

胡鹏飞,唐荣年,胡文锋,等. 基于近红外光谱与双权重竞争特征搜索策略的橡胶树叶片氮素检测[J]. 江苏农业科学,2024,52(18):222-231.
doi:10.15889/j.issn.1002-1302.2024.18.029

基于近红外光谱与双权重竞争特征搜索策略的橡胶树叶片氮素检测

胡鹏飞,唐荣年,胡文锋,李 创

(海南大学机电工程学院,海南海口 570228)

摘要:本研究旨在解决传统近红外光谱分析在橡胶树叶片氮含量(LNC)检测中模型精度和稳定性的局限。通过对180张橡胶树叶片进行定量分析,提出了一种改进的重加权采样算法,即双权重竞争性自适应重加权采样(DWCARS)。该方法综合运用回归系数和变量投影重要性(VIP)2种权重评价标准,并通过竞争性机制优化特征选择。比较分析结果表明,与传统CARS和差分进化(DE)等方法相比,DWCARS能够提取出更少且预测精度更高的波长变量。在测试集上,DWCARS模型展现了显著性能优势,其决定系数(R_p^2)为0.9367,均方根误差($RMSEP$)为0.1215,相比于CARS算法建立的预测模型 $RMSEP$ 值降低了21.66%。表明DWCARS算法在提高橡胶树叶片氮含量检测的准确性和稳定性方面表现卓越,适用于精确监测橡胶树生长阶段的氮素状况。

关键词:近红外光谱;橡胶树;机器学习;光谱波段选择;叶片氮含量;DWCARS

中图分类号:S127 **文献标志码:**A **文章编号:**1002-1302(2024)18-0222-10

天然橡胶是一种重要的生物聚合物,在国民经济和社会发展中发挥重要作用。随着全球对天然橡胶需求的增长,实时监测橡胶树的养分信息已成为植胶区管理的迫切需求^[1]。叶片氮含量(LNC)作为橡胶树养分丰缺情况的敏感指标,通过实时监测橡胶树叶片氮含量,可有效获取橡胶树的生长信息^[2]。因此,建立一种高效准确的橡胶树叶片氮含量检测模型意义重大。

传统的叶氮含量检测方法如凯氏定氮法,由于复杂的化学试剂和繁琐的操作过程,无法满足养分实时监测的需求^[3]。近红外光谱技术因其快速、无损且低成本的特点已在叶片养分含量的检测中广泛应用^[4]。Li等采用偏最小二乘回归模型对湿地松叶片氮含量进行定量分析,选择了与植物组织中氮含量高度相关的5个波段范围^[5];Liu等通过多

种光谱预处理方法相结合,有效消除了光谱数据中的噪声,得到了针对不同干旱胁迫条件下的香椿叶片氮含量估测模型^[6]。然而,随着现代光谱仪器的进步,测量数据量大大提高且具有较高的共线性^[7]。因此,在波长变量与响应浓度之间建立可解释关系,成为提高光谱检测模型性能的关键任务^[8]。

竞争自适应重加权采样(CARS)算法作为一种高效的特征波段选择方法,表现出出色的全局搜索能力,相较于传统的光谱区间选择算法而言更具优势^[9]。在Zhang等的研究中,通过将CARS与随机森林(RF)算法结合,有效地确定了番茄可溶性固形物含量的有效波长,结果表明CARS算法所选波段在性能方面表现最佳且特征数量较少^[10]。然而,CARS算法采用一种竞争性机制来选择和更新模型的权重参数,此机制受到模型中蒙特卡洛采样(MCS)和自适应重加权(ARS)2个随机因素的影响^[11],导致不同特征组合在不同迭代中获胜,可能导致在变量剔除中丢失重要波段^[12]。因此,为了构建更加稳定且具有更高精度的检测模型,需要采用更为有效的特征变量选择策略。

随着特征变量选择方法的不断发展,出现了多种用于测定变量重要性的标准,包括回归系数、变量投影重要性以及选择性比(SR)参数等。由于不

收稿日期:2023-09-20

基金项目:海南省自然科学基金创新研究团队项目(编号:320CXTD431);国家自然科学基金(编号:32060413);海南省重点研发计划(编号:ZDYF2022GXJS008);海南省自然科学基金高层次人才项目(编号:321RC468)。

作者简介:胡鹏飞(1996—),男,山东潍坊人,硕士研究生,主要从事植物养分无损检测研究。E-mail:hpf@hainanu.edu.cn。

通信作者:李 创,博士,教授,博士生导师,主要从事高光谱遥感技术研究。E-mail:lc@hainanu.edu.cn。

同参数对于变量的解释能力各不相同,因此单一评价标准在解释多特征数据方面存在一定局限性^[13]。因此,本研究提出双权重竞争性自适应重加权采样(DWCARS)算法,在同一轮蒙特卡洛采样中引入了2种权重进行波段重要性评价。通过自适应重加权的竞争机制,充分利用这2种权重指标对变量的解释能力,实现了特征波长的有效选择。

本研究探讨了近红外光谱分析技术与化学计量学方法的结合,以实现橡胶树叶片氮含量的快速检测。通过引入DWCARS算法来选择叶片氮素的特征波长,成功地解决了波段选择过程中的冗余波长问题。本研究旨在获取与LNC高度相关的近红外光(NIR)特征波长,并建立高效、准确的橡胶树

LNC快速检测模型。

1 材料与方法

1.1 研究区概况与样本采集

本试验所使用的叶片样本,于2022年4月在海南省儋州市(109.20°~109.70°E,19.40°~19.65°N)的橡胶林试验田内采集,采样区域如图1所示。儋州市属海洋性热带季风气候,雨热资源丰富。年平均降水量1 828.7 mm,年平均气温23.8℃。供试材料为中国热带农业科学院选育的巴西橡胶优良品种热研7-33-97的叶片。随机采集了200份成熟、健康、完整的橡胶树冠层叶片,并进行密封低温保存,以备后续的光谱数据采集和理化分析。

