doi:10.6041/j.issn.1000-1298.2025.04.032

基于中红外光谱与机器学习的生物炭表面碳氧元素及基团含量预测模型研究

曹红亮[1,2]　王　盼[1,2]　王卓超[1,2]　杨争鸣[1,2]　马家敏[1,2]　徐　洋[1,2]

(1.农业农村部智慧养殖技术重点实验室, 武汉 430070; 2.华中农业大学工学院, 武汉 430070)

摘要: 为了实现生物炭表面碳氧元素与活性基团的高精快速预测,基于课题组积累的 120 组生物炭样品,建立了包含生物炭中红外光谱和表面碳氧元素及其赋存形态定量表征信息的数据集;采用支持向量机(SVM)、随机森林(RF)机器学习智能建模方法,结合区间偏最小二乘法(IPLS)和主成分分析法(PCA)等特征筛选策略,构建了 IPLS + RF、IPLS + SVM、PCA + RF、PCA + SVM 共 4 种预测模型,实现了表面碳氧元素含量(S_C、S_O)以及 8 种碳氧赋存形态共计 10 个预测目标的定量快速预测。其中,碳氧赋存形态有来自 C1s 能谱的 C═C、C—C、C—O、C═O、O═C—O 共 5 种形态(C_C═C、C_C—C、C_C—O、C_C═O、C_O═C—O)以及来自 O1s 能谱的 C═O、C—O、O═C—O 共 3 种形态(O_C═O、O_C—O、O_O═C—O)。研究结果表明:生物炭表面碳元素的主要赋存形态为 C_C═C、O_C—O,生物炭表面氧元素的主要赋存形态为 C_C—O、C_C═O、C_O═C—O 以及 O_C═O;特征波段 4 000 ~ 3 464 cm^{-1}和1 588 ~ 650 cm^{-1}均包含与生物炭表面碳氧元素含量及其赋存形态高度相关的特征信息,但 1 588 ~ 650 cm^{-1}蕴含的信息更为丰富;从模型预测精度来看,IPLS + RF、IPLS + SVM、PCA + RF、PCA + SVM 这 4 种预测模型均具有良好的预测能力,IPLS + SVM 和 PCA + SVM 尤为突出,对 10 个预测目标的最优模型决定系数均在 0.93 以上;但从模型稳定性和泛化能力来看,C_C—C、C_O═C—O、O_C═O、O_C—O 还有待进一步提升。

关键词: 生物炭; 碳氧基团; 红外光谱; 机器学习; 定量预测

中图分类号: TP391; O655.4　　**文献标识码**: A　　**文章编号**: 1000-1298(2025)04-0344-09　　**OSID**:

Modeling of Carbon and Oxygen Elements and Groups on Biochar Surface Based on Mid-infrared Spectroscopy and Machine Learning

CAO Hongliang[1,2]　WANG Pan[1,2]　WANG Zhuochao[1,2]　YANG Zhengming[1,2]　MA Jiamin[1,2]　XU Yang[1,2]

(1. *Key Laboratory of Smart Farming Technology for Agricultural Animals, Ministry of Agriculture and Rural Affairs, Wuhan* 430070, *China*

2. *College of Engineering, Huazhong Agricultural University, Wuhan* 430070, *China*)

Abstract: In order to achieve high-precision and rapid prediction of carbon and oxygen elements and active groups on the surface of biochar, a data set containing quantitative characterization information of mid-infrared spectroscopy and surface carbon and oxygen elements and their occurrence forms was established based on 120 groups of biochar samples accumulated by the research group. Using support vector machine (SVM) and random forest (RF) machine learning intelligent modeling methods, combined with interval partial least squares (IPLS) and principal component analysis (PCA) and other feature selection strategies, four prediction models of IPLS + RF, IPLS + SVM, PCA + RF and PCA + SVM were constructed, and the quantitative and rapid prediction of surface carbon and oxygen content (S_C, S_O) and eight carbon and oxygen occurrence forms, a total of ten prediction targets, was realized. Among them, there were five forms of C═C, C—C, C—O, C═O, O═C—O from the C1s energy spectrum and three forms of C═O, C—O, O═C—O from the O1s energy spectrum. The results showed that the main occurrence forms of carbon on the surface of biochar were C_C═C and O_C—O, and the main occurrence forms of oxygen on the surface of biochar were C_C—O, C_C═O, C_O═C—O and O_C═O; the characteristic bands of 4 000 ~ 3 464 cm^{-1} and 1 588 ~ 650 cm^{-1} both contained characteristic information highly related to the content and speciation of carbon and oxygen on the surface

收稿日期: 2024-03-01　修回日期: 2024-04-14

基金项目: 国家自然科学基金项目(31971807)

作者简介: 曹红亮(1982—),男,教授,主要从事生物炭结构调控及其应用研究,E-mail:hongliangcao@mail.hzau.edu.cn

of biochar, but the information contained in 1 588 ~ 650 cm^{-1} was more abundant. From the perspective of model prediction accuracy, the four prediction models of IPLS + RF, IPLS + SVM, PCA + RF and PCA + SVM all had good prediction ability, especially IPLS + SVM and PCA + SVM. The coefficient of determination of the optimal model for ten prediction targets was above 0.93; however, from the perspective of model stability and generalization ability, C_C—C, C_O═C—O, O_C═O, O_C—O still needed to be further improved.

Key words: biochar; carbon and oxygen group; infrared spectroscopy; machine learning; quantitative prediction

0　引言

生物炭具有良好的吸附能力[1-2]，在空气净化、水质净化、土壤改良以及作为炭基缓释肥、缓释农药、微生物菌剂等应用场景，被广泛应用[3-7]。定量表征生物炭空间孔隙、表面元素与活性基团等结构特征，是深入研究生物炭吸附应用中其构效关系必不可少的数据支撑。

目前对生物炭表面元素与活性基团的定量表征方法，主要有 Boehm 滴定法、X 射线光电子能谱（XPS）、零电荷点法（pHPZC）和程序升温脱附（TPD）等[8-11]。上述实验表征方法，能准确获取相关结构参数，但存在实验测试过程工作量大、费时耗力、实时性不强等不足。红外光谱是一种非破坏性、快速、准确的分析方法[12]。但由于红外光谱摩尔吸光系数偏小而致使其相对灵敏度较低，同时还存在吸收带重叠等问题，红外光谱通常只能对物质结构进行定性或半定量分析[13]。但从信息论角度看，红外光谱数据仍然有效包含了被检测对象的结构信息，只是信息量大、维数高、结构复杂，不易定量辨识和提取。

