

王甜甜, 冯国红, 朱玉杰. 近红外光谱结合化学计量学方法快速检测蓝莓可溶性固形物和维生素 C 含量 [J]. 食品工业科技, 2023, 44(16): 297–305. doi: 10.13386/j.issn1002-0306.2022090235

WANG Tiantian, FENG Guohong, ZHU Yujie. Rapid Determination of Soluble Solids and Vitamin C in Blueberry by Near Infrared Spectroscopy Combined with Chemometrics[J]. Science and Technology of Food Industry, 2023, 44(16): 297–305. (in Chinese with English abstract). doi: 10.13386/j.issn1002-0306.2022090235

· 分析检测 ·

近红外光谱结合化学计量学方法快速检测 蓝莓可溶性固形物和维生素 C 含量

王甜甜, 冯国红*, 朱玉杰*

(东北林业大学工程技术学院, 黑龙江哈尔滨 150040)

摘要:采用近红外光谱技术, 对不同贮藏时间的蓝莓营养成分进行定量分析, 以寻求其化学成分与近红外光谱数据的相关性, 实现利用光谱技术对蓝莓营养成分的无损检测。对获取的近红外光谱数据, 运用偏最小二乘回归 (Partial Least Square Regression, PLSR) 和支持向量回归 (Support Vector Regression, SVR) 两种机器学习算法预测蓝莓可溶性固形物 (Soluble Solids Content, SSC) 和维生素 C (Vitamin C, V_C) 含量。为增加预测精度, 采用一阶导数 (First Derivative, 1-DER)、二阶导数 (Second Derivative, 2-DER)、标准正态变换 (Standard Normal Variate Transform, SNV)、多元散射校正 (Multiplicative Scatter Correction, MSC)、Savitzky-Golay 平滑 (S-G) 等一种或几种方法组合对光谱数据进行预处理, 比较分析最佳的预处理方式; 采用竞争适应性重加权采样法 (Competitive Adaptive Reweighted Sampling, CARS) 和随机蛙跳算法 (Random Frog, RF) 及两种算法组合对光谱波长进行降维处理。结果表明, 降维后的 SSC 波长变量分别降到了全光谱变量的 1.7%、4.3% 和 5.6%, V_C 波长变量分别降到了全光谱变量的 2.5%、2.9%、4.8%。在筛选后的光谱波长变量的基础上, 采用 PLSR 建立蓝莓近红外光谱与 SSC 和 V_C 含量的预测模型。对比发现 CARS 结合 RF 算法筛选出的波长变量预测效果更好, 模型校正相关系数分别为 0.9001、0.8707, 校正均方根误差分别为 0.8234、2.9429, 预测相关系数分别为 0.8424、0.8350, 预测均方根误差分别为 0.9613、2.9482。为排除模型性能对预测结果的影响, 建立 SVR 模型将预测结果进行对比, 同样发现 CARS 结合 RF 算法的预测效果更佳, 模型校正相关系数分别为 0.8702、0.8503, 校正均方根误差分别为 0.9549、3.2431, 预测相关系数分别为 0.8269、0.8183, 预测均方根误差分别为 0.8769、2.8818。本研究为蓝莓营养品质监测提供了模型基础, 且选择特征波长的方法可以为更多果蔬营养物质预测模型提供参考。

关键词:近红外光谱, 竞争适应性重加权采样法, 随机蛙跳, 偏最小二乘回归, 支持向量回归

中图分类号: TS255.1

文献标识码: A

文章编号: 1002-0306(2023)16-0297-09

DOI: 10.13386/j.issn1002-0306.2022090235



本文网刊:

Rapid Determination of Soluble Solids and Vitamin C in Blueberry by Near Infrared Spectroscopy Combined with Chemometrics

WANG Tiantian, FENG Guohong*, ZHU Yujie*

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

Abstract: The near-infrared spectroscopy technology was adopted to quantitatively analyze the nutritional components of blueberries given different storage times, so as to determine the correlation between their chemical components and near-infrared spectroscopy data. Besides, spectroscopy technology was applied to perform the nondestructive detection of blueberry nutritional components. As for the obtained near-infrared spectral data, two machine learning algorithms, Partial Least Square Regression (PLSR) and Support Vector Regression (SVR), were used to predict the content of soluble solids