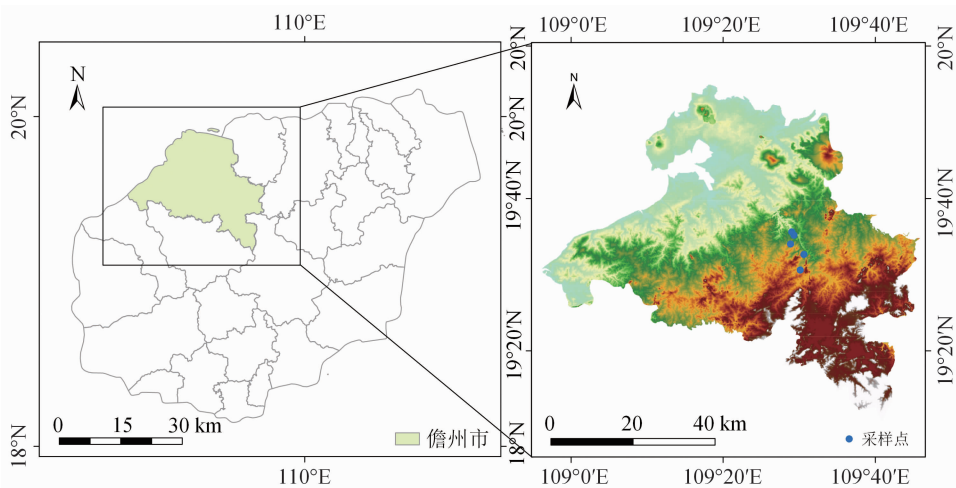


图1 研究区域在海南省儋州市的位置

1.2 光谱数据采集和叶片氮含量分析

1.2.1 光谱数据采集与数据预处理 使用GaiaField-F-N17E高光谱成像仪对叶片样本进行图像采集扫描,波段范围为942~1 701 nm。经清洁处理后,将样本放在光谱仪的传送带上,通过光谱相机捕获样本的反射光,得到一维影像和光谱信息。传送带带动样本进行线扫描,最终计算机软件记录到扫描行程中包括叶片区域和非叶片区域内所有物体的高光谱图像。在室内温度和湿度稳定的条件下,取3次扫描的平均值作为样本的原始光谱。原始光谱的采样点之间的距离平均为3.3 nm,共有224个波长变量。选取原始光谱图像中的叶片区域作为感兴趣区域,计算叶片区域光谱数据的平均值,获得平均光谱。为了去除噪声和增强光谱信息,采用了3种预处理方法:多元散射校正(MSC)、去趋势(DT)和一阶导数变换(D1),同时对它们在模型预测中的提升效果进行分析和对比。经预处理

后的光谱曲线如图2所示。

1.2.2 异常数据剔除 为减少人为误差或仪器误差对模型性能的影响,采用主成分分析(PCA)结合马氏距离(MD)的方法(PCA-MD),以剔除可能对模型产生强影响的极端样本或异常样本。由PCA方法得到光谱的主成分和得分,使用得分数据代替原始数据计算马氏距离,以反映全谱数据信息并压缩计算中涉及的变量数,同时保证 \mathbf{M} 矩阵不存在共线性问题^[14]。其中样品到平均光谱的马氏距离为:

$$D_i = \sqrt{(T_i - T)\mathbf{M}^{-1}(T_i - T)'} \quad (1)$$

式中: D_i 为校正样品*i*的马氏距离; T_i 为校正样品*i*的光谱得分; T 为*n*个样品的平均光谱; \mathbf{M} 为校正样品光谱的马氏矩阵。

根据马氏距离的结果图进行颜色编码,可以得到叶片光谱数据的2个主成分的马氏距离分布图(图3)。在该分布中,距离分布中心 \bar{D} 阈值范围内的样品被认为与样品平均光谱相似。 $D_i - \bar{D}_i$ 值越小

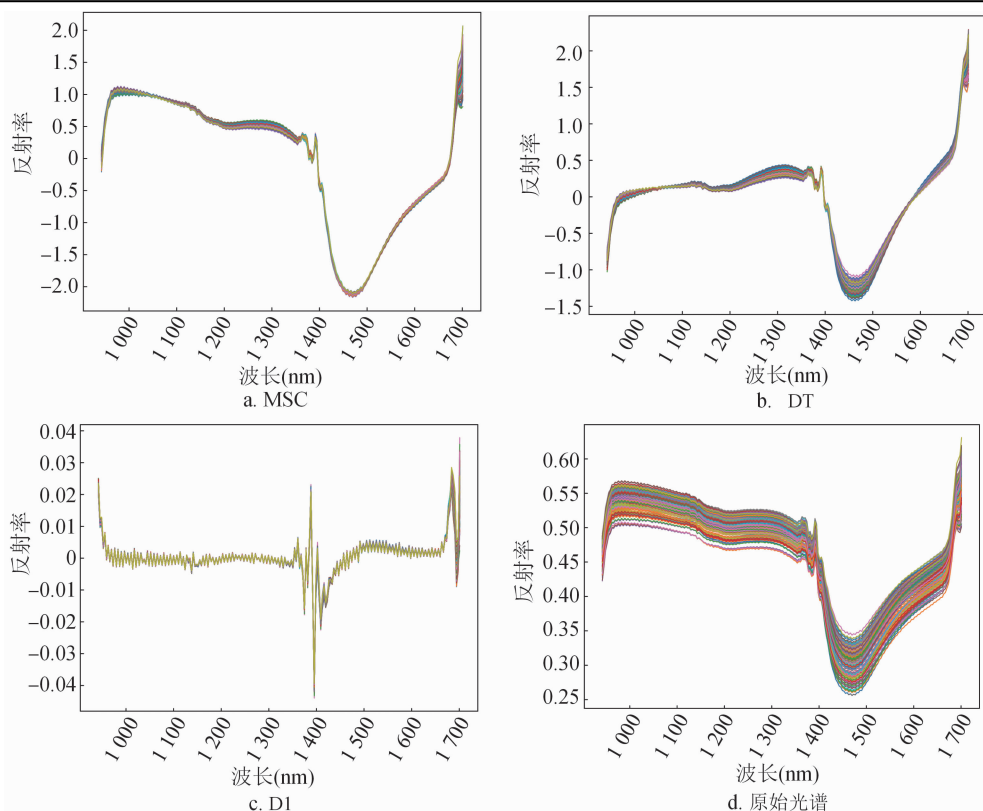


图2 预处理后的光谱曲线

则有越高的相似度,反之则可能为异常样本。根据分布情况,从180个原始样本中剔除了18个异常样本,最终得到162个叶片光谱数据。

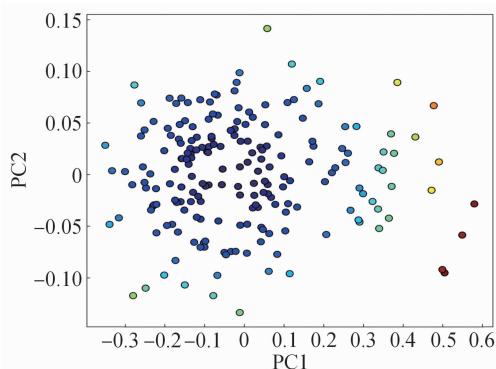


图3 叶片光谱数据的马氏距离分布

1.2.3 叶片氮含量理化分析 凯氏定氮法是用干法测定叶片全氮含量的常规方法,具有较高的准确度。本研究采用凯氏定氮法对橡胶树叶片中氮元素含量进行检测。在完成叶片光谱数据采集后,对样本立即进行烘干研磨,并置于定氮瓶中进行分析测试,最终测得的叶片样本氮元素含量介于2.1%~4.7%之间,呈正态分布。