随着机器学习人工智能技术的快速发展，通过机器学习挖掘光谱数据内在的映射规律，精准构建光谱数据与物质结构信息间的映射关系模型，实现物质相关结构信息的快速定量预测成为可能，并已得到成功应用[14-15]。可见，基于生物炭红外光谱及其结构定量表征实验数据，采用机器学习数据挖掘方法，有望实现其表面元素与活性基团的快速定量建模预测。

基于课题组前期积累的 120 组生物炭，本文建立包含生物炭中红外光谱和表面碳氧元素及其赋存形态定量表征信息的数据集；基于支持向量机（SVM）、随机森林（RF）两类经典的机器学习算法，采用异常值清洗、光谱预处理、特征筛选以及模型预测精度、泛化能力和稳健性综合评价等建模方法与手段，搭建一组能实现生物炭表面碳氧元素含量及其不同赋存形态相对含量的高精快速预测模型，以期为生物炭构效关系解析及吸附应用研究提供有力的模型基础。

1　材料与方法

1.1　光谱数据采集

选取课题组前期积累的包括牛粪、竹屑、水稻秸秆、玉米秸秆、棉花秸秆等不同生物质原料制备的 120 组生物炭样本作为建模对象，采集生物炭的红外光谱和 XPS 光谱数据。红外光谱数据采用 VERTEX 70 型傅里叶红外光谱仪（德国 Bruker 公司）采集。实验样品经过溴化钾压片处理后，采用透射方式获得红外光谱，扫描时仪器参数设定为：波数范围 4 000 ~ 650 cm^{-1}、光谱分辨率 4 cm^{-1}、数据间隔为 2，数据以吸光度形式存储。

生物炭表面碳氧元素（简写为 S_C 和 S_O）原子比含量采用 XPS 光谱法定量表征（AXIS UltraDLD，岛津 KRATOS 公司，英国），其赋存形态借助 Avantage 软件，对 C1s、O1s 进行分峰处理。其中，C1s 能谱按照 C═C、C—C、C—O、C═O、O═C—O 5 种赋存形态进行子峰拟合（简写为 C_C═C、C_C—C、C_C—O、C_C═O、C_O═C—O），相应的分峰结合能分别为 284.8、285.4、286.2、287.4、289.2 eV；O1s 能谱的分峰形态为 C═O、C—O、O═C—O 3 种形态（简写为 O_C═O、O_C—O、O_O═C—O），对应结合能分别为 531.2、532.0、533.3 eV[16]。

1.2　异常样本剔除

采用箱线图和 3σ 原则来识别离群值。对于具有多个波长的光谱数据，首先对每个波长进行离群值检测，然后将判定为离群值的样本剔除。分峰数据异常值则采用学生化残差法剔除[15]。

1.3　光谱数据预处理

为了减少谱图数据本身对建模精度的影响，采用以下光谱预处理方法：用于光程校正的多元散射校正（MSC）算法和标准正态变量变换（SNV）；用于去除噪声的 Savitzky－Golay 卷积平滑法（SG）；用于消除基线漂移的一阶导数（FD）、二阶导数（SD）和自适应迭代加权惩罚最小二乘法（airPLS）。通过上述方法对原始光谱进行预处理

后,以全光谱建立偏最小二乘法(Partial least squares,PLS)定量模型,进而对比不同预处理方法对模型效果的影响。

1.4 特征筛选

生物炭红外光谱测量波长范围 4 000 ~ 650 cm^{-1},共有 6 950 个波长点。红外光谱这种高维数据本身具有稀疏性,不是所有的变量都相互独立,若以全波段进行模型构建,其众多的冗余信息,不仅增加建模难度和复杂程度,还会由于变量多重共线性的出现导致模型过拟合,降低模型精度[17]。为此,采用区间偏最小二乘法(IPLS)算法和主成分分析(PCA)从原始光谱中提取出特征信息[18]。

1.5 机器学习模型选择与超参数优化

1.5.1 预测模型选择

针对生物炭表面碳氧基团红外光谱数据样本量不大且维数高等特点,SVM 和 RF 从不同角度表现良好的针对性和建模潜力,互有利弊[19-20]。为了探究二者在高维小样本建模的表现,选取这两类模型进行深入建模研究。

1.5.2 超参数优化

采用网格搜索法对模型参数进行快速寻优。网格搜索是一种自动化寻找最优超参数的方法,可以生成所有可能的参数组合,针对每一组参数训练模型并进行评估,最后选择表现最好的参数组合。首先确定需要搜索的参数,RF 模型选择了 n_estimators(决策树个数)和 max_depth(决策树最大深度)两个超参数。SVM 模型选择超参数 kernel、c、epsilon、degree、gamma、coef0,c 指的是正则化系数。确定需要搜索的参数后,设定各参数的取值范围,将每个参数的取值范围结合起来,生成参数网络。所有参数组合构建的模型均以均方根误差(RMSE)为评估指标,寻找最佳的超参数组合。

1.6 模型评价及正则化方法

1.6.1 模型评价指标

为了评价模型预测能力与拟合效果,采用决定系数 R^2、RMSE、均方误差(MSE)和平均绝对误差(MAE)作为评价指标。

模型稳健性评价借助于五折交叉验证法进行模型内部检验,计算五次交叉验证的 RMSE 与 MAE,取平均数作为稳健性评价指标,定义为交叉均方根误差($RMSE_{CV}$)和交叉平均绝对误差(MAE_{CV}),进而对模型进行优化与选择。

1.6.2 模型正则化处理

当预测模型出现过拟合时,模型正则化处理是一种行之有效的解决方法[21]。其中 L2 正则化通过 L2 范数添加惩罚项,使模型在决策时考虑更多的特征,而不是强依赖某几个特征,能极大提升模型的鲁棒性和泛化能力,而被广泛采用[22]。

2 结果与分析

2.1 红外光谱与 XPS 分峰数据分析与处理

2.1.1 生物炭表面碳氧元素及基团含量分析

图 1 展示了 120 组生物炭样品的表面碳氧元素含量及其不同基团赋存形态的 25% 分位数、中位数、75% 分位数以及整体数据分布范围(本文含量均指原子百分比含量)。从各参数的分布范围来看,其与文献[23-26]数据基本保持在同一范围,这证明了本文生物炭样本数据集的丰富性和多样性。通过小提琴图的可视化呈现,可以确定存在异常数据。为下一步机器学习异常值剔除以及精确建模奠定了良好的数据基础。

