

基于近红外光谱的草莓多品质参数通用预测模型研究

李博, 朱莉*, 姚庆宇, 姜洪洋

(东北林业大学计算机与控制工程学院, 黑龙江哈尔滨 150040)

摘要: 可溶性固形物 (Soluble Solids Content, SSC) 和硬度 (Firmness, FI) 是影响草莓口感的关键因素。该研究建立了一种基于共同特征的草莓品质参数 (SSC、FI) 通用预测模型。采用竞争性自适应重加权算法 (Competitive Adaptive Reweighted Sampling, CARS)、连续投影法 (Successive Projection Algorithm, SPA) 和无信息变量消除法 (Uniformative Variable Elimination, UVE) 提取光谱特征, 建立了偏最小二乘 (Partial Least Squares Regression, PLSR)、极限学习机 (Extreme Learning Machine, ELM) 和最小二乘支持向量机 (Least Square Support Vector Machines, LS-SVM) 决策模型, 并使用鲸鱼优化算法寻优 LS-SVM 模型的最佳参数。建立了基于 SSC 和 FI 共同特征的通用预测模型。结果表明, 使用 SG 卷积平滑法 (Savitzky-Golay, SG) 进行预处理可有效减少光谱的噪声。CARS-LS-SVM 模型对 SSC 和 FI 的单指标预测效果最好, 预测集相关系数分别为 0.937 和 0.898, 残差预测偏差分别为 2.87 和 2.28; 采用 UVE 方法分别提取的 SSC 和 FI 特征有着最高重合率。基于共同特征建立的 LS-SVM 双指标模型可以对 SSC 和 FI 进行有效预测, 预测集相关系数分别为 0.922 和 0.871, 残差预测偏差为 2.58 和 2.04。利用近红外光谱技术可以同时预测草莓的 SSC 和 FI, 该研究为草莓的多参数通用预测模型提供了理论参考。

关键词: 草莓; 近红外光谱; 鲸鱼优化算法; 最小二乘支持向量机; 通用预测模型

文章编号: 1673-9078(2025)08-227-236

DOI: 10.13982/j.mfst.1673-9078.2025.8.0821

General Predictive Model for Multiple Strawberry Quality Parameters Based on Near-infrared Spectroscopy

LI Bo, ZHU Li*, YAO Qingyu, JIANG Hongyang

(College of Computer and Control Engineering, Northeast Forestry University, Harbin 150040, China)

Abstract: Soluble solids content (SSC) and firmness index (FI) are key factors that affect the taste of strawberries. Accordingly, a general predictive model was established for strawberry quality parameters, SSC and FI, based on common features. First, spectral features were extracted using competitive adaptive reweighted sampling (CARS), the successive projection algorithm (SPA), and uniformative variable elimination (UVE). Subsequently, decision models were constructed using partial least squares regression (PLSR), extreme learning machine (ELM), and least squares support vector machine (LS-SVM). The whale optimization algorithm (WOA) was used to optimize the parameters of the LS-SVM model to ultimately establish a general predictive model based on the common features of SSC and FI. The results indicated that

引文格式:

李博,朱莉,姚庆宇,等.基于近红外光谱的草莓多品质参数通用预测模型研究[J].现代食品科技,2025,41(8):227-236.

LI Bo, ZHU Li, YAO Qingyu, et al. General predictive model for multiple strawberry quality parameters based on near-infrared spectroscopy [J]. Modern Food Science and Technology, 2025, 41(8): 227-236.

收稿日期: 2024-06-11

作者简介: 李博 (1999-), 男, 在读硕士, 研究方向: 水果品质检测, E-mail: 1317860732@qq.com

通讯作者: 朱莉 (1972-), 女, 博士, 副教授, 研究方向: 控制理论与控制工程, E-mail: zhuli-1972@nefu.edu.cn

spectral noise can be effectively reduced using the Savitzky-Golay (SG) convolution smoothing technique for preprocessing. The CARS-LS-SVM model exhibited the best single-indicator predictive performance for SSC and FI, with correlation coefficients of 0.937 and 0.898 and residual predictive deviations of 2.87 and 2.28, respectively. The SSC and FI features extracted using the UVE method had the highest overlap rate. The LS-SVM dual-indicator model that was established based on common features achieved effective predictions of SSC and FI, with correlation coefficients of 0.922 and 0.871 and residual predictive deviations of 2.58 and 2.04, respectively. Near-infrared spectroscopy could simultaneously predict the SSC and FI of strawberries. The results of this study can serve as a theoretical reference for the development of general predictive models capable of determining multiple strawberry quality parameters.

Key words: strawberry; near-infrared spectroscopy; whale optimization algorithm; least-squares support-vector machine; general predictive model

草莓富含有多种抗氧化特性的生物活性物质,如多酚化合物(鞣花酸和阿魏酸)、抗坏血酸,以及各种类黄酮(如儿茶素、花青素等),这些活性物质可有效保障牙龈健康、提神醒脑、美容养颜、滋润喉咙、降低血脂等^[1]。传统的草莓品质因数检测方式虽然能获得相对准确的检测结果,但通常是破坏性的。这些传统检测方法不仅费时费力且无法进行规模性检测,还造成了大量的成本损失,无法满足消费者日益增长的果蔬品质需求。因此,寻求更加快捷、无损且能对大批量水果的内部品质进行实时检测与分析的方法至关重要。

近年来,近红外光谱技术凭借高效、便捷以及无损等优点,已被广泛应用于食品加工工业^[2]、制药^[3]、能源^[4]和工业^[5]等领域。越来越多的国内外研究人员使用近红外光谱技术对水果的品质因数、物理特性以及化学定量指标进行检测分析。通常情况下,光谱仪采集的数据包含了来自仪器、环境以及人员操作不规范带来的各种噪声,且光谱之间的共线性问题较为严重。对光谱变量进行特征选择可以去除大量无关信息^[6],有效减少光谱维数,提高模型的预测精度和稳定性^[7]。朱金艳等^[8]采用联合区间偏最小二乘算法(Synergy Interval Partial Least Square, SiPLS)对蓝莓的可溶性固形物、维生素C和花青素特征波段进行了有效提取,利用极限学习机模型对上述三个指标进行了有效预测。Mancini等^[9]采用主成分分析(Principal Component Analysis, PCA)对草莓光谱进行降维,并使用PLSR模型有效预测了FI和SSC。Liu等^[10]使用连续投影法和稳定性竞争自适应重加权采样法(Stability Competitive Adaptive Reweighted Sampling, SCARS)对“龙香”梨的近红外光谱进行了特征筛选,并建立了PLSR和LS-SVM模型,对可溶性固形物、硬

