基于 IERT 的非线性全光谱复杂水体定量分析算法研究

刘嘉诚[1,2]，胡炳樑[1]，于 涛[1*]，王雪霁[1]，杜 剑[1]，刘 宏[1]，刘 骁[1]，黄琦星[3]

1. 中国科学院西安光学精密机械研究所光谱成像技术重点实验室，陕西 西安 710119
2. 中国科学院大学，北京 100049
3. 深圳市盐田港集团有限公司，广东 深圳 518081

摘 要 水是一种有限的资源，对农业、工业乃至人类的生存都是必不可少的，良好的水环境是可持续发展的重要保障。对水质信息的科学监测，是实现水资源优化配置与高效利用的基础。联合国环境署(UNEP)与世界卫生组织(WHO)指出，应当加强发展中国家的水质监测网络，包括数据质量的保证和分析能力的提高。光谱法作为一种新兴的水质分析方法，相比传统的化学水质监测方法，具有“响应速度快、多参数同步、绿色无污染”的特点。传统单波长、多波长的线性模型依赖于水体对特定波长的吸收特征，不适用于多组分混合溶液且普适性较差。因此，提出了一种基于 IERT 的非线性全光谱定量分析算法，建立适用于多组分混合溶液浓度预测模型，达到利用全光谱信息来预测浓度信息的目的。利用实验室配置的 COD，BOD_5 和 TOC 多组分混合溶液与 NO_3-N、浊度、色度多组分混合溶液作为实验样本，使用光谱仪采集样本的光谱曲线，通过全光谱数据进行浓度预测实验，结果显示，对于 COD，BOD_5 和 TOC 多组分混合溶液，本算法对于三种组分的决定系数(R^2)分别为 0.999 3，0.991 4 和 0.999 3，均方根误差(RMSE)分别为 0.024 4，0.057 7 和 0.000 4；对于 NO_3-N、浊度、色度多组分混合溶液，决定系数(R^2)分别为 0.983 4，0.868 4 和 0.981 0，均方根误差(RMSE)分别为 0.100 5，0.326 4 和 0.120 2。通过对比本算法与偏最小二乘(PLS)、支持向量机回归(SVR)、决策树(DT)、极端随机树(ERT)对于同一组数据的实验结果，表明：在两组多组分混合溶液的实验中，本算法对于其中各组分的决定系数(R^2)均为最优，相比于其他对比算法均方根误差(RMSE)均有大幅减少。本算法可利用光谱信息对多组分混合溶液进行定量分析，在计算时间相当的情况下，可有效的提高浓度预测精度，减少定量分析的均方根误差，可为光谱法水质监测提供一种新的有效途径。

关键词 光谱法水质监测；紫外可见光谱技术；光谱定量分析；多组分混合溶液；极端随机树

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)12-3922-09

引 言

人类的行为活动会对地球生存环境构成重大威胁。据研究，全球所有死亡人数中有 23%是因为环境因素[1]。对人类生存环境的研究、保护和治理刻不容缓。水是自然环境和社会环境中极为重要且活跃的因素，对水质信息的科学监测是实现水资源优化配置与高效利用的基础。水质监测中常用到的指标包括化学需氧量(chemical oxygen demand，COD)、生化需氧量(biochemical oxygen demand，BOD_5)、总有机碳(total organic carbon，TOC)、硝酸盐氮(NO_3-N)、浊度(turbidity)、色度(colority)等。通过这些指标的监测，可以反映当前水体中各种污染物的浓度及变化趋势，从而达到评价水质状况的目的。

常用的水质监测方法多为基于化学检测的方法，包括现场取样进行实验室化学检测和利用基于化学法的仪器进行原位监测。现场取样进行实验室化学检测非原位、周期长，难以实现在线监测；基于化学法的仪器由于其使用化学试剂，存在化学残留，容易导致二次污染。近年来，基于光谱法的水质监测技术由于无需化学试剂、无二次污染、快速准确、成本低的特点，可实现实时在线原位测量，已广泛应用于在线水质监测领域。光谱法水质监测技术是利用水中特定物质

收稿日期：2020-11-03，**修订日期**：2021-02-27

基金项目：中国科学院战略性先导科技专项(A 类)(XDA23040101)，国防科技创新项目(XXX-ZT-00X-014-01)，国家重点研发计划项目(2017YFC1403700)，陕西省重点研发计划项目(2019SF-254)资助

作者简介：刘嘉诚，1993 年生，中国科学院西安光学精密机械研究所助理研究员 e-mail：liujiacheng@opt.ac.cn

*通讯作者 e-mail：yutao@opt.ac.cn

吸收特定波长的光，产生分子吸收光谱，通过建立预测模型，根据测得的吸收光谱来定性定量地分析水质参数。

光谱法水质监测中常用的预测模型主要分为线性模型和非线性模型，线性模型主要包括单波长法、多波长组合法、偏最小二乘(partial least squares，PLS)等。Dogliotti 等利用 645 nm 波段与 859 nm 波段的单波长半分析方法来反演水体浊度[2]；Knaeps 等利用 1 020 nm 波段与 1 071 nm 波段分析水体总悬浮物[3]；Carreres-Prieto 等利用多元线性回归(multivariable linear regression，MLR)预测 COD 等[4]；PLS 是由 Wold 提出的一种多元线性回归方法，它通过不断提取主成分来简化数据，建立回归模型，王莉丽等将 PLS 用于水体化学需氧量的测量并取得了不错的效果[5]；Wang 等使用 PLS 和多种机器学习算法预测水体总氮含量[6]。非线性模型主要包括支持向量机(support vector machines，SVM)，神经网络，决策树等。SVM 是由 Vapnik 提出的一种非线性回归方法，它将低维数据映射到高维空间进行回归，再把高维空间的超平面映射回低维空间，建立回归模型，陈颖等人将 SVM 的改进方法用于水体硝酸盐浓度的预测[7]；Gu 等使用随机森林(random forest，RF)的方法预测河流水体浊度[8]；神经网络是一种仿生的计算方法，用于大规模非线性的系统建模，Charulatha 等将人工神经网络(artificial neural network，ANN)用于地表水的亚硝酸盐检测[9]；Chen 等使用近红外光谱结合卷积神经网络(convolutional neural networks，CNN)检测农业灌溉用水[10]。

