

多光谱数据融合和 GANs 算法的 COD 浓度预测

陈颖¹, 许扬眉¹, 邸远见¹, 崔行宁¹, 张杰¹, 周鑫德¹, 肖春艳², 李少华³

1. 燕山大学电气工程学院, 河北省测试计量技术及仪器重点实验室, 河北 秦皇岛 066004

2. 河南理工大学资源与环境学院, 河南 焦作 454000

3. 河北先河环保科技股份有限公司, 河北 石家庄 050000

摘要 水体中过高浓度的有机污染物含量危害巨大, 不仅会造成严重的环境污染, 而且会危害人类身体健康。化学需氧量(COD)表征了水体中有机污染物的污染程度。提出了一种将紫外(UV)光谱和近红外(NIR)光谱进行多光谱数据级融合(LLDF)和特征级融合(MLDF), 进而构建基于生成对抗式网络(GANs)算法的 COD 浓度定量预测模型。首先按照一定的浓度梯度配制 COD 标准液样本, 分别采集标准液的 UV 光谱(190~310 nm)和 NIR 光谱(830~2 100 nm), 对获取到的 UV 和 NIR 光谱数据进行一阶导数和 Savitzky-Golay (S-G)平滑的预处理, 消除基线漂移和干扰噪声; 基于预处理过的光谱, 直接进行数据级和特征级的数据融合, 结合 GANs 算法搭建 COD 浓度预测模型。并使用评价参数相关系数的平方(R^2)、预测值与真实浓度值的均方根误差(RMSEP)和预测偏差来对模型进行评价。结果表明, 不论是特征级融合模型还是数据级融合模型都不够理想。分析原因可知, 由于 UV 和 NIR 波段数据量不均衡, 导致 NIR 波段掩盖掉了 UV 光谱的模型贡献度, 让光谱融合失去意义。为了避免融合失败, 拟采用归一化的方法处理多光谱数据, 并讨论了标准归一化(SNV)、最大最小归一化(MMN)和矢量归一化(VN)对建模的影响。将经过归一化后的 UV 和 NIR 光谱数据再次进行融合, 分别作为 GANs 模型的输入 X , 将真实测量 COD 值作为输出值 Y , 建立不同归一化方法处理后的 COD 浓度预测模型。建模结果显示, 采用不同归一化方法对多光谱数据融合模型的影响较大, 不论是数据级融合模型还是特征级融合模型的预测精度较未归一化之前有明显的提升, 其中采用最大最小归一化的预测模型效果提升最为明显。与单一谱源的全波长 UV 波段的 GANs 预测模型、全波长 NIR 波段的 GANs 预测模型进行对比来验证多光谱数据融合 GANs 预测模型的精度, 结果表明: 基于 UV 和 NIR 光谱的特征级光谱融合模型的 R^2 为 0.994 7, RMSEP 为 0.976, 比数据级融合的预测模型误差降低了 52.9%, 预测回收率为 98.4%~103.1%, 远好于其他几组, 模型的泛化能力更强, 预测精度也更高。与单一谱源的预测模型相比, 多光谱数据融合能反应更多的水体样品的化学信息, 更加全面揭示水体的污染物程度, 从不同的层面上反应水体中污染物的差异, 为在线监测水体中 COD 浓度提供一定的技术支持。

关键词 紫外光谱; 近红外光谱; 数据融合; GANs 模型; COD 浓度预测

中图分类号: O433.4 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2021)01-0188-06

引言

大量的有机污染物排入到水体中, 导致河流、湖泊和海洋都受到了不同程度的污染^[1]。化学需氧量(chemical oxygen demand, COD)指在一定环境下, 水体中的还原性物质被氧化分解时所消耗氧化剂的量, 单位以耗氧量 $\text{mg} \cdot \text{L}^{-1}$ 表

示, 化学需氧量表征水体受到有机物污染的程度^[2], 因此, 化学需氧量可作为有机污染物监测的综合指标。

近年来, 各国学者进行了大量研究以致力于寻找快速、环保的 COD 检测方法, 紫外光谱(ultraviolet, UV)与近红外光谱(near infrared, NIR)技术具有无损、快速、样品制备简单及可实现在线分析等特点, 因此广泛应用于水体的污染物监测。在国外, Lepot 等^[3]将 UV 光谱与遗传算法引入 COD

