第 14 卷第 4 期
2024 年 4 月

中国无机分析化学
Chinese Journal of Inorganic Analytical Chemistry

Vol. 14, No. 4
410～417

doi:10.3969/j.issn.2095-1035.2024.04.005

水体透射光谱结合主成分分析(PCA)改进化学需氧量(COD)含量估算研究

王彩玲　位欣欣

(西安石油大学 计算机学院,西安 710065)

摘　要　为了解决传统的化学需氧量(Chemical Oxygen Demand,COD)测量方法耗时较长,不利于快速、实时地获取水体中 COD 的信息等问题。通过采集 100 组 COD 水体光谱信息,分别使用 3 种不同的高光谱数据预处理方法对光谱数据进行预处理,并基于不同的预处理方法分别建立高斯过程回归模型(Gaussian Process Regression,GPR)和 BP 神经网络模型,分析不同预处理方法对模型精度的影响,建立了基于透射光谱测量结合主成分分析(Principal Component Analysis,PCA)改进水体 COD 含量估算模型。对各模型结合 PCA 数据降维方法进行模型的改进,通过比较模型的精度选择最优模型进行水体 COD 含量的检测。结果显示:相比于原始光谱数据建立的 GPR 模型和 BP 神经网络模型,数据预处理后的模型精度明显提升;且结合 PCA 对预处理后的数据进一步降维处理后,模型精度得到了进一步的提升。其中,基于标准正态变量变换特征结合 PCA 改进 BP 神经网络模型 R^2 高达 0.994 0,均方根误差 RMSE 为 0.022 540。证明了基于 PCA 数据降维方法对预处理后的光谱数据进行降维处理,有利于去除光谱中的冗余信息,提取特征信息,可以实现 COD 含量估算模型的优化,从而为传统 COD 测量方法存在的问题提出了一种新的解决思路。

关键词　透射光谱法;COD 含量预测;PCA;高斯过程回归;BP 神经网络

中图分类号:O657.39　O433.4　　**文献标志码**:A　　**文章编号**:2095-1035(2024)04-0410-08

Estimation of COD Content by Transmission Spectroscopy Combined with PCA

WANG Cailing, WEI Xinxin

(*School of Computer Science, Xi'an Shiyou University, Xi'an, Shaanxi 710065, China*)

Abstract　COD is an important indicator of organic pollution in water, and the higher the COD, the more serious the degree of water pollution. To solve the traditional method of COD determination is time-consuming, not conducive to rapid, real-time access to COD information in the water and other issues, in this paper, an improved model for COD determination in water on the basis of transmission spectroscopy measurement combined with principal component analysis (PCA) was proposed. Specifically, 100 groups of COD water body spectral information were collected, and three different hyperspectral data preprocessing

收稿日期:2023-07-12　　**修回日期**:2023-12-31

基金项目:陕西省重点研发计划项目(2023-YBSF-437);国家自然科学基金资助项目(31160475,61401439)

作者简介:王彩玲,女,副教授,主要从事高光谱成像中的信息提取技术、深度学习研究。E-mail:azering@163.com

引用格式:王彩玲,位欣欣. 水体透射光谱结合主成分分析(PCA)改进化学需氧量(COD)含量估算研究[J]. 中国无机分析化学,2024,14(4):410-417.

WANG Cailing, WEI Xinxin. Estimation of COD Content by Transmission Spectroscopy Combined with PCA[J]. Chinese Journal of Inorganic Analytical Chemistry, 2024, 14(4): 410-417.

methods were used to preprocess the spectral data, and Gaussian process regression(GPR) and BP neural network models were constructed based on different preprocessing methods to analyze the effects of different preprocessing methods on the accuracy of the models. Compared with GPR model and BP neural network model constructed from original spectrum data, it was found that after data pre-processing, there was a significant improvement in model accuracy, and after further dimension reduction of pre-processing data combined with PCA, there was a further improvement in model accuracy. Among them, the R^2 of the improved BP neural network model based on standard normal variable transformed features combined with PCA was as high as 0.994 0, and the RMSE was 0.022 540. This proved that the dimensionality reduction of the preprocessed spectral data based on the PCA data dimensionality reduction method was helpful to remove the redundant information in the spectral data and extract the feature information, and optimize the COD content estimation model, thereby solving the problems of traditional COD measurement methods. Thus, a new idea for the solution of the problems that exist in the traditional method of COD measurement is proposed.