2.1.2 生物炭表面碳氧元素赋存形态分析

借助箱线图、3σ 原则和学生化残差法,剔除了 9 组异常样品。借助剔除异常值后的 111 组生物炭样品对生物炭表面碳氧元素及赋存形态共计 10 个预测指标进行皮尔森相关性分析(图 2),全面考察和定量表征了生物炭各指标间的相关性和独立性。根据相互性数据可见,生物炭 S_C 含量与 C_C═C 含量(皮尔森相关系数(RCC)为 0.77)、O_C—O 含量(RCC 为 0.73)呈高度正相关,且相关性显著($P<0.001$),说明生物炭表面 C 元素的赋存形态以 C 谱 C═C 和 O 谱 C—O 为主,与 C_C—O(RCC 为 -0.62)、C_C═O(RCC 为 -0.78)、C_O═C—O(RCC 为 -0.73)、O_C═O(RCC 为 -0.79)呈显著性高度负相关,且相关性显著($P<0.001$)。生物炭表面 O 含量与 C_C—O(RCC 为 0.61)、C_C═O(RCC 为 0.77)、C_O═C—O(RCC 为 0.71)、O_C═O(RCC 为 0.75)呈现出高度正相关,相关性显著($P<0.01$)。表明生物炭表面氧元素赋存形态集中在 C 谱 C—O、C 谱 C═O、C 谱 O═C—O 以及 O 谱 C═O 上。生物炭的 S_C 含量与 S_O 含量呈高度负相关(RCC 为 -0.98),与生物炭成炭过程炭结构的演化生长规律吻合,即随着炭化温度升高,生物炭石墨化结构增强(碳元素含量增加),表面活性基团被分解破坏(含 O 基团分解挥发,O 含量降低)[27]。

2.1.3 红外原始光谱分析及预处理

剔除异常值后的生物炭样品中红外光谱如图 3a 所示,样品之间因生物质原料、改性方式、掺杂比例、热解温度等不同,谱线之间存在差异,但是曲线的走势保持一致。

生物炭红外光谱峰位分布如图 3b 所示,在

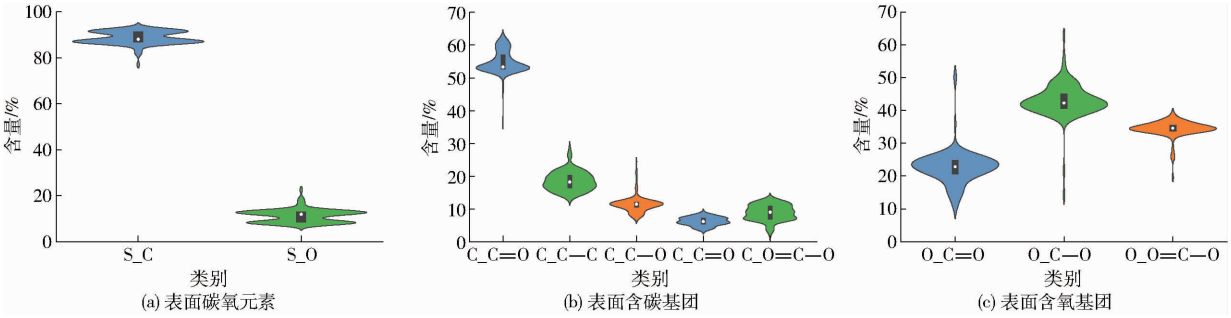

图 1　生物炭表面碳氧元素及基团含量小提琴图

Fig. 1　Violin diagrams of carbon and oxygen elements and group content on biochar surface

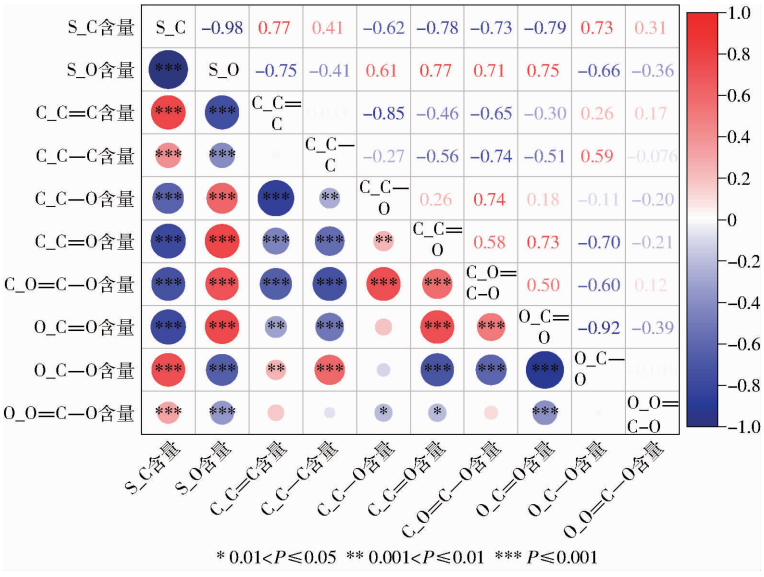

图 2　皮尔森相关性分析

Fig. 2　Pearson correlation analysis

3 447 cm^{-1}左右出现了一个微弱的吸收峰，是醇类以及酚类—OH 伸缩振动的结果；1 747 cm^{-1}附近出现的吸收峰，为饱和羧酸 C ═O 吸收峰，佐证了生物炭表面含有此类的含氧官能团，芳香族基团的吸收峰集中在 1 650 ~ 1 450 cm^{-1}处，其中在 1 541 cm^{-1}处存在一个较强吸收峰，是苯环的骨架震动（V（C ═ C））引起的；在 1 460 cm^{-1}处存在较强峰，该处峰位主要是由—O—CH_2—O—中的—CH 变形振动导致的；1 043 cm^{-1}附近出现了较强的吸收峰，属于纤维素和半纤维素的特征吸收峰，主要是 C—O 和 O—H 键振动引起；875 cm^{-1}处的峰位则是由炭骨架上芳香烃 C—H 变形振动引起的[28]。

从图 3a 可以明显看出，在 1 900 ~ 650 cm^{-1}和 3 800 ~ 3 500 cm^{-1}波段内，光谱存在多个明显吸收峰，为生物炭表面碳氧元素及基团的定量分析提供了丰富的信息。采用多元散射校正（MSC）、标准正态变量变换（SNV）、一阶导数（FD）、二阶导数（SD）、SG 平滑和自适应迭代加权惩罚最小二乘法（airPLS）及其组合方法对原始光谱进行预处理，由于偏最小二乘法是光谱预测领域使用最多的建模算