收稿日期: 2022-09-23

基金项目: 中央高校基本科研业务费专项资金项目 (2572020BL01); 黑龙江省自然科学基金项目 (LH2020C050)。

作者简介: 王甜甜 (1999-), 女, 硕士研究生, 研究方向: 预测与优化, E-mail: 3168055052@qq.com。

* 通信作者: 冯国红 (1980-), 女, 博士, 副教授, 研究方向: 蓝莓质量无损检测, E-mail: fgh_1980@126.com。

朱玉杰 (1964-), 女, 博士, 教授, 研究方向: 工业工程, E-mail: zhuyujie004@126.com。

(SSC) and vitamin C (V_C) in blueberries. In order to improve the accuracy of prediction, one or more of the methods, such as First Derivative (1-DER), Second Derivative (2-DER), Standard Normal Variable Transform (SNV), Multivariate Scatter Correction (MSC), Savitzky Golay smoothing (S-G), were used to preprocess the spectral data, and the best-performing methods were comparatively analyzed. Competitive Adaptive Weighted Sampling (CARS) and Random Frog (RF) were adopted either separately or in combination to reduce the dimensions of spectral wavelengths. Results showed that, after dimension reduction, the SSC wavelength as a variable was reduced to 1.7%, 4.3% and 5.6% of the full spectral variable, while the V_C wavelength as a variable was reduced to 2.5%, 2.9% and 4.8% of the full spectral variable, respectively. With the screened spectral wavelength as a variable, PLSR was used to construct a prediction model of near-infrared spectroscopy for SSC and V_C contents in blueberry. The comparison showed that the wavelength variables screened by CARS in combination with RF algorithm produced a better outcome of prediction. The model correction correlation coefficients were 0.9001 and 0.8707 respectively, the correction root mean square errors were 0.8234 and 2.9429 respectively, the prediction correlation coefficients were 0.8424 and 0.8350 respectively, and the prediction root mean square errors were 0.9613 and 2.9482 respectively. To eliminate the impact of model performance on the prediction results, an SVR model was established to compare the prediction results. It was also discovered that a better prediction result was produced by CARS in combination with RF algorithm. The model correction correlation coefficients were 0.8702 and 0.8503, respectively. The correction root mean square errors were 0.9549 and 3.2431, respectively. The prediction correlation numbers were 0.8269 and 0.8183, respectively. The prediction root mean square errors were 0.8769 and 2.8818, respectively. To sum up, this study provides a model basis for monitoring the quality of blueberry nutrients, and the method proposed to select characteristic wavelength provides a reference for more models of fruit and vegetable nutrients prediction.

Key words: near infrared spectroscopy; competitive adaptive reweighting sampling method; random frog; partial least squares regression; support vector regression

蓝莓学名越橘, 杜鹃花科, 其成熟果实可食率达 100%, 果肉细腻, 有酸甜可口的味道且伴有清爽宜人的香气。富含花色苷、维生素 C、维生素 E、黄酮醇、花青素等多种抗氧化物质, 被誉为“浆果之王”^[1]。果实的营养物质含量与其成熟度、贮藏时间及腐败程度的关联甚密, 是加工产品品质及鲜食口感的重要因素之一。消费者在购买过程中除关心水果的味道外, 营养成分作为重要指标也有着越来越高的关注度^[2]。蓝莓内部品质的传统检测方法多为破坏性检测, 耗时长、效率低, 无法满足生产实践中大批量检测的需求, 亟待发展无损检测技术对蓝莓内部品质进行快速检测。

近年来, 可见/近红外光谱、高光谱成像、电子鼻以及超声波检测等技术已经尝试用于水果内在品质的无损检测^[3-4]。其中, 近红外光谱技术(Near-Infrared Spectroscopy, NIR)因其操作简单、检测速度快、稳定性好、无污染等优点, 逐渐成为农产品无损检测的一项重要技术手段。陈蓓等^[5]利用竞争适应性重加权采样法(Competitive Adaptive Reweighted Sampling, CARS)、无信息变量消除法、随机蛙跳算法(Random Frog, RF)和连续投影算法对 NIR 波长进行降维处理, 定量预测人参皂苷中 Rg1 和 Rc 含量, 所建偏最小二乘模型决定系数均达到 0.94 以上。Agulheiro 等^[6]对光谱进行一阶求导后利用偏最小二乘回归定量分析草莓可溶性固形物含量, 校正集决定系数和预测集决定系数分别达到 0.9277 和 0.8207。姜妍等^[7]通过对光谱进行一阶求导、去趋势算法和标准正太变换组合算法预处理后建立偏最小

二乘模型, 对青饲大豆三种主要品质参数进行定量分析, 结果达到了较高的预测精度。以上研究表明, 光谱预处理方法及分析模型的选择对农作物品质的分析精度具有较大的影响, 不同种类预处理和模型的组合方式会导致不同的分析结果。尽管国内外基于 NIR 预处理和建模的组合方式有很多, 但对于分析不同贮藏时间下的蓝莓果实内部品质的预处理和模型的组合方式还需进一步的研究。

本研究收集了不同贮藏时间的蓝莓果实样品, 基于 NIR 技术, 比较多种预处理方法, 并采用最佳的预处理方法结合竞争适应性重加权采样法和随机蛙跳算法筛选最佳的特征波长, 运用偏最小二乘回归(Partial Least Square Regression, PLSR)和支持向量回归(Support Vector Regression, SVR)两种技术建立蓝莓营养成分含量快速预测模型, 第一种属于线性方法, 被认为是使用 NIR 光谱时的标准技术之一; 第二种属于非线性方法, 可能更适合定量分析。通过比较线性和非线性两种方法, 以期进一步提高蓝莓营养成分的预测精准度, 并为蓝莓收储提供一种快速准确的检测方法。

1 材料与方法

1.1 材料与仪器

蓝莓品种 绿宝石, 挑选大小均匀、无损伤的成熟新鲜蓝莓于采摘当天低温贮藏运回实验室, 为确保实验样品具有代表性梯度范围, 蓝莓以每 6 个为一组分为 300 组装在 PET 专用水果包装盒放置于 10 °C 恒温箱内贮藏, 每次实验取 30 组蓝莓, 待恢复室温后采集其 NIR 光谱, 再进行相关理化指标的测定。