Keywords　transmitted spectrum method; COD content prediction; PCA; Gaussian process regression; BP neural network

化学需氧量(Chemical Oxygen Demand, COD)是表征水体被还原性物质污染程度的指标,该指标作为有机物相对含量的综合指标之一,列入我国主要污染物总量控制指标,根据其排放浓度衡量水体污染程度[1]。传统的 COD 测量方法主要是基于化学分析,耗时较长,操作专业性高,不利于快速、实时地获取水体中 COD 的信息[2]。而高光谱技术结合人工神经网络模型可以快速、准确地估算水体中的 COD 含量,从而为环境监测和水质调控提供了有效手段。

近年来,关于利用高光谱遥感技术评价和监测水资源水质信息状况方面的研究愈发深入[3]。高光谱技术是一种通过对目标物体光谱信息的收集和分析,实现对目标物体性质的识别和定量测量的技术。利用高光谱技术,可以实现对水体中 COD 含量的快速、无损检测。国内外学者利用高光谱技术结合各种算法进行了大量水质检测技术的研究。YES 等[4]应用 UVE-SPA-LS-SUV 的方法实现了对 COD 的建模预测;KIMBERLY 等[5]构建出偏最小二乘最佳高光谱 Chl-a 浓度估算模型;ORTIZ 等[6]利用高光谱技术检测出水体总悬浮固体浓度;曹引等[7]建立偏最小二乘水体浊度高光谱反演模型,为水体浊度大面积遥感检测提供了技术支持;张贤龙等[8]提出高光谱技术水质参数浓度反演模型;蔡建楠等[9]采用 GA 遗传算法实现了基于偏最小二乘法高光谱 COD 检测模型的优化。

本文以水体 COD 含量为研究对象,通过多元散射校正(MSC)、标准正态变换(SNV)、最大最小归一化(MMN)三种不同的高光谱数据预处理方法对采集到的高光谱数据进行预处理,建立相应的高斯过程回归模型(Gaussian Process Regression, GPR)和 BP 神经网络模型,并对模型进行改进。结合主成分分析(Principal Component Analysis, PCA)方法对预处理后的数据进行主成分分析,通过数据降维,保留足以解释 90% 的方差的成分,从预处理后的光谱数据中提取 22 个主成分,筛选出相关性较好的波段,建立改进的 GPR 水体 COD 含量估算模型和 BP 神经网络模型水体 COD 含量估算模型。实验结果表明,基于 PCA 改进的模型的预测精度均明显提高,其中标准正态变量变换特征 PCA-BP 神经网络模型的 R^2 高达 0.994 0,均方根误差为 0.022 540,模型性能最优,能够实现水体中 COD 含量的检测。

1　实验部分

1.1　光谱仪

实验用仪器为 Ocean Optics 公司出品的 OCEAN-HDXXR 微型光纤光谱仪,该光谱仪采用高清晰度光学系统,具有高通量、低杂散光和高热稳定性的特点,适用于精确测量溶液中的分析物,具有体积小,容易集成到许多工业应用的生产过程环境的优势。

1.2　透射光谱数据获取

选择配比溶液为 0.1、0.2、0.3、0.4、0.5、0.6、0.7、0.8、0.9、1.0 mg/L 的 COD 标准溶液,更换光谱仪的狭缝为 10 μm,相同时间间隔各自重复采集 10 次上述标准溶液 200～1 030 nm 的高光谱透射率数据,共得到 100 条数据。

采用白板校正分别得到所采集的三种高光谱数据的光谱透射率值[10],如式(1)所示:

$$R_c=\frac{R_o}{R_w} \quad (1)$$

其中：R_o 为原始光谱数据，R_w 为白板数据。

2 实验结果

2.1 COD 原始透射光谱

图 1(a)为 10 种浓度 COD 原始透射光谱，从图 1 中可以看出，不同浓度溶液的 COD 光谱曲线的趋势类似，在紫外线波段 180.1～400 nm，COD 光谱曲线呈先下降后上升的趋势，这说明随着有机物含量的增加，水体 COD 含量越低，其光谱曲线特征越发明显。

2.2 数据预处理

对于高光谱数据，除了 COD 的特征信息外，还可能有光谱采集过程中产生的背景噪声辐射以及信号转换过程中产生的附加噪声[11]，分别采用不同的预处理方法进行处理，如图 1(b)～1(d)所示。其中，采用多元散射校正有效消除由于散射水平不同导致的光谱数据的差异，增强光谱与数据之间的相关性[12]；采用标准正态变量变换降低固体颗粒大小、表面散射以及光源变换等对光谱信息的影响[13]；采用最大最小归一化在不同程度上消除了光谱散射和背景干扰的影响[13]。