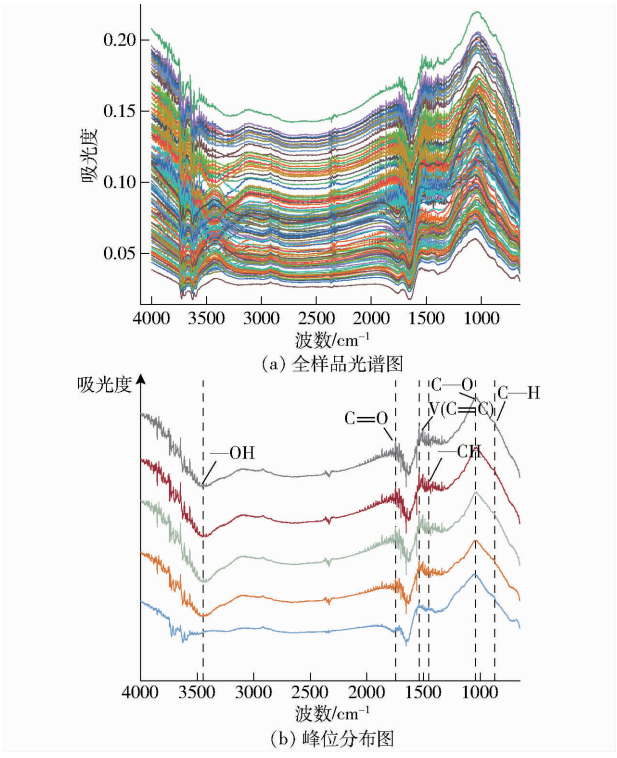

图 3　生物炭中红外光谱

Fig. 3　Mid-infrared spectrum of biochar

法,因此借助偏最小二乘法(PLS)构建生物炭表面C元素的定量分析模型,对比不同预处理组合的建模结果(R^2与RMSE),筛选出最佳的预处理组合,对应的建模结果如表1所示。

表1 红外光谱预处理建模结果对比(表面C元素基于PLSR)

Tab.1 Comparison of infrared spectrum pretreatment modeling results (surface C element based on PLSR)

序号	预处理方法	R^2	RMSE/%
1	无处理	0.525	1.871
2	剔除异常值	0.562	1.778
3	MSC	0.616	1.627
4	SNV	0.628	1.583
5	SG	0.703	1.119
6	FD	0.686	1.227
7	SD	0.698	1.258
8	FD + MSC	0.683	1.307
9	FD + SNV	0.696	1.214
10	SG + FD	0.754	1.104
11	SG + airPLS	0.826	0.727

从表1中可以看出各种预处理方法均对模型预测性能起到了一定程度上的提升作用,且平滑降噪+基线校正组合(序号9、10、11),相较于其他方法,模型精度提升更明显,尤其是SG平滑+airPLS对比原始空白对照组,提升效果十分显著,这是因为原始光谱曲线存在较为严重的噪声干扰和基线漂移,对于光谱数据建模存在较大影响。相较于其他基线校正算法,airPLS还很好地解决了扩展性差等问题,是目前应用最广、效果最好的基线校正算法[29]。因此后续将SG平滑+airPLS处理后数据用于特征筛选和建模。

2.2 基于IPLS与PCA的红外光谱特征筛选

2.2.1 基于IPLS的特征波段筛选

IPLS算法将整个光谱区域(4 000~650 cm^{-1})划分为25个等宽的子区间,由图4可知,在生物炭表面C元素的局部建模结果中,前7区间和后4区间模型MSE普遍较低,对应波段为1 588~650 cm^{-1}和4 000~3 464 cm^{-1},第7个区间由于包含了生物炭中含量最高的C═C键,MSE相较于4~6区间明显降低,第3区间(1 052~918 cm^{-1})对应模型的MSE最小,说明在此区间的信息丰度最高,在此基础上构建的定量回归模型精度最高,从原始光谱图(图3a)可以明显看出,918~650 cm^{-1}波段存在吸收峰,但是强度不高,尖锐峰位出现在1 050 cm^{-1},位于第3区间,与IPLS算法建模结果基本一致。借助IPLS算法,可以有效剔除冗余信息,提取出与目标值最相关的特征波段,从而提升回归模型的精度与速度,便于后续生物炭表面基团快速检测便携装备研发。

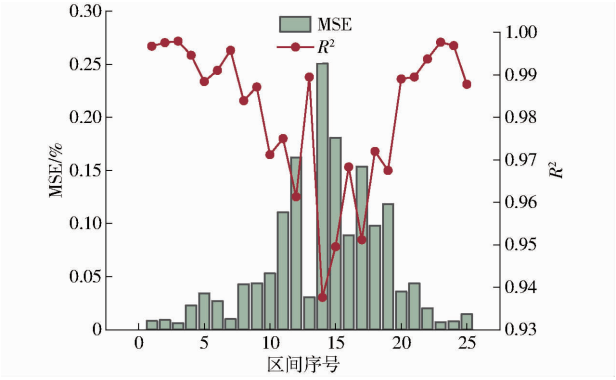

图4 基于IPLS的特征波段分区结果(S_C)

Fig.4 IPLS-based characteristic band partitioning results (S_C)

其他预测目标筛选的特征波段如表2所示,可以发现特征波段基本集中于前4个区间,即1 186~650 cm^{-1},且前7个区间(1 588~650 cm^{-1})与后4个区间(4 000~3 464 cm^{-1})上总体效果都比较优异。从图3中可以看出,在1 900~650 cm^{-1}波段内,光谱存在多个明显的特征吸收峰,结合区间建模结果,可以说明在1 588~650 cm^{-1}范围内的中红外光谱为生物炭表面碳氧元素及基团的定量分析提供了丰富的特征信息。

表2 各预测目标特征波段筛选结果

Tab.2 Characteristic band screening results of each prediction target

预测目标	特征区间序号	特征波段/cm^{-1}
S_C	3	1 052~918
S_O	1	784~650
C_C═C	24	3 866~3 732
C_C—C	1	784~650
C_C—O	4	1 186~1 052
C_C═O	1	784~650
C_O═C—O	1	784~650
O_C═O	1	784~650
O_C—O	2	918~784
O_O═C—O	1	784~650