单波长、多波长的组合方法都依赖于水体对特定波长的吸收特征，同一波长组合建模可能适应于特定应用场景，不具有普适性。PLS 算法虽然利用了全光谱的数据，但只能寻找线性特征进行回归，无法捕捉非线性的特征。SVM 算法对小样本的学习和预测性能较好，但惩罚参数的选择对模型精度影响较大，惩罚参数较大模型容易过拟合，惩罚参数较小模型容易欠拟合。基于神经网络的算法对样本的数量需求较高，在小样本情况下模型泛化能力较差，且模型训练时间长。

1　算法原理

1.1　算法基本原理

为了解决上述问题，引入机器学习中极端随机树的思想，提出了一种基于改进极端随机树(improved extremely randomize trees，IERT)的非线性全光谱浊度定量分析算法。极端随机树(extremely randomize trees，ERT)是由 Pierre Geurts 等学者提出的基于决策树的集成方法，用于解决机器学习中的监督分类和回归问题。该方法对高维特征数据能很好的处理，准确度高，且能够并行计算，执行效率高[11]。由于精细光谱数据的高光谱分辨率导致的数据量大、不同波段之间数据存在冗余等特点，采用核主成分分析(kernal principal component analysis，KPCA)方法进行特征降维，通过非线性函数把吸光度光谱映射到高维空间进行主成分分析，提取数据高维、非线性的特征。之后，正态化降维后的数据，训练基于 IERT 的非线性全光谱浓度预测模型，算法流程图如图 1。

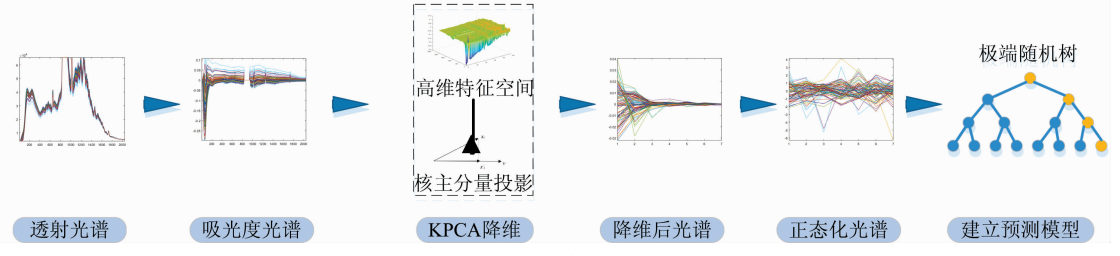


图 1　IERT 算法流程图

Fig. 1　Flowchart of IERT algorithm

1.1.1　吸光度转换

实验室测得的水体光谱为透射光谱，首先应转换为吸光度光谱，转换方法如式(1)所示。

$$A = -\log\left(\frac{I_1}{I_0}\right) \tag{1}$$

式(1)中，I_1 为被测水体的透射光谱，I_0 为标准去离子水的透射光谱，A 为吸光度光谱。

1.1.2　核主成分分析

核主成分分析方法是一种非线性的数据降维方法，它利用投影子空间技术，将信号非线性的映射到特征空间，在特征空间中对转换后的信号运用线性主成分分析进行数据降维，再将降维后的数据投影回输入空间，其中的非线性映射必须是可逆的。

根据 KPCA 算法的原理，将其用于吸光度光谱特征降维的流程如图 2。

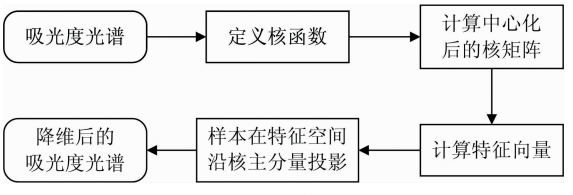


图 2　KPCA 算法流程图

Fig. 2　Flowchart of KPCA algorithm

首先，针对转换后的吸光度光谱，定义核矩阵 $\boldsymbol{K}$。

然后，计算中心化后的核矩阵 $\bar{\boldsymbol{K}}$，如式(2)所示，其中 $\boldsymbol{I}_M \in R^{M\times M}$ 为单位矩阵，$(\boldsymbol{I}_M)_{ij}=1$。

$$\bar{\boldsymbol{K}} = \boldsymbol{K} - \frac{\boldsymbol{I}_M\boldsymbol{K}}{M} - \frac{\boldsymbol{K}\boldsymbol{I}_M}{M} - \frac{\boldsymbol{I}_M\boldsymbol{K}\boldsymbol{I}_M}{M^2} \tag{2}$$

之后，计算核矩阵 $\bar{\boldsymbol{K}}$ 的非零特征值(λ_1，λ_2，λ_3，…，λ_n)和对应的特征向量(α_1，α_2，α_3，…，α_n)，并将特征值矩阵按