度和含水率进行了预测,结果表明,LS-SVM模型的预测效果最好。Xie等^[11]分析了三种颜色豌豆的可见\近红外光谱,使用多元散射校正法和连续投影法建对原始数据进行预处理,结合ELM算法建立了可以预测多个豌豆品种的蛋白质通用检测模型。Włodarska等^[12]测量了草莓的近红外光谱数据,并利用刀切法对不同预处理的样本数据进行了特征筛选,建立了PLSR模型对草莓可溶性固形物的快速检测。以上研究表明,利用不同的特征选择方法可以有效提高近红外光谱模型的精度。在本研究中,以破坏式方法采集草莓样本的理化指标,并结合近红外光谱方法为无损检测模型提供原始数据,以便为该品种的草莓的定量指标建立无损检测模型。

近年来,大多数学者只对草莓的某些定量指标进行单独预测,还没有针对草莓多指标通用预测模型的研究。本研究旨在建立可对SSC和FI进行检测的通用化模型。本文使用多种预处理方法去除光谱噪声,并采用SPA、CARS和UVE算法对光谱数据进行变量选择,建立PLSR、ELM和LS-SVM回归模型,为鲜食草莓的多指标综合预测模型提供参考。

1 材料与方法

1.1 实验仪器与材料

QUEST512近红外光谱仪,英国Ocean optics公司;数显式折光仪,日本爱拓公司,检测范围0~93%;GY-4水果硬度计,韦度电子公司,由测量支架、硬度计表和探头组成。草莓品种为“红颜”,于2024年1月9日从黑龙江省哈尔滨市水果批发市场购入。

1.2 数据采集

挑选出颜色和大小一致，果实饱满且无损伤共 208 颗草莓，平均分为 4 组，贮存在 10 ℃ 和 15% 湿度的条件下。每隔 24 h 采集草莓的 SSC、FI 和近红外光谱数据，分 4 d 完成实验。在 20 ℃ 的室温条件下，将 QUEST512 光谱仪一端通过 USB 与电脑连接，另一端连接卤素光源。卤素光源通过 Y 型光纤探头测量草莓近红外光谱。仪器的工作电压为 5 V，检测的波长范围是 900~1 700 nm，与其配套的 SpectraSuite 软件扫描次数为 32。测试前需将仪器预热 30 min，然后进行黑白板校正以减小仪器本身产生的噪声。测试过程中需将光纤探头垂直对准草莓果腹部分，每隔 120 °测试采集数据，计算 3 次的均值作为最终结果，以减小环境带来的影响。根据水果硬度测定标准^[13]，草莓的硬度采用直径为 8 mm 的探头进行穿刺检测。测试前需将探头垂直对准果实赤道中心位置，当探头压入草莓果肉内部 10 mm 时停止并记录最大值。草莓的糖度使用数显式折光仪测试，测试前需将仪器表面的杂质清除干净，以防对实验结果产生干扰。测试前，将硬度测试区的草莓果肉置于纱布中挤取汁液，反复测试三次取平均值。果实的硬度和可溶性固形物的检测结果如表 1 所示。

表 1 草莓可溶性固形物和硬度指标分布
Table 1 Strawberry soluble solids content and firmness index

指标	样本数	最大值	最小值	均值
SSC/%	208	15.6	8.8	12.2
FI/N	208	8.1	3.0	5.01

1.3 光谱预处理方法

近红外光谱的采集过程中常含有杂散光、电噪声等无关信息及人为干扰，尽管通过合理使用仪器，规范实验步骤等方法可以有效避免多种误差，但环境光强和测量角度的变化仍能导致光谱基线漂移，进而导致数据建模误差变大。因此，本研究使用 SG 卷积平滑法 (Savitzky-Golay, SG)、一阶导数 (First Derivative, FD)、标准正态变量变换 (Standard Normal Variate Transformation, SNV)、多元散射校正 (Multiplicative Scatter Correction, MSC)、SG+FD 和 SG+SNV 六种方法对原始数据进行处理以减小无关信息对模型的干扰。

1.4 特征波段筛选方法

在光谱分析中，有效地选择特征变量直接影响预测模型的性能。在特征选择过程中，需综合考虑目标变量的解释性和自变量之间的冗余性，以提高模型的精度、降低过拟合风险和增强泛化能力。本研究采用 SPA、CARS 和 UVE 三种方法提取光谱数据中的有效信息波段。

SPA 是一种正向特征选择方法，旨在将特征向量之间的共线性问题最小化，能有效地筛选出少数特征波长，消除原始光谱矩阵中的冗余信息^[14]。该算法通过将波长投影到其他波长上比较投影向量的大小，从而选择出最大投影向量的波长，并根据校正模型选择最终的特征波长。该方法原理简单，适用性强，目前被广泛应用于光谱分析领域。

CARS 是一种简单有效的选择关键变量最优组合的方法，它使用偏最小二乘回归模型中回归系数的绝对值作为评估每个光谱变量重要性的指标^[15]，并根据蒙特卡罗采样次数中的交叉验证集均方根误差 (Root Mean Square Error of Cross-Validation, RMSECV) 的最小值来选择最具有代表性的光谱特征组合。

UVE 算法是一种用于多变量统计分析的特征选择方法，根据每个变量的回归系数以及标准误差来衡量特征变量与目标变量之间的相关性。然后设定阈值，将相关性低于阈值的特征剔除，保留重要变量以降低模型复杂度，提高模型的预测精度和稳定性^[16]。

1.5 模型建立与评估

本研究采用 PLSR、ELM 和 LS-SVM 模型实现对草莓 SSC 和 FI 的定量预测。首先，对原始光谱数据进行预处理，根据预测模型的评价指标选出最佳预处理方法。其次，分别采用 SPA、CARS 和 UVE 三种方法对预处理后的光谱数据进行波段特征选择，根据模型评价指标选出最佳特征选择方法以及最优模型组合。最后，根据 SSC 和 FI 的共同特征构建出 SSC 和 FI 的通用预测模型。

PLRS 是一种多元统计数据分析方法，用于复杂数据集的建模^[17]。该方法将变量的高维数据空间投影到相应的低维空间，得到正交特征向量。可以建立自变量与因变量的一元线性回归关系。一般来说，自变量从一组高度相关和共线性的光谱变量中