样品从贮藏当天开始,每天选取 30 组样品进行测定,共收集了蓝莓贮藏 1~10 d 的光谱及理化值数据。

LabSpec 5000 型光谱仪 美国 ASD 公司;UV-1801 紫外可见分光光度计 北京北分瑞利公司;LYT-330 手持式折光仪 上海淋誉贸易有限公司;PX-70B Ⅲ 生化培养恒温箱 天津泰斯特仪器有限公司。

1.2 实验方法

1.2.1 NIR 采集 采用 LabSpec 5000 型光谱仪采集光谱,光谱扫描前,机器预热时间为 30 min,以保证采集结果的稳定,设置采集参数波长范围 350~2500 nm,使用配套软件 Indico ProVersion3.1 采集光谱,采集间隔 1 nm。在样品果腹赤道处每隔约 120°扫描光谱 3 次,以 3 次所得平均结果作为该单果样品的 NIR 光谱,以每组中 6 个样品的平均光谱代表该组样本的光谱数据。

1.2.2 理化值测定 由于果实的化学性质在全贮藏过程中将会发生变化,因此这些性质的无损估计对于销售及健康食用等具有重要的使用价值^[8]。在这方面,本研究考虑的两个化学性质是可溶性固形物(Soluble Solids Content, SSC)和维生素 C(Vitamin C, V_C)。蓝莓样品的这些参数的实际值是在破坏性实验室分析中测量的,在光谱分析之后以下列方式进行。

1.2.2.1 SSC 测定 取一组蓝莓样品,放入研钵捣碎,用纱布包裹,挤出蓝莓汁液,对手持式折光仪进行零点校正后,将汁液滴于折光仪获得读数,测量 3 次取平均值,作为该组样本 SSC 实测值^[9]。

1.2.2.2 V_C 测定 称取 2 g 捣碎的蓝莓加入 1% HCL 研磨均匀,加入蒸馏水定容至 25 mL。取滤液 2 mL 加 0.2 mL 10% HCL,加水稀释至 10 mL。以蒸馏水为空白对照,使用紫外分光光度计在 423 nm 处测待测液吸光度值,并查标准曲线($C=15.12A+0.1222$, $R^2=0.9833$,由抗坏血酸标准液测得)计算得出^[10-11]。

1.3 数据处理

1.3.1 光谱数据预处理 建立 SSC、 V_C 含量的定量识别模型时,采用马氏距离^[12-13]的方法来识别蓝莓样品中的异常样品光谱,剔除马氏距离大于 10 的样本。运用基于 x-y 联合距离的样本集划分方法(SPHY)^[14],按照训练集与验证集样品数量之比 4:1,将剔除异常样品后的全部样品光谱及理化值划分为校正集和预测集,利用校正集数据进行模型的训练,利用预测集数据进行模型的性能测试。

原始光谱包括由于背景和环境光、样品的曲率以及不同大小的蓝莓造成的不需要的信息,为提高光谱分辨率,去除样品表面颗粒不均匀带来的杂散光等,消除或减弱非目标因素的干扰,提高信噪比和判别准确度,实现模型的可靠校准和归一化,建立相对稳定的数学模型,因此有必要对 284 个样本的原始

光谱进行预处理,本研究将在软件 Unscrambler X 10.4 中实现光谱的预处理。光谱预处理主要采用一阶导数(First Derivative, 1-DER)、二阶导数(Second Derivative, 2-DER)、标准正态变换(Standard Normal Variate Transform, SNV)、多元散射校正(Multiplicative Scatter Correction, MSC)、Savitzky-Golay 平滑(S-G)等一种或几种方法组合进行比较分析。其中导数算法可以消除样品背景干扰、区分重叠峰、提高灵敏度;SNV 可以消除固体颗粒大小、表面散射以及光程变化对 NIR 光谱的影响;MSC 主要是消除颗粒分布不均匀及颗粒大小产生的散射影响;S-G 旨在平滑噪声数据并消除具有大障碍的数据点^[15-17]。这几种方法单独或组合使用都会使分析结果不同,因此,需要大量的实验来验证以得到更合适的预处理方法。对光谱进行预处理后,使用偏最小二乘法分别建模。最后,利用校正相关系数(Corrected Correlation Coefficient, R_c)、校正均方根误差(Root Mean Square Error of Correction, RMSEC)、预测相关系数(Prediction Correlation Coefficient, R_p)、预测均方根误差(Root Mean Square Error of Prediction, RMSEP)来定量比较预测效果,计算公式如式(1)和(2)所示。 R_c 和 R_p 取值越接近 1, RMSEC 和 RMSEP 越接近 0,模型的拟合性越好,预测精度越高。

$$R = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}} \quad \text{式 (1)}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad \text{式 (2)}$$

式(1)和(2)中, RMSE 为均方根误差; R 为相关系数; n 为样本个数; $\hat{y}$ 为第 i 个样本的观测值; $\bar{y}$ 为观测值的平均值; y_i 为第 i 个样本的预测值 ($i=1, 2, 3, \dots, n$)。

1.3.2 特征波长选择算法 蓝莓营养成分预测模型的建立,需将所测得的蓝莓样品近红外光谱数据作为预测模型的输入变量,考虑到全光谱的输入会有信号谱带的重叠问题,且降低模型的运行速度^[18];所以将全光谱进行特征提取,以消除原始数据之间存在的共线性关系,提高预测结果稳定性^[19]。采用 CARS 和 RF 两种原理不同的算法,对预处理后的光谱进行特征波长筛选,提取特征变量作为预测模型的输入,此过程及后续的模型建立均在软件 matlab 2018b 中完成。