照特征值降序重新排列特征值向量，确定主特征向量的个数为 N。

最后，计算样本在特征空间上的投影，即为 KPCA 降维后的吸光度光谱。

1.1.3 数据正态化

数据正态化是为了让数据服从标准正态分布。假设 KPCA 降维后的吸光度光谱为 X，它服从均值为 μ、方差为 σ 的正态分布。则数据正态化的计算公式如式(3)所示

$$S = \frac{X-\mu}{\sigma} \tag{3}$$

式(3)中，S 为正态化变换后的数据，服从均值为 0，方差为 1 的正态分布。

1.1.4 极端随机树

决策树(decision tree，DT)是一种分类、回归模型，具有较高的可解释性和鲁棒性[12]。极端随机树是一种基于决策树的集成方法，它由很多棵决策树组成，且每一棵决策树之间没有关联。极端随机树在树节点分割时随机化切割点的选择，随机化的强度可以根据不同问题的需求，通过调节参数的方式来改变[13]。极端随机树使用所有的训练样本得到每棵决策树，组合成为模型，当有一个新的样本输入的时候，让模型中的每一棵决策树分别进行判断。与其他机器学习算法相比，极端随机树除了高准确性之外，还具有高计算效率的优势。

极端随机树算法根据经典的自上而下方法构建一组“自由生长”的决策树或回归树，与其他基于树的集合方法有两点不同：不同于随机森林在一个随机子集内得到最佳分叉属性，它选择分叉的特征属性时是完全随机的；它使用整个学习样本来得到每棵决策树[14]。

极端随机树的实现流程如图 4 所示。

1.2 模型评价标准

1.2.1 决定系数

决定系数(R-Square，R^2)反应了因变量的全部变异能通过回归关系被自变量解释的比例。决定系数越大，自变量对因变量的解释程度越高，自变量引起的变动占总变动的百分比越高，观察点在回归直线附近越密集。如 R^2 为 0.9，表示回归关系可以解释因变量 90%的变异，R^2 越大，表示模型的拟合效果越好。R^2 的计算公式如式(4)

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2} \tag{4}$$

式(4)中，$\hat{y}_i$ 为预测值，y_i 为真实值，$\bar{y}$ 为均值，$R^2 \in [0, 1]$。

1.2.2 均方根误差

均方根误差(root mean squared error，RMSE)是衡量平均误差的方法，可以评价数据的变化程度，均方根误差越小，说明用该预测模型描述实验数据的准确度越高，RMSE 的计算公式如式(5)

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{5}$$

2 实验与结果讨论

2.1 仪器与数据采集

实验室数据采集方法如图 3 所示，实验装置主要由光源、精细光谱分析仪和采集软件组成。

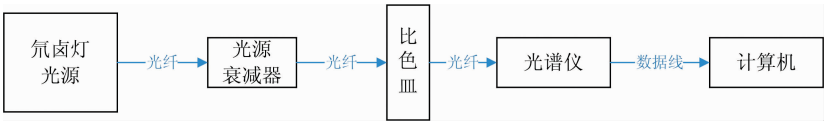


图 3 实验装置示意图

Fig. 3 Schematic diagram of experimental device

其中光源采用氘卤二合一光源，它在一个通道里整合了连续的氘灯和卤素灯宽波段光谱，波长范围 190～2 200 nm，预热时间 40 min。光路采用抗紫外辐照石英光纤，纤芯直径 600 μm，波长范围 185～1 100 nm。比色皿采用石英比色皿，光程 10 mm，适用波长 185～2 500 nm。光谱仪采用项目组自主研发的光谱分析仪，如图 5 所示。

该光谱仪采用了连续谱精细获取技术，整个仪器采用了双光路矫正，采用特征点领域多波长位置实现大量程适应性调节，其光谱范围为 185～1 100 nm，光谱采样间隔为 0.45 nm。

使用本套实验装置，对光源进行 5 h 连续测量，得到本系统稳定性为 2.38%。

2.2 实验数据

使用实验室配置的多组分混合溶液来模拟复杂水体，为了避免本算法只对特定的混合溶液有效，使本算法所建立的非线性全光谱浓度预测模型具有普适性，且克服水体混浊度、色度等对光学测量有严重干扰的影响因子，实验数据选取了两组不同的多组分混合溶液，分别为 200 组 COD，BOD_5 和 TOC 多组分混合溶液与 188 组 NO_3-N、浊度、色度多组分混合溶液。

