摘要：化学需氧量（COD）是反映水体污染程度的重要指标之一。针对紫外可见光谱 COD 测量法数据波段多，易受干扰的问题，提出以局部线性嵌入法（LLE）结合支持向量机回归法（SVR）建立预测模型，来提高预测精度。首先，通过尝试预处理方法与模型分析方法（SVR 和偏最小二乘回归法（PLSR））的不同组合来判断预测模型的效果，结果表明，“小波变换（WT）+SVR”效果较好。为了减少计算复杂度，提高运算效率，分别运用 LLE 和主成分分析算法（PCA）对数据降维，再分别结合 SVR 建立 COD 浓度预测模型。结果表明，利用“LLE+SVR”得到的 COD 浓度预测模型，其训练样本的均方误差为 0.076030，测试样本均方误差为 0.061477，分别小于“PCA+SVR”模型的 0.216076 和 0.317303。这种方法使模型预测精度得到提高，为紫外可见光谱法检测水质 COD 浓度提供了一种可行的分析方法。

关键词：化学需氧量（COD）；支持向量机回归（SVR）；紫外可见吸收光谱；局部线性嵌入（LLE）
中图分类号：X824　　文献标识码：A　　文章编号：1006-883X(2018)09-0011-05
收稿日期：2018-06-06

基于 LLE-SVR 的水质 COD 紫外光谱检测方法研究

康贝　马洁
北京信息科技大学 自动化学院，北京 100192

一、引言

化学需氧量（Chemical Oxygen Demand，COD）是能够反应水体有机污染程度的一项重要的指标 [1]。基于紫外光谱分析的水质监测是通过建立紫外吸光度和有机物以及部分无机物浓度的相关模型来评价水体污染程度，具有环保、低成本、便携等优点，是水质监测仪器的重要发展方向 [2]。在一定程度上，基于紫外可见光谱法测定 COD，其测量精度严重依赖于所建立的数学模型。因此，如何选择合适的数学模型以提高预测精度，成为目前紫外可见光谱 COD 测量研究的重点 [3]。目前，基于紫外可见吸收光谱分析中，模型分析方法主要有人工神经网络（Artificial Neural Network，ANN）[4-5]、主成分回归（Principal Component Regression，PCR）[6-7]、支持向量机回归（Support Vector Regression，SVR）[8-9] 以及偏最小二乘（Partial Least Squares，PLS）[10] 等，在线性分析建模中，比较常用的是 PLS。

张森 [11] 等运用偏最小二乘法结合支持向量机的方法，解决了水质因子多重相关问题，提高了预测精度，预测值与实际值相对误差均低于 1%，最大为 0.7759%，平均相对误差为 0.39%。俞禄 [12] 等以总有机碳（Total Organic Carbon，TOC）、COD 为指标，分别建立 PLS、PCR、偏最小二乘回归（partial least squares regression，PLSR）、最小二乘支持向量机（Least Squares Support Vector Machine，LSSVM）预测模型，结果表明，LSSVM 的预测精度最高。陈武奋 [13] 等以水温、溶解氧、电导率、浊度数据为影响因子，建立基于 SVR 的水质 pH 值预测模型，结果表明，基于 SVR 预测模型训练集决定系数为 0.854、测试集决定系数为 0.897，平均相对误差为 1.419%，该模型为水质评价提供了一定的参考价值。

由于水体具有一定的多样性和复杂性，紫外吸收

光谱信号通常在全波段进行扫描采样。但是全波段光谱信息作为非线性建模输入变量会导致模型的复杂度增加，减低计算速度，使检测的实时性难以保证，同时会导致不确定干扰因素的引入，进而降低准确度[14]。

对原始数据降维是一种有效的消噪并且提取有用信息的方法。流形学习是从高维映射到低维流空间，来达到数据低维、可视的目的，从而找到内在规律[15-16]。局部线性嵌入式算法（Locally Linear Embedding，LLE）是 Roweis 和 Saul[17] 于 2000 年提出的非线性降维方法，其本质是利用局部线性去逼近全局非线性，对原始数据点进行重构，来保持整体的特性。LLE 方法具有低复杂度、少参、高效、容易实现等优点[18-19]。