1.3 特征波长选择策略

1.3.1 基于PLS回归系数与变量投影重要性的变

量评价 偏最小二乘回归(PLSR)是应对近红外光谱中多特征少样本数据问题的常见建模方法。PLSR从输入特征 X 和标签值 Y 中同时提取主成分信息,由此产生了多个变量评价指标,包括回归系数、变量投影重要性、选择性比等。回归系数向量 β 的计算公式如下:

$$\beta = w(q^T q)^{-1} q^T t y^T. \quad (2)$$

式中: q 为载荷矩阵; t 为潜在变量矩阵; y 为响应变量矩阵; w 为权重矩阵。

回归系数向量是模型中所有分解得到的潜在变量的函数,它反映了每个自变量 X_i 对因变量的贡献,较大的回归系数表示相关主成分对因变量有更大的贡献^[15]。变量投影重要性VIP得分由以下公式计算:

$$VIP_j = \sqrt{\frac{m \sum_{k=1}^h \left[q_k^2 t_k^T t_k \left(\frac{w_{jk}}{\|w_{jk}\|} \right)^2 \right]}{\sum_{k=1}^h q_k^2 t_k^T t_k}}. \quad (3)$$

式中: j 代表波长变量的索引; m 表示波长变量的总数; h 为PLS的最佳主因子数量。

VIP得分综合考虑了光谱对构建PLS得分的贡献和PLS得分对浓度变量的解释能力,代表了波长

变量在模型拟合中的重要性^[16]。

通过分析各个波段的得分矩阵、载荷矩阵等信息,可以计算出波段的回归系数 VIP 得分。图 4 展示了光谱各个波段对应的回归系数和 VIP 得分。本研究基于竞争性重加权采样策略,对回归系数与变量投影重要性 2 种用于解释变量重要性的指标进行比较,并提出了双权重竞争的变量选择方法。

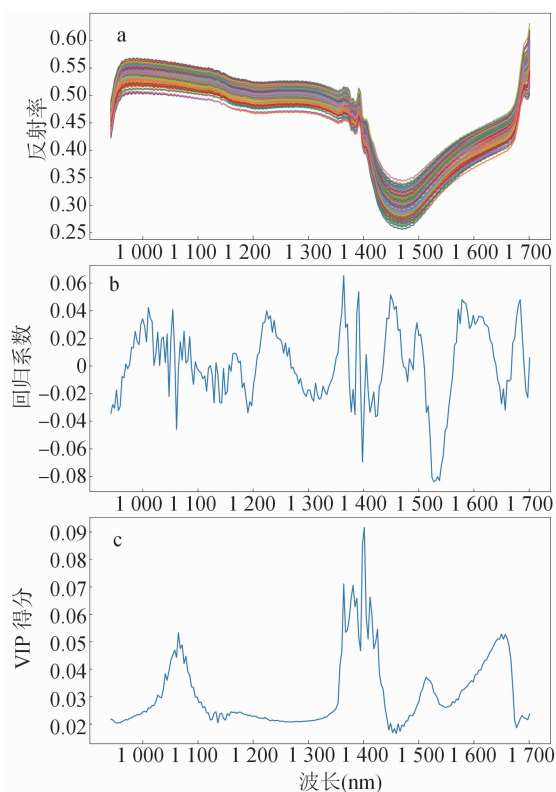


图4 光谱波长变量对应的回归系数与VIP得分

1.3.2 单权重竞争性自适应重加权采样 (VIP - CARS) 方法 尽管 CARS 默认使用回归系数作为变量重要性的评价指标,但本研究引入了一种新的波长选择方法,即 VIP - CARS 方法。该方法将单一权重的回归系数替换为单一权重的变量投影重要性得分,用于波长选择的试验研究。

1.3.3 双权重竞争自适应重加权采样 本研究提出了一种基于双权重竞争的自适应重加权采样方法,用于特征波段选择。DWCARS 的核心思想是每轮蒙特卡洛采样中,通过同时考虑回归系数与 VIP 得分 2 个权重,分别对波段变量进行重要性评估,并生成 2 个不同的权重排序。然后,利用指数衰减函数和自适应重加权采样,对这 2 个不同的排序进行变量选择,最终得到 2 个不同的波段子集,并进行对比选择最佳结果。具体流程如下。

(1) 计算 2 个权重。首先,确定蒙特卡洛采样法的采样次数 N 。每次采样随机选择一定比例的样本进行 PLS 建模,记录每次采样过程中 PLS 模型的回归系数向量 \mathbf{b} 和 VIP 得分向量 \mathbf{v} ,并进行权重计算。其中,回归系数权重 w_{bi} 和 VIP 得分权重 w_{vi} 的计算公式如下:

$$w_{bi} = \frac{|\mathbf{b}_i|}{\sum_{i=1}^p |\mathbf{b}_i|}; \quad (4)$$

$$w_{vi} = \frac{\mathbf{v}_i}{\sum_{i=1}^p \mathbf{v}_i}。 \quad (5)$$

式中: p 表示第 i 次采样后剩余的特征波段个数。

(2) 指数衰减去除波长。通过指数衰减函数 (EDF),去除回归系数绝对值权重相对较小的波长,并根据 EDF 计算得到保留的波长数量的比例。

(3) 自适应加权采样。每次采样时,采用自适应加权采样策略,分别使用回归系数权重 w_{bi} 和 VIP 得分权重 w_{vi} ,从上一次采样选择出的特征波段集合中选择 2 个数量相等的波长变量子集。然后,分别使用这 2 个变量子集进行 PLS 建模,并计算对应的交叉验证均方根误差 ($RMSECV$),分别记为 $RMSECV_b$ 和 $RMSECV_v$ 。

(4) 选择最佳子集。比较 $RMSECV_b$ 和 $RMSECV_v$ 的大小。若 $RMSECV_b < RMSECV_v$,则将 $RMSECV_b$ 对应的波段子集作为本轮的最佳变量子集。反之,若 $RMSECV_b > RMSECV_v$,则将 $RMSECV_v$ 对应的波段子集作为本轮最佳变量。通过二者比较,得到本轮最终最佳子集 V_{new} 和误差 $RMSECV$ 。

(5) 选择最终结果。在 N 次采样结束后,得到 N 个变量子集和对应的 N 个 $RMSECV$ 。选择 $RMSECV$ 最小值所对应的特征波段子集作为最终的结果。

