

基于监督学习的紫外-可见光光谱水质在线异常检测方法研究

尹航, 俞巧君, 侯迪波*, 黄平捷, 张光新, 张宏建

浙江大学控制科学与工程学院, 工业控制技术国家重点实验室, 浙江 杭州 310027

摘要 水资源关系到国计民生, 近年来时有发生的水污染事件使污染物入侵预警得到了广泛的社会关注。针对现有基于紫外-可见光光谱的水质异常检测方法存在的检出下限偏低的问题, 提出一种基于监督学习的紫外-可见光光谱水质异常检测方法。该方法首先获取不同数据集中的正常样本差异性空间, 再使用正交投影方法去除差异性空间中的光谱数据分量, 以达到基线校正的目的; 然后采用偏最小二乘判别分析从校正后的光谱中提取特征, 利用训练集得到的最优阈值确定离群点; 最后采用序贯贝叶斯滚动更新每个时刻上的异常概率, 确定水质报警序列。实验选用苯酚作为模拟污染入侵事件的注入试剂, 采样2周内的紫外-可见光光谱数据, 在实验平台上对提出的方法进行了验证。实验结果表明, 采用的正交投影基线校正方法可以消除不同批次水质光谱的背景差异, 更为充分的利用了光谱信息, 降低了对特征污染物的检出下限。

关键词 紫外-可见光光谱; 水质异常检测; 正交投影; 监督学习; 偏最小二乘判别分析

中图分类号: O657.3 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2019)02-0491-09

引言

水环境安全和饮用水安全关系国计民生, 近年来得到了广泛的社会关注。然而水污染事件仍时有发生, 2005年由于石化企业爆炸事故, 苯类污染物流入松花江造成水质污染; 2011年新安江一辆载有苯酚的槽罐车泄漏, 造成取水水源地污染, 影响下游地区用水安全; 2014年某石化企业管道发生原油泄漏, 污染了供水企业自流沟, 导致自来水苯含量超标。这些突发性的水污染事件对水生态系统造成破坏。

目前污染物入侵事件的预警分析主要基于在线常规水质指标进行异常检测, 采用的方法主要包括: 基于统计学的方法、基于机器学习的方法和基于事件序列的方法。Byer等在早期研究中, 提出一种基于统计学3倍标准差阈值来检测常规单指标水质异常的方法; Yang等通过11种生化污染物实验改良了常规参数传感器的检测阈值, 并研究了多种指标在异常检测方面的关联; 何慧梅等^[1]提出基于自回归和模糊C-均值聚类的多因子融合水质异常检测算法, 显著提高了多指标证据冲突情况下的检测效果; Liu^[2]等结合水质指标间的皮尔逊系数和多维欧式距离进行了异常检测研究, 并且对微弱信号检测效果良好。

基于常规水质指标的水质异常检测方法可以有效检出表征明显的污染物入侵事件, 但其检测过程耗时长、系统投入成本高、易造成二次污染。紫外-可见光光谱检测法具有扫描时间短、响应速度快、免试剂、现场原位等优点, 可作为利用常规指标检测污染物入侵事件的补充。近年来, 有学者基于紫外-可见光光谱法开展了对供水系统中突发污染事件检测方法的研究。一种常见方法是通过紫外-可见光光谱的特征单波长或多个波长时间序列来检测水质异常, 如UV₂₅₄方法等, 但这类方法未能有效利用光谱中的全部有效信息; Langergraber基于统计模型建立了紫外-可见光全光谱与物理指标COD和TSS的关系模型, 通过分析物理指标在时序上的数值变化进行异常检测, 构建了可商用化的污染入侵事件预警系统; Hou^[3-4]等面向未知污染物异常检测问题, 基于多元回归统计模型开展UV/VIS光谱的异常检测研究, 结合Chi-square分布构建统计量确定正常水质的波动范围, 超出波动范围则认为出现异常; 郭冰冰^[5]等提出一种基于主元分析和非对称最小二乘基线校正的方法, 先提取正常水质的特征并利用Q统计量识别正常水质波动外的异常水质, 明显改进了污染物入侵事件的检测性能。

现有基于紫外-可见光光谱法的水质事件检测方法, 虽然可以对当前检测波段内有响应的未知污染物进行异常报

收稿日期: 2017-11-21, 修订日期: 2018-04-19

基金项目: 国家自然科学基金项目(61573313, U1509208), 国家重点研发计划项目(2017YFC1403801), 浙江省重点研发计划项目(2015C03014)资助

作者简介: 尹航, 1993年生, 浙江大学控制科学与工程学院硕士研究生 e-mail: yinhang@zju.edu.cn

* 通讯联系人 e-mail: houdb@zju.edu.cn

警,但在突发污染事件发生初期、污染物浓度偏低时,现有方法仍存在检出下限较高、报警不及时的问题。由于突发性污染事件的特殊性,在上述提及的某些突发事件场景下,污染物的类型可提前通知到污染下游的城市用水地区。因此如果有方法能够充分利用特征污染物的先验知识,将为下游地区的污染物入侵事件进行更有针对性的预警。

监督学习方法是一类利用先验知识完善建模的有效手段。监督学习方法可以从训练资料集中建立一个模型,根据模型预测新出现的输入值的输出。本文拟利用监督学习方法对入侵污染物的全光谱特征信息进行学习,提出一种基于偏最小二乘判别分析(partial least squares discrimination analysis, PLS-DA)的紫外-可见光光谱水质异常检测方法,并结合序贯贝叶斯(Sequential Bayesian)方法滚动更新异常概率检测污染物入侵事件,从而达到利用已知污染物特征信息提升事件检出性能的目的,为及早报告污染物入侵事件提供可能。

1 实验部分

本文提出方法的基本思想是:首先基于训练数据利用 PLS-DA 对特征污染物进行学习建模。在检测开始后,参考训练集正常水质数据,利用正交投影基线校正方法对新采集到的数据进行基线校正和噪声去除,并输入到 PLS-DA 模型获得报警信号(alarm index),再通过阈值判别离群点(outliers)。针对离群点中的单点误报漏报,利用序贯贝叶斯方法滚动更新异常概率,获得连续的水质报警序列。具体方法如图 1 所示。

