为:普通最小二乘法 (Ordinary Least Squares)、Cp 准则、AIC 准则、BIC 准则、Lasso、ALasso、SCAD、ncvreg、岭回归^[9]。汇总各方法的筛选结果,入选的参数用“√”进行标记,按照标记频数从大到小选择八个参数作为模型输入。

2.3 模型建立

本文以优选参数作为输入、吨水电耗作为输出,利用极限梯度提升法建立两者之间的模型关系 (Extreme Gradient Boosting, XGBoost)。XGBoost 算法是一种基于 Boosting 思路的算法,它将损失函数通过泰勒公式展开到二阶导数,并且添加了正则项来调节模型的复杂性,有效减少模型过拟合^[10]。

对于训练数据集 $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, n 为样本数量, x_i 为输入样本, y_i 为对应 x_i 的真实值, $i = 1, 2, \dots, n$ 。使用 XGBoost 算法针对吨水电耗进行建模时,首先定义由损失函数和正则化惩罚项组成的目标函数:

$$Obj(\theta) = L(\theta) + \Omega(\theta) \quad (2)$$

其中 $L(\theta)$ 为损失函数, $\Omega(\theta)$ 为正则化惩罚项。

回归树中的预测结果 \hat{y}_i 为

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (3)$$

其中 K 为回归树的数量, $f_k(x_i)$ 为每一棵树的得分值, θ 为使目标函数最小时所要求出的参数。对式 (2) 进行整理可得:

$$Obj(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (4)$$

其中 $l(y_i, \hat{y}_i)$ 为第 i 个样本的预测误差, $\Omega(f_k)$ 为第 k 棵分类回归树的正则项

$$\Omega(f_k) = \alpha T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 \quad (5)$$

其中 T 为第 t 棵树中总叶子节点的个数; ω_j 为第 j 个叶子结点的权重; α 、 λ 为常数,表示正则惩罚中的参数。

然后输入采集的污水水质参数数据进行模型训练,对于第 t 轮迭代,模型的目标函数可表示为

$$Obj^{(t)} = \sum_{i=1}^n l[y_i, \hat{y}_i^{(t-1)} + f_t(x_i)] + \sum_{k=1}^K \Omega(f_k) + C \quad (6)$$

其中 $f_t(x_i)$ 表示加入的第 t 棵分类回归树; 常数 C 表示前 $t-1$ 棵树的复杂程度,即 $C = \sum_{i=1}^{t-1} \Omega(f_i)$ 。

当误差函数 l 不是平方误差时,用泰勒展开的前三项来近似原始损失函数。泰勒展开式:

$$f(x + \Delta x) \simeq f(x) + f'(x)\Delta x + \frac{1}{2}f''(x)\Delta x \quad (7)$$

这里，将 $\hat{y}_i^{(t-1)}$ 作为 $f(x)$ ，将 $f_t(x_i)$ 作为 Δx ，将第 t 轮迭代的目标函数用泰勒级数展开，可以得到：

$$Obj^{(t)} \simeq \sum_{i=1}^n \left[l(y_i, y_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \sum_{k=1}^K \Omega(f_k) + C \quad (8)$$

其中 $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, y^{(t-1)})$ ， $h_i = \partial_{\hat{y}^{(t-1)}}^2 l(y_i, y^{(t-1)})$

式中的 $l(y_i, \hat{y}_i^{(t-1)})$ 和 C 均为常数项，舍去。进一步化简可以得到最终的目标函数：

$$Obj^{(t)} \simeq \sum_{j=1}^T \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T \quad (9)$$

其中， $G_j = \sum_{i \in I_j} g_i$ ， $H_j = \sum_{i \in I_j} h_i$ ， I_j 表示在第 j 个子结点上的样本集合。

显然，此时只需寻找一个最优的权重值，即可获得最优的目标函数值。因此，令目标函数 $Obj^{(t)}$ 对 ω_j 的偏导数等于 0，可以求出最优权重值为：

$$\omega_j^* = -\frac{G_j}{H_j + \lambda} \quad (10)$$

将式(10)代入式(9)可得最优的目标函数值为

$$Obj^{(t)} = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T \quad (11)$$

使用网格搜索优化 XGBoost 相关参数，结果如表 1 所示。

Parameter	learning_rate	maxdepth	n_estimators	min_child_weight	gamma	subsample	colsample_bytree	alpha
Value	0.06	6	22	1	0	0.9	0.6	0

