第 6 卷◆第 11 期◆版本 1.0◆2025 年
文章类型：论文|刊号（ISSN）：2737-4580(P) / 2737-4599(O)

机器学习算法在化工废水水质预测模型中的构建与应用

莫智高　王斌　陈梁良
浙江新鸿检测技术有限公司
DOI:10.12238/jpm.v6i11.8545

[摘　要] 以广东省某炼化企业 2023 年出水水质数据为基础，开发适配高波动性炼化污水的 COD 预测模型，采用 Bootstrap 加上随机插值法填补缺失之处，采用 Hampel 算法清除离群数值，构造 BP－NN 以及 SVR 模型，继而引入 MPSO 与 GS 实现参数优化，MPSO－BP－NN 模型展现出最优的预测性能。测试集的 R^2 为 0.81 这一数值，MAPE 为 2.58 个百分点，具备突出的泛化实力，该方法能为污水处理系统的调控提供相关理论支撑。

[关键词] 炼化污水；COD 预测；机器学习；BP 神经网络；MPSO 优化

Construction and Application of Machine Learning Algorithms in Chemical Wastewater Quality Prediction Model

Mo Zhigao　Wang Bin　Chen Liangliang
Zhejiang Xinhong Testing Technology Co.，Ltd.

[Abstract] Based on the effluent water quality data of a refining enterprise in Guangdong Province in 2023，a COD prediction model adapted to high volatility refining wastewater was developed. Bootstrap and random interpolation were used to fill in the gaps，and Hampel algorithm was used to remove outliers. BP-NN and SVR models were constructed，and MPSO and GS were introduced to achieve parameter optimization. The MPSO-BP-NN model showed the best prediction performance. The R^2 of the test set is 0.81，and the MAPE is 2.58 percentage points，demonstrating outstanding generalization ability. This method can provide relevant theoretical support for the regulation of sewage treatment systems.

[Key words] Refining wastewater；COD prediction；machine learning BP neural network；MPSO optimization

引言：

伴随数据驱动方案的成长，把机器学习技术引入污水处理过程建模内，能挖掘污染物浓度时间上的演化规律，强化预测水平，鉴于炼化行业背景的复杂性，要引入结构及参数的优化算法，提升预测模型在准确性与适应性方面的表现[1]。

1.材料与方法

1.1 模型构建及优化算法

1.1.1BP-NN 模型结构

BP 神经网络由输入层、隐藏层与输出层组成，信号从输入层正向传播至输出层，误差通过反向传播以梯度下降法形式调整权重与阈值。输入层节点数设为 5，对应前五日 COD 值，输出层节点为 1，对应预测值。隐藏层节点数设为 6，激活函数为 Sigmoid 函数，结构优化通过试验设定。神经元输出为：

$$H_j = f\left(\sum_{i=1}^{n} w_{ij}\, x_i - b_j\right)$$

$$Y_k = \sum_{j=1}^{l} w_{jk} H_j - b_k$$

其中，w_{ij}、w_{jk}为权重，b_j、b_k为偏置项，$f(\cdot)$为 Sigmoid 函数。训练采用 L-M（Levenberg-Marquardt）算法，设置迭代次数 1000 次，误差阈值 0.000001，学习率 0.01。

1.1.2SVR 模型结构

SVR 模型通过核函数将输入变量映射至高维空间，寻找最优回归超平面，以拟合输出变量。定义预测函数为：

$$f(x) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

其中，α_i、α_i^*为拉格朗日乘子，$K(x_i, x)$为核函数，本研究选择高斯核函数（GaussianKernel）形式：

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

模型目标函数为最小化正则化风险：

$$R(\omega, \xi, \xi^*) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

在满足以下约束条件下：

$$\begin{aligned} y_i - \omega \cdot \phi(x_i) - b &\le \varepsilon + \xi_i \\ \omega \cdot \phi(x_i) + b - y_i &\le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* &\ge 0 \\ C &\ge 0 \end{aligned}$$

其中，C 为惩罚因子，ε 为不敏感损失函数，$\phi(x)$为非线性映射函数。

1.1.3MPSO 及 GS 算法

MPSO（MutantParticleSwarmOptimization）在传统粒子群算法基础上引入随机变异机制，采用 rand 函数对粒子维度值进行扰动，提升种群多样性以避免陷入局部最优[3]。MPSO 中，粒子更新公式如下：

$$v_{id}^{t+1} = \omega v_{id}^{t} + c_1 r_1 (p_{id} - x_{id}^{t}) + c_2 r_2 (g_d - x_{id}^{t})$$

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1}$$

其中，x_{id}为位置，v_{id}为速度，p_{id}为个体最优，g_d为全局最优。用于 BP-NN 时，每个粒子维度等于网络中所有权值与偏置之和；用于 SVR 时，每个粒子维度为 2，对应惩罚因子 C 与核函数参数 g。GS（GridSearch）算法对 C 与 g 进行遍历组合搜索，每组参数均进行模型训练并以最小 MSE 值为选择标准[4]。

1.1.4 模型预测结果评价

采用五折交叉验证（K-foldCrossValidation）对模型预测结果进行评估。通过五次分组交替训练与验证获得平均性能指标。模型评价指标包括平均绝对误差（MAE）、均方根误差（RMSE）、决定系数（R^2）、相关系数（r）、平均偏差误差（MBE）及平均绝对百分比误差（MAPE），其定义如下：