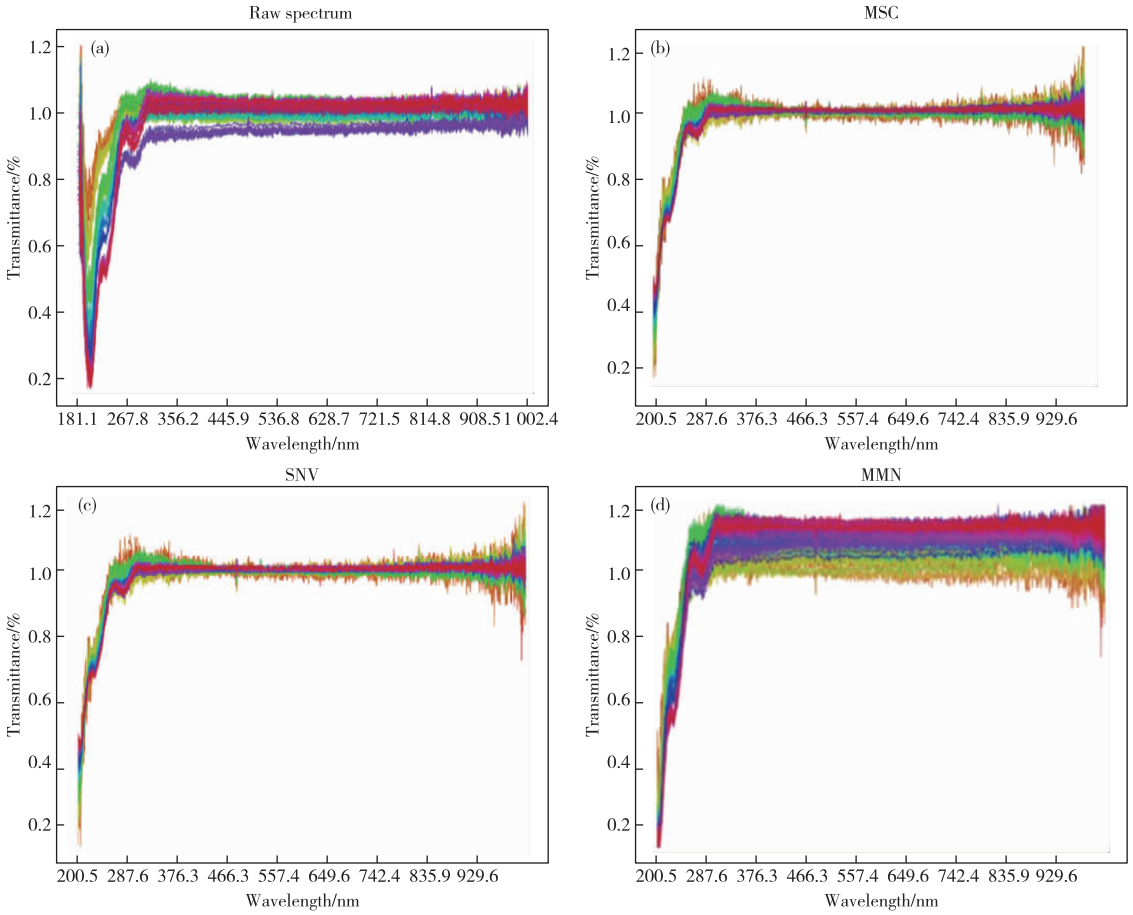

图 1 透射光谱图

Figure 1 Transmission spectrograms.

2.3 模型的建立

采用高斯过程回归模型和 BP 神经网络模型以上述预处理后的高光谱数据为自变量，将不同浓度的 COD 样本与光谱数据进行拟合，为了防止在模型的训练过程中出现过拟合的现象，采用五折交叉验证方法。输入为光谱数据，输出为 COD 样本的浓度。然后分别建立各类自变量的高斯过程回归模型和 BP 神经网络模型。

2.3.1 高斯过程回归模型建立

高斯过程回归(GPR)是一种建立在贝叶斯框架下的统计学习方法，模型性质完全由均值函数和协方差函数确定[14]。它有严格的统计学理论基础，对处理高维数、小样本、非线性等复杂回归问题具有良好的适应性[14]；该算法还具有容易实现，参数自适应获取，输出结果具有概率意义等优点[14]。

将预处理后的透射光谱数据作为模型的输入,建立高斯过程回归模型。使用 MATLAB 中自带的 Quadratic Rational Gaussian Process Regression 算法对高斯过程回归模型进行学习训练。本次实验中将该算法的基函数设置为常量,核函数选用二次有理函数,同时在训练过程中对高光谱数据进行标准化,优化数值参数,以达到最优效果。模型输出结果如图 2 所示。