DWCARS 方法通过引入 VIP 得分作为额外权重,增加了回归系数与 VIP 得分之间的竞争机制,从而更全面地考虑了波段的重要性。详细的算法流程见图 5。

2 结果与分析

2.1 数据集划分

在建立光谱数据和叶片氮含量建模之前,将 162 个样本按 3 : 1 的比例划分为训练集 (121 个样本) 和测试集 (41 个样本)。表 1 呈现了橡胶树叶片氮含量的训练集和测试集统计信息,其中训练集和

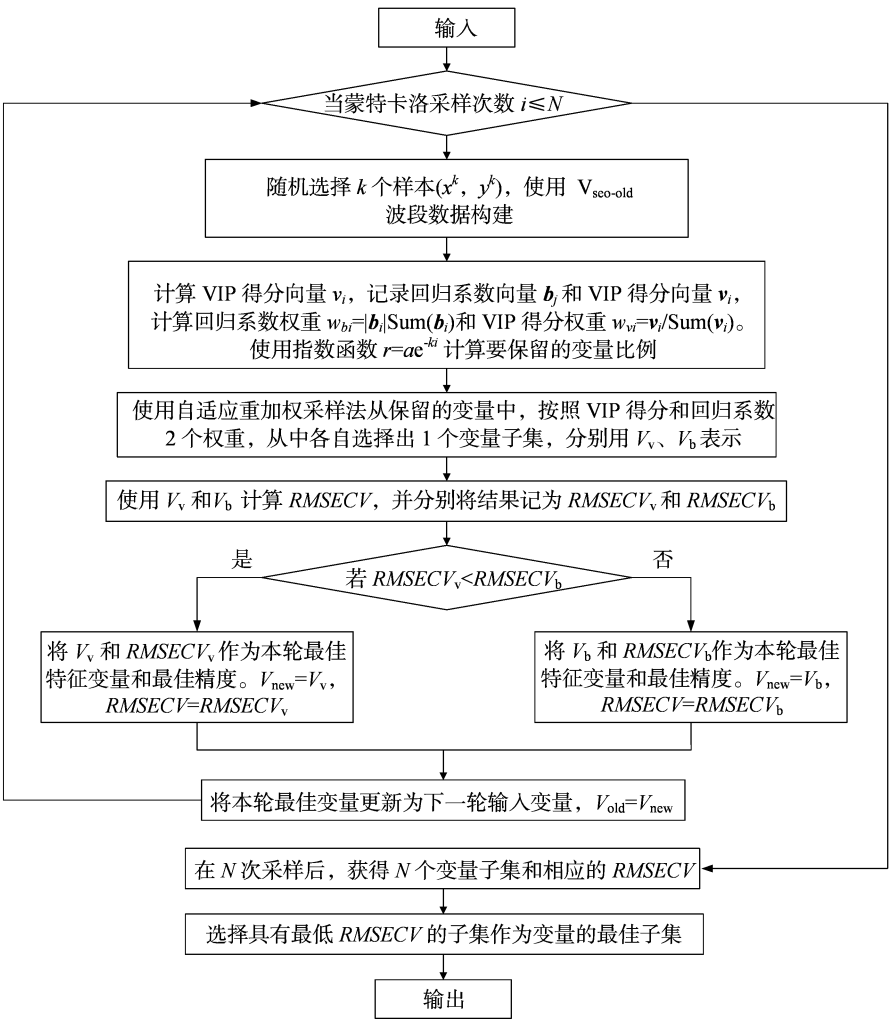


图5 DW-CARS 算法流程

测试集的氮含量标准差分别为 0.415% 和 0.483%，变异系数分别为 12.97% 和 14.55%。表 1 中的数据表明,训练集和测试集的样本分布相对均匀,因此,这种样品划分方法适合光谱分析,可用于建立橡胶树叶片氮含量回归校正模型。

表 1 数据集划分情况

数据集划分	样本数量 (个)	氮含量 (%)				
		最小值	最大值	平均值	标准差	变异系数
总数	162	2.056	4.281	3.229	0.436	13.50
训练集	121	2.283	4.153	3.199	0.415	12.97
测试集	41	2.056	4.281	3.319	0.483	14.55

2.2 光谱预处理结果分析

应用 PCA - MD 方法剔除异常样本后,模型精度明显提高。从表 2 可以看出,DT 方法预处理后的光谱模型精度较低,可能因为 DT 方法增强光谱数据的同时也干扰了与氮素相关的重要信息。相比

其他的预处理方法,MSC 处理后的光谱模型效果最佳,表现为 $R_p^2 = 0.696\ 3$, $RMSE_p = 0.260\ 5$ 。这表明 MSC 预处理方法有效地减少了原始光谱中仪器基线漂移、散射和信号噪声等干扰信息。因此,在后续的数据分析中,采用经过 MSC 预处理的光谱数据。