本文对室内光谱仪测量的水样紫外光谱进行研究，光谱信息量庞大、维数过高，存在着噪声干扰，需要对数据降维。首先通过 LLE 对紫外可见吸收光谱数据进行非线性降维，然后建立基于 SVR 的预测模型，由此结合 LLE 和 SVR 的优点，建立了基于水质 COD 预测模型。结果表明，LLE-SVR 方法建立的预测模型效果显著。

二、实验材料

实验中，共获得 54 组样本，取自于某市生活废水、河流地表水以及工业排放废水，本实验采用的是 BIM-6002A 光谱探测器（杭州 Brolight 公司生产）采用交叉非对称 C-T 光路结构，光学分辨率高达 0.35nm~1nm，光源选择 LS~3000 型 50W 卤素灯，工作波长范围为 200nm~900nm。根据 GB11914-89《重铬酸盐法水质化学需氧量的测定》来获得每个样本的 COD 真值[20]。图 1 为 54 组样本的光谱曲线，光谱采集的范围为 190nm~400nm。

三、结果与性能评价

本实验总共获得 54 个实验样本，将这些样本划分为训练集和预测集，其中训练集样本 42 个，用来建立数学模型，预测集样本 12 个，用来检验模型的精度和预测能力。

1、光谱数据预处理与初步建模分析

在数据采集的过程中，由于仪器设备的高频噪音、人员操作、外界环境等因素，往往会导致一些无关因素参与到模型的建立，从而影响所建模型预测精度，选择合适的预处理方法能够提高模型的预测精度。对光谱数据分别进行 S-G 平滑滤波（Savitzky-Golay smoothing filter，SG）、标准正态变换（Standrd Normal Variate，SNV）、一阶微分（First Derivative，FD）以及小波变换（Wavelet Transform，WT）的预处理，再通过 PLSR 和 SVR 两种方法进行建模分析，来分析比较不同预处理方法对模型预测结果的影响。

本实验采用均方误差（Mean Square Error，MSE）作为性能指标评价模型的建模和预测能力。

通过表 1 可以得出，通过不同预处理的光谱数据得出了不同的预测效果，通过比较训练样本和测试样本，WT+SVR 获得的预测模型精度最高。WT 可以有效的抑制无用噪声，并保留有用信息，采用 SVR 建模在总体上优于 PLSR，其原因可能是 SVR 能够有效利用光谱信息中隐含的与水质 COD 浓度相关的非线性关系。

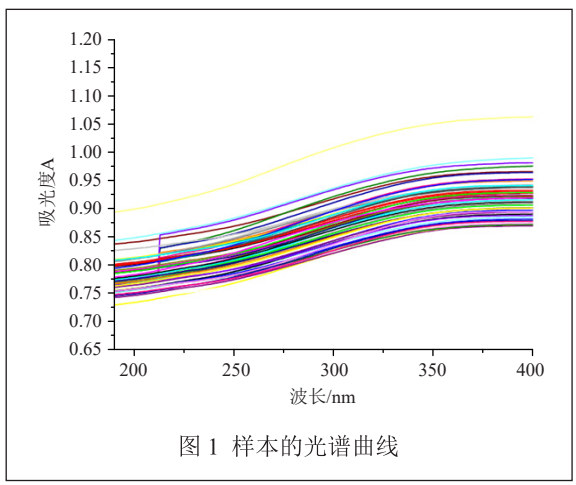

图 1 样本的光谱曲线

表 1 不同预处理方法误差

	PLSR		SVR	
预处理方法	训练误差	测试误差	训练误差	测试误差
SNV	8.2	8.3	8.7	9.2
S-G	8.4	12.6	9.0	9.5
FD	9.4	9.2	10.9	11.3
WT WT	7.8	6.5	8.5	7.7