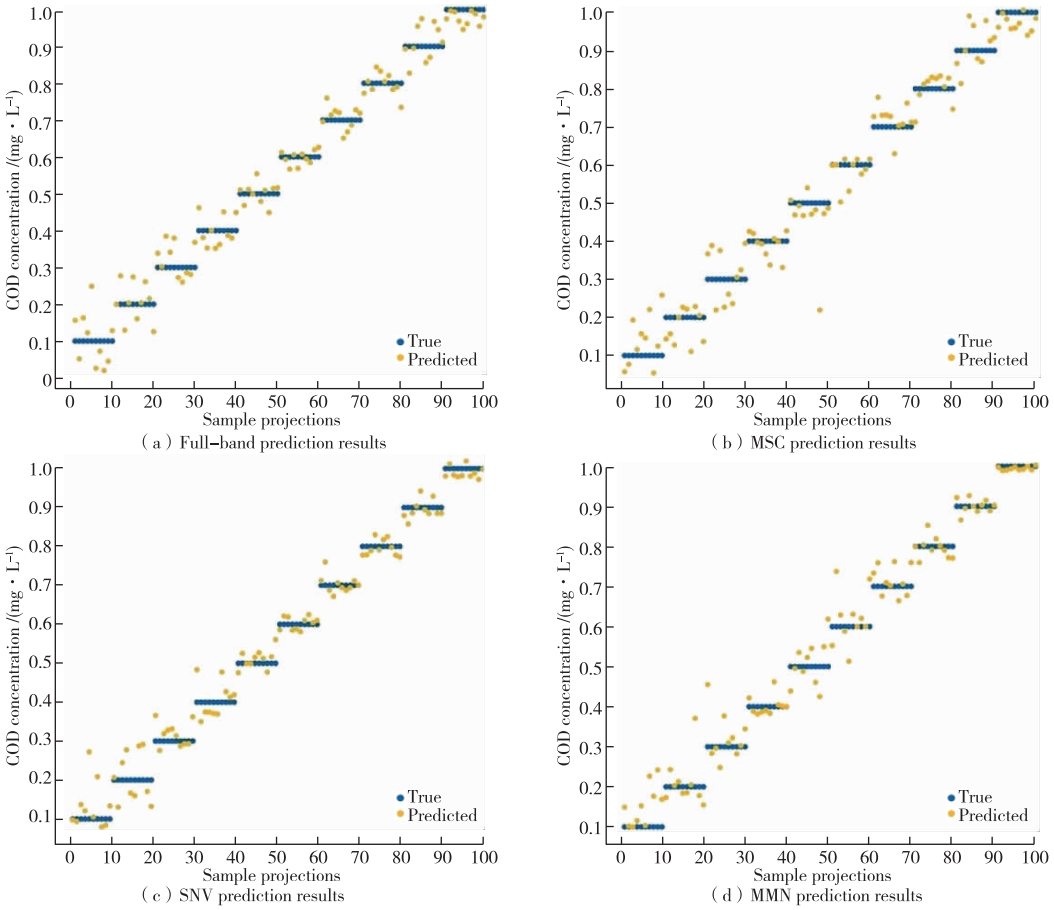

图 2　高斯过程回归模型预测结果

Figure 2　The prediction results of Gaussian process regression model.

2.3.2　BP 神经网络模型建立

使用 MATLAB 中自带的 Scaled Conjugate Gradient Backpropagation 算法对 BP 模型进行学习训练。该算法根据缩放共轭梯度法更新权重和偏差值,同时占用更少的内存,适用于高光谱数据,选择三层神经网络模型进行训练,第一层神经元个数设置为 20,第二、三层设置为 10,该算法中迭代次数(Epoch)阈值为 1 000,激活函数设置选用 Sigmoid 函数,探究不同预处理方法对 BP 网络模型回归准确率影响。模型输出结果如图 3 所示。

2.3.3　模型结果评估

以均方根误差 RMSE 和决定系数 R^2 为标准对所建立的各个模型进行精度检验与比较。其中:均方根误差 RMSE 越小,说明模型选择和拟合更好;决定系数 R^2 越接近 1,说明模型拟合的效果越好。检验结果如表 1 所示。