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

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

$$R^2 = \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

上述指标用于全面衡量预测精度、拟合能力与误差分布情况，为模型性能优劣提供判定标准。

2.结果与分析

2.1 参数相关性分析

采用皮尔逊相关系数（PCC）与斯皮尔曼等级相关系数（SCC）对六个输入变量与目标输出 COD 之间的相关性进行分析。结果显示，所有输入参数与 COD 之间的相关系数均低于 0.2，缺乏显著线性或单调相关性。其中 PCC 分析结果中，TOC 与 COD 相关系数为-0.15，TN 为-0.10，HS 为-0.086，pH 为-0.097，BOD_5为-0.037；SCC 结果中，TOC 为-0.11，TN 为-0.10，HS 为-0.058，pH 为-0.074，BOD_5为-0.028。六个参数之间也未表现出相互依赖关系，交叉相关性系数普遍低于 0.1，说明该类数据具有高维稀疏特性。以 2023 年 4 月 3 日至 8 日的监测数据为例（见表 1），HS 在 17.5 至 5.7 mg/L 间变化，COD 在 38 至 19 mg/L 间波动，未表现出协同变化趋势。部分污染物如硫化物与多环芳烃等难降解组分具有毒性积累效应，干扰其他参数对 COD 的表征能力，导致统计意义上的相关性显著下降，传统的多参数回归模型难以从中提取稳定的预测特征。

2.4 模型搭建

多参数相关性弱，直接构建回归模型难以捕捉污染物间复杂交互关系，故采用时间序列建模策略增强模型对 COD 变化趋势的学习能力，以炼化企业监测池出水 COD 历史数据为输入，基于 5 日历史值预测第 6 日 COD 值，按污水处理工艺水力停留时间构建输入输出序列，企业上游设两个 10，000 m^3 均质罐与一个 2，000 m^3 均质池，下游设 2，000 m^3 缓冲池，通过水量调节减缓污染物浓度突变，为历史数据建模提供工程可行性。数据重构后形成 339 组样本，前 309 组用于训练，后 30 组为测试集。BP-NN 模型优化结果显示，最优网络结构

为输入层 5 节点、1 层 6 节点隐藏层、输出层 1 节点，激活函数为 Sigmoid，训练函数为 Levenberg-Marquardt，学习率 0.01，最大迭代 1000 次，用 MPSO 算法优化神经网络权重与偏置，设粒子种群数 15，单种群迭代 100 次，SVR 模型比较多种核函数后，选 Gaussian 核函数构建非线性支持向量机，c 与 g 参数在（2^{-5}，1）与（1，2^{5}）区间搜索，损失函数 ε 设为 0.1，MPSO 用于优化 c 与 g 值。

表 1　4 月现场部分水质数据

日期	HS（mg/L）	TN（mg/L）	TOC（mg/L）	BOD_5（mg/L）	pH	COD（mg/L）
4 月 3 日	17.5	0.005	7.8	16.2	4.4	38
4 月 4 日	5.7	0.005	7.7	14.2	9.4	44
4 月 5 日	11.7	0.007	8.2	14.4	3.7	19
4 月 6 日	15	0.005	7.8	14.3	26	33
4 月 7 日	8.5	0.003	7.7	9.8	3.6	33
4 月 8 日	11.2	0.005	7.8	17.6	20	29

2.5 模型准确性验证

完成 BP-NN、MPSO-BP-NN、GS-SVR、MPSO-SVR 四类模型在训练集和测试集的性能评估，测试集结果（如表 2 所示）。MPSO 优化使两种模型在数据突变时拟合能力显著提升。2024 年 1 月实际出水数据泛化验证表明，MPSO-BP-NN 模型在污染物缓变时保持高拟合度，剧烈波动时能跟踪趋势变化，最大相对误差 14.82%，平均误差 2.71%。极端波动如 1 月 10 日 COD 突降和 1 月 25 日上升阶段，模型预测值与真实值变化方向一致，展现动态响应能力与泛化性能。

表 2　四类模型在测试集上的预测性能比较

模型	R^2	r	RMSE（mg/L）	MAPE（%）	MAE（mg/L）	MBE（mg/L）
BP-NN	0.65	0.81	2.40	3.73	1.59	-0.28
MPSO-BP-NN	0.81	0.89	1.63	2.58	1.10	-0.25
GS-SVR	0.67	0.82	2.07	3.19	1.29	0.49
MPSO-SVR	0.78	0.88	1.65	2.60	1.11	0.31

3.结论

基于广东省某炼化企业 2023 年出水水质监测数据，结合炼化污水处理系统复杂性与波动特性，构建适用于高复杂度工业废水的水质预测模型，参数相关性分析显示 HS、TN、TOC、BOD_5、pH 与 COD 之间相关性均低于 0.2，传统多变量回归建模方法难以提取有效特征信息，模型结构选择上，采用时间序列方式将历史 COD 数据作为输入建模，构建 BP-NN 与 SVR 两种模型，并分别引入 MPSO 与 GS 优化算法调整参数，实验结果表明，MPSO-BP-NN 模型在预测精度、误差控制和趋势识别能力方面均优于其他模型，测试集中 R^2 为 0.81，RMSE 为 1.63 mg/L，MAPE 为 2.58%。2024 年 1 月现场数据验证显示该模型在原料切换和负荷扰动等工况变化下仍具较强趋势跟踪能力与稳定性，具备良好泛化性能。

[参考文献]

[1]汪锐，余雅丹，潘志成，等.模拟预测模型在污水处理中的应用：现状与挑战[J].水处理技术，2022，48(06)：20-23+29.

[2]陈霖，刘浩威，王庆宏，等.基于机器学习算法的炼化污水厂出水水质预测模型研究[J/OL].工业水处理，1-23[2025-05-21].

[3]陈霖，晏欣，李巨峰，等.基于优化机器学习的炼化企业污水场均质池出水水质预测研究[J].给水排水，2024，60（10）：159-168.

作者简介：莫智高（1993—），身份证号码：330424199304011615，男，汉族，浙江嘉兴，本科，中级工程师，从事工作为化工安全技术与管理。