2.3 特征波段选择

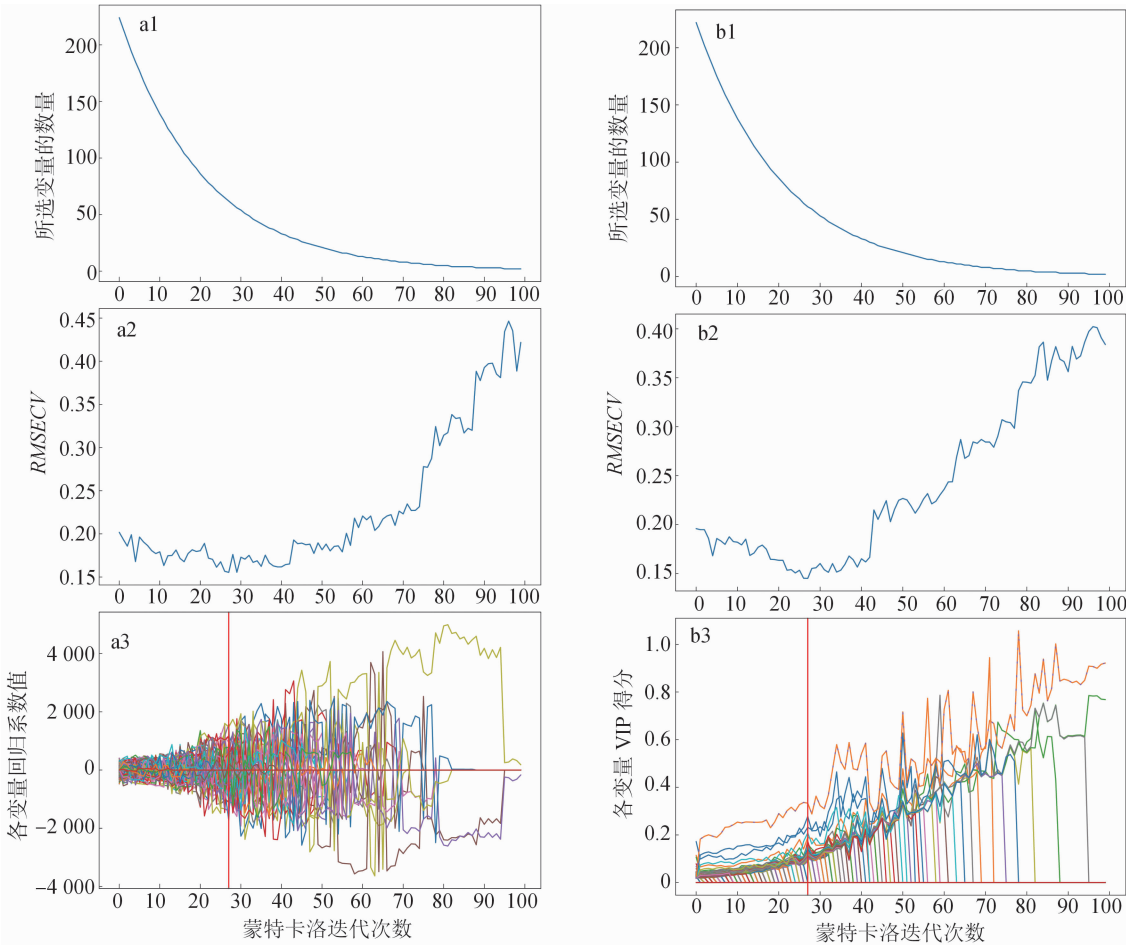
2.3.1 CARS 与 VIP - CARS 波段选择 图 6 展示了 CARS 算法和 VIP - CARS 算法进行波长选择的迭代过程。图 6 - a1 显示了在 100 次蒙特卡洛采样中,采样变量数量的变化。图 6 - a2 呈现了 RMSECV 值的演变,图 6 - a3 展示了每个变量的系数或得分的变化情况。从图 6 - a1 和图 6 - b1 可以看出,由于采用相同的递减函数,二者以相同的递减速度减少波段数量。在图 6 - b3 中,某些波段在迭代过程中 VIP 得分权重始终高于其他波段。且图 6 - b3 中垂直红线标记的具有最低 RMSECV 值的最

表 2 预处理对模型的影响

异常样本检测方法	预处理方法	训练集		预测集	
		$RMSE_c$	R_c^2	$RMSE_P$	R_P^2
无样本剔除	无预处理	0.287 1	0.589 1	0.305 7	0.581 8
	D1	0.258 8	0.666 1	0.283 5	0.640 2
	DT	0.290 2	0.580 0	0.325 0	0.527 2
	MSC	0.251 1	0.685 6	0.276 5	0.657 8
PCA - MD	无预处理	0.280 5	0.621 2	0.270 8	0.616 6
	D1	0.260 0	0.652 8	0.238 6	0.682 9
	DT	0.295 0	0.581 3	0.296 2	0.541 2
	MSC	0.178 6	0.841 0	0.260 5	0.696 3

佳子集,当 $RMSECV$ 值出现下降变化时,部分变量的 VIP 得分同时出现跃升的情况,该现象证实了关键变量的存在。对比图 6 - a2 和 6 - b2 可以看出,

尽管 $VIP - CARS$ 更早达到最低 $RMSECV$ 位置,但整体而言, $CARS$ 算法的最低 $RMSECV$ 值更小一些,说明回归系数权重相对于 VIP 得分效果更好。



a. CARS; b. VIP-CARS
图6 CARS 算法和 VIP-CARS 算法的迭代过程

2.3.2 DWARS 特征波段选择 使用 DWARS 算法选取叶片氮素含量的特征波长时,每轮迭代中利用 PLS 模型得到的回归系数与 VIP 得分的 2 个权

重排序。然后,通过自适应重加权采样选择 2 个不同的变量子集,并以交叉验证均方差 ($RMSECV$) 作为评价指标来竞争选取最优子集。图 7 为 DWARS 算

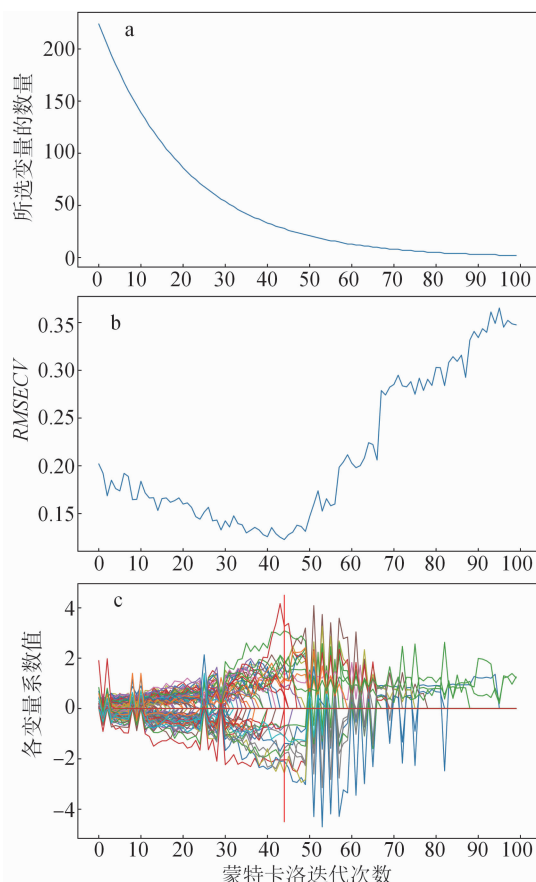
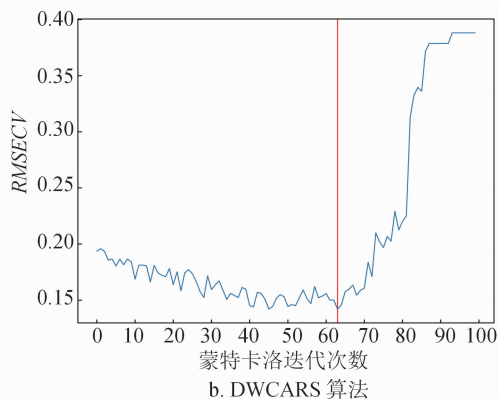
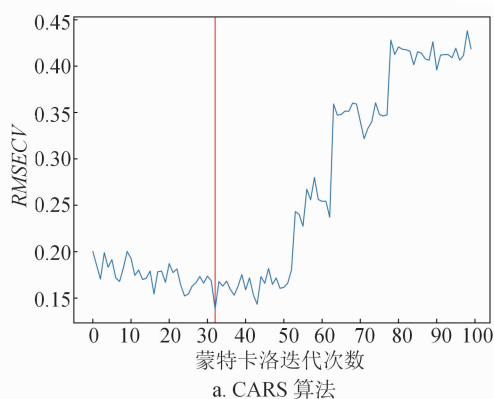
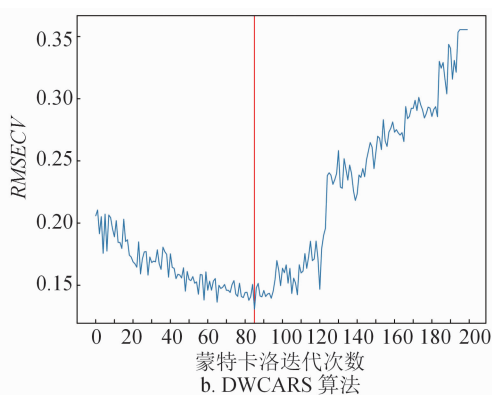
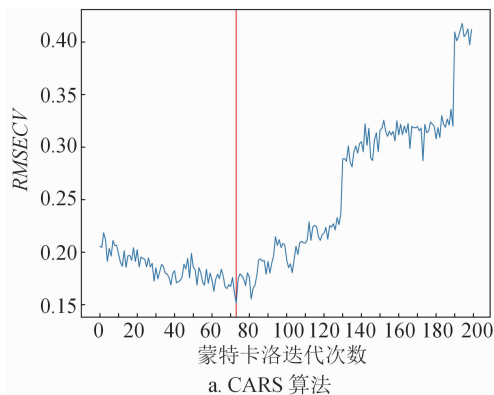