2、光谱数据的再处理

在水质检测中，利用全光谱参与模型建立会增加模型复杂度，计算时间长，效率低，不利于模型的应用和推广。由于全光谱数据中可能含有一些无关信息参与模型建立，运用以上预处理方法仍旧不能很好的改善模型预测精度。通过数据降维，一方面可以降低维数，减小复杂度，另一方面可以更好地提取有用信息。下面分别运用 LLE 和主成分分析（Principal Component Analysis，PCA）算法对数据降维，再分别结合 SVR 建立 COD 浓度预测模型。

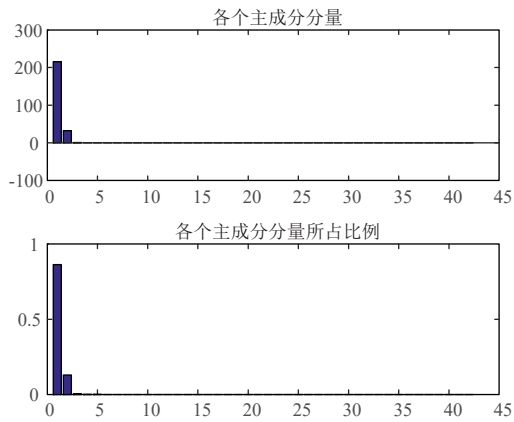

图 2 各主成分分量及所占比例

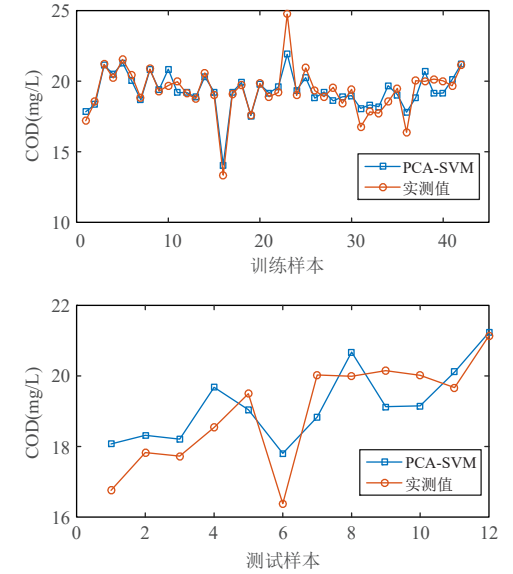

图 3 PCA-SVR 模型预测值与实际值比较曲线

（1）PCA-SVR

PCA 算法能够在数据空间中发掘出能尽可能使数据从高维降到低维的向量，以此来得到原始数据空间对应的最低维空间的算法。通过对预处理后的光谱数据进行 PCA 降维，得到其各主成分分量以及所占比例，如图 2 示。

取累计贡献率达到 85% 以上的前 14 维数据，采用 SVR 进行建模分析，得出训练样本 *MSE*=0.216076，测试样本 *MSE*=0.317303，模型预测值与实际值比较曲线如图 3 所示。

（2）LLE-SVR [21]

假设有 n 个水样紫外光谱数据样本，$\boldsymbol{X}=\{x_1, x_2,\ldots, x_n\}$ 为初始光谱样本，且 $x_i=R^p$，原始光谱维数为 p。映射到低维空间的样本为 $\boldsymbol{Y}=\{y_1, y_2,\ldots, y_n\}$，$y_i \in R^d$，$d$ 为降维后数据维数（$d<p$），d 为预先设定的值。

LLE 算法的步骤如下：

LLE 方法是映射数据集 $\boldsymbol{X}=\{x_1, x_2,\ldots, x_n\}$，$x_i=R^p$ 到数据集 $\boldsymbol{Y}=\{y_1, y_2,\ldots, y_n\}$，$y_i \in R^d$（$d<p$），主要包括 3 步：

第 1 步，局部邻域，计算出每个样本点 x_i 与其他 n-1 个样本之间的欧氏距离，选取 x_i 的 k 个近邻点，k 为预设值；

第 2 步，重新计算对每个样本点 x_i 以及它的 k 个近邻点的权值；

$$\varepsilon(W)_{\min}=\sum_{i=1}^{n}\left\|x_i-\sum_{j=1}^{n}\omega_{ij}\cdot x_j\right\|^2 \quad (1)$$