1.1 基于正交投影校正的预处理方法

设 i 时刻观测的光谱数据用 $X_{i,obs}, i=1, \dots, N$ 表征, N 为观测时长。可以假设 $X_{i,obs}$ 由两部分叠加组成,纯净的光谱 X_i , 以及很难估计的、由噪声及基线漂移等为主引起的光谱差异部分 h_i ^[6]。如果令空间 ϵ^D 表征光谱的差异性,则可认为 ϵ^D 是由 $\{h_i\}, i=1, \dots, N$ 张成的。根据正交投影定理,校正后的光谱 $X_{i,corr}$ 可以被式(1)表征,其中,矩阵 P 为空间 ϵ^D 的一组基

$$X_{i,corr} = X_{i,obs} P^{\perp} = X_{i,obs} [I - P^T (PP^T)^{-1} P] \quad (1)$$

可以证明,由 $\{h_i\}, i=1, \dots, N$ 张成的空间 ϵ^D 可通过如下方法估计。首先,随机在训练集中抽取一定数量的正常光谱,由于可以保证检测开始时水质为正常,故可采集相同数量的正常水质光谱,依时序保存至测试集。将两数据集中正常光谱数据合并成为一个数据表,在数据表上做主成分分析(principal component analysis, PCA),提取到的光谱主成分代表了上述的两数据集光谱的差异性,故主成分分析得到的荷载矩阵 \hat{P} 可作为 P 的估计。则由 PCA 求解过程中的约束条件可知,式(1)可简化为式(2)

$$X_{i,corr} = X_{i,obs} (I - \hat{P}^T \hat{P}) \quad (2)$$

在实际应用中,因为正常水质本身是具有一定波动差异的,剔除的差异越多越容易改变光谱中与污染物相关的信息,不可以将光谱所有的主成分去掉。所以利用正交投影校正的方法,需要确定一个自然数参数 p , 只取主成分分析中

贡献率大的前 p 维矩阵来代替上述方法中的 \hat{P} 。

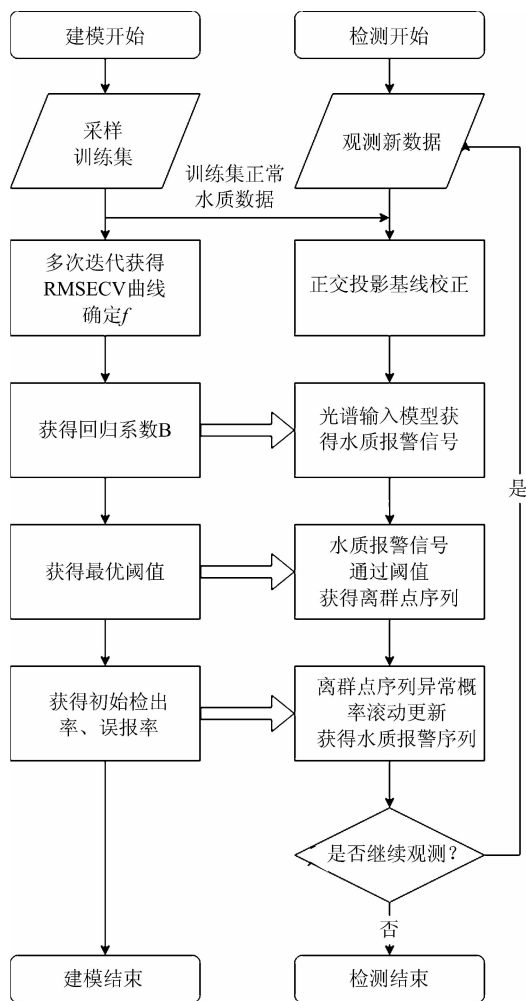


图 1 建模及检测流程图

Fig. 1 Modelling and testing workflow chart

1.2 基于 PLS-DA 的水质离群点检测方法

偏最小二乘法(partial least squares, PLS)回归方法是化学计量学与光谱分析中的常用方法,具体计算步骤和使用方法详见参考文献[7-8],这里仅说明 PLS 方法中的一个重要参数——成分选取个数 f 的确定。在建模过程中, f 选取过大容易过拟合,选取过小泛化能力不足^[8],本文通过 K 折交叉验证法选取 f 。每次 PLS 迭代后,可以获得一个新成分,并更新了反映光谱特征的回归系数向量 B 。记录不同回归系数向量 B 时测试所得 RMSECV 的值,作 RMSECV 随 f 增加时的曲线。利用 F 检验方法,自动判断 RMSECV 首次下降不显著时,取该处对应的 f 作为模型参数。

PLS 方法本质上是一种基于特征变量的回归方法。若将其中的因变量变为哑变量,由 0 或 1 进行替代,则可为提取光谱信息并进行分类的方法,通常称为 PLS-DA^[9]。PLS-DA 在本文建模过程中 0 作为正常时刻,1 作为异常时刻,利用上述迭代方法建模。得到模型后,根据训练数据 ROC 曲线确定最优阈值 $\text{Threshold}_{\text{best}}$ 。设 PLS-DA 模型用 $F_f(\cdot)$ 表示,在 i 时刻输出的报警信号用 y_i 表示,即 $y_i = F_f(X_{i,obs})$ 。

通过 y_i 与 $\text{Threshold}_{\text{best}}$ 对比, 得到最终的离群点序列 Y_i 。对于 i 时刻输入到模型的光谱 $X_{i,\text{obs}}, i=1, \dots, N$, 其离群点判别方法由式(3)描述

$$Y_i = \begin{cases} \text{Outliers,} & y_i = F_f(X_{i,\text{corr}}) > \text{Threshold}_{\text{best}} \\ \text{Normal,} & y_i = F_f(X_{i,\text{corr}}) \leq \text{Threshold}_{\text{best}} \end{cases} \quad (3)$$