2.2.1 COD，BOD_5，TOC 多组分混合溶液数据集

采用国标方法，用邻苯二甲酸氢钾、谷氨酸和葡萄糖配置 200 组不同浓度的 COD，BOD_5，TOC 混合溶液，如表 1 所示。

将 200 组样本随机分配，取其中 40 组为测试集，其余 160 组为训练集。

2.2.2 NO_3-N、浊度、色度多组分混合溶液数据集

实验室采用国标方法，用硝酸钾、硫酸肼、六次甲基四胺、六氯铂酸钾和六水氯化钴配置 188 组不同浓度的 NO_3-N、浊度、色度混合溶液，如表 2 所示。

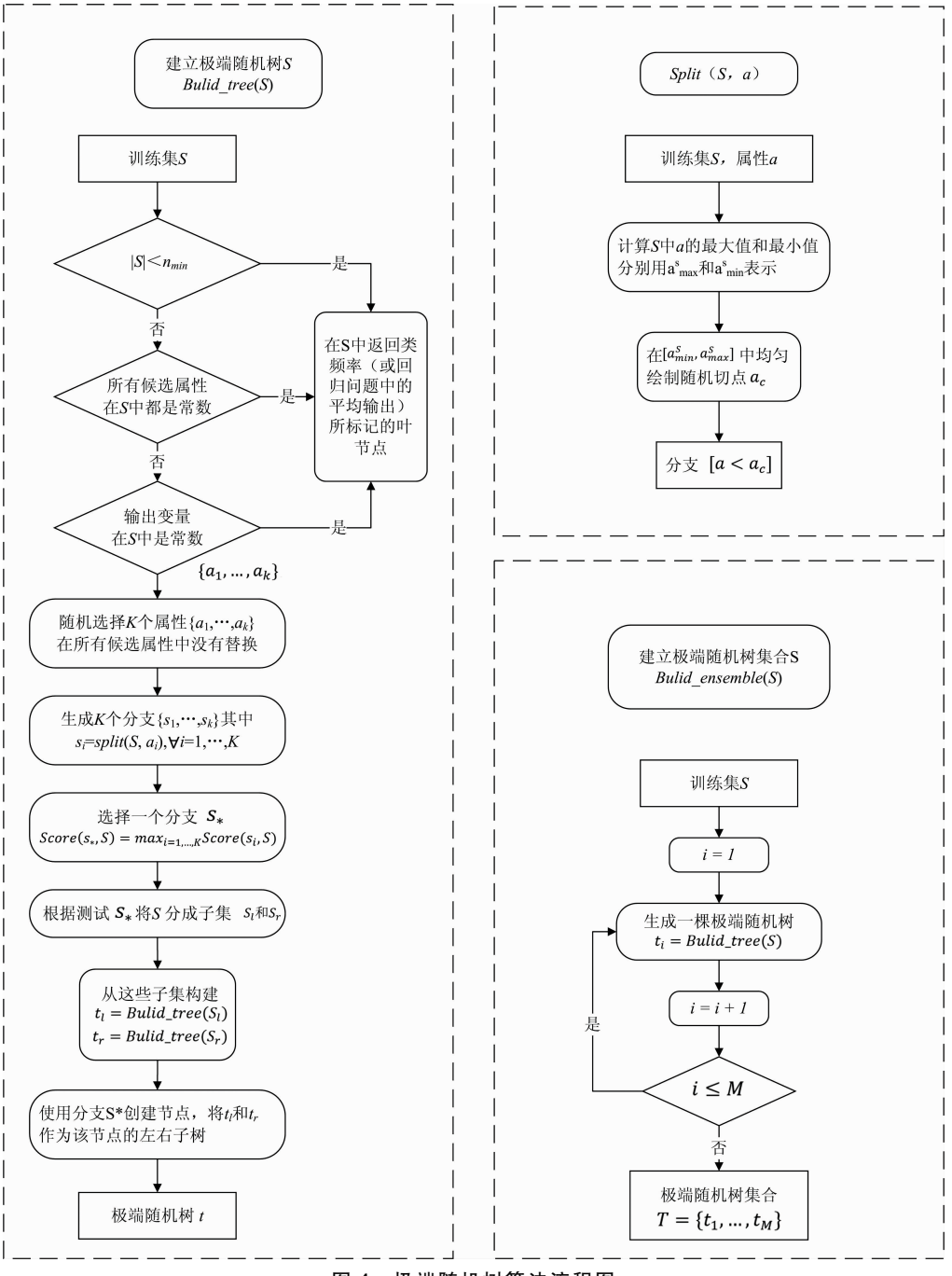


图 4　极端随机树算法流程图

Fig. 4　Flowchart of extremely randomize trees algorithm

图 5　光谱仪模型图及实物图

Fig. 5　Model diagram and physical diagram of the spectrometer

将 188 组样本随机分配，取其中 38 组为测试集，其余 150 组为训练集。

2.2.3　多组分混合溶液光谱数据采集

采用上述实验装置，在室温[(25±1) ℃]条件下对配置的样本进行透射光谱的测量，每个样本扫描 10 次，取平均值作为该样本的测量值。COD，BOD_5 和 TOC 混合溶液的透射光谱如图 6(a)所示，NO_3-N、浊度、色度混合溶液的透射光谱如图 6(b)所示。

表 1 COD，BOD_5，TOC 多组分混合溶液数据集样本特性表

Table 1 The sample characteristics table of multi-component mixed solution dataset of COD, BOD_5 and TOC

组分	COD/(mg·L^{-1})	BOD_5/(mg·L^{-1})	TOC/(mg·L^{-1})
最小值	0.3	0	0.1
下四分位数	3.95	2.78	1.58
上四分位数	14	6.8	5.6
最大值	20	9.7	8
平均数	9.57	4.93	3.83
标准偏差	5.81	2.57	2.32

表 2 NO_3-N、浊度、色度多组分混合溶液数据集样本特性表

Table 2 The sample characteristics table of multi-component mixed solution dataset of NO_3-N, turbidity and colority

组分	NO_3-N/(mg·L^{-1})	浊度/(mg·L^{-1})	色度/PCU
最小值	7	0.5	7
下四分位数	9	1.5	9
上四分位数	14	4	13
最大值	15	5	15
平均数	11.43	2.7	10.76
标准偏差	2.74	1.52	2.43