其中，ω_{ij}—x_i 和 x_j 之间的权值，且当 x_j 不属于 x_i 的近邻时，$\omega_{ij}=0$；

第 3 步，根据重建权值，将所有样本数据点映射到低维空间中，得到低维输出，且尽量保持高维空间中的局部线性特征，使重构误差函数最小。

$$\varepsilon(\boldsymbol{Y})_{\min}=\sum_{i=1}^{n}\left\|y_i-\sum_{j=1}^{n}\omega_{ij}\cdot y_j\right\|^2 \quad (2)$$

要求满足下面两个条件，即：

$$\begin{cases}\sum_{i=1}^{n} y_i=0\\ \frac{1}{n}\sum_{i=1}^{n} y_i y_i^{\mathrm{T}}=1\end{cases} \quad (3)$$

k 和 d 为 LLE 算法的两个可调参数。k 和 d 的大

小不同，训练样本和测试样本也会得到不同的预测结果，k 的选择受样本个数的影响，本实验为小样本，选取的 k 值较小；而 d 的选择受光谱数据中干扰因素大小及多少的影响，维数过小导致提取信息不够充分，维数过高则加大噪声等无用信息对模型建立的影响。

本文运用六折交叉验证的方法，k 从 10 取到 20，d 从 10 取到 30，得到最佳的 k 为 13，最佳的 d 为 27。图 4 为参数选择结果图。

SVR 不敏感损失系数 ε 取 0.021、惩罚系数 C 取 10000、RBF 核函数的宽度系数 γ 取 7.2，训练样本 MSE=0.076030，测试样本 MSE=0.06147，模型预测值与实际值比较曲线如图 5 所示。

从图 3、图 5 可以看出，LLE-SVR 模型预测结果的拟合精度相对于 PCA-SVR 有很大的提高，以 MSE 为评价标准，进一步对比两个模型的预测性能，相对 PCA-SVR 模型训练样本 MSE=0.216076，测试样本 MSE=0.317303，LLE-SVR 模型训练样本 MSE=0.076030，测试样本 MSE=0.061477，模型精度显著提高，由此可见，LLE-SVR 模型有效提取了光谱中的非线性特征，预测效果优于 PCA-SVR 模型。

四、结束语

由于水体成分复杂，无关因素干扰比较多，若以全波段作为输入，对于所建模型精度必将有很大地影响。本文采用局部线性降维（LLE）和支持向量机回归（SVR）相结合的方法，建立了水样紫外可见光谱吸光度与 COD 浓度之间的预测模型。得到以下结论：

（1）分别用不同的预处理方法结合 SVR 和 PLSR，发现运用 WT 结合 SVR 建立的模型效果最好；

（2）预处理后的光谱数据结合 LLE 非线性降维工具，并与 PCA 降维进行比较，采用 LLE 降维后的预测效果更理想；

（3）本文利用 LLE 这一非线性降维工具结合 SVR 建立预测模型，揭示了水质 COD 浓度和吸光度之间的非线性关系，提高了模型预测精度，为紫外可见光谱法检测水质 COD 浓度提供了一种可行的分析方法。

参考文献

[1] 李婉茹 , 洪世杰 , 王向工 . 黑河水体 COD 浓度与重点污染源相关性研究 [J]. 漯河职业技术学院学报 (综合版), 2005(02): 1-3.

[2] 曾甜玲 , 温志渝 , 温中泉 , 张中卫 , 魏康林 . 基于紫外光谱分析的水质监测技术研究进展 [J]. 光谱学与光谱分析 , 2013, 33(04): 1098-1103.

[3] 包鑫 , 戴连奎 . 汽油多参数拉曼光谱分析仪中的稳健支持向量机方法 [J]. 仪器仪表学报 , 2009, 30(9): 1829-1835.