1.3 基于序贯贝叶斯的异常事件判定

由于原始数据的噪声干扰和其他因素的影响, 在去掉基线漂移之后, 仍然会存在单个的离群点。一般而言, 污染物入侵事件会持续多个检测周期, 故连续多个离群点有更大概率为水质异常。利用序贯贝叶斯更新概率的方法, 可将连续多个水质异常点检出为事件。在时序分析中, 当前点是否是异常点的概率是由历史的观测点决定的。序贯贝叶斯方法得到的 $t+1$ 时刻的状态由贝叶斯公式(4)决定

$$\pi_1(t+1) = \frac{\pi_1(t)P(y_{t+1} | \theta_1)}{(1 - \pi_1(t))P(y_{t+1} | \theta_0) + \pi_1(t)P(y_{t+1} | \theta_1)} \quad (4)$$

式(4)中, $\pi_1(t)$ 为 t 时刻为异常的概率, 初始时刻可认为发生概率较小, 故 $\pi_1(0)$ 为一个很小的值, 实验中取 10^{-5} ; 通过式(4)得到一个概率来表征为异常的可能性, 传递给下一个时刻; 当概率 $\pi_1(t+1)$ 超过 0.95 时认为当前是异常点, 设定为报警, 低于 0.95 时认为是正常情况, 设定为不报警。式(4)中 θ_1 为异常, θ_0 为正常, $P(y_{t+1} | \theta)$ 可根据历史数据或训练数据的最佳阈值点位置处的检出率、误报率进行估计^[10]。

1.4 ROC 曲线与性能评价指标

受试者工作特征曲线^[11] (receiver operating characteristic curve, ROC curve) 是以真阳性率(检出率)为纵坐标, 假阳性率(误报率)为横坐标绘制的, 直观判断检测模型性能的曲线^[12]。在同一光谱数据集中测试, 利用 ROC 曲线下的面积 (area under receiver operating characteristic, AUROC) 对两种或两种以上检测模型进行性能比较, 一般 $0.5 < \text{AUROC} < 1$, 且 AUROC 越大, 则分类器效果越好; 在同一种水质检测模型中, 一般认为检出率越高、误报率越低, 检测结果越好, 故使用 ROC 曲线上最靠近左上角的点处所代表的阈值是一个平衡点, 可认为是该模型的最优阈值 $\text{Threshold}_{\text{best}}$ 。

本文使用的性能指标包括如下三种: 查准率、查全率及 F-Score^[12]。查准率 (precision) 是正确预报的异常光谱样本数占预测为异常的样本总数的比率; 查全率 (recall) 是正确预报的异常光谱样本数占样本中异常点总数的比率。一般希望二分类算法的查准率和查全率都能接近 1, 但查准率和查全率两个指标会出现此消彼长的矛盾情况, 因此需要综合考虑, 最常见的方法是使用 F-Score 参数, 即上述两参数的加权调和平均值, 如式(5)

$$F_\alpha = \frac{(\alpha^2 + 1)\text{Precision} \times \text{Recall}}{\alpha^2(\text{Precision} + \text{Recall})} \quad (5)$$

当 $\alpha=1$ 时, 即为 F_1 参数。 F_1 参数越接近 1, 性能越好。

1.5 实验平台简介及参数描述

城市饮用水管网检测实验装置是由接入实际管网系统的测量主管道以及光谱传感器组成, 如图 2 所示。实验装置中的开环调节流量为 $600 \text{ L} \cdot \text{h}^{-1}$, 连接实际管网。所配的污染物标准浓度溶液通过脉冲泵及橡胶管注入到静态混合管道,

注入的流速通过一个闭环反馈系统控制脉冲泵, 根据主管道流量进行微调, 保证加药流量稳定在 $0.2 \text{ L} \cdot \text{min}^{-1}$ (注入污染物的流量与主管道流量之比为 1:50)。静态混合管道将主管道流量与污染物溶液进行充分混合, 以确保污染物入侵事件在管道横截面方向的浓度分布均匀。

主管道外接下导管, 连接浸入式 10 mm 光程的 UV/VIS 光谱浸入式传感器 (S: CAN analysis : lyser)。污染物从注入入口到下导管的水力学时间大概是 11.67 s。光谱仪每隔 30 s 进行一次测量, 其中 20 s 为测量时间, 10 s 为延迟时间。光谱仪选取的波长范围为 240~400 nm, 分辨率为 2.5 nm。光谱仪监测数据可通过 USB 接口导出分析。

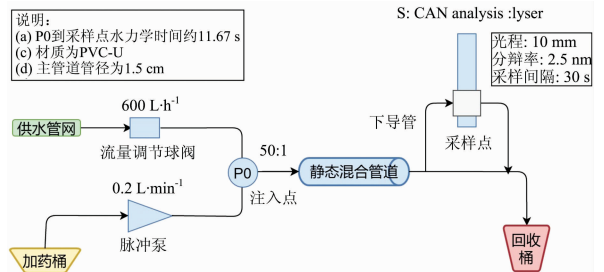


图 2 城市饮用水管网检测装置结构图

Fig. 2 Detection device for urban water distribution system

1.6 数据采集

本文选用苯酚 (Phenol) 作为注入试剂。苯酚等芳香酚类物质是 2012 年国家饮用水健康标准中列出的重要污染物, 可溶于水, 不慎摄入一定量会出现急性中毒症状, 饮用含有苯酚的水会抑制中枢神经系统, 损害肝、肾功。它们被广泛使用在化工生产中, 近年来苯酚等芳香酚类物质在管网污染事浸中曾被检出。因此, 对于管网水中出现的苯酚, 需做出快速、及时、有效的反应, 并积极采取措施减小事故危害。