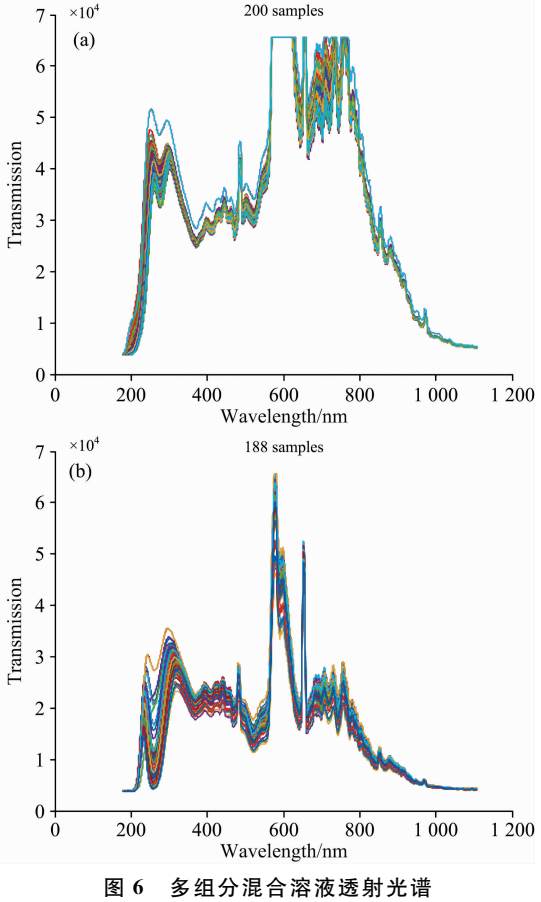


图 6 多组分混合溶液透射光谱

Fig. 6 Transmission spectra of multi-component mixed solutions

2.3 IERT 算法相关参数确定

2.3.1 核函数的选取

在 1.1.2 节核主成分分析中介绍了核函数，常用的核函数包括线性核函数、多项式核函数、高斯核函数、余弦核函数、sigmoid 核函数等，实验选取上述 5 类作为 IERT 算法中的核函数，在 COD，BOD_5 和 TOC 多组分混合溶液数据集上进行实验，通过结果中的决定系数值，初步选取 IERT 算法中的核函数。实验结果如表 3 所示，其中 R^2(ave)表示 COD，BOD_5 和 TOC 的平均决定系数。

表 3 IERT 算法不同核函数的实验结果

Table 3 Experimental results of IERT algorithm using different kernel functions

核函数	linear	polynomial	rbf	cosine	sigmoid
R^2(COD)	0.970 9	0.973 4	0.951 6	0.944 6	0.969 0
R^2(BOD_5)	0.372 3	0.383 5	0.328 0	0.413 2	0.408 6
R^2(TOC)	0.970 5	0.974 8	0.958 4	0.948 4	0.968 6
R^2(ave)	0.771 2	0.777 2	0.746 0	0.768 7	0.782 1

由表 3 可知，对于 COD，BOD_5 和 TOC 这 3 种指标，最合适的核函数不尽相同，但 sigmoid 核函数的平均决定系数最大，即选择 sigmoid 函数作为 IERT 算法中的核函数，可同时满足 COD，BOD_5 和 TOC 这 3 种指标的需求。

2.3.2 核函数的参数选择

上节中选择了 sigmoid 函数作为 IERT 算法的核函数，sigmoid 函数的计算公式为式(6)所示。

$$k(x,y) = \tanh(\alpha x^t y + c) \tag{6}$$

其中 α 通常取特征数的倒数，此时取 $\alpha=\frac{1}{2\ 048}$。式中 c 为一个常数，同时还需确定主成分的个数 n。实验选取主成分的个数 n 为 2 到 10，参数 c 为 0 到 10 进行实验，选择最合适的核函数参数。表 4 为 sigmoid 核函数在选取主成分数为 n 的条件下的决定系数。

由表 4 可知，当 sigmoid 核函数在主成分数选取 5 和 8 时，均有两个指标的决定系数最大，当主成分数 n 选取 5 时，平均决定系数最大，因此选择 sigmoid 核函数的主成分数 n 为 5。

表 5 为 sigmoid 核函数主成分数 n 为 5 时，在不同参数 c 下的决定系数。

由表 5 可知，当 sigmoid 核函数在主成分数 n 为 5、参数 c 为 6 时，平均决定系数最大，因此 IERT 算法选取 sigmoid 核函数的主成分数 n 为 5、参数 c 为 6。

2.3.3 极端随机树中的参数选择

极端随机树中树的数量为 m，表 6 为 IERT 算法在不同 m 下的实验结果。

由表 6 可知，当树的个数 m 取 320 时，平均决定系数最大，因此 IERT 算法选取极端随机树中树的个数 m 为 320。

2.4 实验结果

2.4.1 COD，BOD_5 和 TOC 多组分混合溶液实验结果

选取COD，BOD_5 和TOC多组分混合溶液数据集进行

表 4　sigmoid 核函数在选取主成分数为 *n* 时的实验结果

Table 4　The experimental results of the sigmoid kernel function when the number of principal components is selected as *n*

n	2	3	4	5	6	7	8	9	10
R^2(COD)	0.998 4	0.999 1	0.999 1	0.999 3	0.998 9	0.998 9	0.999 3	0.999 2	0.998 9
R^2(BOD_5)	0.970 1	0.984 1	0.985 1	0.990 3	0.982 8	0.985 1	0.985 2	0.970 1	0.970 9
R^2(TOC)	0.998 6	0.999 1	0.998 9	0.999 2	0.998 9	0.999 2	0.999 3	0.999 0	0.999 1
R^2(ave)	0.989 0	0.994 1	0.994 4	0.996 3	0.993 5	0.994 4	0.994 6	0.989 4	0.989 6

表 5　sigmoid 核函数在不同参数 *c* 下的实验结果

Table 5　Experimental results of sigmoid kernel function under different parameters *c*

c	0	1	2	3	4	5	6	7	8	9	10
R^2(COD)	0.999 2	0.999 3	0.999 1	0.999 0	0.999 3	0.999 4	0.999 3	0.999 3	0.999 4	0.999 3	0.999 3
R^2(BOD_5)	0.988 2	0.988 4	0.987 1	0.989 7	0.990 3	0.988 7	0.991 4	0.988 8	0.988 2	0.989 9	0.988 8
R^2(TOC)	0.999 3	0.999 3	0.998 9	0.999 2	0.999 3	0.999 3	0.999 3	0.999 2	0.999 3	0.999 3	0.999 1
R^2(ave)	0.995 6	0.995 7	0.995 0	0.996 0	0.996 3	0.995 8	0.996 7	0.995 8	0.995 6	0.996 2	0.995 7