[4]Raed Jafar, Isam Shahrour, Ilan Juran. Application of Artificial Neural Networks (ANN) to model the failure of urban water mains[J]. Mathematical and Computer Modelling, 2010, 51(9):1170-1180.

[5] 李颖 , 李耀辉 , 王金鑫 , 张成才 . SVM 和 ANN 在多光谱遥感影像分类中的比较研究 [J]. 海洋测绘 , 2016, 36(05): 19-22.

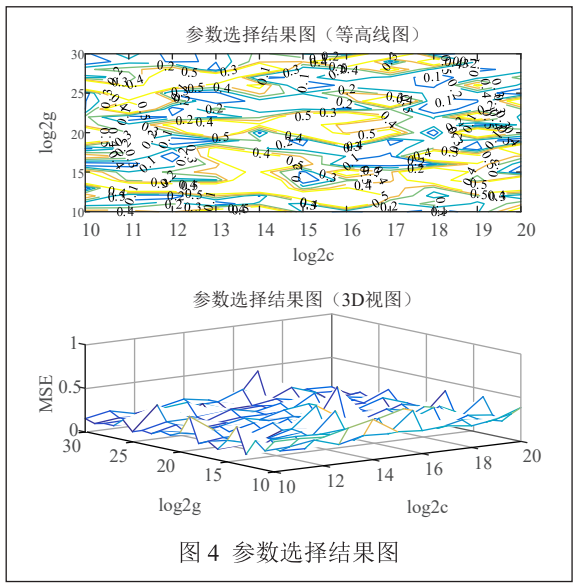

图 4 参数选择结果图

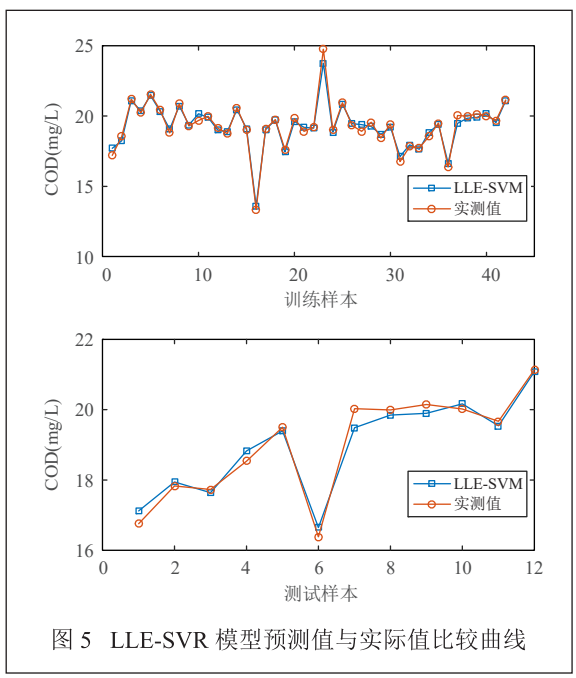

图 5 LLE-SVR 模型预测值与实际值比较曲线

[6]Mengli Fan, Xiuwei Liu, Xiaoming Yu, Xiaoyu Cui, Wensheng Cai, Xueguang Shao. Near-infrared spectroscopy and chemometric modelling for rapid diagnosis of kidney disease[J]. Science China(Chemistry), 2017, 60(02): 299-304.
[7] 曲楠 . 近红外光谱技术在药物无损非破坏定量分析中的应用研究 [D]. 长春 : 吉林大学 , 2008.
[8] 刘双印 , 徐龙琴 , 李振波 , 李道亮 . 基于 PCA-MCAFA-LSSVM 的养殖水质 pH 值预测模型 [J]. 农业机械学报 , 2014, 45(05): 239-246.
[9] 龚怀瑾 , 毛力 , 杨弘 . 基于变尺度混沌 QPSO-LSSVM 的水质溶氧预测建模 [J]. 计算机与应用化学 , 2013, 30(03): 315-318.
[10] 胡益 , 马贺贺 , 侍洪波 . On-Line Batch Process Monitoring Using Multiway Kernel Partial Least Squares[J]. Journal of Donghua University(English Edition), 2011, 28(06): 585-591.
[11] 张森 , 石为人 , 石欣 , 郭宝丽 . 基于偏最小二乘回归和 SVM 的水质预测 [J]. 计算机工程与应用 , 2015, 51(15): 249-254.
[12] 俞禄 . 几种建模方法在光谱水质分析中的应用和比较 [A]. 见 : 中国自动化学会控制理论专业委员会 . 中国自动化学会控制理论专业委员会 B 卷 [C]. 北京 : 中国自动化学会控制理论专业委员会 , 2011:4.
[13] 陈武奋 , 张倩华 , 黄钲武 , 黄立 , 林年旺 . 基于回归支持向量机的温泉水质 pH 值预测研究 [J]. 人民珠江 , 2016, 37(07): 94-97.
[14] 杨辉华 , 覃锋 , 王义明 , 罗国安 . NIR 光谱的 Isomap-PLS 非线性建模方法 [J]. 光谱学与光谱分析 , 2009, 29(02):322-326.
[15] 徐伟 . 基于多信息融合的流形学习方法研究 [D]. 扬州 : 扬州大学 , 2013.
[16] 冷亦琴 . 局部线性嵌入流形学习及其应用研究 [D]. 苏州 : 苏州大学 , 2014.
[17]S T Roweis, L K Saul. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500):2323-2326.
[18]Aimin MIAO, Zhihuan SONG, Zhiqiang GE, Le ZHOU, Qiaojun WEN. Nonlinear fault detection based on locally linear embedding [J]. Journal of Control Theory and Applications, 2013, 11(04): 615-622.
[19] 段志臣 , 芮小平 , 张立媛 . 基于流形学习的非线性维数约简方法 [J]. 数学的实践与认识 , 2012, 42(08):230-241.
[20] Wu Yuanqing, Du Shuxin, Yan Yun. Ultraviolet spectrum analysis methods for detecting the concentration of organic pollutants in Water [J]. Spectroscopy and Spectral Analysis, 2011, 31(1):233-237.