本文数据采集历时 2 周, 包括正常和异常样本。每次加药持续 10 min 模拟污染物入侵事件, 每个浓度的污染物入侵事件持续 20 个采样周期。训练集的浓度由 3 个不同浓度梯度组成 ($50, 100$ 和 $200 \mu\text{g} \cdot \text{L}^{-1}$)。为了进一步测试检测方法的效果, 测试集设计为 5 个浓度梯度 ($30, 40, 50, 100$ 和 $200 \mu\text{g} \cdot \text{L}^{-1}$)。两个批次的水质数据在不同的时间段采集。

2 结果与讨论

2.1 水质基线校正

本文采用正交投影基线校正方法, 对采集到测试集光谱数据进行校正, 以减小用于建模的训练数据与新采集的测试数据之间的基线漂移。用训练集随机选取 100 个正常水质的数据、与依时序采集的前 100 个测试集光谱数据提取正交投影向量 \hat{P} 。由于两个批次的数据在不同时间段采集, 受外部因素影响, 实验的训练集谱线较为分散, 而依时序取样的测试集谱线较为集中, 如图 3(a) 所示。经正交投影基线方法校正之后的光谱如图 3(b) 所示 (取正交投影校正的参数 $p=2$), 可见减小了两个数据集之间的基线漂移。图 3(b) 中校正后的

曲线污染物峰的位置更加明显，在 270 nm 特征峰处可见测试集、训练集光谱数据分层，校正方法对有效信号进行了放大。将(b)图测试集的数据去均值，可在时序上见 5 次事件，如图 3(c)所示。

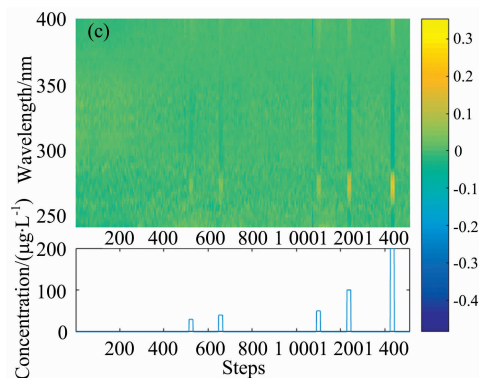
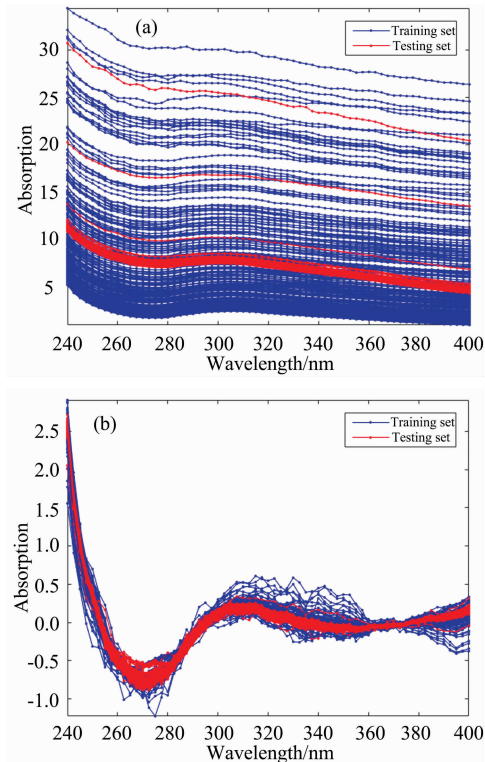


图 3 苯酚紫外可见光谱 (特征波长 270 nm)

(a): 训练集和测试集原始数据; (b): 训练集和测试集校正后;
(c): 测试集校正后去均值的时序展示

Fig. 3 UV/VIS Spectra of Phenol (Characteristic wavelength: 270 nm)

(a): Original data of training and testing set; (b): After correction;
(c): Corrected and normalized testing set

在正交投影校正中， p 参数的选择是关键。如果 p 选择过大，那么被去掉的成分越多，越可能将含有污染物信息的成分去掉；如果 p 参数选择过小，那么可能会导致基线校正不足。本文选择 p 参数秉承两个原则，其一是消除不同批次水质光谱的背景差异，其二是尽可能多的保持原光谱的信息，由于不同数据间的差异形式和种类不同， p 值会有所不同。针对本文实验，一种经验的选取参数 p 的方法是：令 p 从 0 开始自增，观察被去掉 p 维主成分后的正常水质光谱主