减少到 20 个以内潜在变量 (LVs)^[18]。PLSR 对变量的约束较少, 其灵活性使其适用于传统多变量校正方法无法解决的许多情况, 例如某些观测数据少于预测变量的数量。

ELM 是一种单隐层前馈神经网络算法, 与传统的神经网络算法不同, 该算法通过随机初始化输入层权重和隐含层偏置快速计算隐含层的输出, 使用最小二乘法计算隐含层到输出层的权重, 使得隐含层输出矩阵和目标输出之间的差距最小化^[19]。该算法训练速度快, 泛化能力强, 广泛应用于分类和回归等任务^[20]。

LS-SVM 是支持向量机的改进方法, 也是分类问题和模式识别中一种强大的机器学习方法^[21]。该算法将高维特征空间中的点积运算转化为原始空间核函数。选择径向基函数作为 LS-SVM 模型的核函数, 该函数对光谱的非线性建模过程具有自适应稳定性和鲁棒性。参数 γ 和 σ 是径向基核函数的两个主要参数。参数 γ 反映了建模过程的训练误差 (即校准集的模型精度), 参数 σ 是径向基核的宽度。鲸群优化算法 (Whale Optimization Algorithm, WOA) 是近几年提出的元启发式算法, 该算法受座头鲸捕猎行为的启发^[22], 可以对多参数进行高效搜索优化。本研究通过鲸鱼优化算法同时调整 γ 和 σ , 以实现 LS-SVM 模型的优化。

1.6 数据分析

评估校正模型的预测能力可使用校正集相关系数 (Correlation Coefficient of Calibration, R_c)、校正集均方根误差 (Root Mean Square Error of Calibration Set, $RMSEC$)、预测集相关系数 (Correlation Coefficient of Prediction, R_p) 和预测集均方根误差 (Root Mean Square Error of Prediction Set, $RMSEP$), 以及残差预测偏差 (Residual Predictive Deviation, RPD)。其中, RPD 介于 2 至 2.5 之间意味着模型可进行有效的定量预测, 介于 2.5 至 3 或以上时, 预测精度分别为良好和优异^[23]。

2 结果与讨论

2.1 样本光谱分析

将 208 颗草莓的数据随机列在 Excel 表格中, 使用联合 x-y 距离的样本集划分法 (Sample Set Partitioning Based on Joint X-Y Distance, SPXY) 以 7:3 比例划分为 145 个校正集样本和 63 个预测集样

本。所有草莓样本的原始光谱曲线如图 1 所示。

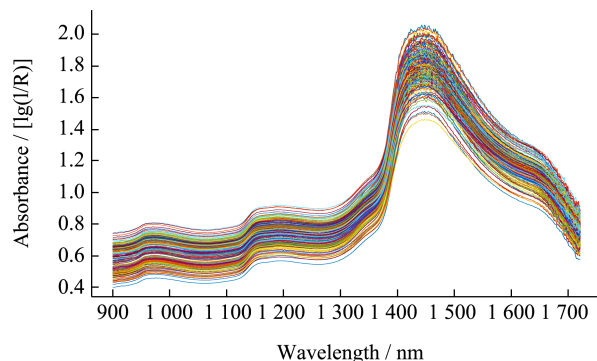


图 1 样本原始光谱

Fig.1 Original spectra of samples

2.2 鲸鱼优化算法的参数选择

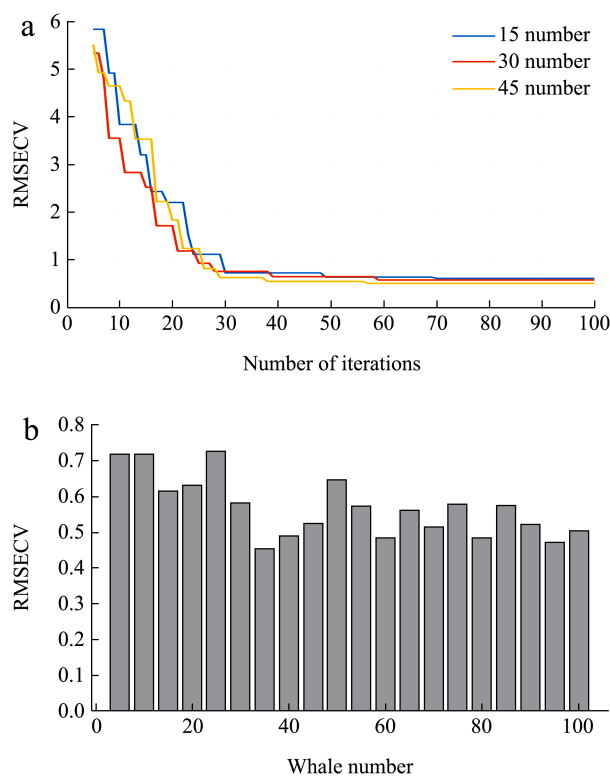


图 2 可溶性固形物的 RMSECV 随迭代次数 (a) 和鲸鱼数量 (b) 的变化趋势

Fig.2 The trend of RMSECV of soluble solids content with the number of iterations (a) and the number of whales (b)

鲸鱼优化算法的迭代次数 (T) 和鲸鱼数量 (N) 是影响 LS-SVM 模型性能的两个重要参数。其中 T 和 N 过小时, 模型的收敛性能会变差, 导致回归结果不准确; T 和 N 过大时, 会增加模型的复杂度以及运行时间, 导致模型收敛速度变缓, 影响模型的运算效率。因此, 有必要找到合适的 T 和 N 来保证模型预测性能同时降低模型的运行时间。在 LS-SVM

模型中使用原始光谱数据确定 T 和 N 的最佳值。

在 SSC 的预测中, T 的取值范围为 5~100, 在每次迭代中计算 LS-SVM 模型的 $RMSECV$, 分别记录鲸鱼数为 15、30 和 45 时的 $RMSECV$ 的变化曲线。由图 2a 所示, 三条曲线趋势相近, 随着迭代次数的增加, 模型的 $RMSECV$ 值迅速下降, 然后趋于稳定。考虑到模型的预测精度和运算效率, SSC 定量预测的最优迭代次数为 100 次。 N 的起始值为 5, 每次计算模型的 $RMSEC$ 后种群数加 5, 直至 100。模型的 $RMSECV$ 随鲸鱼数的变化如图 2b 所示。随着种群数量增加, $RMSECV$ 总体呈下降趋势, 变化幅度趋于稳定, 当 $N=35$ 时, $RMSECV$ 达到最小。FI 指标也有相似规律, 因此 WOA 算法的最优迭代次数为 100 次, 鲸鱼种群数量为 35。

