基于 PSO-PLS 混合算法的水体 COD 紫外吸收光谱检测研究

郑培超，赵伟能，王金梅*，赖春红，王小发，毛雪峰

重庆邮电大学光电工程学院，光电信息感测与传输技术重庆重点实验室，重庆 400065

摘 要 化学需氧量(COD)是反映水体受有机物污染程度的重要指标。紫外吸收光谱法是目前水体 COD 检测研究中应用最为广泛的方法，具有样品无需预处理，成本低，无污染，测定速度快等优点。但是，原始光谱数据维数高，光谱信息中包含大量冗余变量，直接将全光谱数据进行建模存在精度低，计算复杂等问题。针对紫外吸收光谱全光谱建模精度低，光谱数据存在大量共线性的问题，提出了一种基于粒子群算法(PSO)结合偏最小二乘(PLS)优选特征波长建立预测模型的方法，以提高紫外吸收光谱预测模型的精度和适用性，简化模型。利用搭建的紫外吸收光谱装置，采集 29 份不同浓度的 COD 标准溶液的紫外光谱数据，每份标准溶液采集 5 次取平均值并对其进行平滑处理，减少仪器和环境带来的误差。考虑到标准溶液在 200～310 nm 的光谱范围内存在吸收，故选取该波段范围内 246 个波长点作为建模数据，每个波长点下的吸光度数据作为一个粒子并按照顺序编号，以 PLS 为建模方法，相关系数 r 和均方根误差(RMSE)为评价指标，设置粒子群算法适应度函数 $f(x)=\min(\mathrm{RMSE})$，取粒子初始种群数为 20 个，惯性权重 $w=0.6$，自我学习因子 $c1=1.6$，群体学习因子 $c2=1.6$，最大迭代次数为 200 次，算法终止条件为达到最大迭代次数。算法输出全局最优变量取值为 168，94，181，183，175，209，106 和 142。采用粒子群算法优选的 8 个波长点建立 PLS 预测模型的相关系数 r 和预测均方根误差 RMSE 分别为 0.999 98 和 0.155 1。为了验证 PSO-PLS 建立的预测模型效果，建立了 PLS，iPLS 和 SVR 三种预测模型进行对比。验证结果表明，PSO-PLS 模型的相关系数 r 和均方根误差 RMSE 均优于其他三种预测模型，说明粒子群算法能有效的提取用于 PLS 建模的特征波长，消除子区间变量的共线性，提高预测模型的精度。该方法为实现水体 COD 实时在线监测提供了一种有效途径。

关键词 粒子群算法；紫外吸收光谱；COD 测量；PLS 回归

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)01-0136-05

引 言

随着经济的迅速发展和城市化进程的加快，水污染问题日益严重，已经成为制约经济发展的瓶颈。为实现水环境的治理和管控，需要大力发展水质检测设备。在水质检测中，化学需氧量(chemical oxygen demand，COD)是评价水体受有机物污染程度的重要指标。针对 COD 的检测常采用传统的化学法，如高锰酸钾法和重铬酸钾法，由于这些方法存在分析时间长，有大量二次污染的化学试剂等问题[1]，近年来，紫外吸收光谱法逐渐应用到 COD 的检测，它通过建立紫外吸光度和有机物浓度的预测模型来反演 COD 值，该方法不需要任何化学预处理，具有检测快速，操作简单，不会对环境构成二次污染等优点[2-3]。

由于紫外吸收光谱法采集的光谱信号的光谱范围为 200～800 nm，光谱信息量庞大，光谱数据中存在大量共线信息，如何选取有效的波长吸光度建立回归模型是提高预测模型精度的主要问题。目前，紫外吸收光谱法检测 COD 在模型算法选择上主要有偏最小二乘回归(partial least squares regression，PLSR)、支持向量机回归(support vector regression，SVR)、人工神经网络(artificial neural network，ANN)和机器学习(machine learning，ML)等[4-8]。在算法模型选择方面，毕卫红等运用偏最小二乘法建立不同谱区的校正模型，得到最好预测模型的 $r=0.995\ 8$，RMSEC=16.186 5。杨鹏程[9]等利用间隔偏最小二乘法(iPLS)建立了海水硝酸盐浓度模型，校正集均方根误差 RMSECV 降到了 9.83。Li[10]