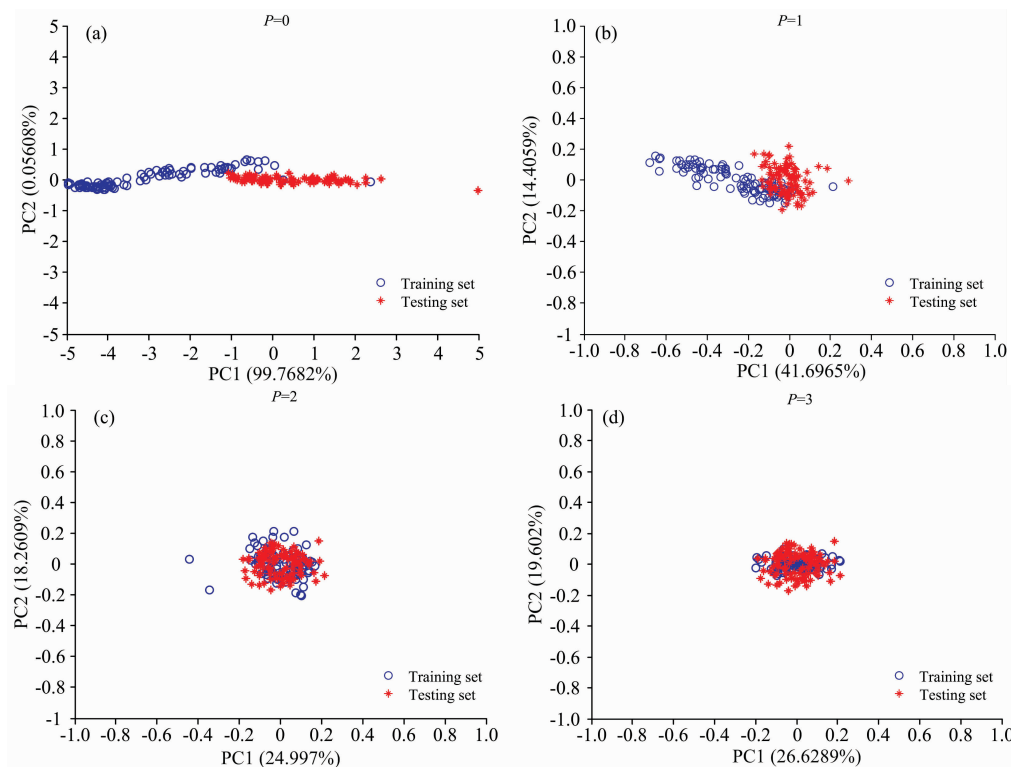


图 4 不同参数 p 下的正常光谱主成分 PC1, PC2 (p 从 0 变化到 3)

Fig. 4 First and second principle component among different p in normal spectrum (p Ranged from 0 to 3)

成分。当去掉 p 维主成分后的两个数据集中正常水质光谱主成分刚好接近重合时,选择此时的 p 。图 4 显示了 p 从 0 自增到 3 时,两个数据集随机挑选的正常水质光谱的前二维主成分。可见 $p=2$ 时已经重合,故选取的最佳 p 值不应大于 2。

2.2 异常检测结果分析

2.2.1 无监督异常检测方法

在无监督异常检测的研究中,利用 PCA 提取新采集的光谱数据主成分,通过统计量查看报警信号是否超过阈值,

超过阈值则被判为离群点。图 5 所示为基于无监督的水质异常检测效果图。图 5(a)所示事件强度分别为 30, 40, 50, 100 和 $200 \mu\text{g} \cdot \text{L}^{-1}$ 的苯酚污染事件。利用 PCA 提取每条谱线的主成分,根据主成分分析原理,对按列服从高斯分布的原始矩阵进行主成分变换,则可以根据 Chi-square 分布构建统计量^[4]进行统计检验,若统计量超过阈值,则认为当前监测数据为离群点。阈值 T_α 需要根据历史数据及经验确定。图 5(b)为计算的统计量指标作为报警信号,统计量指标通过与阈值 T_α 对比,得到离群点序列如图 5(c)。

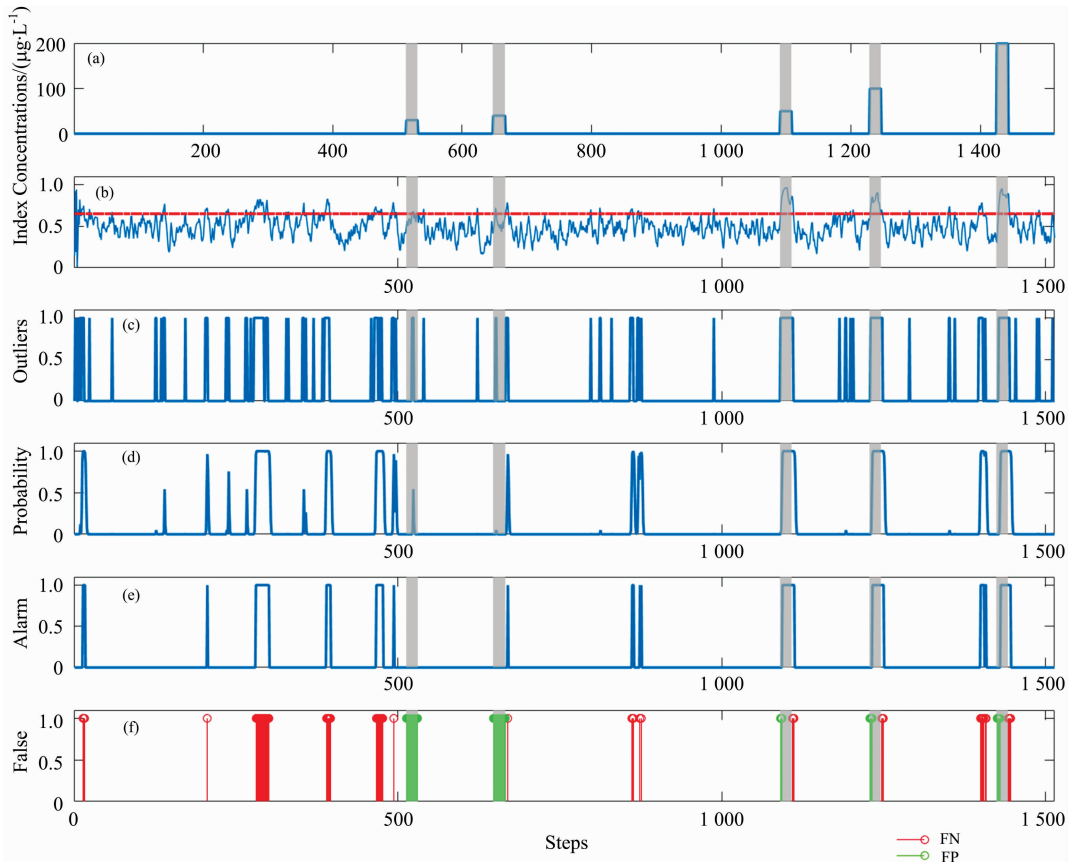


图 5 非监督方法检测效果

(a): 入侵事件实际浓度; (b): 统计量作为报警信号, 阈值 $T_\alpha=0.65$; (c): 离群点;

(d): 异常概率; (e): 水质报警序列; (f): 与真实情况对比

Fig. 5 Detecting results of unsupervised method

(a): Concentrations of intrusion events; (b): Statistic as alarm signal $T_\alpha=0.65$; (c): Outliers;

(d): Anomaly probability; (e): Alarm sequences; (f): Compared results with ground truth