收稿日期: 2018-12-27, 修订日期: 2020-02-09

基金项目: 国家重点研发计划项目(2016YFC1400601-3), 河北省重点研发计划项目(19273901D, 20373301D), 河北省自然科学基金项目(F2020203066), 中国博士后基金项目(2018M630279), 河北省博士后择优资助项目(D2018003028), 河北省高等学校科学技术研究项目(ZD2018243)资助

作者简介: 陈颖, 女, 1980年生, 燕山大学电气工程学院教授 e-mail: chenying@ysu.edu.cn

建模, 效果良好; Abedinzadeh 等^[4]利用 NIR 光谱对造纸厂废水的 COD 进行了在线监测, 所得的预测误差大约为被测水样 COD 浓度的 1/10; Martelo-Vidal 等^[5]采用 UV 光谱联合 NIR 光谱, 搭建了神经网络预测模型, 取得了不错的预测精度。在国内, 赵友全等^[6]利用主成分分析结合欧式距离分析 UV 光谱, 并基于偏最小二乘法建模, 实现了水样的分类预测; 仲洋等^[7]基于 UV 和 NIR 光谱进行多光谱融合对水质 COD 进行检测, 证明了多光谱融合建模可有效提高 COD 的预测精度。

本文提出了一种基于 UV 光谱和 NIR 光谱的多光谱信息融合 COD 浓度预测模型, 通过一阶导数和 S-G 平滑对原始 UV 和 NIR 光谱预处理; 为了避免光谱融合过程中数据量不一致导致的融合失败, 提出并分析了标准归一化 (standard normal variation, SNV)、最大最小归一化 (max-min-nor, MMN) 和矢量归一化 (vector normalization, VN) 处理光谱的效果, 筛选出最合适的归一化方法; 采用数据级数据融合 (data level fusion, LLDF) 和特征级数据融合 (feature level fusion, MLDF)^[6]对归一化后的光谱进行融合; 基于生成对抗网络 (generative adversarial networks, GANs) 算法建立最终的多光谱 COD 浓度预测模型。

1 方法和原理

1.1 GANs 算法

生成对抗网络是机器学习的模型, 它包含两个学习网络: G (Generator) 网络和 D (Discriminator) 网络。其中 G 网络是一个生成模型的学习网络, 它接受样本集的数据进行训练, 并生成一个预测模型, 记作 $G(z)$; D 是一个判别网络, 判别预测模型的结果是不是“准确的”。它的输入参数是 x , x 代表准确值, 输出 $D(x)$ 代表 x 为准确值的概率, 如果为 1, 就代表预测准确性为 100%, 依次类推^[8]。

$$\min_G \max_D V(D, G) = E_{x \sim p_{\text{data}}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1)$$

其中, x 为真实值, z 为输入的训练样本, 而 $G(z)$ 表示 G 网络生成的预测模型; $D(G(z))$ 是 D 网络判断 G 生成的预测模型是否准确的概率, G 应该希望自己生成的模型“越接近准确值越好”。 G 取 $D(G(z))$ 最大值, $V(D, G)$ 变小, 在式 (1) 标记为 \min_G 。同理, D 的能力越强, $D(x)$ 越大, $D(G(x))$ 越小, $V(D, G)$ 会变大, 此时标记为 \max_D 。

2 实验部分

2.1 光谱采集

实验室中, 用 U251UV 分光光计和 NIR 光谱测试装置 (由激光器、待测样品装置、光谱仪、电脑以及相应的连接光纤组成) 完成 UV 光谱和 NIR 光谱检测, 德国 Brand (1.5 mL) 数字可调精密移液器、50 mL 比色管, 邻苯二甲酸氢钾标准液, 配制 1 000 mg·L⁻¹ 的邻苯标准液, 用蒸馏水定容至标线, 摇匀, 分别稀释成 1~500 mg·L⁻¹ 浓度备用。

采用透射光谱法, 将石英比色皿作为样品池, 空气作参

比, 采集 COD 标准液的 UV 和 NIR 吸收光谱, 如图 1 所示, UV 波段采集范围为 190~310 nm, NIR 波段范围为 830~2 100 nm (12 500~830 cm⁻¹), 分辨率为 1 nm, 积分时间为 3 ms, 每个样本重复测量 10 次, 结果取平均值。