收稿日期：2019-11-25，**修订日期**：2020-03-20

基金项目：国家自然科学基金项目(61805030，61705028)，重庆市基础与前沿技术研究专项(cstc2018jcyjA0585)资助

作者简介：郑培超，1980 年生，重庆邮电大学光电工程学院教授 e-mail：zhengpc@cqupt.edu.cn

*通讯作者 e-mail：wangjm@cqupt.edu.cn

等采用组合区间偏最小二乘(siPLS)对南京钱湖水样进行建模分析，得到最优相关系数 $r=0.833\ 4$，RMSE＝2.63。Pan[11]等利用傅里叶变换红外光谱(FTIR)快速测量废水中的 COD。结果表明，采用变参数移动窗偏最小二乘(mw-PLS)选择谱区有效的提高了预测精度。汤斌[12]等采用粒子群算法联合最小二乘支持向量机(PSO_LSSVM)，最大相对误差仅为 5.83%。

基于以上研究存在的预测精度不高，算法模型复杂等问题。本研究以 PLS 算法来建立预测模型，结合粒子群算法收敛速度快，全局搜索最优解的特点对特征波长进行筛选，建立一种快速选择特征波长，建立预测模型的方法。与 PLS，iPLS 和 SVR 三种预测模型对比，PSO-PLS 模型预测效果最佳。

1　实验部分

1.1　装置

本实验所采用的实验装置如图 1 所示，氙-卤钨灯光源(Avantes，AvaLight-DH-S-BAL)发出的紫外-可见-近红外光通过可变光程(为获取合适的吸光度，本实验设置的光程差 10 mm)的反射式浸入式探头的入射光纤(Avantes，FDP-7xx200-VAR)，将探头浸入到溶液中，光源发出的光被溶液吸收后反射到探头的出射光纤上，出射光纤与便携式光谱仪(Ocean Optics，Maya Pro 2000)连接，采集的光谱数据通过数据线传输到计算机进行处理。

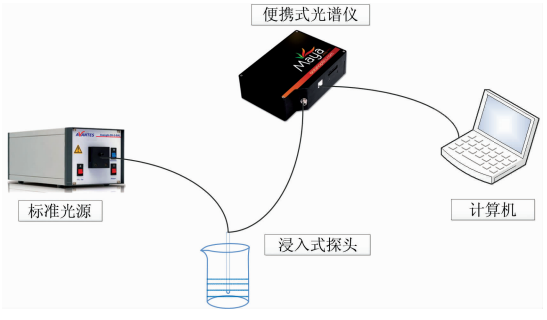

图 1　实验装置示意图

Fig. 1　Schematic diagram of the experimental setup

1.2　样品制备

准确称量 0.425 1 g 邻苯二甲酸氢钾(分析纯，川东化工)溶于去离子水，配置成 1 000 mg·L^{-1} 的 COD 标准溶液，通过逐步稀释配制成不同浓度的 COD 标准溶液，采用预先搭建的实验装置，设置光谱仪积分时间为 100 ms，平均次数 5 次，平滑度为 3，首先以空白溶液为参比，采集其吸收光谱，然后对 29 份不同浓度的待测样品，进行光谱的采集。

如图 2 所示，标准溶液在近紫外区域有明显的吸收峰，尤其波长在 220～310 nm 区间内，紫外吸光度与溶液浓度之间具有很好的相关性，而大于 310 nm 的波长点几乎无吸收。且可以看出标准溶液在浓度低时线性关系良好，浓度高时，光谱重叠比较严重。

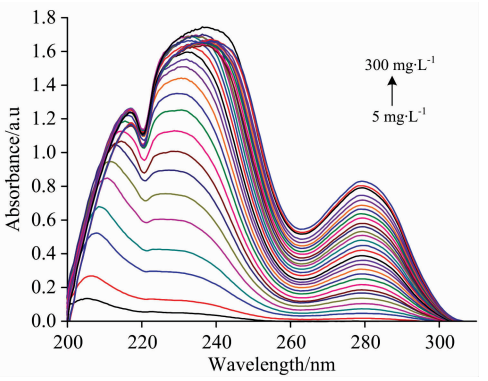

图 2　标准溶液紫外吸收光谱图

Fig. 2　Ultraviolet absorption spectrum of standard solution

1.3　方法

1.3.1　偏最小二乘原理

偏最小二乘(partial least squares regression，PLSR)由化学界的 Wold 和 Albano 等在 1983 年提出，这种方法集主成分分析、典型相关分析和多元线性回归分析三种方法的优点于一身，很好地解决了自变量间多重共线性的问题[14]。本研究以标准溶液的浓度为因变量，不同波长下的吸光度的值为自变量，建立 PLS 回归预测模型。建立自变量和因变量的特征向量间的一元线性回归关系。提取变量的终止条件为交叉有效性验证。