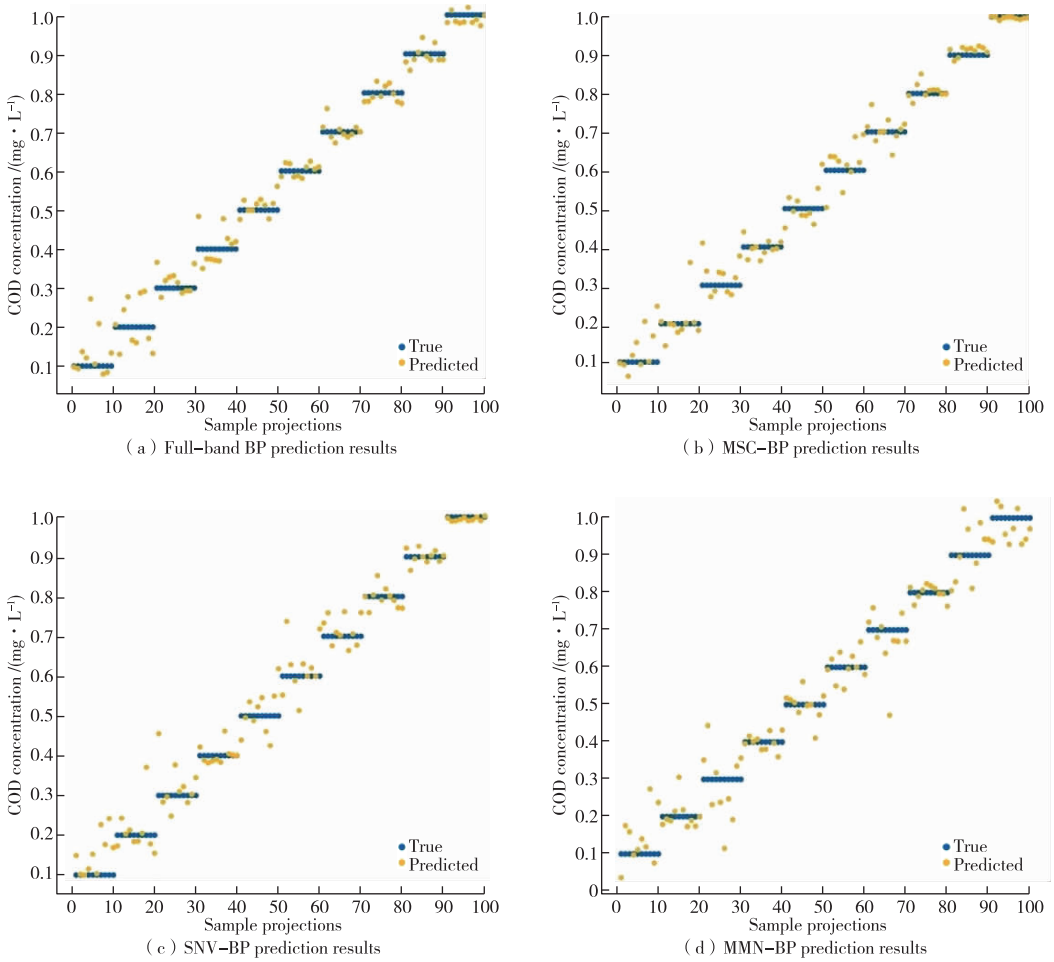

图 3 BP 神经网络模型预测结果

Figure 3 The prediction results of BP neural network model.

表 1 未改进模型精度检验结果

Table 1 Testing results of unimproved model accuracy

Model	R^2	*RMSE*
Full-band GPR	0.938 6	0.064 553
MSC-GPR	0.982 6	0.038 168
SNV-GPR	0.980 3	0.040 752
MMN-GPR	0.961 9	0.056 248
Full-band BPNN	0.940 1	0.063 416
MSC-BPNN	0.979 3	0.041 567
SNV-GPR	0.972 8	0.047 891
MMN-BPNN	0.958 5	0.058 702

由表 1 可知，与全波段的模型相比，经过预处理后的二次有理 GPR 模型和 BP 神经网络模型的性能均有所提高。其中，预处理后的二次有理 GPR 模型其 R^2 最高达 0.982 6；其 RMSE 最低为 0.038 168；预处理后的 BP 神经网络其 R^2 最高达 0.979 3，比全波段 R^2 高出 0.039 2，其 RMSE 最低为 0.041 567；与全波段的模型相比，预测精度均比原数据较高。说明采用预处理方法对数据进行处理可以有效提取有效光谱信息，排除干扰信息，从而提高光谱数据与 COD 浓度之间的相关性，使得模型的性能提高，预测效果更好。

2.4 基于 PCA 改进模型的建立

利用主成分分析法（PCA）对模型进行改进，建立基于 PCA 的 BP 神经网络定量估算模型以及二次有理 GPR 的定量估算模型。PCA 是一种使用最广泛的基于线性映射的特征提取技术，该算法通过一定的变换将高维数据映射到一个新的低维空间，使得任何数据投影的第一大方差在第一个坐标（称为第一主成分）上，第二大方差在第二个坐标（第二主成分）上，依此类推，这些主成分能够反映绝大部分的变量信息[15]。本文实验中设置 PCA 保留足以解释 90% 方差的成分。模型训练后，提取 22 个主成分。每成分的解释方差（顺序排列）：37.0%、18.4%、9.1%、4.3%、3.0%、2.0%、1.7%、1.5%、1.4%、1.3%（隐藏最不重要成分的方差）。