表 1 XGBoost 模型构建参数

3. 结果与讨论

3.1 输入筛选结果

各参数利用 9 种方法筛选结果的入选频数如表 2 所示。从此可见，进水流量、进水氨氮浓度、进水 TP 浓度、进水 TN 浓度、出水 TP 浓度、好氧池末端 DO、厌氧池末端 ORP 与氨氮负荷等八个参数在频数上超过 7，故将此八个参数作为建模输入。其中进水流量与水量有关，出水 TN 浓度、进水氨氮浓度、进水 TP 浓度、出水 TP 浓度与水质有关，好氧池末端 DO 浓度、氨氮负荷、厌氧池末端 ORP 与运维调控有关。

参数代号	x ₁	x ₂	x ₃	x ₄	x ₅	x ₆	x ₇	x ₈	x ₉	x ₁₀	x ₁₁	x ₁₂	x ₁₃	x ₁₄
OLS	√	√			√			√	√		√	√		√
Cp 准则	√	√			√			√	√		√	√		√
AIC	√	√						√	√			√		√
BIC	√	√			√			√	√		√	√		√
Lasso(Python)	√		√							√				

参数代号	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
Lasso(R)	√	√		√	√	√		√	√	√		√		√
ALasso	√	√			√	√		√	√		√	√		√
SCAD	√	√			√			√	√		√	√		√
ncvreg	√	√			√			√	√		√	√		√
岭回归	√	√	√	√	√	√	√	√	√	√	√	√	√	√
入选次数	10	9	2	2	8	3	1	9	9	3	7	9	1	9
选取结果	*	*			*			*	*		*	*		*

表 2 参数入选频数

为了验证筛选结果的合理性，对所有输入与处理吨水电耗开展 Pearson 相关性分析，结果如图 2 所示。从中可见，所筛选的参数与吨水电耗的相关系数绝对值排在前八位。在处理水量方面，进水流量与吨水电耗成较强的负相关关系，属规模效应的典型范例。当污水厂的处理规模增大，相较于小型污水处理厂提高了管理和资源利用效率，因此吨水所需电耗较低 [11]。

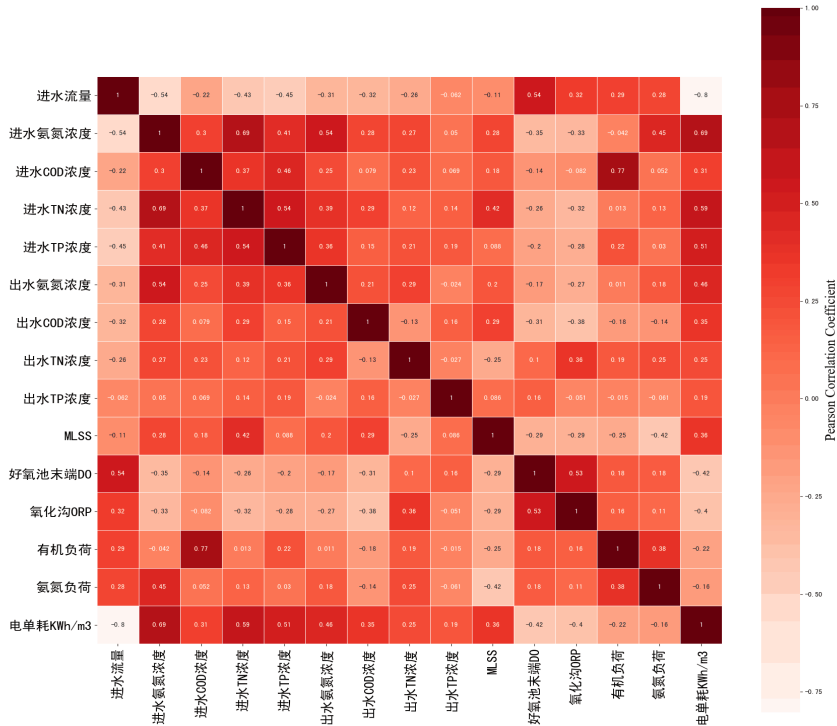


图 2 各参数间的 Pearson 相关性

在处理水质方面，出水 TN 浓度与污水处理中的脱氮过程密切相关。脱氮过程由硝化与反硝化两部分组成。硝化反应需要耗电曝气，硝化越充分耗电也越高。但在硝化反应中 TN 浓度不下降，去除 TN 主要发生在反硝化过程。当吨水电耗增加硝化作用充分时，过度曝气会消耗大量 COD 导致反硝化不充分，造成出水 TN 浓度上升 [12]；进水氨氮浓度与脱氮过程密切相关，在硝化反应中曝气越充分氨氮浓度下降越快。进水氨氮浓度越高，相应曝气量需求也越高，吨水电耗也将随之增加；在污泥浓度较低时，磷浓度的变化主要与聚磷菌表现有关，聚磷菌除磷可分为放磷与吸磷两部分。吸磷过程需要电子受体（例如氧气），故进水