1.3.2　PSO 原理

粒子群算法(particle swarm optimization，PSO)最早是由 Eberhart 和 Kennedy 于 1995 年提出，它的基本思想源于模拟鸟群觅食过程中的行为而提出来的一种基于群体智能的全局随机搜索算法[14]。首先在所有解空间中初始化一群粒子，用位置、速度、适应度三项指标来表示该粒子特征。粒子在解空间中以一定速度运动，通过个体极值 Pbest 和种群极值 Gbest 更新速度和位置，粒子每更新一次位置，就计算一次适应度值，并且通过比较新粒子的适应度值和个体极值、群体极值的适应度值更新个体极值和群体极值位置。不断迭代，更新速度和位置，直到得到满足最终条件的最优解。速度更新公式

$$V_{id}^{k+1}=\omega V_{id}^{k}+c_1 r_2(P_{id}^{k}-X_{id}^{k})+c_2 r_2(P_{gd}^{k}-X_{id}^{k}) \quad (1)$$

ω 为非负数，称为惯性因子，体现的是粒子继承先前速度的能力，较大的惯性权重有利于全局搜索，较小的惯性权重有利于局部搜索。C_1 叫自我认知，是粒子跟踪自己历史最优值的权重系数，表示粒子自身的认知。C_2 叫社会认知，是粒子跟踪群体最优值的权重系数，表示粒子对整个群体知识的认识，r_1 和 r_2 是[0，1]区间内均匀分布的随机数，赋予算法一定的空间搜索能力。

1.3.3　PSO-PLS 特征波长选择方法

偏最小二乘和区间偏最小二乘的主要思想是采用全光谱或者是将全光谱的划分成一定数量的子区间来建立预测模型，其主要目的都是为了剔除冗余变量，提高预测模型精度。但是不可避免的是各个子区间仍然存在一些共线性的冗余变量，提出的 PSO-PLS 优选特征波长的方法可以解决上

述问题，将采集的光谱数据每个波长点下的吸光度数据作为一个粒子，按照位置顺序编号为 1—246，首先采用 PLS 建立 k 个变量的回归模型，以模型输出的 RMSE 为粒子群算法的适应度函数，粒子在整个谱区根据最小的 RMSE 更新速度和位置，最后找出满足条件的最优变量取值。

2 结果与讨论

2.1 PSO-PLS 建模

定义粒子群算法适应度函数为 $F(x)=\min(\text{RMSE})$，其中均方根误差(RMSE)为特征波长处 COD 真实值与测量值之间的均方根误差。设置初始种群个数为 20，惯性权重 $w=0.6$，自我学习因子 $c1=1.6$，群体学习因子 $c2=1.6$，位置参数限制为[1—246]，速度限制为[0—1]，最大迭代次数为 200。粒子群算法的具体流程如图 3。

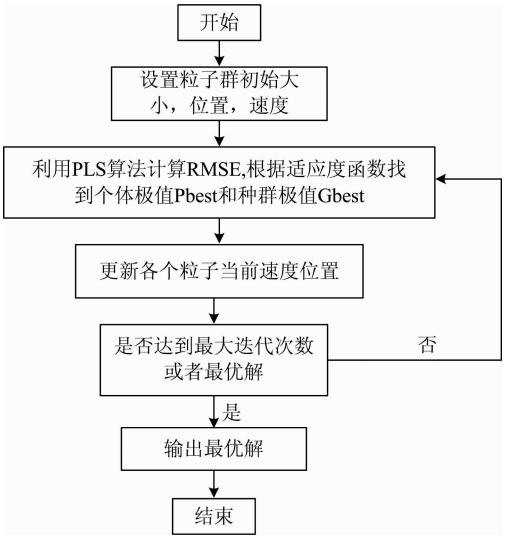

图 3 粒子群算法流程图

Fig. 3 Particle swarm algorithm flowchart

采用粒子群算法筛选的最优波长数为 8 个，位置分别 168，94，181，183，175，209，106 和 142。对应波长分别为 256.7，230.6，271.9，272.8，269.0，285.1，236.3 和 253.4 nm。图 4 是 PLS 利用粒子群算法筛选出的特征波长建立的预测模型，相关系数达到了 0.999 98，RMSE=0.155 1。