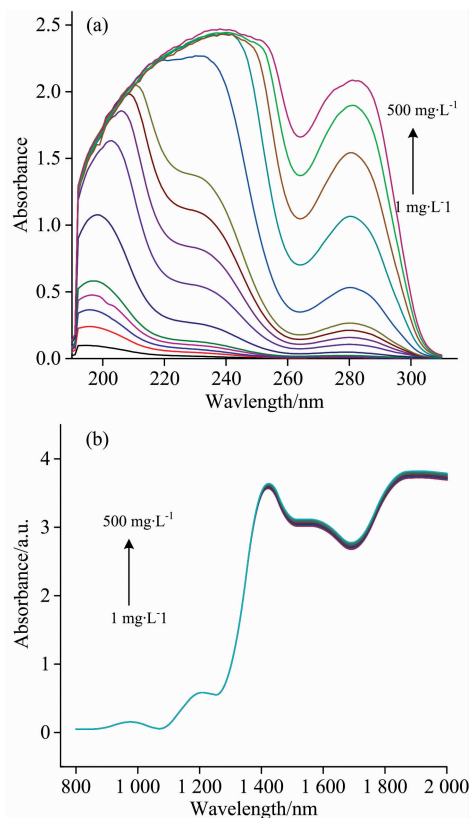


图 1 COD 水样光谱

(a): UV 光谱; (b): NIR 光谱

Fig. 1 COD water sample spectra

(a): UV spectrum; (b): NIR spectrum

由图 1 可知, 不同浓度的 COD 标准液的吸收谱在 UV 波段具有两个典型特征吸收峰。从官能团的角度来分析, 第一个峰是羟基和羧基共同作用形成, 且随着邻苯标准液浓度的升高, 该官能团吸收带出现明显红移, 并最终吸收趋于饱和; 第二个峰是苯环官能团作用形成, 且随着标准液浓度的上升, 该官能团吸收带的吸收明显增强^[9]。

水中有机物对 NIR 光的吸收很弱, 常被水的强吸收峰掩盖, 不易从原始光谱中直接观察到^[10]。因此, 在 NIR 波段建立模型之前, 需要将 NIR 吸收光谱进行处理以突出样品中污染物主要官能团的光谱特性。将吸收光谱进行一阶导数谱, 在 954, 1 286, 1 447 和 1 753 nm 附近有明显的吸收峰, 这是污染水体含有的芳香烃化合物的 C—H 伸缩振动的一、二级倍频和羰基二、三级倍频的吸收带^[10]。UV 波段与 NIR 波段相比可知, UV 波段的吸收峰的信息含量并不高, 但稳定较好; NIR 波段覆盖范围广, 反映污染物的种类也更加丰富。

2.2 光谱数据预处理

光谱数据的采集受到外界环境影响, 通过对 UV 和 NIR 原始光谱进行一定预处理, 可以有效地降低外界环境的影

响,提高最终定量预测模型的预测精度。采用一阶导数、Savitzky-Golay 平滑^[11]等方法对光谱数据进行预处理。一阶导数谱可以有效消除基线漂移、旋转以及背景干扰,然而,在放大信息的同时,噪声也被放大。为了消除噪声影响,采用 S-G 滤波,对一阶导数谱进行滤波。

将所有的水体样本按照浓度随机打乱,依照 3:1 的比例分为校正集和预测集,测量值的最大最小样本归为校正集。

3 结果与讨论

3.1 多光谱数据融合的归一化方法选择

基于 UV 和 NIR 光谱建立数据级和特征级的融合预测

模型,对经过预处理之后的 UV 和 NIR 光谱数据进行融合。

在特征级融合的过程中,将经过预处理之后的 UV 光谱数据和 NIR 光谱数据分别采用反向区间偏最小二乘法^[12]对其特征区间进行挑选,其中反向区间偏最小二乘法划分的信息区间数量为 15 个,最大因子数设为 8,在设定子区间数下,具体筛选波段结果如表 1 所示。将筛选出的特征区间吸光度-浓度数据矩阵 \mathbf{X} ,作为 GANs 模型的输入,将真实测量 COD 值作为输出值 \mathbf{Y} ,获得特征级融合 UV-NIR 的 COD 浓度预测模型。