CARS 算法是通过蒙特卡罗采样建立 PLSR 模型,基于自适应重加权采样和一个指数递减函数,获得多个波长子集。然后,为每个波长子集建立一个交叉验证模型。最后根据交叉验证均方根误差(Cross Validation Root Mean Squared Error, RMSECV)值最小原则选择最终的波长^[20-21]。由于自适应重加权采样和指数递减函数的随机性,每次的筛选结果都不

同,因此需要多次执行 CARS 算法,将 RMSECV 值最小时对应的波长变量作为最终的选择。

RF 算法类似于可逆跳跃马尔可夫链蒙特卡罗,通过迭代进行,在随机选取的波长子集的基础上生成一个新的波长子集,并以一定的概率被接受。循环该步骤,直到完成 N 次迭代并选择区间子集,计算每次迭代中每个变量被选中的概率^[22-23]。概率值越高,变量的重要性越大;具有最高概率的变量被选为特征变量。

1.3.3 预测模型 随着光谱技术的发展,PLSR 也成为其分析时的常用方法。在通过 PLSR 方法创建模型之前,需要确定潜在变量的数量。本研究利用留一法交叉验证优化潜在变量的数量。潜在变量的最佳数量对应于最低的 RMSECV 值^[24-27]。应用此过程后,可以得到最佳的潜在变量,从而可得到更好的预测效果。

SVR 用于解决机器学习的问题,被认为是功能强大且能灵活应用的一种计算算法。目前,被广泛应用于具有非线性性质的、数据量大且繁杂的工作中。与 PLSR 一样,SVR 也需要在校准过程之前设置一些参数。参数 ε 决定了回归曲线的平坦程度,且 $0 < \varepsilon < 1$;核函数参数 g 和惩罚因子参数 c 对支持向量回归模型的精度和泛化能力的影响是最主要的,因此需要确定这两个参数以提高模型的效果^[28-29]。本文利用网格搜索技术的全局搜索能力,通过设定网格搜索的 c 和 g 值的搜索范围和步长,自动确定 c 和 g 参数的取值,然后进行交叉验证,找出精确度最高的(c, g)对,从而减少参数选择的盲目性,提高参数选择的精度^[30-31]。

2 结果与分析

2.1 样本数据处理结果

图 1 所示为蓝莓样本第 1~10 d 的平均原始光谱曲线图,不同颜色的曲线代表不同的储藏期,图中第 1~10 d 的光谱反射率由高到低依次下降。由于 350~500 nm 和 2300~2500 nm 的范围内的光谱噪声

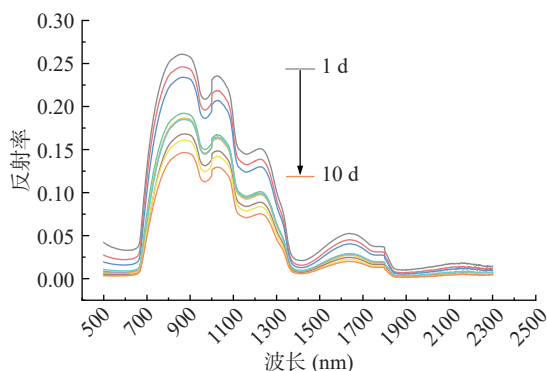


图 1 不同贮藏期蓝莓原始平均光谱曲线图

Fig.1 Original average spectral curve of blueberries in different storage periods

注:曲线图高度随贮藏天数的增加依次下降,最高曲线对应贮藏期 1 d,最低曲线对应贮藏期 10 d。

太大,因此截取 500~2300 nm 的光谱作为本研究的分析范围。在 700~1300、1440~1830 nm 范围,不同贮藏期蓝莓样品之间的反射率差异较大,因为该区域主要承载了 -OH、-NH、-CO 和 C-H 官能团信息,而 SSC 和 V_C 的化学结构主就是由这类化学键组成,因此,利用近红外光谱可用于检测蓝莓的 SSC 和 V_C 。

图 2 为 300 个样本的马氏距离,黑色直线以上的部分是马氏距离大于 10 的样本,剔除马氏距离大于 10 的样本共 16 个,将其余的 284 个样本按 4:1 划分为校正集和预测集。SSC 和 V_C 分布范围、校正集及预测集的平均值如表 1 所示。SSC、 V_C 校正集的分布范围大于预测集的分布范围,证明 SPXY 算法划分后的校正集具有代表性,可以为模型提供更准确的结果。

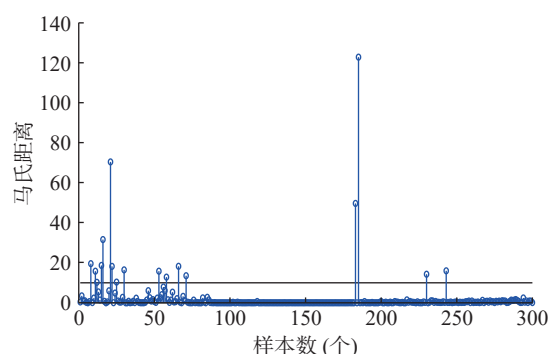


图 2 所有样本马氏距离

Fig.2 Mahalanobis distance for all samples

表 1 SPXY 算法划分校正集和预测集结果统计

Table 1 SPXY algorithm divides calibration set and prediction set result statistics