2.4.1 基于 PCA 改进的高斯回归模型

将 COD 数据集作为 PCA-二次有理 GPR 模型的输入。模型输出结果如图 4 所示。

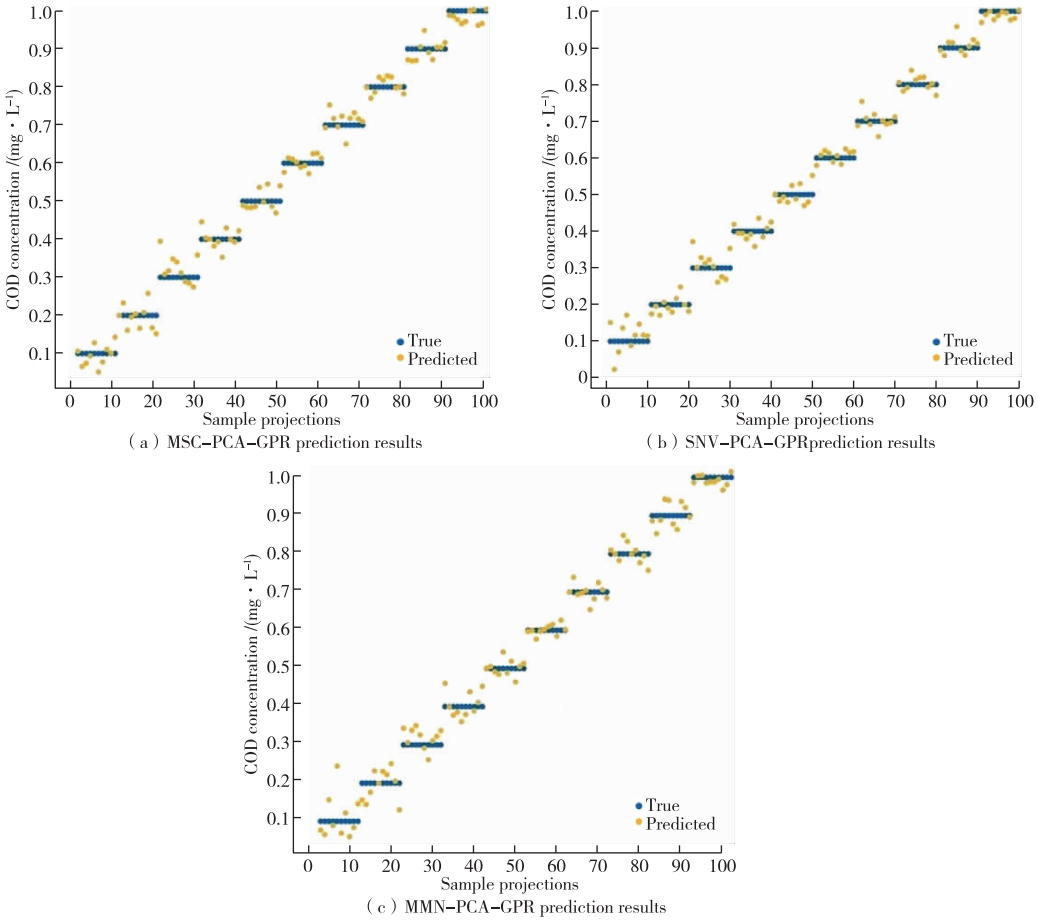

图 4　改进的高斯回归模型预测结果

Figure 4　The prediction results of improved Gaussian regression model.

2.4.2　基于 PCA 改进的 BP 神经网络模型

将 COD 数据集作为 PCA-BP 神经网络模型的输入。模型输出结果如图 5 所示。

2.4.3　基于 PCA 改进的模型结果评估

从输出的结果可以看出,预测值与真实值差异较小,具有很好的相关性。对所建立的各个改进的二次有理 GPR 模型以及 BP 神经网络模型进行精度检验并进行比较。改进模型检验结果如表 2 所示。

由表 2 可知,与未改进的模型相比,基于 PCA 改进模型的预测精度均有所提高。其中,多元散射校正特征 PCA-二次有理 GPR 模型的 R^2 增长为 0.990 9,多元散射校正特征 PCA-BP 神经网络模型的 R^2 增长为 0.990 8,其 RMSE 均有所减少;标准正态变换特征 PCA-二次有理 GPR 模型的 R^2 增长为 0.992 0,标准正态变量变换特征 PCA-BP 神经网络模型的 R^2 增长为 0.994 0,可以发现改进后的标准正态变量变换的 R^2 更接近于 1,且 RMSE 均明显减少,精度较为提高;最大最小归一化特征 PCA-二次有理 GPR 模型和最大最小归一化特征 PCA-BP 神经网络模型的 R^2 增长为 0.988 3 和 0.984 4;其 RMSE 减少为 0.031 195 和 0.036 048,预测精度相比未改进的模型也有所提升。说明采用 PCA 对预处理后的数据进行数据降维,可以实现 COD 含量估算模型的优化。