[21] 姜伟 , 杨炳儒 . 基于流形学习的维数约简算法 [J]. 计算机工程 , 2010, 36(12): 25-27.

Study of UV Visible Spectrum-Based COD Detection Method for Water Quality Monitoring Based on LLE-SVR

KANG Bei, MA Jie

(*School of Automation, Beijing Information Science & Technology University, Beijing* 100192, *China*)

Abstract: Chemical Oxygen Demand (COD) is one of the important indicators reflecting the degree of water pollution. In view of the problems of multispectral data and easy to be disturbed in UV visible spectrum-based COD measurement, a method combining Locally Linear Embedding (LLE) with Support Vector Regression (SVR) is proposed to build prediction model to improve prediction accuracy. Firstly, different combinations of preprocessing methods and model analysis methods (SVR and Partial Least Square Regression (PLSR)) are tried to estimate the effect of the prediction model. The results show that "Wavelet Transform(WT)+ SVM" is better in the model effect. Then, in order to reduce computational complexity and improve computational efficiency, LLE and Principal Component Analysis (PCA) are used respectively to reduce the dimensionality of the data and establish the COD concentration prediction models combined with SVR. The results show that the mean square errors of the training samples and the test samples are 0.076030 and 0.061477 in "LLE+SVR" prediction model, which are less than 0.216076 and 0.317303 respectively in the "PCA +SVR" model. This method improves the prediction accuracy of the model and provides a feasible analysis method for the UV visible spectrum-based COD concentration determination in water quality monitoring.

Key words: Chemical Oxygen Demand (COD); Support Vector Regression (SVR); ultraviolet visible absorption spectrum; Locally Linear Embedding (LLE)

作者简介

康贝，北京信息科技大学自动化学院，硕士研究生，研究方向为生态与环境检测技术。

通讯地址：北京市海淀区北京信息科技大学小营校区

邮编：100192

邮箱：815355364@qq.com

马洁，北京信息科技大学，教授，研究方向为复杂系统建模、分析与控制。