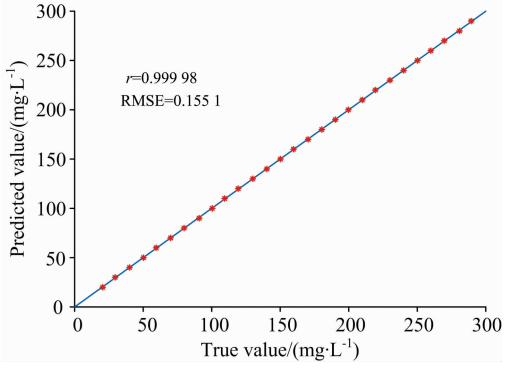

图 4 预测值和真实值相关关系

Fig. 4 Correlation between predicted value and true value

图 5 为粒子群算法运行过程中，适应度随迭代次数的变化函数，从图中可以看出，随着迭代次数的增加，RMSE 由最开始的 0.954 7 逐渐减小至 0.155 1，直到达到最大迭代次数。

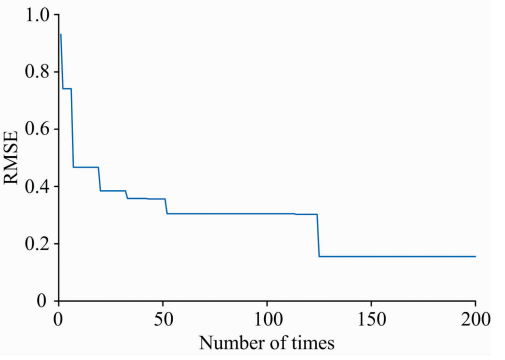

图 5 适应度变化曲线图

Fig. 5 Trend graph of fitness function

2.2 建模效果对比

采用 PLS，iPLS，SVR 以及 PSO-PLS 对采集的 29 份不同浓度的标准溶液光谱数据进行分析。为验证 PSO-PLS 建立的预测模型优劣，另外建立 PLS，iPLS 和 SVR 三种预测模型与本文提出的 PSO-PLS 进行对比，图 6 为 4 种预测模型预测浓度值与真实值的相对误差，由图 6 可知，低浓度时，4 种算法建立的预测模型相对误差上下波动较大。随着浓度的升高，相对误差趋于平稳。整体来看，PSO-PLS 方法的相对误差在 0～300 mg·L^{-1} 浓度范围内波动最小，标准溶液浓度为 30 mg·L^{-1} 时，PSO-PLS，PLS，iPLS，SVR 的相对误差绝对值分别为 0.009 6，0.021 8，0.021 9 和 0.016 8。PSO-PLS 模型的预测效果优于其他 3 种模型。

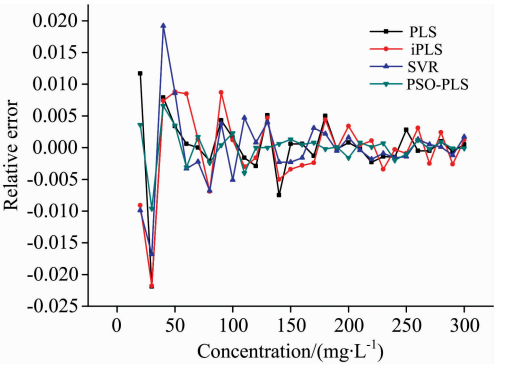

图 6 四种预测模型相对误差对比

Fig. 6 Comparison of relative error between four prediction models

表 1 为 4 种建模方法的相关系数和均方根误差。从表 1 可以看出，4 种算法中，SVR 建立的预测模型相关系数最小，均方根误差最大。PSO-PLS 建立的预测模型的相关系数最大，均方根误差 RMSE 仅为 0.155 1，远低于其他 3 种预测模型。比较可知 PSO-PLS 预测模型效果优于其他 3 种预测模型。

表 1　4 种建模方法效果对比

Table 1　Comparison of four modeling methods

建模方法	相关系数 r	均方根误差
iPLS	0.999 97	0.622 6
PLS	0.999 90	1.180 9
SVR	0.999 32	1.725 3
PSO-PLS	0.999 98	0.155 1

的方法，建立了 COD 浓度预测回归模型，模型适用于 COD 浓度低于 500 mg·L^{-1}的水体。并且将该模型与 PLS，iPLS 和 SVR 所建立的预测模型进行对比，实验数据表明，使用粒子群算法与偏最小二乘回归相结合的方法建立的模型能够有效减少建模波长数量，提高预测模型的精度。在实际水质监测中，使用粒子群算法选择特征波长，不仅能大大节约成本，还能在模型满足精度足够高的情况下快速进行建模，为快速无污染的紫外光谱水质监测提供便捷有效的算法依据。

3　结　论

采用粒子群算法与偏最小二乘回归相结合选择特征波长

References

[1] Kolb Marit, Bahadir Mufit, Teichgraber Burkhard. Water Research, 2017, 122(19): 645.