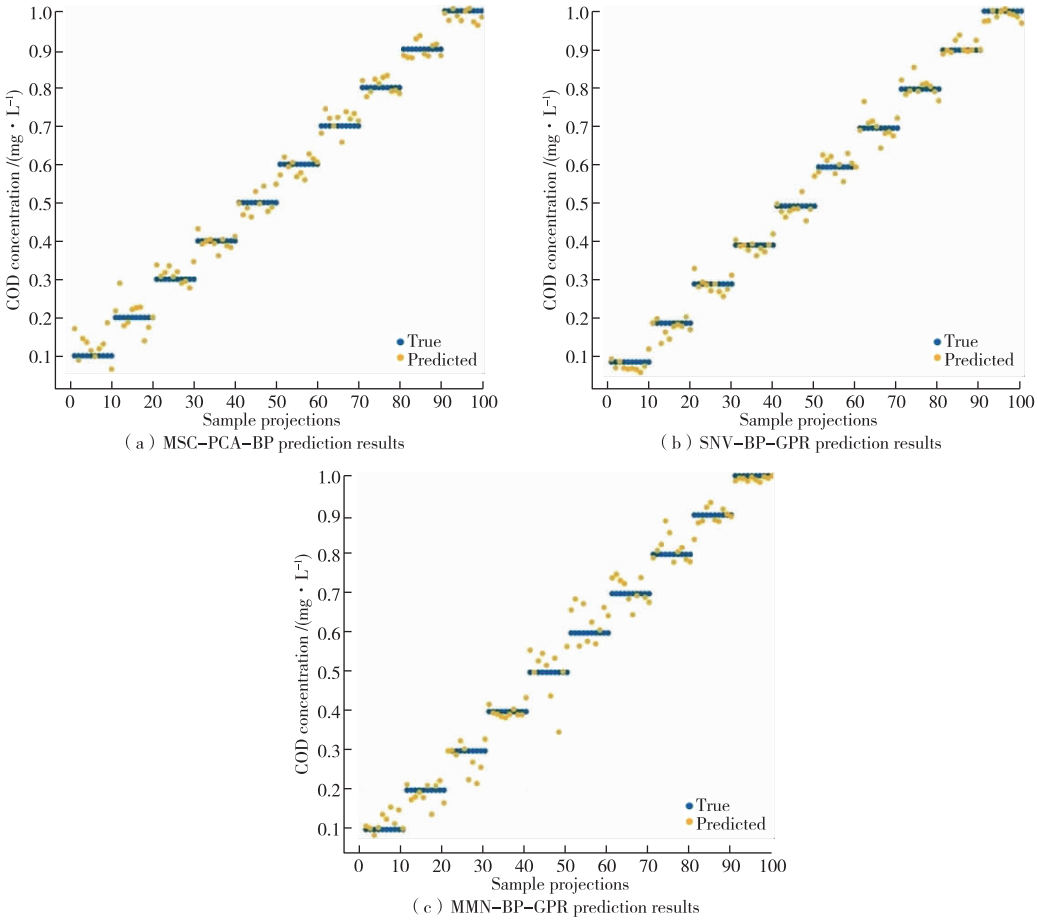

图 5 改进的 BP 神经网络预测结果

Figure 5 The prediction results of improved BP neural network model.

表 2 改进的模型精度检验表

Table 2 Testing results of improved model accuracy

Model	R^2	*RMSE*
MSC-PCA-GPR	0.990 9	0.027 545
SNV-PCA-GPR	0.992 0	0.026 005
MMN-PCA-GPR	0.988 3	0.031 195
MSC-PCA-BPNN	0.990 8	0.027 678
SNV-PCA-BPNN	0.994 0	0.022 540
MMN-PCA-BPNN	0.984 4	0.036 048

3 结论

分别采用多元散射校正、标准正态变量变换、最大最小归一化对光谱透射率数据进行预处理，并建立二次有理高斯回归模型和 BP 神经网络模型，对于不同的模型，探究不同特征输入对模型精度的影响，结果表明：3 种预处理方法可以有效降低噪音对数据的干扰，且二次有理 GPR 模型相比 BP 神经网络模型有较好的预测精度；基于 PCA 对各预处理后的透射光谱数据进行数据降维，筛选出相关性较好的波段，从而建立改进的二次有理 GPR 模型和 BP 神经网络模型。其中，标准正态变量变换特征 PCA-BP 神经网络模型决定系数达到了 0.994 0，均方根误差为 0.022 540，依据 R^2 最大、RMSE 最小原则，采用 PCA 改进的标准正态变量变换特征 BP 神经网络模型可以建立精度较好的 COD 定量估算模型。

参考文献

[1] 杨璟爱，关玉春，韩少强，等. 电导抑制-梯度淋洗离子色谱法测定化工废液中 4 种有机酸[J]. 中国测试，2021，47(8)：71-76.
YANG Jingai，GUAN Yuchun，HAN Shaoqiang，et al. Determination of four organic acids in chemical industry waste liquid by gradient elution ion chromatography with suppressed conductivity detection [J]. China Measurement & Test，2021，47(8)：71-76.