同理,在数据级融合过程中,将经过预处理之后的 UV 区间光谱数据和 NIR 区间数据直接串联形成一个新的吸光度-浓度数据矩阵 \mathbf{X} ,获得数据级融合 UV-NIR 的 COD 浓度预测模型。

表 1 未归一化处理的数据级和特征级融合 GANs 模型

Table 1 Data fusion and feature level fusion GANs models without normalized treatment

融合方法	预处理	特征提取	筛选波段/nm	校正集			验证集		
				R^2	RMSECV	偏差	R^2	RMSEP	偏差
LLDF	Der-1S-G	—	—	0.978	2.356	0.318	0.915	3.928	0.764
MLDF	Der-1S-G	BiPLS	820~952 1 719~1 836	0.984	1.659	0.061	0.933	2.013	0.874

由表 1 可知,不论是数据级融合还是特征级融合,预测模型的精度都不够令人满意,其中数据级融合模型不论是校正集还是验证集的误差都较大,其 R^2 分别为 0.978 和 0.915, RMSECV 分别达到了 2.356 和 1.659,偏差分别为 0.318 和 0.764,这说明模型的预测精度不高,泛化能力也不能让人满意;对比数据级融合模型,虽然直接融合的特征级融合模型的 R^2 , RMSEC, RMSEP 和偏差较数据级融合有一定提高,但是模型评价参数也不够理想,不能达到精准预测的要求。

由于实验中得到的 UV 光谱(190~310 nm)和 NIR 光谱(830~2 100 nm)的数据量不均匀,数据量占有绝对优势的 NIR 光谱可能在光谱融合的过程中掩盖掉 UV 光谱的贡献,从而主导最终融合模型的结果。经过反向区间偏最小二乘法筛选过的优选区间也刚好处于 NIR 波段,波段分别为 820~952 和 1 719~1 836 nm,而 UV 波段被掩盖住,这就验证了在 UV 光谱和 NIR 光谱融合的过程中,确实存在两种光谱

对预测模型贡献度不平衡的问题,其中 NIR 的贡献度远远大于 UV 光谱的贡献度,从而导致 UV 波段被掩盖,相应的多光谱的预测模型的精度也达不到要求,让光谱融合失去了意义。为了解决这一问题,采用归一化方法处理 UV 和 NIR 光谱,并讨论不同归一化方法对建模的影响,分别采用标准归一化、最大最小归一化和矢量归一化对光谱数据进行归一化处理,将经过归一化后的 UV 和 NIR 光谱数据进行融合,并使用反向区间偏最小二乘法进行特征区间选择,分别作为 GANs 模型的输入 \mathbf{X} ,将真实测量 COD 值作为输出值 \mathbf{Y} ,建立不同归一化方法处理后的 COD 预测模型,具体如表 2 所示。

通过表 2 可知,不论是数据级融合模型还是特征级融合模型的预测精度较未归一化之前都有明显的提高,其中基于最大最小归一化进行数据处理后得到的多光谱的预测模型的性能最好,不论从相关系数 R^2 还是 RMSECV, RMSEP 和偏差,都表现出很高的预测精度。

表 2 不同归一化方法的 GAN 预测模型结果统计

Table 2 Statistics of the result of GAN prediction model using different normalization methods

特征提取方法	归一化方法	校正集			检验集		
		R^2	RMSECV	偏差	R^2	RMSECV	偏差
BiPLS	SNV	0.989	1.477	0.014	0.973	1.811	0.156
		0.989	1.467	0.055	0.973	2.012	0.259
	Min-Max-Nor	0.993	1.136	0.023	0.982	1.655	0.211
		0.994	1.069	0.031	0.985	1.814	0.185
	VN	0.987	1.653	0.063	0.935	2.928	0.851

3.2 LLDF 和 MLDF 预测模型

基于上面的研究进行数据级融合,对 UV 区间光谱数据

和 NIR 区间预处理后的数据进行归一化处理(采用 Max-Min-Nor 预处理),直接串联形成一个新的吸光度-浓度数据

矩阵 X , 并获得最终的 UV-NIR 的 COD 浓度预测模型。通过验证集实验验证模型的预测精度情况, 如图 2(a) 所示。

同理, 进行特征级数据融合, 对 UV 区间光谱数据和 NIR 区间预处理后的数据进行归一化处理(Max-Min-Nor 预处理), 通过反向区间偏最小二乘法筛选出特征区间, 并进行特征级数据(MLDF)融合, 采用 GANs 算法建立 MLDF 多光谱融合的 COD 浓度预测模型。通过验证集实验验证模型的预测精度情况, 如图 2(b) 所示。