由于背景水质波动、工况变化或噪声干扰等情况也会产生离群点,故对离群点序列采用了序贯贝叶斯方法,进行异常概率更新,得出该离群点是否属于水质报警序列的概率。选取滑动窗长为 200,考虑当前时刻以前 200 个点的数据,在滑动窗内根据序贯贝叶斯公式,可计算得到当前时刻的异常概率为图 5(d)所示,其中的检出率、误报率参数,是根据历史数据通过经验得到的。根据小概率事件原理,设置概率阈值为 0.95,当超过该阈值时认为污染入侵事件一定发生,将结果记录到水质预警序列,如图 5(e)所示。图 5(f)为报警序列图 5(e)与实际情况图 5(a)的对比,得到的误报漏报

结果。

表 1 无监督方法检出效果

Table 1 Detecting results based on unsupervised method

检测效果指标	Precision	Recall	F_1
非监督方法	0.45	0.37	0.40

综上所述,在当前 $T_\alpha=0.65$ 情况下,前两个浓度较低的 30 和 $40 \mu\text{g} \cdot \text{L}^{-1}$ 事件漏报。 $50 \mu\text{g} \cdot \text{L}^{-1}$ 以上浓度的污染,在水质基线稳定的情况下能够检出。如果减小阈值 T_α ,会增

加误报 FP, 增加阈值 T_o , 会增加漏报 FN, 很难使用现有方法在低浓度事件检出时获得一个高检出低误报的效果。表 1 统计了在该数据集下的检测效果指标。

2.2.2 监督学习异常检测模型建立

基于 2.2.1 提出的异常检测研究思路, 针对训练集的先验信息增加有监督学习环节。利用本文提出的基于 PLS-DA 监督学习异常检测方法, 按图 1 中的建模流程, 首先利用全部训练集数据进行建模。建模基于 $K=5$ 的交叉验证, 通过多次迭代, 获取不同 f 下的 RMSECV。如图 6 所示, 为迭代 25 次 RMSECV 随 f 的变化, 由 F 检验得出 f 变化显著的位置为 $f=3$ 。

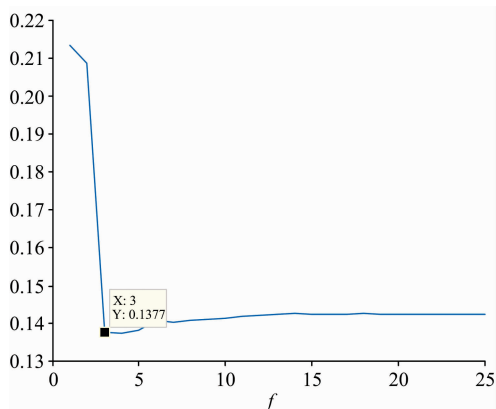


图 6 RMSECV 变化 (f 从 1 到 25)

Fig. 6 RMSECV curve changes (parameter f ranged from 1 to 25)

选取 $f=3$, 可得 PLS-DA 的回归系数向量 B , 向量长度同光谱的波数点一致, 如图 7 所示。通过观察可见, 在苯酚特征波长 270 nm 附近出现了较大的数值。由于使用 PLS-DA 方法在提取自变量矩阵和因变量矩阵主成分的同时, 能够使提取到的自变量主成分对因变量主成分有很强的解释能力, 所以在光谱数据分层的波长上回归系数的数值会偏大。由此可见, 选用 PLS-DA 有监督方法对光谱进行特征提取, 具有一定的物理解释。

获取系数 B 之后, 可以得到训练集的测试结果。在表 2 列了训练集的 AUROC 结果。可见结果接近于 1, 说明该模型对自身的解释能力较强。通过训练集的 ROC 结果, 可以

选取最接近左上角的阈值, 作为模型的最优阈值 $Threshold_{best}$ 。

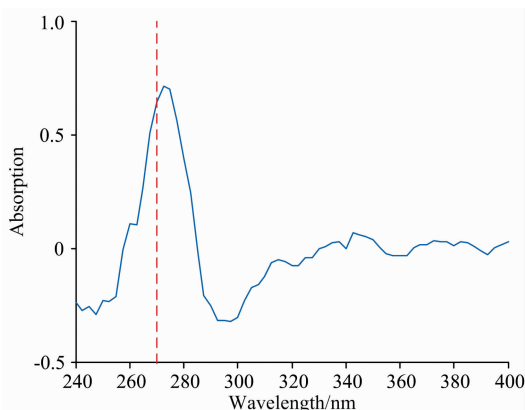


图 7 PLS-DA 回归系数 (标出苯酚特征波长 270 nm)

Fig. 7 PLS-DA coefficient of regression (labelled the characteristic wavelength 270 nm)