[2] Yang L, Shin H S, Jin H. Sensors, 2014, 14(1): 1771.

[3] Pigani L, Simone G V, Foca G, et al. Talanta, 2018, 178: 178.

[4] Bleyen N, Albrecht A, De Canniere P, et al. Applied Geochemistry, 2019, 100: 131.

[5] Hu Yingtian, Wen Yizhang, Wang Xiaoping. Sensor and Actuators B: Chemical, 2016, 227: 393.

[6] Hou D, Liu S, Zhang J, et al. Journal of Spectroscopy, 2014, 2014: 1.

[7] Li Z, Guan A, Ge H, et al. Microchemical Journal, 2017, 132: 185.

[8] Brito R S, Pinheiro H M, Ferreira F, et al. Urban Water Journal, 2014, 11(4): 261.

[9] YANG Peng-cheng, DU Jun-lan, CHENG Chang-kuo(杨鹏程，杜军兰，程长阔). Marine Environmental Science(海洋环境科学), 2016, 35(6): 943.

[10] Li Jingwei, Tong Yifei, Guan Li, et al. Optik, 2018, 174: 591.

[11] Pan T, Ji Q, Chen J M, et al. Key Engineering Materials, 2012, 500: 820.

[12] TANG Bin, ZHAO Jing-xiao, WEI Biao, et al(汤　斌，赵敬晓，魏　彪，等). China Environmental Science(中国环境科学), 2015, 35(2): 478.

[13] Chen Baisheng, Wu Huanan, Li Sam Fong Yau. Talanta, 2014, 120: 325.

[14] Mojtaba Shourian, Jamshid Mousavi S. Water Resources Management, 2017, 31: 4835.

Detection of COD UV Absorption Spectra Based on PSO-PLS Hybrid Algorithm

ZHENG Pei-chao, ZHAO Wei-neng, WANG Jin-mei*, LAI Chun-hong, WANG Xiao-fa, MAO Xue-feng

Chongqing Municipal Level Key Laboratory of Photoelectronic Information Sensing and Transmitting Technology, College of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing　400065, China

Abstract　Chemical oxygen demand (COD) is an important indicator of the degree of water pollution by organic matter. Ultraviolet absorption spectroscopy is the most widely used method for COD detection in water. It has the advantages of no pretreatment of samples, low cost, no pollution, and fast measurement speed. However, the original spectral data has high dimensions, and the spectral information contains a large number of redundant variables. Modeling the full spectral data has problems such as low accuracy and complicated calculations. Aiming at the low accuracy of UV absorption full-spectrum modeling and a large amount of collinearity in spectral data, this paper presents a method based on particle swarm optimization (PSO) and partial least squares (PLS) to select characteristic wavelengths to establish a prediction model. Improve the accuracy and applicability of the UV absorption spectrum prediction model and simplify the model. The UV spectrum data of 29 different concentrations of COD standard solutions were collected. Each standard solution was collected 5 times and averaged and smoothed to reduce the errors caused by the instrument and the environment. Taking into account the absorption of the standard solution in the spectral range of 200～310 nm, 246 wavelength points in this wavelength range were selected as modeling data, and the

absorbance data at each wavelength point was used as a particle and numbered in order. PLS was used as the model Method, the correlation coefficient r and the root mean square error (RMSE) are used as evaluation indicators. The particle swarm algorithm fitness function $f(x)=\min$ (RMSE) is set. The initial population of particles is 20, the inertia weight $w=0.6$, and the self The learning factor $c1=1.6$, the group learning factor $c2=1.6$, the maximum number of iterations is 200, and the algorithm termination condition is to reach the maximum number of iterations. The output value of the optimal global variable of the algorithm is 168, 94, 181, 183, 175, 209, 106, 142. The correlation coefficient r and the predicted root mean square error RMSE of the PLS prediction model established by the eight wavelength points selected by the particle swarm optimization algorithm were 0.999 98 and 0.155 1, respectively. In order to verify the effectiveness of the prediction model established by PSO-PLS, three prediction models of PLS, iPLS and SVR were established for comparison. The verification results show that the correlation coefficient r and the root mean square error RMSE of the PSO-PLS model are better than those of the other three prediction models, which shows that the particle swarm algorithm can effectively extract the characteristic wavelengths used for PLS modeling and eliminate the common of sub-interval variables Linear, improving the accuracy of the prediction model. This method provides an effective way for real-time online monitoring of COD in water bodies.

Keywords Particle swarm optimization; UV absorption spectroscopy; COD measurement; PLS regression

(Received Nov. 25, 2019; accepted Mar. 20, 2020)

* Corresponding author