图7 DW-CARS 算法迭代过程

图8 CARS 算法和 DW-CARS 算法 100 次迭代 $RMSECV$ 变化图9 CARS 算法和 DW-CARS 算法 200 次迭代 $RMSECV$ 变化

法的迭代过程图,从图 7 - c 中可以看出,在达到最佳量子集的迭代次数之前,VIP 得分在与回归系数的竞争中有 7 次占优。图 8 显示了 CARS 算法和 DW-CARS 算法在经过 100 次蒙特卡洛采样过程中 $RMSECV$ 的变化。如图 8 所示,CARS 算法的 $RMSECV$ 值随着采样次数的增加而交替增大和减小,而 DW-CARS 算法则因为引入了同轮迭代之间的竞争机制,导致 $RMSECV$ 值在迭代到 30 次左右时连续下降,进一步降低了 $RMSECV$ 值,相较于 CARS 算法表现更佳。鉴于 DW-CARS 算法的最佳 $RMSECV$ 位置较 CARS 算法靠后,本研究将采样次数调整为 200 次以考虑其影响。图 9 展示了 200 次采样的 $RMSECV$ 变化,结果与 100 次采样情况基本一致,DW-CARS 算法的最佳迭代次数出现在 CARS 算法靠后的位置,表明 DW-CARS 在减少特征波段数量方面表现更佳,同时剔除了冗余信息。另外,针对不同的采样次数(N 分别为 50、100、200 和 500)进行了 50 次 DW-CARS 重复运行并记录 $RMSECV$ 值,统计箱线图。结果(图 10)显示,除了在采样次数过低时出现较大偏差值外,增加采样次数对 DW-CARS 性能影响不大。因此,本研究在建模分析中默认采用 100 次采样。

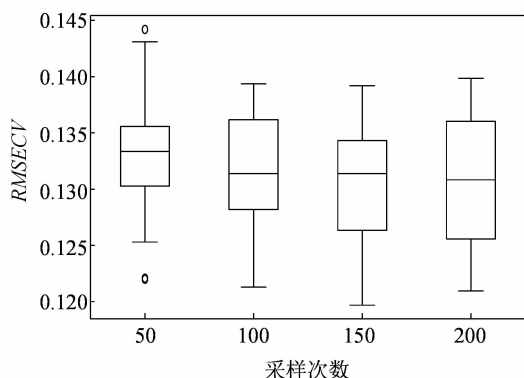


图10 DW-CARS 算法 50、100、200、500 次采样结果箱线图

2.4 建模结果分析

为了评估 SACARS 算法的建模性能,引入差分进化(DE)算法进行对比。分别使用全波长(用 FULL 表示)以及 CARS、VIP-CARS、DW-CARS、DE

算法选择特征波长,建立了橡胶树叶片中氮含量的不同定量分析模型。图 11 显示了全波段光谱和不同波段选择算法建立的预测模型之间的对比。其中,通过 DW-CARS 进行特征选择的预测模型表现最佳。从表 3 中的评价指标可以得知,使用 MSC 预处理的 DW-CARS 模型的 $RMSEP$ 和 R_p^2 分别为 0.121 5 和 0.936 7。这些 R^2 和 $RMSE$ 值表明 DW-CAR 模型在定量检测叶片氮含量方面具有出色的预测能力,模型的预测值与实测值大致呈对角线分布,真值线与拟合线基本重合,预测集的回归性能充分体现了模型的稳健性。相比于分别使用 CARS、VIP-CARS 和 DE 算法建立模型,使用 DW-CARS 模型能够明显提升模型的评价指标。综上所述,基于 MSC 预处理的 DW-CARS 算法进行波段选择的预测模型是橡胶树叶片氮含量预测的最佳模型。