表 2 PLS-DA 建模参数

Table 2 PLS-DA modelling parameters

污染物	f	RMSECV	训练集 AUROC	$Threshold_{best}$
苯酚	3	0.137 7	0.985 3	0.216 8

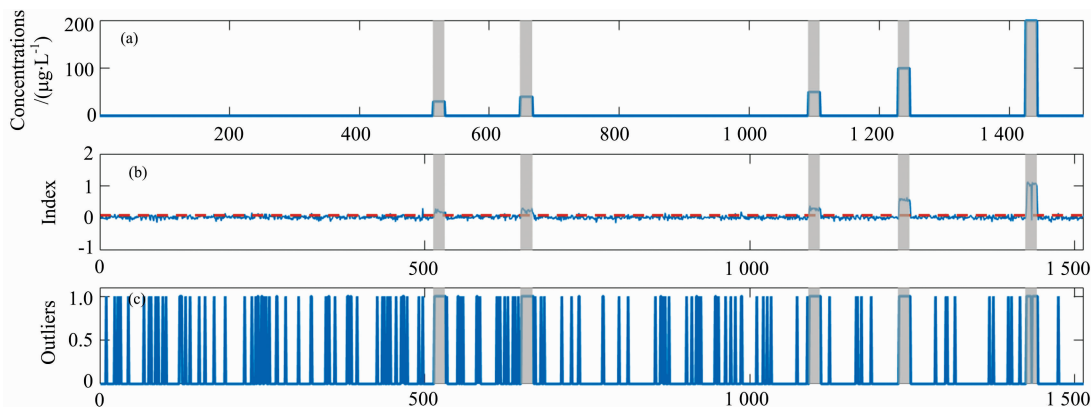
2.2.3 监督学习异常检测方法

针对本文提出的方法, 按图 1 中的检测流程, 将与 2.2.1 中同样的数据集作为观测所得的测试集, 利用正交投影校正方法进行基线校正 (取 $p=2$), 输入到 2.2.2 得到的 PLS-DA 模型中, 输出报警信号, 如图 8(b) 所示, 虚线部分为 $Threshold_{best}$, 通过 $Threshold_{best}$ 及式 (3) 判断报警信号。得到的水质离群点如图 8(c) 所示。同样利用了 2.2.1 中用于滚动更新异常概率的序贯贝叶斯方法, 取滑动窗长为 200, 得到每个时刻离群点为异常的概率和最终的水质报警序列, 分别如图 8(d) 和图 8(e)。对检测效果指标进行统计, 结果如表 3。

表 3 有监督方法检出效果

Table 3 Detecting results based on supervised method

检测效果指标	Precision	Recall	F_1
有监督方法	0.78	0.76	0.77



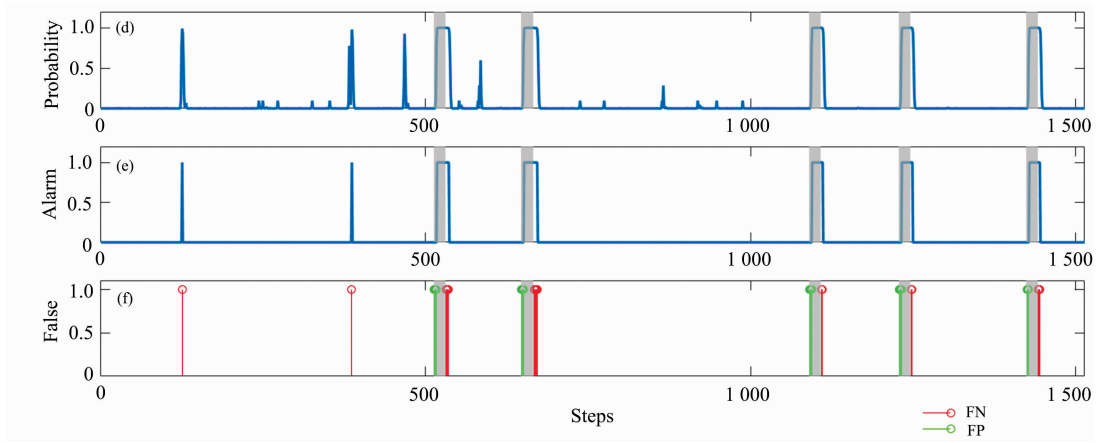


图 8 监督学习方法检测效果

(a): 入侵事件浓度; (b): 报警信号, 阈值为 0.216 8; (c): 离群点; (d): 异常概率; (e): 水质报警序列; (f): 与真实情况对比

Fig. 8 Detecting results of unsupervised method

(a): Concentrations of intrusion events; (b): Alarm signal when threshold=0.216 8; (c): Outliers;

(d): Anomaly probability; (e): Alarm sequences; (f): Compared results with ground truth

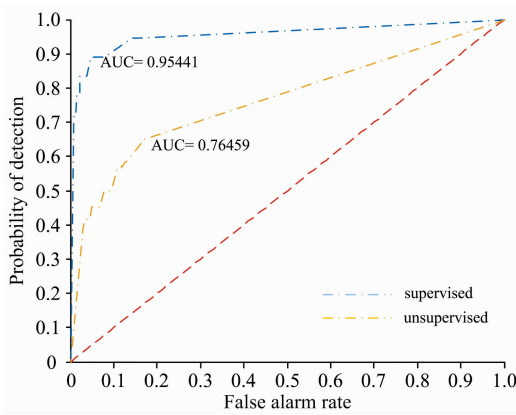


图 9 有监督 PLS-DA 方法, 和无监督的方法, ROC 比较及 AUROC

Fig. 9 ROC curve and value of AUROC from PLS-DA method and unsupervised method

根据测试数据集上的结果, 得到 ROC 曲线如图 9。得到基于 PLS-DA 的监督学习方法和基于 PCA 的非监督方法, AUROC 分别为 0.95 和 0.76。

由表 1、表 3 及图 5、图 8、图 9 所得结果比较可知, 有监督方法不仅可有效检出浓度较低的 30 和 40 $\mu\text{g} \cdot \text{L}^{-1}$ 事件, 也可以检出无监督方法有效的高浓度事件。在当前含有低浓度情况的数据集下, 有监督方法的查准率、查全率、 F_1 以及 AUROC 均优于无监督方法。由此可见, 有监督的 PLS-DA 方法能够在具有一定物理意义的情况下对预处理后的特征污染物光谱进行针对性的特征提取, 能够利用特征污染物的先验信息降低检出限。在已知待测污染物类型的情况下, 有监督方法可作为非监督的方法的补充, 提升了在低浓度情况下的水质异常事件检测效果。