2.2.2 基于PCA的特征波长筛选

采用PCA筛选特征波长结合RF与SVM模型的预测结果如图5所示(以O谱O═C—O为例),图5a、5b分别为RF模型、SVM模型训练集与测试集的决定系数与均方根误差随特征维数变化结果。

两个模型特征维数均从25到1依次递减,R^2与RMSE都呈现出较大波动,并且在特征维数降到1时,模型预测性能出现断崖式下跌,究其原因主要是主成分过少,缺失了过多的重要特征光谱信息导致

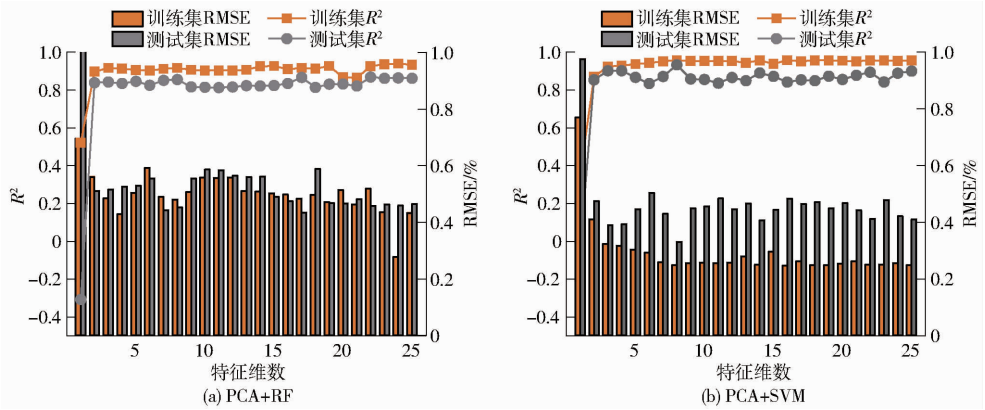

图 5　基于 PCA 结合 RF 与 SVM 模型的 O_O═C—O 预测模型性能

Fig. 5　Performance of O_O═C—O prediction model based on PCA combined with RF and SVM models

模型无法得到充分训练，拟合效果极差。从整体上看，SVM 模型 R^2 普遍高于 RF 模型，RMSE 明显低于 RF 模型。说明 SVM 模型对 O 谱 O═C—O 的定量预测明显优于 RF 模型。RF 模型特征维数取值为 17 时达到最优，R^2 达到最大，为 0.867，RMSE 达到最小，为 0.434。SVM 模型特征维数取值为 8 时，R^2 取得最大值，为 0.933，RMSE 取得最小值，为 0.330。同时从图 5b 中可以看出，在最优特征维数 8 两侧，R^2 出现波动式降低，RMSE 呈现出波动式升高，并且测试集的 R^2 明显低于训练集，RMSE 显著高于训练集，证明了存在一定的过拟合情况。而当特征维数超过 10 时，基于 PCA 特征波长提取的模型性能趋于稳定，说明在获取的特征信息充足稳定的情况下，模型的性能指标与特征数量无明显相关关系，这与前人的研究结论基本一致[30]。

2.3　红外光谱定量建模与评价

2.3.1　基于 IPLS 的中红外光谱建模效果与评价

利用 IPLS 算法提取的特征波段建立 10 个指标的 RF 和 SVM 模型之前，依照 1.5.2 节中的步骤，对所有超参数进行网格化寻优，对应模型的超参数优化结果如表 3 所示。各预测目标的 RF 与 SVM 模型优化、预测与交叉验证结果如表 4 所示。

表 3　基于 IPLS 的中红外光谱预测模型超参数优化

Tab. 3　Hyper-parameter optimization of mid-infrared spectral prediction model based on IPLS

预测目标	模型							
	RF			SVM				
	n_estimators	max_depth	kernel	c	epsilon	degree	gamma	coef0
S_C	34	7	rbf	1.947	0.183		scale	
S_O	99	7	rbf	1.169	0.118		scale	
C_C═C	97	7	poly	1.941	0.149	4	scale	0.925
C_C—C	36	5	rbf	1.995	0.127		scale	
C_C—O	41	7	rbf	1.013	0.115		scale	
C_C═O	75	6	rbf	1.211	0.114		scale	
C_O═C—O	99	9	rbf	1.820	0.123		scale	
O_C═O	68	30	rbf	1.054	0.125		scale	
O_C—O	91	10	rbf	1.471	0.100		scale	
O_O═C—O	69	8	rbf	1.667	0.114		scale	

基于 IPLS 的生物炭表面碳氧元素及碳氧官能团的模型对比与选择：由表 4 可知，S_C 选择 IPLS + RF 模型，其 RF 模型（R^2 为 0.975，RMSE 为 0.339%）相较于 SVM 模型（R^2 为 0.983，RMSE 为 0.277%）预测精度略低，但交叉验证后模型的平均绝对误差与均方根误差均大幅度低于后者，说明 RF 模型的鲁棒性与泛化性能更佳。从模型预测性能与稳健性的综合考虑出发，最终模型选择前者。同理，其余预测指标最优模型依次：S_O 选择 IPLS + RF 模型，C_C═C 选择 IPLS + SVM 模型，C_C—C 选择 IPLS + RF 模型，C_C—O 选择 IPLS + RF 模型，C_C═O 选择 IPLS + SVM 模型，C_O═C—O 选择 IPLS + SVM 模型，O_C═O 选择 IPLS + RF 模型，O_C—O 选择 IPLS + SVM 模型，O_O═C—O 选择 IPLS + SVM 模型。

各指标最优模型评价从模型预测性能与稳健性两个方面着手。从预测性能来看，除了 O 谱的 O═C—O（R^2 为 0.839，精度稍低，有待优化提升），其余

表 4 基于 IPLS 的生物炭表面碳氧元素及基团的定量回归结果（L2 正则化前）

Tab 4 Quantitative regression results of carbon and oxygen elements and groups on biochar surface based on IPLS (before L2 regularization)