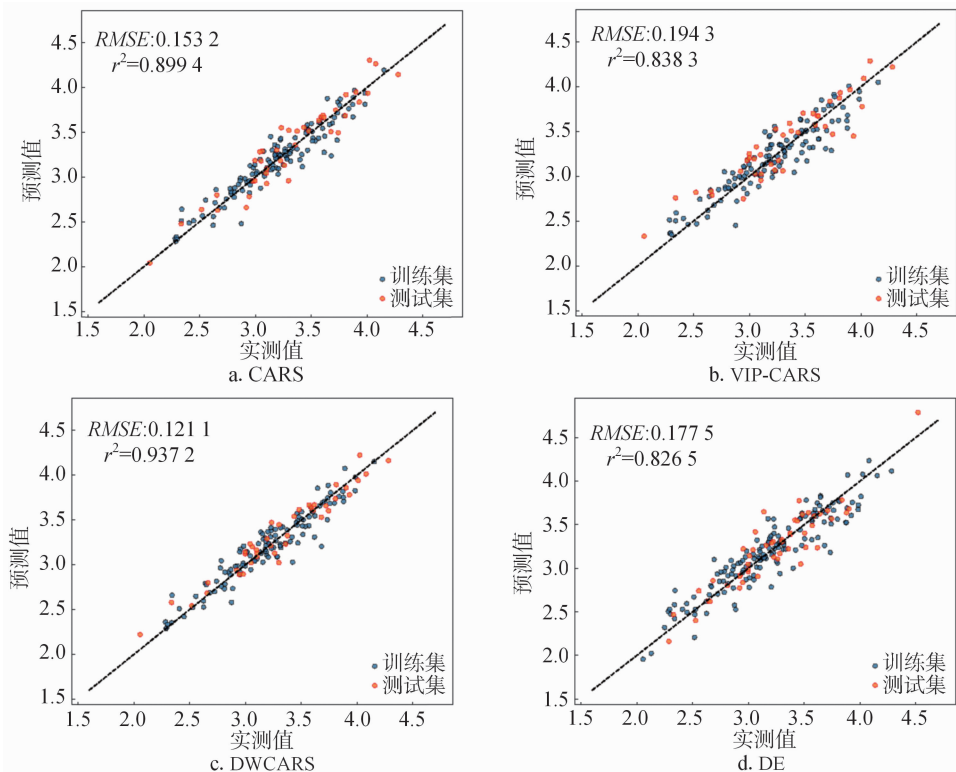


图11 使用不同波段选择算法建模的预测结果评估的散点图

2.5 变量筛选结果

不同的算法会影响波段的筛选位置和数量,较少的特征波段数量有助于减少冗余信息^[17]。在橡胶树叶片光谱数据上,使用 MSC 预处理的 CARS、VIP-CARS、DW-CARS 和 DE 算法筛选出来的变量个数分别为 56、94、20 和 105。

图 12 直观地展示了不同的变量选择方法在橡

胶树叶片光谱数据中所选波段在全谱范围的分布。可以观察到,VIP-CARS 算法选出的波段分布最集中,主要在 1 270 ~ 1 440、1 570 ~ 1 700 nm 范围内,CARS 和 DE 算法选择出的波段则在整个谱段中分布相对均匀。其中,DW-CARS 算法选择的特征波长数量最少,主要集中在 1 330 ~ 1 380、1 470 ~ 1 520 nm 区域,这 2 个区域对应于光谱数据中明显的吸收峰。

表 3 所有模型的预测结果

预处理方法	波段选择算法	训练集		预测集	
		$RMSE_c$	R_c^2	$RMSE_p$	R_p^2
无预处理	FULL	0.280 5	0.621 2	0.270 8	0.616 6
	CARS	0.138 0	0.908 3	0.161 6	0.863 9
	VIP - CARS	0.150 3	0.869 2	0.198 8	0.830 7
	DWCARS	0.121 9	0.914 0	0.153 1	0.899 6
	DE	0.181 2	0.810 8	0.242 9	0.769 0
D1	FULL	0.260 0	0.652 8	0.238 6	0.682 9
	CARS	0.123 3	0.921 7	0.145 7	0.881 7
	VIP - CARS	0.153 3	0.890 6	0.177 5	0.826 5
	DWCARS	0.116 4	0.934 5	0.132 1	0.909 1
	DE	0.165 9	0.867 4	0.203 2	0.787 4
DT	FULL	0.295 0	0.581 3	0.296 2	0.541 2
	CARS	0.147 3	0.895 4	0.144 0	0.892 0
	VIP - CARS	0.139 5	0.906 2	0.195 8	0.800 3
	DWCARS	0.115 2	0.923 2	0.144 9	0.910 1
	DE	0.160 4	0.880 1	0.187 6	0.806 3
MSC	FULL	0.178 6	0.841 0	0.260 5	0.696 3
	CARS	0.118 0	0.919 4	0.155 1	0.896 9
	VIP - CARS	0.151 2	0.867 7	0.194 7	0.837 5
	DWCARS	0.114 9	0.938 2	0.121 5	0.936 7
	DE	0.187 6	0.836 1	0.180 3	0.821 0

通过分析所选的特征波长,可以得出 CARS、VIP - CARS 和 DWCARS 所选择的波长范围相对一致,但 DWCARS 选择的特征波段中剔除了大部分相关性较低的波段。

3 讨论

本研究 DWCARS 波段选择方法具有以下特点:首先,通过引入 VIP 得分权重,有效提高了对变量的解释能力并减少了特征波长的冗余信息。其次,将 2 种权重整合到每一轮蒙特卡洛迭代中进行竞争,充分发挥了不同变量评价系数的特性,使其避免了算法陷入局部最优。

通过与其他变量选择方法的比较,DWCARS 在选择与 LNC 密切相关的特征波段方面表现最佳。相比之下,DE 算法的回归精度较低且计算时间成本较高。因为 DE 进化算法的选择过程没有考虑变量的重要性信息,导致搜索效率下降^[18]。VIP - CARS 模型精度略低于 CARS 模型,可能因 VIP 得分过于突出高重要性特征波段,没有很好地区分权重相对中等的波长,从而在剔除波长时损失了关键特征。