2. 2. 4 关于正交投影校正预处理对检出结果的影响

考虑实际使用情况中, 训练数据和测试数据之间的基线漂移现象, 2. 2. 3 中采用正交投影方法校正基线。为了验证

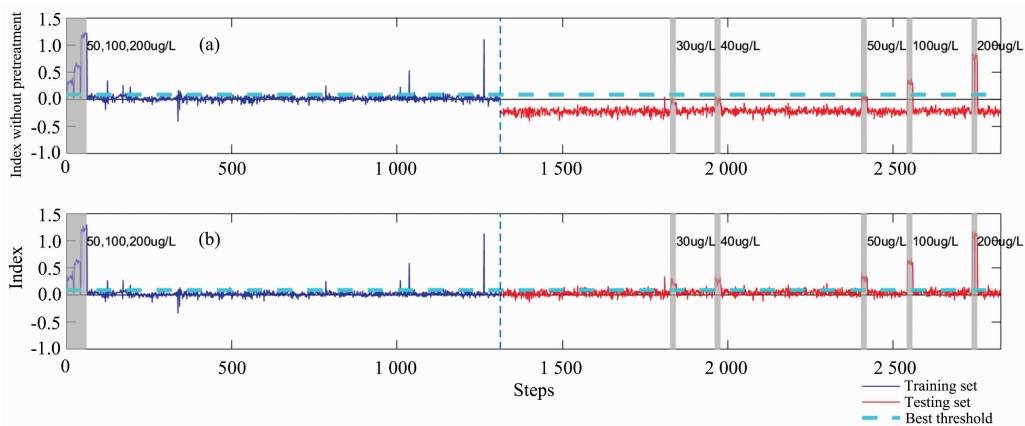


图 10 基线校正对训练集(左)和测试集(右)报警信号的影响

(a): 未基线校正处理的报警信号, $p=0$; (b): 基线校正后的报警信号, $p=2$

Fig. 10 Baseline correction impact on training set (left) and testing set (right)

(a): Alarm signal without baseline correction, $p=0$; (b): Alarm signal with baseline correction, $p=2$

正交投影方法可以有效校正基线,提高监督学习方法对水质基线变化的适应性,在训练集(左侧蓝色)和测试集(右侧红色)上的未经基线校正处理($p=0$)的报警信号展示如图 10(a)。由于水质基线的漂移,可以看出未经基线校正得到的 $\text{Threshold}_{\text{best}}$ 测试集报警信号,会发生一定程度的漂移,从而使训练集中得到的在测试集中失效,当前算法中无法有效检出测试集的低浓度事件。作为对比,图 10(b)补充了图 8(b)中未显示的训练集报警信号,可见经过基线校正后的测试集报警曲线可用最优阈值判定低浓度情况下的异常。

酚特征污染物的异常检测。针对不同批次的光谱数据,采用正交投影空间校正研究了如何消除不同批次数据间的基线漂移;对比了无监督 PCA 与有监督的 PLS-DA 特征提取方法的检出效果;结合序贯贝叶斯方法,对异常概率进行了更新,消除了大量不属于异常的水质离群点。实验结果表明,与无监督方法相比,本文提出的方法能有效降低特征污染物的检出下限,更为充分利用了光谱信息,可作为无监督方法的补充,进一步提升污染物入侵事件检出效果。

3 结 论

采用基于监督学习方法的 UV/VIS 光谱进行了针对苯

References

- [1] HE Hui-mei, HOU Di-bo, ZHAO Hai-feng(何慧梅, 侯迪波, 赵海峰). Journal of Zhejiang University(浙江大学学报·工学版), 2013, (4): 735.
- [2] Liu S, Smith K, Che H. Water Research, 2015, 80: 109.
- [3] Hou D, Liu S, Zhang J, et al. Journal of Spectroscopy, 2014, 2014: 150636.
- [4] Hou D, Zhang J, Yang Z, et al. Optics Express, 2015, 23(13): 17487.
- [5] GUO Bing-bing, HOU Di-bo(郭冰冰, 侯迪波). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2017, 37(5): 1460.
- [6] Jean-Claude Boulet, Jean-Michel Roger, Chemometrics and Intelligent Laboratory Systems, 2012, 117: 61.
- [7] Tormod Nzs, Tomas Isaksson, Tom Fearn, et al. A Userfriendly Guide to Multivariate Calibration and Classification. NIR Publication, 2004, 27.
- [8] WANG Hui-wen(王惠文). Partial Least Squares Regression Method and Application(偏最小二乘回归方法及其应用). Beijing: National Defense Industry Press(北京: 国防工业出版社), 1999. 140.
- [9] Brereton R G, Lloyd G R. Journal of Chemometrics, 2014, 28: 213.
- [10] Perelman L, Arad J, Housh M, et al. Environmental Science & Technology, 2012, 46(15): 8212.
- [11] CHU Xiao-li(褚小立). Molecular Spectroscopy Analytical Technology Combined with Chemometrics and Its Applications(化学计量学方法与分子光谱分析技术). Beijing: Chemical Industry Press(北京: 化学工业出版社), 2011. 106.
- [12] ZHOU Zhi-hua(周志华). Machine Learning(机器学习). Beijing: Tsinghua University Press(北京: 清华大学出版社), 2016. 28.

In-Situ Detection of Water Quality Anomaly with UV/Vis Spectrum Based on Supervised Learning

YIN Hang, YU Qiao-jun, HOU Di-bo*, HUANG Ping-jie, ZHANG Guang-xin, ZHANG Hong-jian

State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China

Abstract Water resources are related to national economy and people's livelihood, detection of water quality anomaly has attracted more attention because of the water pollution events happened in recent years. In this paper, the detection method of water quality anomaly with UV-Vis Spectrum based on supervised learning was proposed to solve the problems of existing methods, which behaved as a high detection limit and poor adaptability method. The pretreatment of orthogonal projection was used to correct the gap between different batches of spectral data. Afterwards the Partial Least Squares Discriminatory Analysis was adopted to extract the features from the data set. Outliers were found by comparing the alarm signal with the best threshold from the training set. Finally, Sequential Bayesian Method was used to update the probability of Contaminate Intrusion Events and to get the alarm sequence. The results showed that the proposed method had the lower detection limit than unsupervised method and the pretreatment of orthogonal projection improved the adaptability of detection method based on supervised learning for baseline changing.

Keywords UV-Vis spectrum; Water quality anomaly detection; Orthogonal projection; Supervised learning; Partial least squares discriminatory analysis

(Received Nov. 21, 2017; accepted Apr. 19, 2018)

* Corresponding author