2.3 预处理结果与分析

分别采用原始光谱数据 (RAW) 和六种预处理后的光谱数据建立 PLSR、LS-SVM 和 ELM 模型。其中 PLSR 模型采用 5 折交叉验证, 并选取最佳主成分因子数进行 SSC 和 FI 预测; ELM 模型的激活函数为 Sigmoid 函数; 隐含层神经元的初始个数为 10, 步长为 5, 等差增加至 100, 以模型的最小 $RMSECV$ 值确定最优神经元个数^[20]; LS-SVM 模型采用径向基核函数, 并根据 WOA 方法选择的最优核函数参数进行 SSC 和 FI 的预测。

SSC 的预测结果如表 2 所示。在 PLSR 模型中, SG+FD 预处理方法取得了最好的预测效果, R_p 、 $RMSEP$ 和 RPD 分别为 0.924、0.395 和 2.58。该方法和 SG 方法取得的效果相近。在 LS-SVM 模型中, SG 预处理方法的预测效果最好, R_p 、 $RMSEP$ 和 RPD 分别为 0.925、0.388 和 2.64。在 ELM 模型中, SG 预处理方法的预测效果最好, R_p 、 $RMSEP$ 和 RPD 分别为 0.917、0.437 和 2.51。在三个模型中, LS-SVM 对 SSC 的预测效果最好。

FI 的预测结果如表 3 所示。在 PLSR 模型中, SG 方法的预测效果最佳, 且 RPD 超过了 2, 可以进行有效预测。FD、MSC、SNV、SG+FD 和 SG+SNV 的表现性能不如原始光谱。在 LS-SVM 模型中, SG 方法取得了最好的预测效果, 其 R_p 、 $RMSEP$ 和 RPD 分别为 0.883、0.441 和 2.13。其中, SG、SG+FD 和 SG+SNV 方法的 RPD 值均超过了 2, 可以进行有效预测。在 ELM 模型中, SG+FD 方法取得了最好的预测结果, RPD 为 2.11。在三种模型中, LS-SVM 的预测精度最高, 预测集稳定性最佳。

表 2 不同预处理方法下的SSC预测结果对比
Table 2 Comparison of SSC prediction results under different

pretreatment methods						
模型	预处理方法	R_C	$RMSEC$	R_p	$RMSEP$	RPD
PLSR	RAW	0.958	0.370	0.892	0.471	2.21
	SG	0.957	0.374	0.920	0.404	2.56
	FD	0.683	1.142	0.563	1.526	1.21
	MSC	0.934	0.427	0.893	0.450	2.22
	SNV	0.957	0.381	0.908	0.418	2.39
	SG+FD	0.962	0.372	0.924	0.395	2.58
	SG+SNV	0.964	0.386	0.907	0.418	2.38
LS-SVM	RAW	0.943	0.431	0.881	0.483	2.12
	SG	0.957	0.373	0.925	0.388	2.64
	FD	0.732	1.052	0.589	1.494	1.24
	MSC	0.952	0.395	0.896	0.442	2.27
	SNV	0.956	0.383	0.904	0.405	2.34
	SG+FD	0.952	0.377	0.923	0.382	2.60
	SG+SNV	0.964	0.354	0.915	0.433	2.47
ELM	RAW	0.948	0.424	0.887	0.513	2.16
	SG	0.963	0.414	0.917	0.437	2.51
	FD	0.558	1.472	0.428	1.687	1.11
	MSC	0.928	0.457	0.902	0.483	2.32
	SNV	0.946	0.412	0.892	0.435	2.21
	SG+FD	0.954	0.394	0.907	0.423	2.37
	SG+SNV	0.951	0.377	0.985	0.406	2.24

总结而言, 除了 FD 方法外, 其他预处理方法均能有效地提高模型的预测性能, 其中 SG 方法在大部分情况下可以提供更稳定以及较好的预测结果。经 FD 方法处理里的光谱不满足建模条件, 这是因为原始光谱重合度较高, 分辨率较低, 且噪声偏多, 而 FD 方法使得光谱中的噪声被放大, 导致模型无法区分原始光谱中的重叠峰。而 SG+FD 方法在消除大部分噪声后可以增加光谱分辨率。比如在 PLSR 模型对 SSC 预测时, SG+FD 方法的预测效果优于其他方法; 在 ELM 模型对 FI 预测时, SG+FD 方法的预测效果优于其他方法。这是因为 SG 方法消除了原始光谱中的大部分噪声, 大大降低了噪声对 FD 方法的干扰, 增加了光谱曲线的区分度。在大部分情况下, SG 和 SG+FD 方法在对 SSC 和 FI 的预测中有着相近的预处理效果, 但是 SG+FD+ELM (SSC) 和 SG+FD+PLSR (FI) 模型的预测效果稍差。这是因为 SG 方法不能消除原始