指标	样本集	数量(个)	最大值	最小值	平均值	标准差
可溶性固形物(%)	校正集	227	20.80	7.87	13.47	1.93
	预测集	57	15.5	8.93	13.04	1.30
维生素C(mg/100 g)	校正集	227	61.56	17.89	34.68	6.10
	预测集	57	44.78	24.84	35.66	4.54

为了筛选最佳的光谱预处理方法,在 Unscrambler X 10.4 软件中得出 12 种预处理组合,其中 1-DER 采用间隔点数 7, 2-DER 采用间隔点数 6, S-G 采用窗口宽度 13,多项式导数为 3。用 PLSR 建立模型,通过比较每种组合所得出的 R_c 、RMSEC、 R_p 、RMSEP 来衡量预处理的效果。光谱预处理结果如表 2 所示,通过 MSC+2-DER(图 3)组合处理的光谱数据用于预测可溶性固形物和维生素 C 含量均表现出相对较好的效果,且相较于原始光谱预测效果也大幅提升,说明该预处理方法可提供更加有效的信息,可用于进一步波长筛选及建模分析。

2.2 特征波长筛选及效果评价

将经过光谱预处理的原始光谱利用 CARS 算法和 RF 算法提取特征波长,以下是 SSC 和 V_C 特征波长提取过程。

CARS 的提取中,蒙特卡罗采样设置为 15 次,

表 2 不同组合预处理结果
Table 2 Different combinations of preprocessing results

预处理方法	R_c	RMSEC	R_p	RMSEP
原始光谱	0.6860	1.4447	0.5202	0.9026
MSC	0.6660	1.4892	0.4427	0.9851
SNV	0.6665	1.4883	0.4458	0.9818
1-DER	0.7381	1.3050	0.5334	1.0571
2-DER	0.7811	1.2155	0.6473	0.9826
S-G+1-DER	0.6715	1.4413	0.5313	0.9200
S-G+2-DER	0.7652	1.2549	0.7279	0.8757
SNV+1-DER	0.6375	1.5016	0.7318	0.9064
SNV+2-DER	0.5607	1.6061	0.6057	1.0591
MSC+1-DER	0.6432	1.4930	0.6910	0.9943
MSC+2-DER	0.7796	1.2052	0.7376	0.9673
MSC+S-G+2-DER	0.6971	1.3663	0.6323	1.1569
SNV+S-G+2-DER	0.7678	1.2410	0.7113	1.0037
原始光谱	0.6758	4.4561	0.5750	3.0163
MSC	0.6718	4.5356	0.6648	3.4054
SNV	0.6751	4.5171	0.6549	3.4271
1-DER	0.7399	4.1244	0.7461	3.0205
2-DER	0.7929	3.7446	0.6720	3.1022
S-G+1-DER	0.6788	4.5284	0.6956	2.8505
S-G+2-DER	0.7973	3.7003	0.6587	3.3886
SNV+1-DER	0.6857	4.4916	0.5805	3.7441
SNV+2-DER	0.8161	3.5434	0.7070	3.4925
MSC+1-DER	0.6865	4.4860	0.5383	3.9248
MSC+2-DER	0.8089	3.5839	0.7021	3.6885
MSC+S-G+2-DER	0.6902	4.3124	0.6697	3.7040
SNV+S-G+2-DER	0.7825	3.8575	0.6762	3.4760

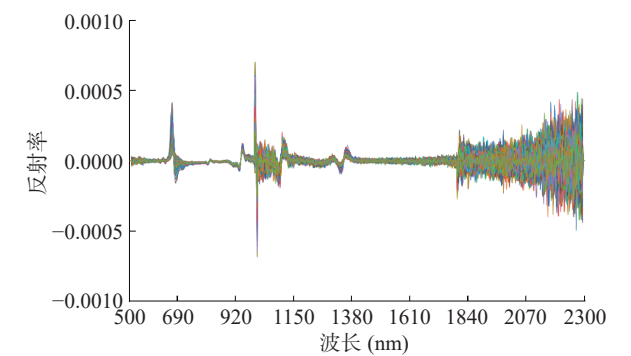


图 3 MSC+2-DER 预处理后的近红外光谱
Fig.3 NIR spectra after MSC+2-DER pretreatment

9 折交叉验证。300 次迭代过程中 SSC 和 V_C 的 RMSECV 变化如图 4(a)和图 4(b)所示,RMSECV 值随着重复选择次数的增加均呈现先缓慢下降,然后再整体上升的趋势,最后趋于平缓,当 RMSECV 值达到最小值时,为所建 PLSR 模型的最优结果,此时对于 SSC 和 V_C 采样分别运行了 178 次和 162 次, RMSECV 值分别为 1.427 和 4.262。所选波长数量分别为 32 个和 46 个,对应于全波段的位置分别如图 5(a)和图 5(b)所示。

基于 RF 算法,对于 SSC 和 V_C , 每个变量被选择的概率如图 6(a)和图 6(b)所示,选择概率大于 0.3 的变量为最终特征波长,SSC 和 V_C 得到满足条件的特征波长分别有 79 个和 53 个,如图 7(a)和图 7(b)所示。