预测目标	模型	测试集		验证集	
		R^2	RMSE/%	$RMSE_{CV}$/%	MAE_{CV}/%
S_C	IPLS + RF	0.974	0.345	0.404	0.501
	IPLS + SVM	0.978	0.313	0.380	0.480
S_O	IPLS + RF	0.981	0.286	0.610	0.638
	IPLS + SVM	0.981	0.277	0.522	0.591
C_C═C	IPLS + RF	0.972	0.440	1.504	0.915
	IPLS + SVM	0.996	0.181	0.201	0.314
C_C—C	IPLS + RF	0.915	0.674	2.502	1.344
	IPLS + SVM	0.918	0.658	3.475	1.574
C_C—O	IPLS + RF	0.959	0.262	0.579	0.568
	IPLS + SVM	0.912	0.361	0.664	0.553
C_C═O	IPLS + RF	0.904	0.339	0.332	0.454
	IPLS + SVM	0.982	0.160	0.104	0.270
C_O═C—O	IPLS + RF	0.940	0.448	1.106	0.835
	IPLS + SVM	0.973	0.302	0.795	0.652
O_C═O	IPLS + RF	0.970	0.507	3.386	1.350
	IPLS + SVM	0.941	0.669	7.015	1.914
O_C—O	IPLS + RF	0.913	0.786	3.162	1.541
	IPLS + SVM	0.971	0.499	4.803	1.763
O_O═C—O	IPLS + RF	0.784	0.549	1.098	0.837
	IPLS + SVM	0.862	0.444	0.664	0.576

9 个预测指标最优模型决定系数 R^2 均大于 0.9，其中以表面 C 元素与表面 O 元素的预测精度最高；在误差方面，所有模型的均方根误差 RMSE 都小于 1，尤其是表面 C、表面 O、C 谱 C═C、C 谱 C—O、C 谱 C═O、C 谱 O═C—O 都低于 0.3。由于 RMSE 是先对误差进行平方的累加再开方，放大了最大误差之间的差距，因此 RMSE 越小，其意义越大。模型 RMSE 整体都比较小，说明模型对数据的拟合程度比较好，预测误差较低。

模型稳健性能方面，对所有模型进行五折交叉验证。验证结果表明：表面 C（$RMSE_{CV}$ 为 0.388%，MAE_{CV} 为 0.503%），表面 O（$RMSE_{CV}$ 为 0.610%，MAE_{CV} 为 0.638%），C 谱 C—O（$RMSE_{CV}$ 为 0.549%，MAE_{CV} 为 0.561%），C 谱 C═O（$RMSE_{CV}$ 为 0.144%，MAE_{CV} 为 0.315%）交叉验证后，交叉均方根误差和交叉平均绝对误差依旧保持在较低的水平，证明模型稳健性与泛化能力很高；但是 C 谱 C═C（$RMSE_{CV}$ 为 2.311%，MAE_{CV} 为 1.082%），C 谱 C—C（$RMSE_{CV}$ 为 2.656%，MAE_{CV} 为 1.351%），C 谱 O═C—O（$RMSE_{CV}$ 为 1.285%，MAE_{CV} 为 0.827%），O 谱 C═O（$RMSE_{CV}$ 为 3.502%，MAE_{CV} 为 1.338%），O 谱 C—O（$RMSE_{CV}$ 为 6.895%，MAE_{CV} 为 2.119%），O 谱 O═C—O（$RMSE_{CV}$ 为 1.027%，MAE_{CV} 为 0.759%），交叉验证前后，误差差异较大，模型的稳健性较差。究其原因，主要是因为：①建模前数据集划分采用训练集占比 70% 的随机划分方式，相比于 KS 与 SPXY 划分而言，误差会较大，但能反映数据真实分布情况，构建的模型更具有泛化性。②所用数据集偏小。③稳健性较差的模型，训练集与测试集中的输入变量具有不同的数据分布，协变量偏移较大，训练集与测试集的数据分布不满足独立分布，导致在不同的划分数据集中，模型性能存在较大区别[31]。

在无法增加训练数据的情况下，选择对上述过拟合的 6 个预测目标的模型进行 L2 正则化。正则化后结果如表 5 所示，对比表 4、5 中 C 谱 C═C、C 谱 C—C、C 谱 O═C—O、O 谱 C═O、O 谱 C—O、O 谱 O═C—O 正则化前后的评价指标，可发现，经过 L2 正则化，模型的严重过拟合得到了较大地缓解，但是过拟合的现象没有得到完全解决。

2.3.2 基于 PCA 的中红外光谱建模效果与评价

经过 PCA 特征波长筛选、超参数优化、交叉验证，得到基于 PCA 的生物炭表面碳氧元素及碳氧基团的模型效果如表 5 所示。从整体模型效果来看，除了 C_C═C 外，其余预测目标的 PCA + SVM 模型都要优于 PCA + RF 模型，前者不仅预测精度更高，而且交叉均方根误差、交叉平均绝对误差更小，说明 PCA + SVM 模型稳定性与泛化性能更好。

2.3.3 各类方法最优预测精度对比

为了筛选出 10 个预测目标的最优模型，汇总各类方法的最优性能指标及相关参数如表 5 所示，其中 C 谱 C═C、C 谱 C—C、C 谱 O═C—O、O 谱 C═O、O 谱 C—O、O 谱 O═C—O 这 6 个预测目标的所有模型由于过拟合已经全部进行了 L2 正则化处理。对 S_C 而言，基于 IPLS 特征波段筛选的两种模型，不仅预测性能优异，均方根误差在测试集与验证集中也没有明显差异，说明模型有很好的泛化能力。同样地，S_O 的 4 个模型中，基于 PCA 特征波长筛选的两个模型在整体性能上要更为突出，其中 PCA + SVM 模型效果最佳。其余预测目标的最优模型均如表 5 所示。

从表 5 中可知，所有最优模型均集中在 IPLS + SVM 和 PCA + SVM 之间，说明相较于 RF 模型，SVM 模型性能更高，证实了 SVM 模型在小样本建模中更具优势。同时可发现，除了 S_C、S_O、C_C═C、C_C—O、C_C═O，其余 5 个预测目标的交叉均方根误差在测试集均方根误差的基础上明显增大，表明模

表 5　生物炭表面碳氧元素及基团的定量回归结果

Tab. 5　Quantitative regression results of carbon and oxygen elements and groups on biochar surface