光谱中的全部噪声，导致 SG+FD 方法在模型预测中的效果不稳定。因此在后续的特征选择过程中，采用 SG 方法处理的光谱数据进行建模。

表 3 不同预处理方法下的FI预测结果对比

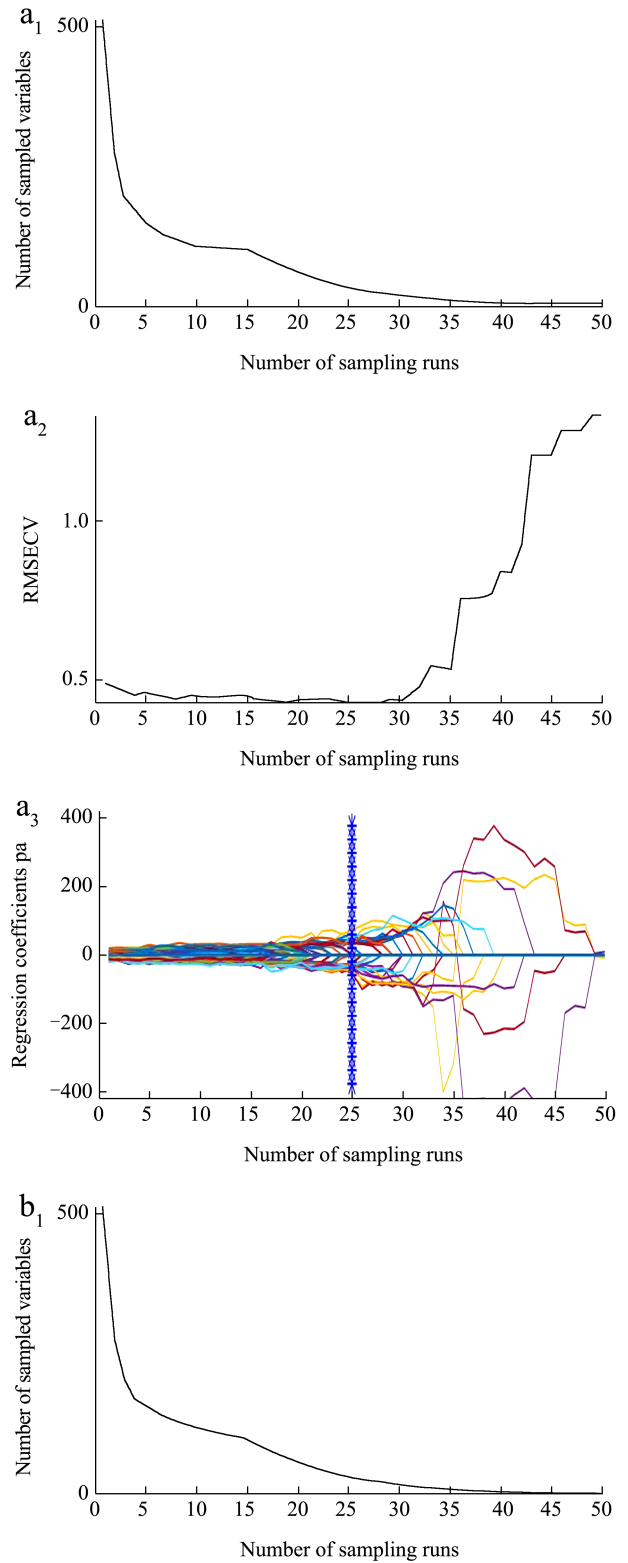
Table 3 Comparison of FI prediction results under different pretreatment methods						
模型	预处理方法	R_c	$RMSEC$	R_p	$RMSEP$	RPD
PLSR	RAW	0.962	0.315	0.859	0.438	1.96
	SG	0.951	0.336	0.869	0.430	2.02
	FD	0.527	1.344	0.473	1.672	1.13
	MSC	0.945	0.335	0.837	0.477	1.83
	SNV	0.944	0.327	0.852	0.397	1.91
	SG+FD	0.948	0.402	0.848	0.445	1.89
	SG+SNV	0.939	0.394	0.857	0.466	1.94
LS-SVM	RAW	0.947	0.385	0.815	0.473	1.72
	SG	0.952	0.346	0.883	0.387	2.13
	FD	0.646	1.284	0.572	1.441	1.22
	MSC	0.956	0.311	0.861	0.403	1.96
	SNV	0.948	0.343	0.855	0.395	1.95
	SG+FD	0.963	0.372	0.875	0.385	2.07
	SG+SNV	0.951	0.367	0.868	0.394	2.01
ELM	RAW	0.942	0.415	0.864	0.434	1.99
	SG	0.964	0.372	0.868	0.381	2.01
	FD	0.689	1.182	0.611	1.354	1.26
	MSC	0.964	0.384	0.851	0.417	1.90
	SNV	0.961	0.357	0.854	0.386	1.92
	SG+FD	0.955	0.364	0.880	0.379	2.11
	SG+SNV	0.944	0.401	0.859	0.421	1.95

2.4 模型结果

2.4.1 光谱特征波长选择

在 CARS 方法中，设置蒙特卡洛采样次数为 50 次，计算循环次数为 100 次，采用五折交叉验证法，根据蒙特卡洛采样过程中的最小 $RMSECV$ 选择光谱变量。图 3a 和图 3b 分别为 SSC 和 FI 的 CARS 运行过程。在采样开始时，被选取的波长数量迅速减少，随后下降趋势变缓。随着包含重要信息的波段被剔除， $RMSECV$ 逐渐增大。在针对 SSC 的特征波长选择中，蒙特卡洛采样次数为 25 时， $RMSECV$

的值达到最小，此时筛选的波长数量为 34 个；在针对 FI 的特征波长选择中，蒙特卡洛采样次数为 28 个时， $RMSECV$ 的值达到最小，此时筛选的波长数量为 18 个。



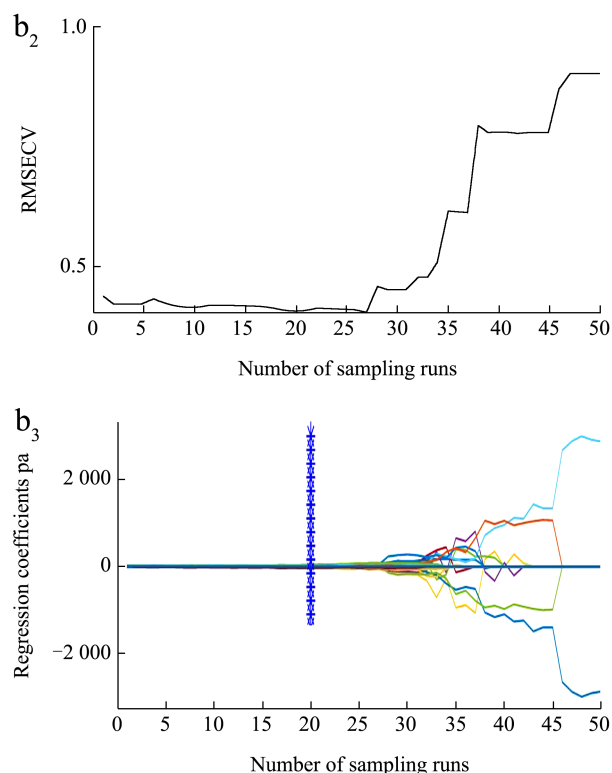


图3 草莓可溶性固形物 (a) 和硬度 (b) 的 CARS 特征波段提取过程

Fig.3 The CARS characteristic band extraction process of strawberry soluble solids content (a) and hardness (b)