TP 浓度越高需要的曝气电耗越大；在污泥浓度较高时，则主要依靠同化作用除磷，故出水 TP 浓度相关性比进水 TP 浓度有所减弱。

在运维调控方面，好氧池末端 DO 与曝气量直接成明显正相关关系，曝气量越大残留的 DO 越多，则吨水电耗也越大；在污泥浓度较低时，氨氮负荷与进水 TN 浓度正相关，此时脱氮以硝化反应为主。在污泥浓度较高时，微生物同化作用去除氨氮占比更为显著。该厂污泥浓度较高，故氨氮负荷与电单耗成较弱的负相关趋势^[13]；ORP 可作为表征厌氧池厌氧环境效果的参数，ORP 越低厌氧环境越好。为了达到较好的除磷效果，需要创造较低的 ORP，故因此厌氧池末端 ORP 与电单耗呈负相关关系。

3.2 模型拟合结果

将实际采集的前 300 天参数数据输入并训练模型后，再使用剩余的 53 天的参数数据进行验证。训练和验证的效果如图 3 所示。在测试集上，XGBoost 表现较好，R2 为 0.805，在训练集表现略差，R2 为 0.631。但测试集上预测结果已能较好的与原始数据增降趋势吻合，证明模型已具有较好的泛化能力。李畅等利用 XGBoost 进行出水的指标的软测量，可用于实时预测出水氨氮和 TP 浓度。但其缺少变量优选过程，模型泛化能力较低^[14]。廉小亲等利用 SOM 算法优化神经网络结构，实现了对污水中 COD 的高精度预测。提出的神经网络结构较为复杂，运算速率较慢，实用性较差^[15]。

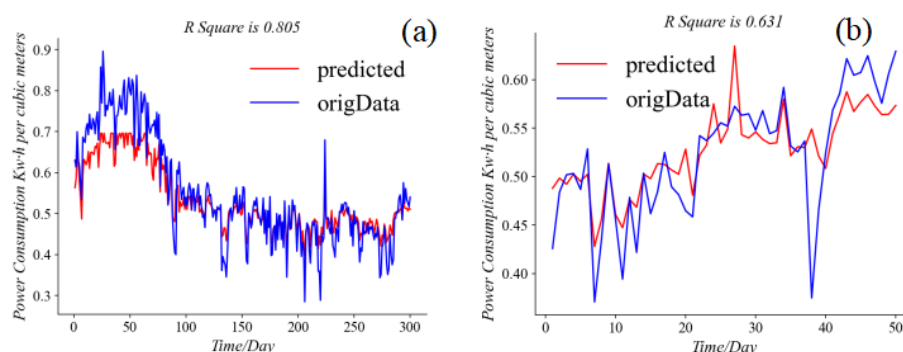


图 3 XGBOOST 模型拟合结果 (a) 训练集；(b) 测试集

3.3 参数影响权重

通过 XGBoost 中的 F score 考察各输入对吨水电耗的影响大小，结果如图 4 所示。从中可见各输入参数对吨水电耗的影响由强到弱依次为：进水流量、出水 TN 浓度、进水氨氮浓度、好氧池末端 DO、进水 TP 浓度、氨氮负荷、厌氧池末端 ORP、出水 TP 浓度。发现其影响顺序与通过 Pearson 相关性分析得到的结果相似。在不可控的因素中，影响较大的是进水流量、进水氨氮浓度、进水 TP 浓度。这说明在处理工艺不变的情况下，处理规模、处理水质等外在条件是决定吨水电耗的根基。在可控的因素中，影响较大的是出水 TN 浓度、好氧池末端 DO、氨氮负荷等，且出水 TN 低浓度与吨水电耗成正比。这意味着在好氧池内通过减少曝气适当控制低溶解氧状态，不仅有助于节能，而且还会强化脱氮。另外，维持高污泥浓度强化同化作用脱氮，也有助于降低运行电耗。以上结果，对碳中和发展战略下污水处理厂的提标降耗运行均具有借鉴意义。