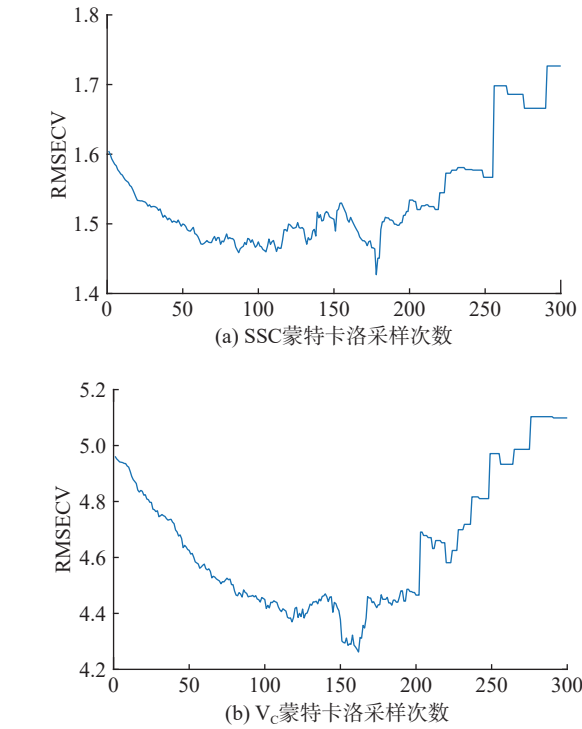


图 4 RMSECV 随采样次数变化过程
Fig.4 Process of RMSECV changing with the number of samples
注: RMSECV 指交叉验证均方根误差。

将两种方法所选波长变量结合起来,去掉重复的波长,SSC 和 V_C 分别有 100 个和 87 个波长变量。以上优选出的三种波长组合将分别作为模型的输入,通过比较建模结果分析 CARS 和 RF 两种算法的适用性。

2.3 建模结果与评价

基于预处理之后选出的特征变量作为输入量,分别建立 PLSR 和 SVR 模型,对蓝莓中 SSC 和 V_C 含量进行预测。

建立 SVR 模型时,利用网格法寻找最佳的参数

表 3 PLSR 和 SVR 模型预测效果
Table 3 PLSR and SVR model prediction effect

建模方法 波长筛选方法 筛选波长数量			R_c	RMSEC	R_p	RMSEP
SSC	PLSR	全波段	0.6971	1.3663	0.6323	1.1569
		CARS	0.8202	1.0832	0.8104	0.8292
		RF	0.8841	0.8836	0.8302	0.9456
		CARS+RF	0.9001	0.8234	0.8424	0.9613
	SVR	全波段	0.6671	1.7312	0.6152	1.4336
		CARS	0.8310	1.0652	0.8019	0.8393
		RF	0.8841	1.0199	0.8160	0.8889
		CARS+RF	0.8702	0.9549	0.8269	0.8769
	PLSR	全波段	0.8089	3.5839	0.7021	3.6885
		CARS	0.8574	3.0865	0.8468	3.3039
RF		0.8470	3.2039	0.8061	3.1132	
CARS+RF		0.8707	2.9429	0.8350	2.9482	
V_C	SVR	全波段	0.7786	3.2895	0.6308	3.8733
		CARS	0.8510	3.1694	0.8394	2.9012
		RF	0.8371	3.3612	0.8148	2.7946
		CARS+RF	0.8503	3.2431	0.8183	2.8818