表 6　IERT 算法在不同参数 *m* 下的实验结果

Table 6　Experimental results of IERT algorithm under different parameters *m*

m	20	50	80	120	150	180	220	250	280	320	350
R^2(COD)	0.998 6	0.999 2	0.999 2	0.999 6	0.999 2	0.999 4	0.999 4	0.999 3	0.999 5	0.999 3	0.999 5
R^2(BOD_5)	0.980 1	0.984 8	0.989 3	0.987 3	0.986 9	0.988 8	0.985 8	0.989 8	0.988 0	0.991 4	0.990 8
R^2(TOC)	0.999 1	0.998 9	0.999 1	0.999 1	0.999 3	0.999 3	0.999 5	0.999 2	0.999 3	0.999 3	0.999 3
R^2(ave)	0.992 6	0.994 3	0.995 9	0.995 3	0.995 1	0.995 8	0.994 9	0.996 1	0.995 6	0.996 7	0.996 5
m	380	420	450	480	520	550	580	620	650	680	720
R^2(COD)	0.999 1	0.999 3	0.999 4	0.999 4	0.999 3	0.999 3	0.999 4	0.999 3	0.999 2	0.999 2	0.999 3
R^2(BOD_5)	0.988 0	0.990 1	0.990 4	0.990 0	0.987 8	0.988 6	0.987 7	0.990 0	0.990 2	0.989 3	0.989 3
R^2(TOC)	0.999 3	0.999 2	0.999 3	0.999 2	0.999 2	0.999 1	0.999 0	0.999 2	0.999 3	0.999 3	0.999 3
R^2(ave)	0.995 5	0.996 2	0.996 4	0.996 2	0.995 4	0.995 7	0.995 4	0.996 2	0.996 2	0.995 9	0.996 0

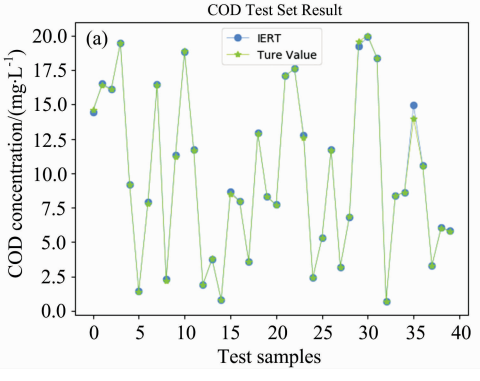
实验，IERT 算法的核函数为 sigmoid 函数，主成分数 *n* 为 5，参数 *c* 为 6，*m* 为 320。图 7 为其测试集的实验结果，包括三种组分的真实值与 IERT 算法预测值对比和 IERT 算法的相对误差。由图 7 可以看出，在测试集中 IERT 算法可准确的在多组分混合溶液中预测 COD，BOD_5 和 TOC 的含量。对于测试集的 40 个样本，其中 38 个样本 COD 的相对误差在 2% 以内，37 个样本 TOC 的相对误差在 2% 以内，最大不超过 7%；38 个样本 BOD_5 的相对误差在 10% 以内。

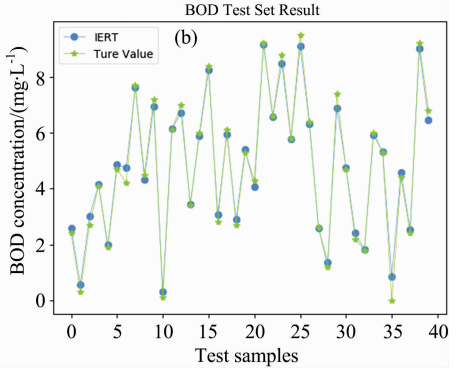
表 7 为 IERT 算法与 4 种对比算法的决定系数与均方根误差。

由表 7 可以看出，对于 COD，BOD_5 和 TOC 这三种指标，IERT 算法可同时预测混合溶液中的三种指标，而其他 4 中对比算法对多组分混合溶液中 BOD_5 这一指标的效果均较差。同时 IERT 算法的决定系数均大于 4 种比较算法，均方根误差均小于 4 种比较算法。

2.4.2　NO_3-N、浊度、色度多组分混合溶液实验结果

选取 NO_3-N、浊度、色度多组分混合溶液数据集进行实验，由于篇幅原因，此处不在列出预测值与真实值的对比图，只列出与几种对比算法的模型评价参数对比。由表 8 可以看出，对于 NO_3-N、浊度、色度这三种指标，IERT 算法也可同时预测混合溶液中的三种指标，而其他 4 中对比算法对多组分混合溶液中浊度这一指标的效果均较差。