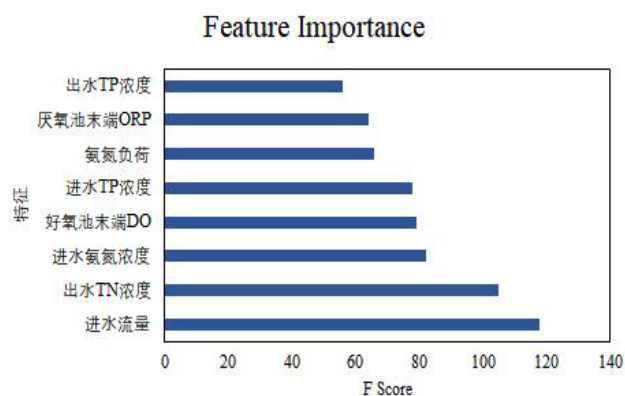


图 4 XGBoost 的 F score

4. 结论

1.通过参数筛选,可知,在污水处理厂日常监测的十四种参数中。进水流量、进水氨氮浓度、进水 TP 浓度、进水 TN 浓度、出水 TP 浓度、好氧池末端 DO、厌氧池末端 ORP 与氨氮负荷八项参数与电单耗具有较为明显的相关性;

2.通过 XGBoost 将筛选后的八项参数作为自变量,电单耗作为因变量构建模型,拟合程度较为良好,可在一定程度上证明参数筛选的有效性;

3.利用 XGBoost 建模的同时,并可按照自变量对因变量影响权重大小进行排序,其结果对水厂运维中节能降耗方案的制定提供参考。

参考文献:

- [1]杨庆,王亚鑫,曹效鑫,刘秀红,章世勇.污水处理碳中和运行技术研究进展[J/OL].北京工业大学学报,2022(03):1-14[2022-03-14].<http://kns.cnki.net/kcms/detail/11.2286.T.20220215.1708.014.html>
- [2]陈治池,何强,蔡然,罗华瑞,罗南,宋忱馨,程鸿.碳中和趋势下数学模拟在污水处理系统中的发展与综合应用[J/OL].中国环境科学:1-23[2022-03-14].DOI:10.19674/j.cnki.issn1000-6923.20220208.003.
- [3]孙慧,王佳伟,吕竹明,吕泽瑜,蒋彬,陈晨,刘小杏,于澜.北京某大型城市污水处理厂节能降耗途径和效果分析[J].中国给水排水,2019,35(16):31-34.DOI:10.19853/j.zgjsps.1000-4602.2019.16.006.
- [4]高建磊,闫怡新,吴建平. 河南省城市污水处理厂能耗物耗统计与影响因素分析[C]//.全国给水排水技术信息网 2009 年年会论文集.,2009:80-85.
- [5]荆玉姝,牟润芝,姜怡名,刘长青,杨延栋.曝气精确控制实现污水处理厂节能降耗的应用研究[J/OL].环境工程:1-7[2022-03-14].<http://kns.cnki.net/kcms/detail/11.2097.X.20211119.1604.004.html>
- [6]魏全源.污水处理厂节能降耗技术措施和途径探讨[J].资源节约与环保,2015(04):6-7.DOI:10.16317/j.cnki.12-1377/x.2015.04.007.
- [7]张晓霞.污水处理厂节能降耗的有效措施分析[J].中国资源综合利用,2021,39(07):157-159.
- [8]许少新,杜渐,郑坚钦,李舒丹,黄丽燕,梁永图.基于人工神经网络的油库出库量预测研究[J].油气田地地面工程,2021,40(12):52-56.
- [9]曾津,周建军.高维数据变量选择方法综述[J].数理统计与管理,2017,36(04):678-692.DOI:10.13860/j.cnki.sltj.20170329-001.
- [10] CHEN T, GUESTRIN C. XGBoost: A Scalable Tree Boosting System [Z]. Proceeding s of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data

Mining. San Francisco, California, USA; Association for Computing Machinery. 2016: 78 5 – 94.10.1145/2939672.2939785

[11]王志国. 华北 XX 污水处理厂能耗分析与节能措施研究[D].河北工程大学,2016.

[12]唐霞,李江鹏,申云,李碧清,吴学伟,郭欢.基于 A²O—MBBR 工艺的污水厂提标改造中试研究[J].中国给水排水,2021,37(11):52-57.DOI:10.19853/j.zgjsps.1000-4602.2021.11.009.

[13]刘佳东,李彦澄,温馨,龚本洲,周健.氨氮负荷对垃圾渗滤液与城镇污水协同处理效能影响研究[J].工业水处理,2015,35(02):30-33.

[14] 李畅, 潘丰. 基于 XGBoost 的出水氨氮和总磷浓度的软测量研究 [J]. 计算机与数字工程, 2021, 49(05): 1003-8.

[15] 廉小亲, 王俐伟, 安飒, 等. 基于 SOM-RBF 神经网络的 COD 软测量方法 [J]. 化工学报, 2019, 70(09): 3465-72.