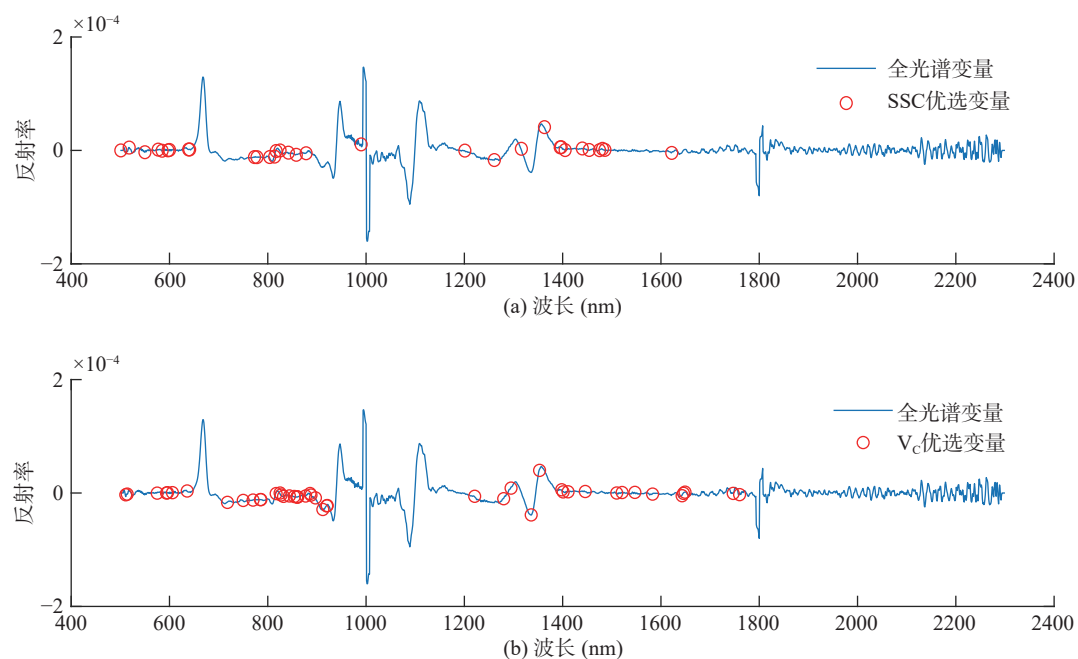


图5 优选波长分布

Fig.5 Preferred wavelength distribution

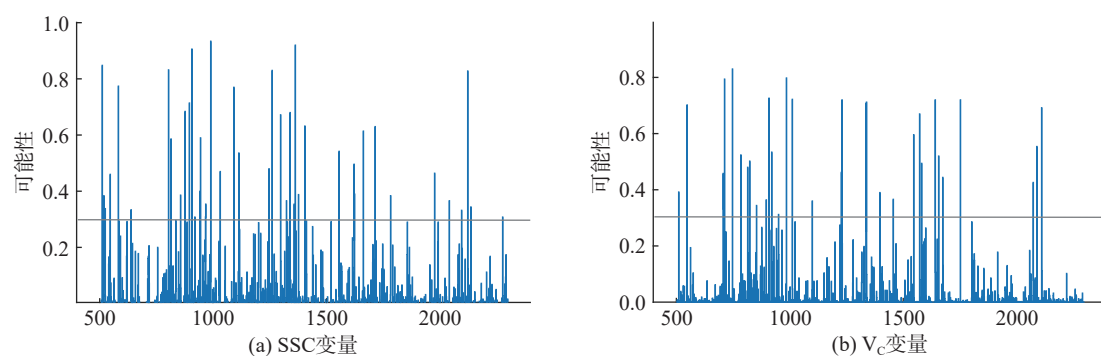


图6 变量选择概率

Fig.6 Variable selection probability

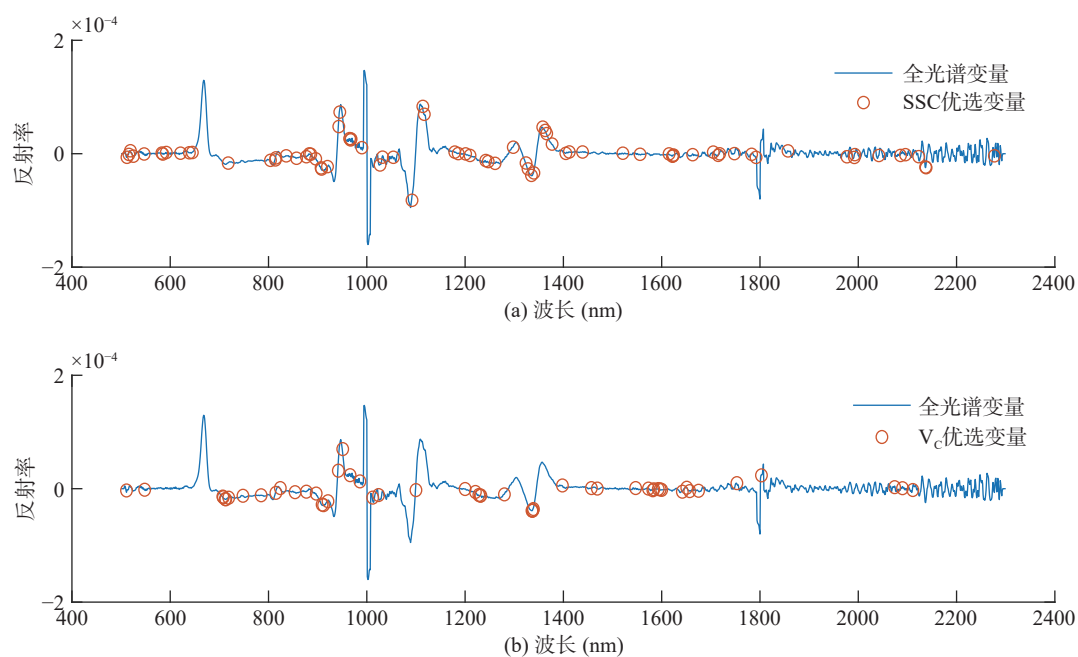


图7 优选波长分布

Fig.7 Preferred wavelength distribution

c 和 g。本文设定 $\varepsilon=0.01$, 设定网格搜索的 c、g 值的搜索范围和步长, 这里设置为 $c \in [0, 10]$, 步长为 0.5, $g \in [-10, 0]$, 步长为 1, 将模型运行 50 次的结果取平均值作为最终预测结果进行分析。两种模型的预测结果见表 3 所示。

波长选择后, 2 种优化算法消除了大量与预测内容无关的波长, 用于建模的光谱数据量显著减少。将 CARS 算法和 RF 算法提取到的特征波长及组合波长建立 PLSR 模型, 相较于全波段所建立的 PLSR 模型, 优选出的特征波长所得到的 R_c 和 R_p 均大幅提高, RMSEC 和 RMSEP 也相对减少, SSC 波长变量分别降到了全波段变量 1.7%、4.3%、5.6%, V_C 波长变量分别降到了全波段变量的 2.5%、2.9%、4.8%。说明, 利用 CARS 和 RF 特征波长提取方法均可提高模型对蓝莓中 SSC 和 V_C 含量的预测精度。