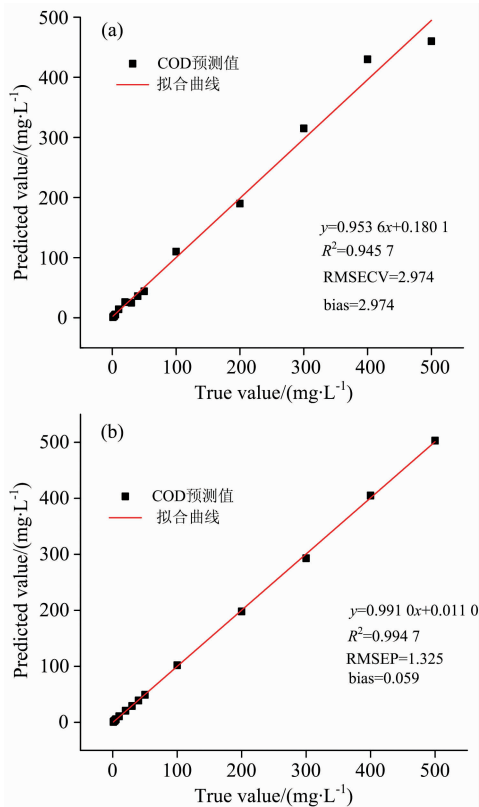


图 2 基于 GANs 算法建立 COD 预测模型

(a): 数据级融合; (b): 特征级融合

Fig. 2 COD prediction model based on GANs algorithms

(a): Data level fusion; (b): Feature level fusion

3.3 不同预测模型的评价

根据光谱类别以及融合方式的差异性, 建立不同的基于

GANs 算法的 COD 浓度预测模型, 包括单一谱源的 UV 全波段 GANs 模型、单一谱源的 NIR 全波段 GANs 模型、UV + NIR 数据级融合的 GANs 模型和 UV + NIR 特征级融合的 GANs 模型, 并进行预测性能对比, 结果如图 3 和表 3 所示。

依据模型校正集和验证集的评价参数: 决定系数 R^2 、交叉均方根误差 RMSECV, RMSEP 和加标回收率, 可以得到以下结论: 对于校正集的水质 COD 预测模型而言, NIR 吸收光谱的模型效果最差; UV 吸收光谱模型的预测效果也较差; 而基于 UV + NIR 的数据级融合(LLDF)模型和特征级融合(MLDF)模型的校正集评价指标较单一谱源的预测模型提升明显, 其中 MLDF 模型的 RMSECV 较单一谱源 UV 模型、NIR 模型和 LLDF 提高 189%, 261% 和 91%, R^2 提高 5.4%, 9.3% 和 1.1%。

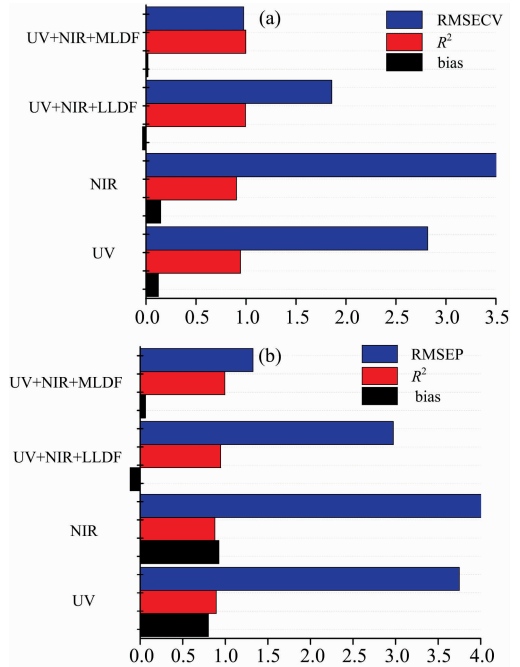


图 3 不同定量预测模型评价参数的对比图

(a): 校正集评价参数; (b): 验证集评价参数

Fig. 3 Comparison of evaluation parameters of different quantitative prediction models