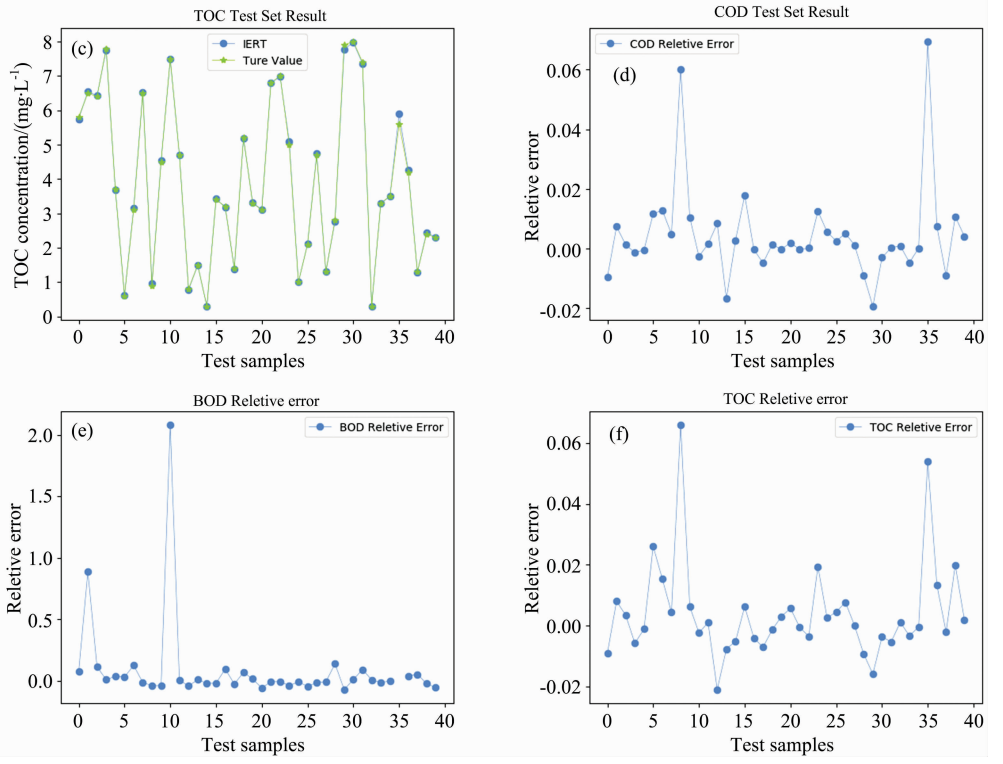


图 7 IERT 算法在 COD，BOD_5 和 TOC 多组分混合溶液测试集实验结果

(a)：COD 真实值与预测值对比；(b)：BOD_5 真实值与预测值对比；(c)：TOC 真实值与预测值对比

(d)：COD 预测值相对误差；(e)：BOD_5 预测值相对误差；(f)：TOC 预测值相对误差

Fig. 7 Test set experimental results of IERT algorithm for COD, BOD_5, TOC multi-component mixed solutions

(a): Comparison of the true value and the prediction value of COD;

(b): Comparison of the true value and the prediction value of BOD_5;

(c): Comparison of the true value and the prediction value of TOC;

(d): Relative error of COD prediction value; (e): Relative error of BOD_5 prediction value;

(f): Relative error of TOC prediction value

表 7 COD，BOD_5 和 TOC 多组分混合溶液中 IERT 算法与 4 种预测算法的评价参数对比

Table 7 Comparison of evaluation parameters between IERT algorithm and 4 prediction algorithms for COD, BOD_5, TOC multi-component mixed solutions

评价标准	指标	PLS	SVR	DT	ERT	IERT
R^2	COD	0.918 5	0.907 3	0.968 2	0.998 6	0.999 3
	BOD_5	0.057 7	0.126 8	0.358 2	0.900 8	0.991 4
	TOC	0.918 9	0.935 2	0.965 1	0.998 1	0.999 3
RMSE	COD	2.866 8	3.262 4	1.118 3	0.046 6	0.024 4
	BOD_5	6.316 1	5.852 8	4.301 8	0.664 8	0.057 7
	TOC	0.453 8	0.362 2	0.195 0	0.010 6	0.000 4

表 8 NO_3-N、浊度、色度多组分混合溶液中 IERT 算法与 4 种预测算法的评价参数对比

Table 8 Comparison of evaluation parameters between IERT algorithm and 4 prediction algorithms for NO_3-N, turbidity, colority multi-component mixed solutions

评价标准	指标	PLS	SVR	DT	ERT	IERT
R^2	NO_3-N	0.244 7	0.528 4	0.913 0	0.938 0	0.983 4
	浊度	0.005 0	0.163 6	0.448 0	0.731 7	0.868 4
	色度	0.622 0	0.671 6	0.933 5	0.961 0	0.981 0
RMSE	NO_3-N	4.571 4	2.854 6	0.526 3	0.375 0	0.100 5
	浊度	2.466 9	2.073 4	1.368 4	0.665 2	0.326 4
	色度	2.392 6	2.079 0	0.421 1	0.246 6	0.120 2

2.4.3 算法计算时间实验结果

实验选择 COD，BOD_5 和 TOC 多组分混合溶液数据集 1 和 NO_3-N、浊度、色度多组分混合溶液数据集 2 在同一硬件配置的计算机上，对 5 种算法所需的计算时间进行比较。实验采用的计算机硬件配置如下，处理器型号为 IntelCorei7，主频为 1.99 GHz，内存为 16 G。表 9 给出了 5 种算法在两组数据集上所需的计算时间。

表 9 IERT 算法与 4 种预测算法的计算时间对比

Table 9 Comparison of calculation time between IERT algorithm and 4 prediction algorithms

算法	PLS	SVR	DT	ERT	IERT
混合数据集 1 计算时间/s	3.1	3.2	3.5	3.8	4.3
混合数据集 2 计算时间/s	3	3.1	3.1	3.6	4.1

由表 9 可以看出，IERT 算法有着不错的计算速度，与传统算法在同一量级。

3 结 论

提出了一种基于改进极端随机树的非线性全光谱定量分析算法，利用多组分混合溶液数据集进行实验，并与传统的算法进行比较，得出以下结论：

传统的光谱定量分析算法大多只适用于单组分的水质分析，在多组分混合溶液上表现较差，IERT 算法通过全光谱数据进行非线性分析，相比传统的算法，具有更高的决定系数和更低的均方误差，对多组分混合溶液的预测效果很好。