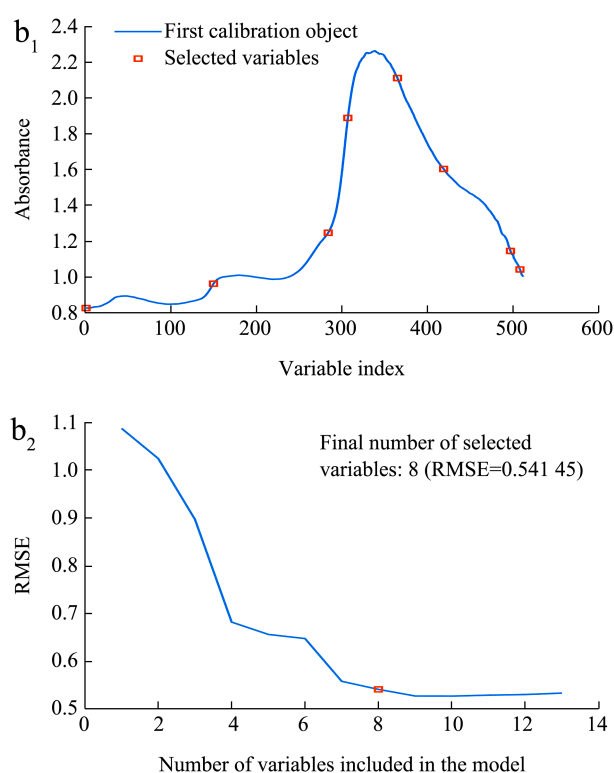
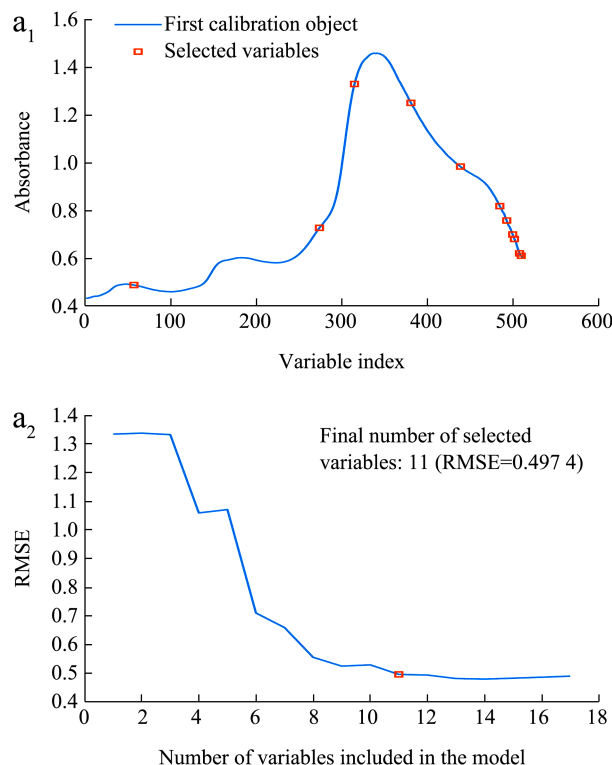


图4 草莓可溶性固形物 (a) 和硬度 (b) 的 SPA 特征波段提取过程

Fig.4 SPA characteristic band extraction process of strawberry soluble solids content (a) and hardness (b)

在 SPA 方法中, 根据均方根误差 (Root Mean Square Error, $RMSE$) 的最小值来确定筛选的波段。如图 4 所示, 随着选择的波段数量不断增加, $RMSE$ 迅速降低, 然后趋于平缓。当选择的波段个数继续增加时, $RMSE$ 缓慢变大。经筛选的 SSC 和 FI 的最优波段数量分别是 11 个和 8 个。

在 UVE 方法中, 设置噪声随机变量为 512 个, 与光谱波段数相等, 设置稳定性剔除阈值为 99%。图 5a 为 SSC 的光谱波段选择过程, 图 5b 为 FI 的光谱波段选择过程。图中, 两条蓝色水平虚线以内包含的光谱波段为被剔除的无关变量, 蓝色水平虚线以外的光谱波段为所选的有用变量。

2.4.2 单指标模型结果

分别将经过 SG 预处理的光谱变量与 CARS、SPA 和 UVE 选择的特征波段输入到 PLSR、ELM 和 LS-SVM 模型中。对 SSC 和 FI 进行预测时, PLSR 模型最优 LVs 数、ELM 隐含层神经元个数 (N) 和 LS-SVM 模型最优参数如表 4 所示。

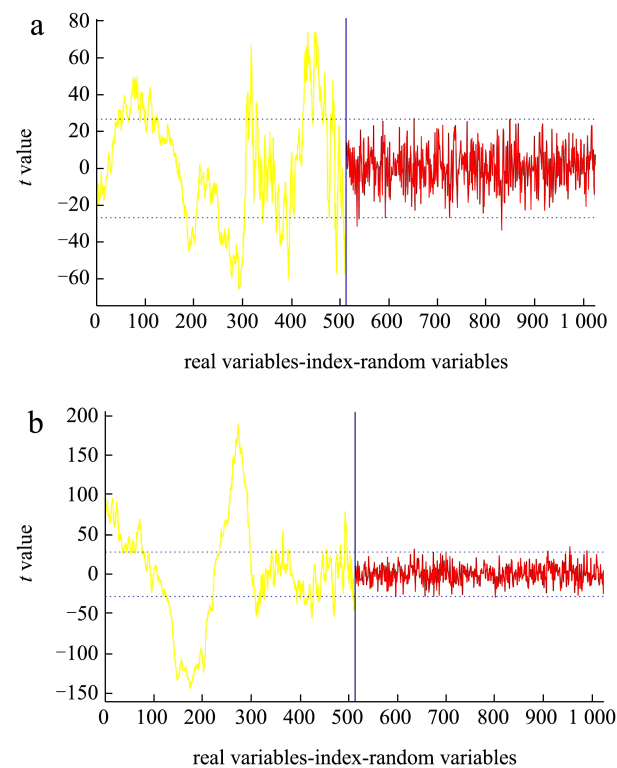


图5 草莓可溶性固形物 (a) 和硬度 (b) 的 UVE 特征波段提取过程

Fig.5 UVE characteristic band extraction process of strawberry soluble solids content (a) and hardness (b)

表 4 不同特征选择方法下的PLSR、ELM和LS-SVM模型参数

Table 4 PLSR, ELM and LS-SVM model parameters under different feature selection methods

模型	特征选择方法	PLSR	ELM	LS-SVM	
		LVs	N	γ	σ^2
SSC	CARS	9	35	14 372	2 369.2
	SPA	8	55	10 894	3 824.8
	UVE	11	40	20 687	3 052.1
FI	CARS	9	50	10 145	413.6
	SPA	7	35	37 821	391.5
	UVE	9	45	15 566	2 695.8