(a): Correction sets; (b): Verification sets

表 3 不同定量预测模型评价参数表

Table 3 Evaluation parameters of different quantitative prediction models

最优模型分类	校正集			验证集			加标回收率
	R^2	RMSECV	偏差	R^2	RMSEP	偏差	
UV 全波段	0.944 1	2.817	0.123	0.894 1	3.749	0.801	90.2~121.6
NIR 全波段	0.904 6	3.522	0.147	0.877 4	4.110	0.925	82.1~130.4
UV+NIR(LLDF)融合	0.994 7	1.858	-0.035	0.945 7	2.974	-0.128	93.2~110.4
UV+NIR(MLDF)融合	0.997 7	0.976	0.016	0.994 7	1.325	0.059	98.4~103.1

而对于验证集而言, 单一谱源的 UV 和 NIR 预测模型的评价参数较差, 不能满足监测的精度要求; UV + NIR 的数

据级融合模型的评价参数下降、回收率不佳, 也不能满足实际的水体环境监测; UV + NIR 的特征级融合则预测效果最

佳,模型的 R^2 达到了 0.994 7、RMSEP 为 1.325,回收率达到了 98.4~103.1,都是所有模型中效果最佳的,其中 MLDF 模型的 RMSEP 较单一谱源 UV 模型、NIR 模型和 LLDF 模型提升了 183%,210%和 124%, R^2 提升了 10.1%,11.8%和 4.9%。说明在实际日常监测中,基于特征级信息融合的多光谱 COD 模型能够取得最佳的预测效果。

4 结 论

提出了一种基于多光谱融合和 GANs 算法的 COD 浓度预测模型,该模型是基于 UV 光谱和 NIR 光谱进行信息融合,结合 GANs 算法建立的非线性多光谱 COD 预测模型。

(1)基于未经归一化的 UV 和 NIR 光谱,直接搭建数据

级和特征级数据融合的 COD 浓度预测模型,模型的整体效果不理想。分析得知 UV 和 NIR 光谱的数据量不均衡,让 NIR 光谱掩盖了 UV 光谱的贡献度。

(2)采取归一化方法处理 UV 光谱和 NIR 光谱,克服光谱数据量不均衡的问题。并讨论了不同归一化方法对建立 COD 浓度预测模型精度的影响。

(3)实验验证表明该模型相关系数的平方为 0.994 7,预测均方根误差为 1.325,比数据级融合的预测模型误差降低了 52.9%,预测回收率为 98.4%~103.1%,远低于其他几组,模型的泛化能力更强,预测精度也更高。与单一谱源的监测模型相比,多光谱数据融合能反映更多的水体样品的化学信息,更加全面揭示水体的污染物程度,提高最终的 COD 预测准确率。

References

- [1] 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.
- [2] Brito R S, Pinheiro H M, Ferreira F, et al. Applied Spectroscopy, 2016, 70(3): 443.
- [3] Lepot M, Torres A, Hofer T, et al. Water Research, 2016, 101: 519.
- [4] Abedinzadeh N, Shariat M, Monavari S M, et al. Process Safety and Environmental Protection, 2018, 18(1): 82.
- [5] Maria J, Martelo Vidal, Manuel Vázquez. CyTA—Journal of Food, 2015, 13(1): 32.
- [6] ZHAO Quan-you, LI Xia, LIU Xiao, et al(赵友全,李 霞,刘 潇,等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2016, 36(11): 3592.
- [7] ZHONG Yang, XIA Feng-yi, LIAN Ji-yao(仲 洋,夏凤毅,廉继尧). Chinese Journal of Environmental Engineering(环境工程学报), 2017, 11(2): 1300.
- [8] YAO Sen, LI Tao, LIU Hong-gao, et al(姚 森,李 涛,刘鸿高,等). Food Science(食品科学), 2018, 39(8): 212.
- [9] Rischbeck P, Cardellach P, Mistele B, et al. Journal of Agronomy & Crop Science, 2017, 203(6): 483.
- [10] Barmeier G, Mistele B, Schmidhalter U. Crop and Pasture Science, 2016, 67(12): 1215.
- [11] YANG Shuo, CHEN Li-fang, SHI Yu, et al(杨 朔,陈丽芳,石 瑀,等). Journal of Computer Applications(计算机应用), 2018, 38(6): 1554.
- [12] Zhu J Y, Zheng W S, Lai J H, et al. IEEE Transactions on Information Forensics & Security, 2017, 9(3): 501.