IERT 算法具有挖掘数据深度特征的能力，这弥补了水质在线测量中光谱定量分析算法对海浪光谱信息利用不足的劣势，使得光谱定量分析对数据的挖掘能力得到提升，有效的提升了光谱法水质在线监测的能力。

IERT 算法在多组分混合溶液数据集中对浊度的检测均方根误差为 0.326 4，虽为 5 种算法中的最优结果，但误差仍偏高，下一步工作将继续优化 IERT 算法对浊度的检测能力，和更多种组分混合溶液的检测能力。

References

[1] Prüss-üstün A, Wolf J, Corvalán C, et al. Preventing Disease Through Healthy Environments. World Health Organization, 2016.
[2] Dogliotti A, Ruddick K, Nechad B, et al. Remote Sensing of Environment, 2015, 156: 157.
[3] Knaeps E, Ruddick K, Doxaran D, et al. Remote Sensing of Environment, 2015, 168: 99.
[4] Carreres-Prieto D, Garcia J, Cerdan-Cartagena F, et al. Sensors, 2020, 20: 1.
[5] Wang Lili, Liu Xianhua, Shi Xiaoxuan, et al. Advanced Materials Research, 2013, 726-731: 1534.
[6] Wang Jingzhe, Shi Tiezhu, Yu Danlin, et al. Environmental Pollution, 2020, 266: 115412.
[7] CHEN Ying, HE Lei, CUI Xing-ning, et al(陈 颖, 何 磊, 崔行宁, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(5): 1489.
[8] Gu Ke, Zhang Yonghui, Qiao Junfei, et al. IEEE Transactions on Instrumentation and Measurement, 2020, 69(11): 9028.
[9] Charulatha G, Srinivasalu S, Uma Maheswari O, et al. Arabian Journal of Geoscience, 2017, 10: 128.
[10] Chen Huazhou, Chen An, Xu Lili, et al. Agricultural Water Management, 2020, 240.
[11] JIN Kang-rong, YU Dong-jun(金康荣, 於东军). Journal of Nanjing University of Aeronautics & Astronautics(南京航空航天大学学报), 2018, 50(5): 619.
[12] Rivera-Lopez R, Canul-Reich J. IEEE Access, 2018, 6: 5548.
[13] Xia Bin, Zhang Hong, Li Qianmu, et al. IEEE Transactions on Nanobioscience, 2015, 14(8): 882.
[14] HUANG Cong-wu, CHEN Bao-zhang, MA Chao-qun, et al(黄丛吾, 陈报章, 马超群, 等). Acta Meteorologica Sinica(气象学报), 2018, 76(5): 779.

Nonlinear Full-Spectrum Quantitative Analysis Algorithm of Complex Water Based on IERT

LIU Jia-cheng[1,2], HU Bing-liang[1], YU Tao[1*], WANG Xue-ji[1], DU Jian[1], LIU Hong[1], LIU Xiao[1], HUANG Qi-xing[3]

1. Key Laboratory of Spectral Imaging Technique, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China
2. University of Chinese Academy of Sciences, Beijing 100049, China
3. Shenzhen Yantian Port Group Co., Ltd., Shenzhen 518081, China

Abstract Water is a finite resource, essential for agriculture, industry and even human existence. A good water environment is an important guarantee for sustainable development. The scientific monitoring of water quality information is the basis for optimal allocation and efficient use of water resources. The United Nations Environment Program (UNEP) and the World Health Organization (WHO) pointed out that national water quality monitoring networks in developing countries should be strengthened, including improving analytical capabilities and data quality assurance. As an emerging water quality analysis method, spectral method has the characteristics of "fast response, synchronization of multiple parameters, environmental protection and pollution-free" compared with traditional chemical water quality monitoring methods. The traditional single-band, multi-band linear model, relies on the absorption characteristics of water at specific bands, and it cannot be used for multi-component mixed solutions and has poor universality. Therefore, this paper proposes a non-linear full-spectrum quantitative analysis algorithm based on IERT. The concentration prediction model suitable for multi-component mixed solution is established

to use full spectrum information to predict concentration information. We use the COD, BOD_5, TOC multi-component mixed solution and NO_3-N, turbidity, chroma multi-component mixed solution configured in the laboratory as the experimental sample, use the spectrometer to collect the spectral curve of the sample, and conduct the concentration prediction experiment through the full spectrum data. The experimental results show that for COD, BOD_5, TOC multi-component mixed solutions, the determination coefficients (R^2) of this algorithm for the three components are 0.999 3, 0.991 4 and 0.999 3. The root means square error (RMSE) is 0.024 4, 0.057 7 and 0.000 4. For the multi-component mixed solution of NO_3-N, turbidity, and colority, the coefficient of determination (R^2) is 0.983 4, 0.868 4 and 0.981 0. The root means square error (RMSE) is 0.100 5, 0.326 4 and 0.120 2. By comparing the experimental results of this algorithm with partial least squares (PLS), support vector regression (SVR), decision tree (DT), and extreme random tree (ERT) for the same set of data, the results show that in the experiment of mixed solution, this algorithm is the best alternative to the coefficient of determination (R^2) of each component. The root means square error (RMSE) has been greatly reduced compared with other comparison algorithms. This algorithm can use spectral information to analyze the multi-component mixed solution quantitatively. It can effectively improve the concentration prediction accuracy and reduce the root-mean-square error of the quantitative analysis in the case of equivalent calculation time. Moreover, this algorithm can provide a theoretical basis for spectral methods on water quality monitoring.

Keywords Spectroscopic water quality monitoring; Ultraviolet visible spectroscopy technology; Spectral quantitative analysis; Multi-component mixed solution; Extreme random trees

(Received Nov. 3, 2020; accepted Feb. 27, 2021)

* Corresponding author