CARS 选择了 56 个特征波长,占全部变量的

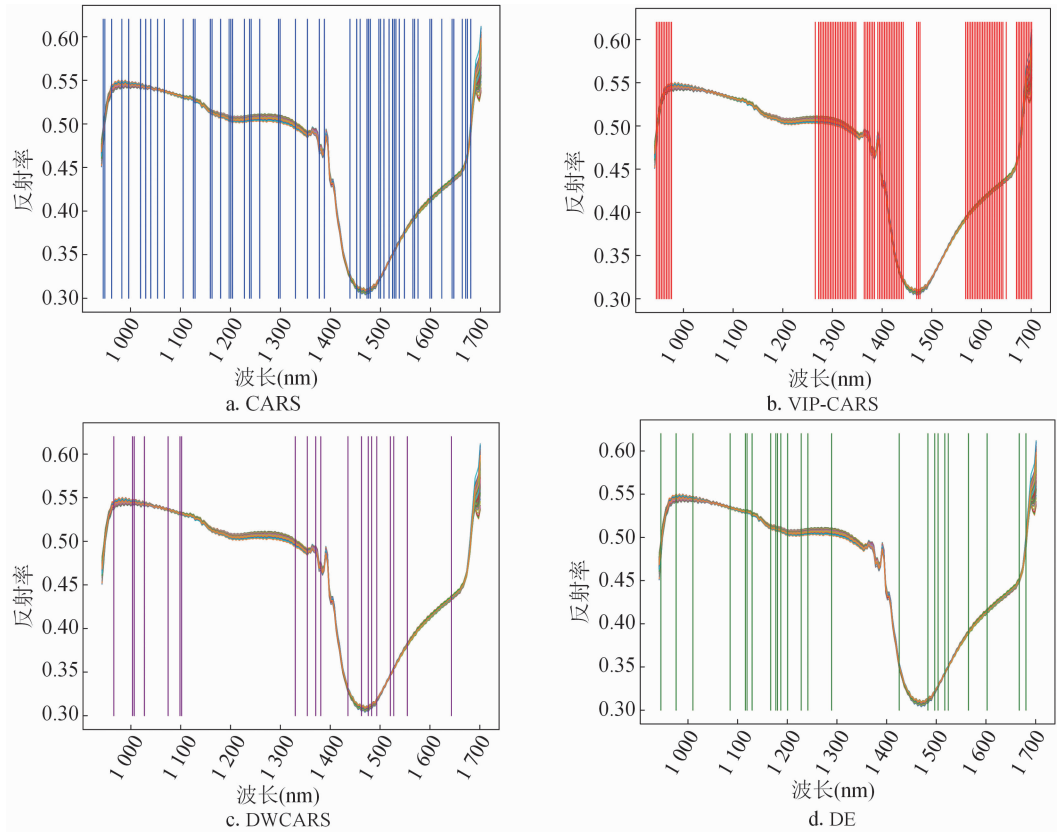


图12 CARS、VIP-CARS、DWCARS 和 DE 选择波长变量的比较

25.0%。DWCARS 选择了 20 个特征波长,占全部变量的 8.9%, $RMSE_p$ 值较 CARS 降低了 21.66%,其评价指标均好于其他模型,这表明 DWCARS 算法能够有效提取关键特征,去除冗余信息,建立了鲁棒性和可扩展性的模型。特别是,DWCARS 算法仅使用 20 个特征波长构建了氮含量的高性能检测模型,少量具有高相关性的特征波长更有利于集成到在线检测设备中,并可用于连续校准回归模型^[19]。

尽管 DWCARS 算法模型表现出良好的性能,仍有约 13% 的 LNC 变异未能被解释,可能因训练数据的非均质性和数据采集环境因素影响了模型性能。未来的研究需要开发化学计量学方法,提高模型的质量和稳定性,以实现快速检测。

4 结论

本研究旨在快速检测橡胶树叶片中的氮含量,采用了近红外光谱结合化学计量学方法。为提高模型精度,采用多种特征波长选择方法,其中 DWCARS 模型表现最佳, R_p^2 值为 0.936 7, $RMSE_p$ 值为 0.121 5。DWCARS 综合了回归系数与 VIP 得分的权重,充分利用了重加权竞争机制,有效解决了特征选择中的冗余信息问题。该研究为在线监测橡胶树叶片氮含量提供了理论支持,为实际应用提供了有效方法。

参考文献:

- [1] Warren - Thomas E, Dolman P M, Edwards D P. Increasing demand for natural rubber necessitates a robust sustainability initiative to mitigate impacts on tropical biodiversity[J]. Conservation Letters, 2015, 8(4): 230 - 241.
- [2] 黎 舟,杨思林,刘云根,等. 基于微分变换的湿地植物高光谱全氮反演[J]. 环境科学研究, 2022, 35(5): 1268 - 1276.
- [3] Ali M M, Al - Ani A, Eamus D, et al. Leaf nitrogen determination using non - destructive techniques: a review[J]. Journal of Plant Nutrition, 2017, 40(7): 928 - 953.
- [4] 冯 镇,刘 馨,张 震,等. 基于近红外光谱技术对小麦中毒死蜃农药残留测定方法的研究[J]. 食品工业科技, 2022, 43(4): 271 - 277.
- [5] Li Y J, Sun H G, Tomasetto F, et al. Spectrometric prediction of nitrogen content in different tissues of slash pine trees[J]. Plant Phenomics, 2022, 2022: 9892728.
- [6] Liu W J, Li Y J, Tomasetto F, et al. Non - destructive measurements

- of *Toona sinensis* chlorophyll and nitrogen content under drought stress using near infrared spectroscopy[J]. Frontiers in Plant Science, 2022, 12: 809828.
- [7] 郭 拓,梁小娟,马晋芳,等. 基于可扩展的自表示学习波段选择算法在近红外光谱回归建模中的影响研究[J]. 分析测试学报, 2022, 41(8): 1214 - 1220.
- [8] Yun Y H, Li H D, Deng B C, et al. An overview of variable selection methods in multivariate analysis of near - infrared spectra[J]. TrAC Trends in Analytical Chemistry, 2019, 113: 102 - 115.
- [9] Wu X H, Zeng S P, Fu H J, et al. Determination of corn protein content using near - infrared spectroscopy combined with A - CARS - PLS[J]. Food Chemistry: X, 2023, 18: 100666.
- [10] Zhang D Y, Yang Y, Chen G, et al. Nondestructive evaluation of soluble solids content in tomato with different stage by using Vis/ NIR technology and multivariate algorithms[J]. Spectrochimica Acta(Part A: Molecular and Biomolecular Spectroscopy), 2021, 248: 119139.
- [11] 王偲刚,万幼川,王明威,等. 改进蚁群算法及其在高光谱影像分类中的研究[J]. 计算机工程与应用, 2018, 54(1): 196 - 203.
- [12] Bin J, Ai F F, Fan W, et al. An efficient variable selection method based on variable permutation and model population analysis for multivariate calibration of NIR spectra[J]. Chemometrics and Intelligent Laboratory Systems, 2016, 158: 1 - 13.
- [13] Zhang J, Cui X Y, Cai W S, et al. A variable importance criterion for variable selection in near - infrared spectral analysis[J]. Science China Chemistry, 2019, 62(2): 271 - 279.
- [14] 刘翠玲,胡玉君,吴胜男,等. 近红外光谱奇异样本剔除方法研究[J]. 食品科学技术学报, 2014, 32(5): 74 - 79.
- [15] Huang X, Xia L. Improved kernel PLS combined with wavelength variable importance for near infrared spectral analysis[J]. Chemometrics and Intelligent Laboratory Systems, 2017, 168: 107 - 113.
- [16] Wang Z Z, Wu Q Y, Kamruzzaman M. Portable NIR spectroscopy and PLS based variable selection for adulteration detection in quinoa flour[J]. Food Control, 2022, 138: 108970.
- [17] Kumar B, Dikshit O, Gupta A, et al. Feature extraction for hyperspectral image classification: a review[J]. International Journal of Remote Sensing, 2020, 41(16): 6248 - 6287.
- [18] Liang L, Wei L L, Fang G G, et al. Prediction of holocellulose and lignin content of pulp wood feedstock using near infrared spectroscopy and variable selection[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2020, 225: 117515.
- [19] Bruning B, Liu H J, Brien C, et al. The development of hyperspectral distribution maps to predict the content and distribution of nitrogen and water in wheat (*Triticum aestivum*) [J]. Frontiers in Plant Science, 2019, 10: 1380.