COD Concentration Prediction Model Based on Multi-Spectral Data Fusion and GANs Algorithm

CHEN Ying¹, XU Yang-me¹, DI Yuan-jian¹, CUI Xing-ning¹, ZHANG Jie¹, ZHOU Xin-de¹, XIAO Chun-yan², LI Shao-hua³

1. Hebei Province Key Laboratory of Test/Measurement Technology and Instrument, School of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China
2. School of Resources and Environment, Henan University of Technology, Jiaozuo 454000, China
3. Hebei Sailhero Environmental Protection Hi-Tech Co., Ltd., Shijiazhuang 050000, China

Abstract Excessive concentration of organic pollutants in water is harmful, which causes not only serious environmental pollution but also endangers human health. Chemical oxygen demand (COD) can be used to characterize the pollution degree of organic pollutants in water. A quantitative prediction model of COD concentration based on generative adversarial networks (GANs) algorithm is proposed, which combines ultraviolet (UV) and Near Infrared (NIR) spectra with data-level fusion (DLDF) and feature level data fusion (FLDF). In this paper, firstly, COD standard samples are prepared according to a certain concentration gradient, and the ultraviolet spectrum (190 ~ 310 nm) and near-infrared spectrum (830 ~ 2 100 nm) of the standard sample are collected respectively. The first derivative and Savitzky-Golay (S-G) smoothing pretreatment of the obtained ultraviolet and near-infrared spectrum data are carried out to eliminate the baseline drift of the spectrum and the interference noise. Then, the data fusion of data level and featural level are carried out directly basing on the pretreated ultraviolet and near-infrared spectra, and the COD concentration prediction model is constructed by GANs algorithm. The model is evaluated by using the square of the correlation coefficient of the evaluation parameters (R^2), the mean square root error of the predicted value and the real concentration value (RMSEP) and the prediction deviation. The results show that neither FLDF model nor DLDF model is not ideal. The analysis shows that the model contribution of the ultraviolet spectrum is concealed in the near-infrared band due to the unbalanced data in the ultraviolet and near-infrared bands, which makes the spectral fusion meaningless. In order to avoid the problem of fusion failure, the normalization method is proposed to deal with the mixed spectrum in the text. The effects of standard normal variation (SNV), maximum and minimum normalization (MMN) and vector normalization (VN) on the modeling are discussed. Then the normalized ultraviolet and near-infrared spectral data are fused again under the given sub-interval number, the input X of GAN model is taken as the input X , and the real measured COD value is taken as the output Y . The prediction models of COD concentration are established after different normalization methods. The modeling results show that different normalization methods have a great influence on the hybrid spectral data fusion model, and the prediction accuracy of the data-level fusion model and the feature-level fusion model is significantly improved before it is normalized, among which the prediction model with the maximum and minimum normalization is the most obvious. Finally, in order to verify the accuracy of the multi-spectral data fusion GANs Prediction model, the GANs prediction model of the full wavelength ultraviolet band of a single spectral source and the GANs prediction model of the full wavelength near-infrared band of a single spectral source are established. The experimental results show that the correlation coefficient of the characteristic level spectral fusion model basing on the ultraviolet and near-infrared spectra is 0.994 7, the prediction mean square root error is 0.976, the prediction model error is reduced by 52.9% comparing with the data level fusion, and the predicted recovery rate is 98.4%~103.1%, which is much better than the other groups. The generalization ability of the model is strong and the prediction accuracy is high. Compared with the monitoring model of single spectral source, the data fusion of mixed spectra can reflect more the chemical information of water samples, and reveals the pollutant degree of a water body more comprehensively, reflects the difference of pollutants in a water body from different levels, provides some technical support for on-line monitoring of COD concentration in water.

Keywords UV spectrum; NIR spectrum; Data fusion; GANs model; COD concentration prediction

(Received Dec. 27, 2018; accepted Feb. 9, 2020)