SSC 的模型预测结果如表 5 所示。在 PLSR 模型中，CARS 方法表现的预测性能更好，其预测集稳定性较为突出， R_p 、 $RMSEP$ 和 RPD 分别为 0.924、0.368 和 2.62。在 ELM 模型中，UVE 方法取得了最好的预测效果， R_p 、 $RMSEP$ 和 RPD 分别为 0.928、0.372 和 2.68。在 LS-SVM 模型中，CARS 方法的预测效果在三个模型中最好， R_p 、 $RMSEP$ 和 RPD 分别为 0.937、0.354 和 2.87。在三个模型中，采用 CARS 方法和 UVE 方法可以取得较为不错的预测

精度，两个方法的预测效果接近。除 SPA 方法外，另外两种特征选择方法可以有效提取光谱中的重要信息。

表 5 不同特征处理方法下的模型预测SSC结果对比

Table 5 Comparison of model prediction SSC results under different feature processing methods

模型	特征选择方法	R_c	$RMSEC$	R_p	$RMSEP$	RPD
PLSR	SG	0.957	0.374	0.918	0.404	2.52
	CARS	0.949	0.324	0.924	0.368	2.62
	SPA	0.940	0.398	0.894	0.424	2.23
	UVE	0.950	0.362	0.922	0.376	2.58
ELM	SG	0.963	0.414	0.917	0.437	2.51
	CARS	0.956	0.361	0.925	0.387	2.65
	SPA	0.946	0.393	0.884	0.417	2.14
	UVE	0.962	0.358	0.928	0.372	2.68
LS-SVM	SG	0.957	0.373	0.925	0.388	2.64
	CARS	0.958	0.332	0.937	0.354	2.87
	SPA	0.951	0.388	0.896	0.412	2.25
	UVE	0.956	0.325	0.928	0.384	2.71

表 6 不同特征处理方法下的模型预测FI结果对比

Table 6 Comparison of model prediction FI results under different feature processing methods

模型	特征选择方法	R_c	$RMSEC$	R_p	$RMSEP$	RPD
PLSR	SG	0.951	0.336	0.869	0.430	2.02
	CARS	0.935	0.383	0.885	0.402	2.15
	SPA	0.942	0.363	0.868	0.415	2.01
	UVE	0.939	0.357	0.879	0.391	2.10
ELM	SG	0.964	0.372	0.868	0.381	2.01
	CARS	0.947	0.364	0.886	0.383	2.16
	SPA	0.958	0.354	0.859	0.384	1.95
	UVE	0.948	0.370	0.882	0.396	2.12
LS-SVM	SG	0.952	0.346	0.883	0.441	2.13
	CARS	0.941	0.334	0.898	0.375	2.28
	SPA	0.928	0.373	0.832	0.421	1.81
	UVE	0.956	0.357	0.891	0.382	2.21

FI 的模型预测结果如表 6 所示。在 PLSR 模型中，四种方法的 RPD 值均大于 2，可以进行有效预测，其中 CARS 方法取得的预测效果最佳。在 ELM 模型中，采用 SPA 方法的 RPD 值为 1.95，不能对 FI 进行有效预测。在 LS-SVM 模型中，CARS 方法取得了三个模型中最高的预测效果， R_p 、 $RMSEP$ 和

RPD 分别为 0.898、0.375 和 2.28。LS-SVM 模型在大部分情况下都有着更低的预测误差和更高的相关系数，说明其在预测 SSC 和 FI 方面更为可靠。特别是在使用 CARS 和 UVE 这两种变量选择方法时，LS-SVM 模型有着更好的性能。

在对 SSC 和 FI 预测时，除 SPA 方法外，CARS 和 UVE 均可以有效提取光谱中的重要信息，减少无关变量对模型的干扰。

2.5 通用预测模型结果

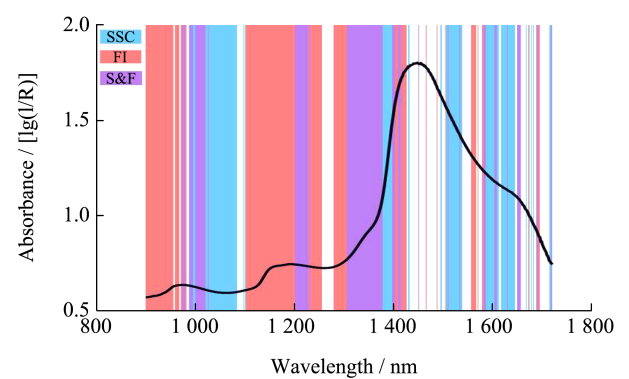


图 6 SSC 和 FI 共同特征波段分布
Fig.6 The common characteristic band distribution of SSC and FI

采用 CARS 方法所筛选出来的 SSC 和 FI 共同特征重合率较低，二者的共同特征不足以建立通用预测模型。而使用 UVE 方法提取的特征波段重合率较高，特征分布如图 6 所示，重合波段主要分布在 980~1 020、1 200~1 225 和 1 310~1 380 nm 之间。经 UVE 提取的 SSC 和 FI 共同特征波段数为 118 个，分别占各自特征总数的 52.21% 和 41.4%。本文的通用预测模型利用 SSC 和 FI 的共同特征建模，以实现一种模型对 SSC 和 FI 两种指标的预测。

表 7 LS-SVM模型最优参数以及SSC和FI的预测结果
Table 7 The optimal parameters of LS-SVM model and the prediction results of SSC and FI

特征选择方法	指标	γ	σ^2	R_c	$RMSEC$	R_p	$RMSEP$	RPD
UVE	SSC	17 241 3	152.8	0.953	0.348	0.922	0.384	2.58
	FI	33 517	901.8	0.948	0.362	0.871	0.398	2.04

由 2.4 知，LS-SVM 模型对 SSC 和 FI 预测中有着较好的预测效果，因此本研究采用该模型进行预测。预测结果如表 7 所示，SSC 预测结果的 R_p 、 $RMSEP$ 和 RPD 分别为 0.922、0.384 和 2.58；FI 预测结果的 R_p 、 $RMSEP$ 和 RPD 分别为 0.871、0.398