预测目标	模型	最优特征维度	测试集		验证集	
			R^2	RMSE	$RMSE_{CV}$	MAE_{CV}
S_C	IPLS + RF		0.974	0.345	0.404	0.501
	IPLS + SVM		0.978	0.313	0.380	0.480
	PCA + RF	2	0.915	0.636	1.114	0.673
	PCA + SVM	6	0.975	0.410	1.08	0.718
S_O	IPLS + RF		0.981	0.286	0.610	0.638
	IPLS + SVM		0.981	0.277	0.522	0.591
	PCA + RF	11	0.976	0.334	0.464	0.478
	PCA + SVM	11	0.981	0.301	0.385	0.387
C_C═C*	IPLS + RF		0.962	0.627	1.439	0.731
	IPLS + SVM		0.994	0.187	0.198	0.316
	PCA + RF	6	0.964	0.513	1.113	0.731
	PCA + SVM	4	0.968	0.504	1.081	0.722
C_C—C*	IPLS + RF		0.903	0.968	2.372	1.113
	IPLS + SVM		0.905	0.971	3.384	1.356
	PCA + RF	7	0.869	1.107	3.764	1.579
	PCA + SVM	7	0.938	0.874	2.015	0.914
C_C—O	IPLS + RF		0.959	0.262	0.579	0.568
	IPLS + SVM		0.912	0.361	0.664	0.553
	PCA + RF	5	0.937	0.313	0.327	0.448
	PCA + SVM	6	0.986	0.172	0.162	0.269
C_C═O	IPLS + RF		0.904	0.339	0.332	0.454
	IPLS + SVM		0.982	0.160	0.104	0.270
	PCA + RF	7	0.964	0.216	0.152	0.300
	PCA + SVM	8	0.964	0.218	0.145	0.299
C_O═C—O*	IPLS + RF		0.934	0.531	1.117	0.858
	IPLS + SVM		0.966	0.372	0.781	0.643
	PCA + RF	8	0.934	0.554	1.137	0.683
	PCA + SVM	7	0.962	0.435	1.074	0.620
O_C═O*	IPLS + RF		0.957	0.686	3.027	1.273
	IPLS + SVM		0.932	0.774	3.524	1.803
	PCA + RF	5	0.943	0.818	3.146	1.230
	PCA + SVM	9	0.958	0.985	2.057	1.122
O_C—O*	IPLS + RF		0.907	0.925	3.014	1.461
	IPLS + SVM		0.954	0.837	3.167	1.643
	PCA + RF	5	0.895	1.246	2.214	1.037
	PCA + SVM	7	0.962	0.967	1.661	0.829
O_O═C—O*	IPLS + RF		0.771	0.656	1.034	0.631
	IPLS + SVM		0.846	0.562	0.627	0.438
	PCA + RF	17	0.854	0.513	0.557	0.436
	PCA + SVM	8	0.931	0.357	0.461	0.425

注：* 表示该预测目标的所有模型均进行了 L2 正则化处理。

型的泛化能力较差。模型的训练需要大量的样本数据，而本文的样本总量仅为 111，属于小样本建模。而小样本建模的可靠性与泛化能力一直是机器学习领域共同关注的问题[32]。

从图 5 以及表 5 中，可以发现某些模型训练集 R^2 远大于测试集，尤其是 C_C—C、C_O ═C—O、O_C ═O和 O_C—O 的最优模型经过 L2 正则化还是存在过拟合的问题。在排除模型复杂度过高，特征数据维度过高的原因后，基本可以确定是建模数据量较小的问题。根据以往的同类文献的研究结果，当样本数量在数百个时有望达到较好的效果。

3　结论

（1）生物炭表面 C 元素的赋存形态以 C_C ═C 和 O_C—O 为主，表面 O 元素赋存形态集中在 C_C—O、C_C ═O、C_O ═C—O 以及 O_C ═O 上。

（2）由 IPLS 建模结果可知，特征波段 4 000 ~ 3 464 cm^{-1}和 1 588 ~650 cm^{-1}均包含了与生物炭表面碳氧元素含量及其赋存形态高度相关的特征信息，但 1 588 ~650 cm^{-1}囊括的信息丰度更高。

（3）对 10 个预测指标构建了 4 种模型，包括 IPLS + RF、IPLS + SVM、PCA + RF、PCA + SVM，其中 IPLS + SVM 和 PCA + SVM 的整体性能尤为突出。从模型的预测精度来看，所建最优模型对 S_C、S_O、C_C ═C、C_C—C、C_C—O、C_C ═O、C_O ═C—O、O_C ═O、O_C—O、O_O ═C—O 都具有良好的定量分析能力（$R^2 > 0.93$），但是 C_C—C、C_O ═C—O、O_C ═O、O_C—O 的最优预测模型交叉验证前后，RMSE 的变化较大，说明模型稳健性与泛化能力较差。

参　考　文　献

[1] CHEN X, ZHAO Y, YANG L, et al. Identifying the specific pathways to improve nitrogen fixation of different straw biochar during chicken manure composting based on its impact on the microbial community[J]. Waste Management, 2023, 170: 8 - 16.

[2] 沈秀丽，燕海朋，曾剑飞，等. 畜禽粪便生物炭内源重金属在酸性土壤中的迁移转化[J]. 农业工程学报，2022，38(8)：209 - 217.
SHEN Xiuli, YAN Haipeng, ZENG Jianfei, et al. Migration and transformation of endogenous heavy metals from animal manure biochar in acid soil[J]. Transactions of the CSAE, 2022, 38(8): 209 - 217. (in Chinese)

[3] 付强，石净，李天霄，等. 不同调控模式下寒区土壤物理结构与水力特性改良研究[J]. 农业机械学报，2023，54(9)：374 - 385.
FU Qiang, SHI Jing, LI Tianxiao, et al. Improvement of soil physical structure and hydraulic characteristics in cold regions by different regulation modes[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(9): 374 - 385. (in Chinese)

[4] 马腾，郝彦辉，姚宗路，等. 秸秆水热生物炭燃烧特性评价[J]. 农业机械学报，2018，49(12)：340 - 346.
MA Teng, HAO Yanhui, YAO Zonglu, et al. Evaluation on combustion characteristics of straw hydrothermal bio-char[J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(12): 340 - 346. (in Chinese)

[5] 徐洋. 生物炭基油菜抗倒生长调节剂成粒与缓释协同调控研究[D]. 武汉：华中农业大学，2022.
XU Yang. Synergistic regulation of granulation and sustained release of biochar-based rape lodging-resistant growth regulator [D]. Wuhan: Huazhong Agricultural University, 2022. (in Chinese)

[6] QI X, XIAO S, CHEN X, et al. Biochar-based microbial agent reduces U and Cd accumulation in vegetables and improves rhizosphere microecology[J]. Journal of Hazardous Materials, 2022, 436: 129147.