[2] 范丽华，施玉格，达莉芳，等. 酸化吹气-重铬酸钾法测定高氯地表水中化学需氧量[J]. 中国无机分析化学，2021，11(5)：97-101.
FAN Lihua，SHI Yuge，DA Lifen，et al. Determination of chemical oxygen demand in high chloride surface water

by adaptability of acidified blow-acidizing-dichromate[J]. Chinese Journal of Inorganic Analytical Chemistry, 2021, 11(5):97-101.

[3] 刘长宇,董绍俊. 水质生化需氧量快速检测新方法研究进展——现场、实时和就地监测[J]. 中国科学:化学, 2018,48(8):956-963.
LIU Changyu, DONG Shaojun. Advances in research on new methods for rapid detection of biochemical oxygen demand in water-in-situ, real-time, and in-place monitoring[J]. Scientia Sinica(Chimica), 2018, 48(8): 956-963.

[4] YES F, WANG D, MIN S G. Successive projections algorithm combined with uninformative variable elimination for spectral variable selection[J]. Chemo Metrics and Intelligent Laboratory Systems, 2008, 91(2):194-199.

[5] KIMBERLY R, KHALID A. Application of a partial least-squares regression model to retrieve chlorophyll-a concentrations in coastal waters using hyper-spectral data[J]. Ocean Science Journal, 2016, 51(2):209-221.

[6] ORTIZ A. Multivariate approach for chlorophyll-a and suspended matter retrievals in case Ⅱ type waters using hyperspectral data[J]. Hydrological Sciences Journal, 2016, 61(1):200-213.

[7] 曹引,冶运涛,赵红莉,等. 基于离散粒子群和偏最小二乘的水源地浊度高光谱反演[J]. 农业机械学报, 2018, 49(1):173-182.
CAO Yin, YE Yuntao, ZHAO Hongli, et al. Satellite hyperspectral retrieval of turbidity for water source based on discrete particle swarm and partial least squares[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(1):173-182.

[8] 张贤龙. 基于激光诱导击穿光谱和高光谱技术的水质指标定量研究[D]. 乌鲁木齐:新疆大学, 2019.
ZHANG Xianlong. Quantitative analysis of water quality indexes based on laser induced breakdown spectroscopy and hyperspectral technology[D]. Urumqi: Xinjiang University, 2019.

[9] 蔡建楠,刘海龙,姜波,等. 基于 GA-PLS 算法的河网水体化学需氧量高光谱反衍[J]. 灌溉排水学报, 2020, 39(9):126-131.
CAI Jiannan, LIU Hailong, JIANG Bo, et al. Using hyperspectral imagery and ga-pls algorithm to estimate chemical oxygen demand concentration of water in river network[J]. Journal of Irrigation and Drainage, 2020, 39(9):126-131.

[10] 朱亚东,何鸿举,王魏,等. 高光谱成像技术结合线性回归算法快速预测鸡肉掺假牛肉[J]. 食品工业科技, 2020,41(4):184-189.
ZHU Yadong, HE Hongju, WANG Wei, et al. Quick detection of beef adulteration using hyperspectral imaging technology combined with linear regression algorithm[J]. Science and Technology of Food Industry, 2020, 41(4):184-189.

[11] AMIGO J M, SANTOS C. Preprocessing of hyperspectral and multispectral images[M]. Data Handling in Science and Technology, 2020, 32:37-53.

[12] 陶培峰,王建华,李志忠,等. 基于高光谱的土壤养分含量反演模型研究[J]. 地质与资源, 2020, 29(1):68-75, 84.
TAO Peifeng, WANG Jianhua, LI Zhizhong, et al. Research of soil nutrient content inversion model based on hyperspectral data[J]. Geology and Resources, 2020, 29(1):68-75, 84.

[13] 李尚科,李跑,杜国荣,等. 基于近红外光谱技术和优化预处理方法的不同品牌燕麦无损鉴别分析[J]. 食品安全质量检测学报, 2019, 10(24):8204-8210.
LI Shangke, LI Pao, DU Guorong, et al. Non-destructive identification of different brands of oats based on near-infrared spectroscopy and optimized pretreatment methods[J]. Journal of Food Safety & Quality, 2019, 10(24):8204-8210.

[14] 何志昆,刘光斌,赵曦晶,等. 高斯过程回归方法综述[J]. 控制与决策, 2013, 28(8):1121-1129, 1137.
HE Zhikun, LIU Guangbin, ZHAO Xijing, et al. Overview of Gaussian process regression[J]. Control and Decision, 2013, 28(8):1121-1129, 1137.

[15] 张亮. 基于 PCA 和 SVM 的高光谱遥感图像分类研究[J]. 光学技术, 2008, 34(增刊 1):184-187.
ZHANG Liang. Study on the hyperspectral remote sensed image classify based on PCA and SVM[J]. Optical Technique, 2008, 34(Suppl. 1):184-187.