和 2.04，二者的 *RPD* 值均超过了 2，表明模型预测有效。

3 结论

本文基于近红外光谱技术构建了草莓的 SSC、FI 单指标预测模型以及 SSC 和 FI 双指标通用预测模型。采用 SG 预处理后的光谱数据能够有效减少原始光谱中的冗余变量以及无关信息。与 SPA 方法相比，UVE 和 CARS 可有效降低光谱变量之间的共线性问题，有效提升了模型的预测效果。相比于 PLSR 和 ELM 模型，LS-SVM 模型在 SSC 和 FI 中的预测中具有一定优势。其中 CARS-LS-SVM 模型在 SSC (R_p =0.937、 $RMSEP$ =0.354、 RPD =2.87) 和 FI (R_p =0.898、 $RMSEP$ =0.375、 RPD =2.28) 的预测中取得了最好的效果。在通用预测模型中，经过 UVE 方法处理的 SSC 和 FI 的特征波段重合率比较高，采用两者的共同波段建立的 LS-SVM 模型均可有效预测 SSC 和 FI。

本文论证了利用近红外光谱技术同时对草莓品质因数预测的可行性，为草莓的多参数通用预测模型研究提供了理论依据和思路。

参考文献

[1] NORIEGA F, MARDONES C, FISCHER S, et al. Seasonal changes in white strawberry: Effect on aroma, phenolic compounds and its biological activity [J]. Journal of Berry Research, 2021, 11(1): 103-118.

[2] POREP J U, KAMMERER D R, CARLE R. On-line application of near infrared (NIR) spectroscopy in food production [J]. Trends in Food Science & Technology, 2015, 46(2): 211-230.

[3] PEDERSEN T, RANTANEN J, NAELAPÄÄ K, et al. Near infrared analysis of pharmaceutical powders with empirical target distribution optimization (ETDO) [J]. Journal of Pharmaceutical and Biomedical Analysis, 2020, 181: 113059.

[4] TOSCANO G, RINNAN Å, PIZZI A, et al. The use of near-infrared (NIR) spectroscopy and principal component analysis (PCA) to discriminate bark and wood of the most common species of the pellet sector [J]. Energy & Fuels, 2017, 31(3): 2814-2821.

[5] IYAKWARI S, GLASS H J, ROLLINSON G K, et al. Application of near infrared sensors to preconcentration of hydrothermally-formed copper ore [J]. Minerals Engineering, 2016, 85: 148-167.

[6] GUO Z, WANG M, SHUJAT A, et al. Nondestructive monitoring storage quality of apples at different temperatures

- by near-infrared transmittance spectroscopy [J]. Food Science & Nutrition, 2020, 8(7): 3793-3805.
- [7] KUTSANEDZIE F Y H, CHEN Q, HASSAN M M, et al. Near infrared system coupled chemometric algorithms for enumeration of total fungi count in cocoa beans neat solution [J]. Food Chemistry, 2018, 240: 231-238.
- [8] 朱金艳,朱玉杰,冯国红,等.基于近红外光谱技术联合极限学习机的蓝莓贮藏品质定量模型建立[J].食品与发酵工业,2022,48(16):270-276.
- [9] MANCINI M, MAZZONI L, GAGLIARDI F, et al. Application of the non-destructive NIR technique for the evaluation of strawberry fruits quality parameters [J]. Foods, 2020, 9(4): 441.
- [10] LIU D, WANG E, WANG G, et al. Nondestructive determination of soluble solids content, firmness, and moisture content of “Longxiang” pears during maturation using near-infrared spectroscopy [J]. Journal of Food Processing and Preservation, 2022, 46(3): e16332.
- [11] XIE C, QIAO M, YANG L, et al. Establishment of a general prediction model for protein content in various varieties and colors of peas using visible-near-infrared spectroscopy [J]. Journal of Food Composition and Analysis, 2024, 127: 105965.
- [12] WŁODARSKA K, SZULC J, KHMELINSKII I, et al. Non-destructive determination of strawberry fruit and juice quality parameters using ultraviolet, visible, and near-infrared spectroscopy [J]. Journal of the Science of Food and Agriculture, 2019, 99(13): 5953-5961.
- [13] 北京市农林科学院林业果树研究所,农业部果品及苗木质量监督检验测试中心.水果硬度的测定:NY/T 2009-2011[S].北京:中华人民共和国农业部, 2011.
- [14] LI B, HOU B, ZHANG D, et al. Pears characteristics (soluble solids content and firmness prediction, varieties) testing methods based on visible-near infrared hyperspectral imaging [J]. Optik, 2016, 127(5): 2624-2630.
- [15] LI Y, YANG X. Quantitative analysis of near infrared spectroscopic data based on dual-band transformation and competitive adaptive reweighted sampling [J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2023, 285: 121924.
- [16] GAO S, XU J. Hyperspectral image information fusion-based detection of soluble solids content in red globe grapes [J]. Computers and Electronics in Agriculture, 2022, 196: 106822.
- [17] MEHMOOD T, SAEBO S, LILAND K H. Comparison of variable selection methods in partial least squares regression [J]. Journal of Chemometrics, 2020, 34(6): e3226.
- [18] LEE Y, HAN S H. Feasibility of nondestructive sugar content analysis of Korean pears by using near-infrared diffuse-reflectance spectroscopy [J]. Bulletin of the Korean Chemical Society, 2016, 37(2): 166-173.
- [19] WANG H, CHU X, CHEN P, et al. Partial least squares regression residual extreme learning machine (PLSRR-ELM) calibration algorithm applied in fast determination of gasoline octane number with near-infrared spectroscopy [J]. Fuel, 2022, 309: 122224.
- [20] 倪鸿飞,斯乐婷,黄家鹏,等.近红外光谱结合遗传算法优化的极限学习机实现银杏叶纯化过程有效成分快速测定[J].中国中药杂志,2021,46(1):110-117.
- [21] ZHENG H, LU H. A least-squares support vector machine (LS-SVM) based on fractal analysis and CIELab parameters for the detection of browning degree on mango (*Mangifera indica* L.) [J]. Computers and Electronics in Agriculture, 2012, 83: 47-51.
- [22] MIRJALILI S, LEWIS A. The whale optimization algorithm [J]. Advances in Engineering Software, 2016, 95: 51-67.
- [23] LI J, ZHANG H, ZHAN B, et al. Nondestructive firmness measurement of the multiple cultivars of pears by Vis-NIR spectroscopy coupled with multivariate calibration analysis and MC-UVE-SPA method [J]. Infrared Physics & Technology, 2020, 104: 103154.