[7] ALHAJERI N S, TAWFIK A. Integrating biochar and microbial community for detoxification of wastewater industry containing analgesics[J]. Journal of Water Process Engineering, 2024, 58: 104767.
[8] SZEWCZUK-KARPISZ K, WIŚNIEWSKA M, NOWICKI P, et al. Influence of protein internal stability on its removal mechanism from aqueous solutions using eco-friendly horsetail herb-based engineered biochar[J]. Chemical Engineering Journal, 2020, 388: 124156.
[9] CAO D, JI Y, LIU L, et al. A facile and green strategy to synthesize N/P co-doped bio-char as VOCs adsorbent: through efficient biogas slurry treatment and struvite transform[J]. Fuel, 2022, 322: 124156.
[10] SBIZZARO M, CÉSAR SAMPAIO S, RINALDO DOS REIS R, et al. Effect of production temperature in biochar properties from bamboo culm and its influences on atrazine adsorption from aqueous systems[J]. Journal of Molecular Liquids, 2021, 343: 117667.
[11] RINCÓN PRAT S, SCHNEIDER C, KOLB T. Determination of active sites during gasification of biomass char with CO_2 using temperature-programmed desorption. Part 2: influence of ash components[J]. Fuel, 2020, 267: 117179.
[12] SHARD A G. Practical guides for x-ray photoelectron spectroscopy: quantitative XPS[J]. Journal of Vacuum Science & Technology A, 2020, 38(4): 041201.
[13] 王巧云,单鹏. 分子光谱检测及数据处理技术[M]. 北京:科学出版社, 2019.
[14] MUNAWAR A A, ZULFAHRIZAL, MEILINA H, et al. Near infrared spectroscopy as a fast and non-destructive technique for total acidity prediction of intact mango: comparison among regression approaches[J]. Computers and Electronics in Agriculture, 2022, 193: 106657.
[15] 杨芳, 刘朝霞, 牛文娟, 等. 基于 FT-MIR 的秸秆炭热值快速检测方法[J]. 华中农业大学学报(自然科学版),2017, 36(6): 121-126.
YANG Fang, LIU Zhaoxia, NIU Wenjuan, et al. Rapid detection of calorific value in straw biochar based on FT-MIR[J]. Journal of Huazhong Agricultural University, 2017, 36(6): 121-126. (in Chinese)
[16] MA J, GARG A, ZHONG F, et al. Coupling behavior and enhancement mechanism of porous structure, graphite microcrystals, and oxygen-containing groups of activated biochar for the adsorption of phenol[J]. Environmental Science: Water Research & Technology, 2023, 9(7): 1944-1957.
[17] ZHENG X, CHEN L, LI X, et al. Non-destructive detection of meat quality based on multiple spectral dimension reduction methods by near-infrared spectroscopy[J]. Foods, 2023, 12(2): 300.
[18] CRUZ-TIRADO J P, AMIGO J M, BARBIN D F. Determination of protein content in single black fly soldier (*Hermetia illucens* L.) larvae by near infrared hyperspectral imaging (NIR-HSI) and chemometrics[J]. Food Control, 2023, 143: 109266.
[19] WANG M, YAN Z, LUO J, et al. A band selection approach based on wavelet support vector machine ensemble model and membrane whale optimization algorithm for hyperspectral image[J]. Applied Intelligence, 2021, 51(11): 7766-7780.
[20] GHOSH D, CABRERA J. Enriched random forest for high dimensional genomic data[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022, 19(5): 2817-2828.
[21] LI Y, LI M, ZHANG L. Evolutionary polynomial regression improved by regularization methods[J]. PLOS ONE, 2023, 18(2): e0282029.
[22] CAI M, WAN M, YANG G, et al. Structure preserving projections learning via low-rank embedding for image classification[J]. Information Sciences, 2023, 648: 119636.
[23] 牛文娟, 任鲁娜, 邓继猛, 等. 活化氧化竹基多孔活性炭理化结构与电容性能[J]. 农业工程学报, 2023, 39(19): 221-31.
NIU Wenjuan, REN Lu'na, DENG Jimeng, et al. Physicochemical structure and capacitive properties of activated oxidized bamboo-based porous activated carbon[J]. Transactions of the CSAE, 2023, 39(19): 221-231. (in Chinese)
[24] CAO S, DUAN F, WANG P, et al. Biochar contribution in biomass reburning technology and transformation mechanism of its nitrogen foundational groups at different oxygen contents[J]. Energy, 2018, 155: 272-280.
[25] FENG Y, QIU X, TAO Z, et al. Oxygen-containing groups in cellulose and lignin biochar: their roles in U(VI) adsorption[J]. Environmental Science and Pollution Research, 2022, 29(51): 76728-76738.
[26] LI C, FENG Y, ZHONG F, et al. Optimization of microwave-assisted hydrothermal carbonization and potassium bicarbonate activation on the structure and electrochemical characteristics of crop straw-derived biochar[J]. Journal of Energy Storage, 2022, 55: 105838.
[27] YU S, WU L, NI J, et al. The chemical compositions and carbon structures of pine sawdust- and wheat straw-derived biochars produced in air-limitation, carbon dioxide, and nitrogen atmospheres, and their variation with charring temperature[J]. Fuel, 2022, 315: 122852.
[28] YANG F, ZUO X, YANG H, et al. Ionic liquid-assisted production of high-porosity biochar with more surface functional groups: taking cellulose as attacking target[J]. Chemical Engineering Journal, 2022, 433: 133811.
[29] ZHU C, JIANG H, CHEN Q. High precisive prediction of aflatoxin B1 in pressing peanut oil using raman spectra combined with multivariate data analysis[J]. Foods, 2022, 11(11): 1565.
[30] 宋彦, 汪小中, 赵磊, 等. 基于近红外光谱技术的眉茶拼配比例预测方法[J]. 农业工程学报, 2022, 38(2): 307-315.
SONG Yan, WANG Xiaozhong, ZHAO Lei, et al. Predicting the blending ratio of Mee Tea based on near infrared spectroscopy[J]. Transactions of the CSAE, 2022, 38(2): 307-315. (in Chinese)
[31] FDEZ-DÍAZ L, GLEZ-TOMILLO S, MONTAÑÉS E, et al. Improving importance estimation in covariate shift for providing accurate prediction error[J]. Expert Systems with Applications, 2022, 193: 116376.
[32] CAI B, SHENG C, GAO C, et al. Artificial intelligence enhanced reliability assessment methodology with small samples[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(9): 6578-6590.