三种优选变量组合中, CARS+RF 组合的波长数量最多, 建模结果最好。其余两种波长组合中, 对于 SSC 而言, RF 组合的建模效果比 CARS 组合更好; 对于 V_C 而言, CARS 组合的建模效果比 RF 组合更好。总之, 采用 CARS 和 RF 算法选出的光谱特征波长, 能够保留有效的光谱信息, 在此基础上将两者相结合, 进一步提高了模型的预测效果, 实现了算法之间的优势互补。

图 8 为利用 CARS 结合 RF 算法提取的特征波长下, SSC 和 V_C 含量真实值与预测值的变化趋势, (a) 和 (c) 为 PLSR 模型对 SSC 和 V_C 含量的预测结果, (b) 和 (d) 为 SVR 模型对 SSC 和 V_C 含量的预测结果。两种模型进行对比发现, 非线性的 SVR 算法与线性的 PLSR 算法的建模效果相差不大, 但 SVR 算法并没有表现出更好的预测效果。

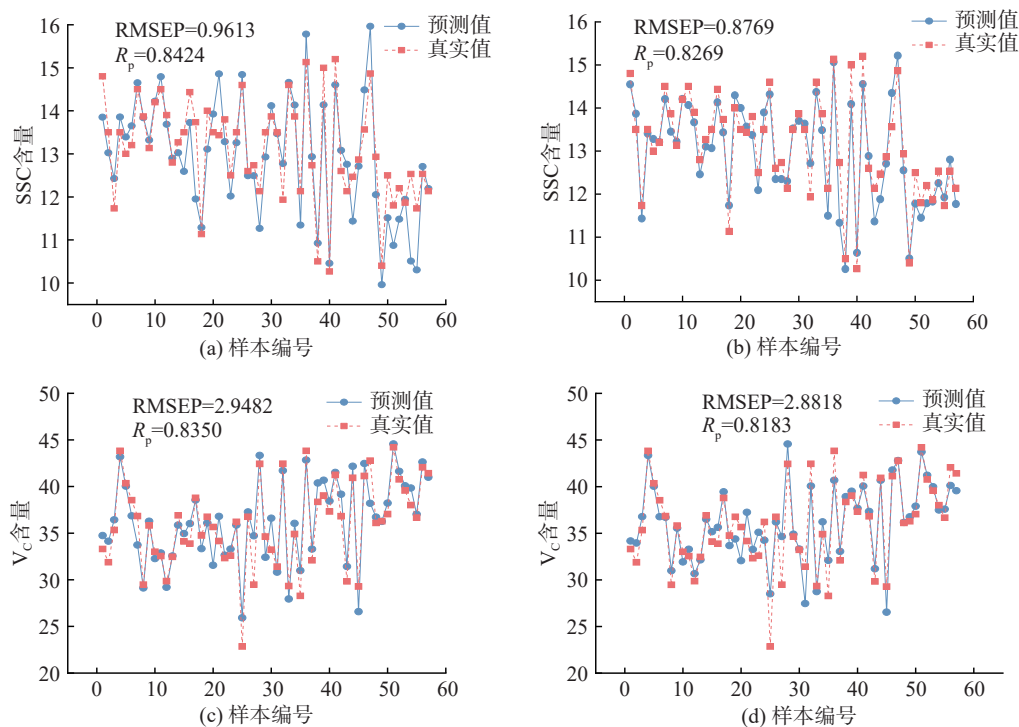


图 8 SSC 与 V_C 预测值与真实值对比图

Fig.8 Comparison chart of predicted values and real values of SSC and V_C

注: (a)、(c): PLSR 模型预测结果; (b)、(d): SVR 模型预测结果。

3 结论

本研究结合 CARS 和 RF 算法筛选特征波长, 利用 PLSR 和 SVR 算法对蓝莓中的 SSC 和 V_C 含量进行快速定量预测。采用近红外光谱仪采集 300 组蓝莓的原始光谱, 利用马氏距离剔除 16 个异常样本, 比较五种不同的光谱预处理方法 (1-DER、2-DER、SNV、MSC、S-G), 并结合各种预处理方法对蓝莓样品的原始光谱进行处理, 并建立 PLSR 模型进行对比, 选择 MSC 结合 2-DER 作为蓝莓原始光谱的最佳预处理方法, SSC 和 V_C 含量预测的 R_c 分别为 0.7796、0.8089, RMSEC 分别为 1.2052、3.5839, R_p 分别为 0.7376、0.7021, RMSEP 分别为 0.9673、

3.6885。利用 CARS 和 RF 算法及两种算法组合对预处理后的全光谱波段进行特征波长筛选, 对于 SSC 和 V_C , CARS 算法分别筛选出 32 和 46 个特征波长, RF 算法分别筛选出 79 和 53 个特征波长, 组合筛选特征波长分别为 101 和 87 个。比较筛选出的三种波长组合对 PLSR 模型预测性能的影响, 发现两种波长筛选算法均能有效选取 SSC 和 V_C 的特征波长, CARS 结合 RF 算法提取的特征波长预测结果最好。为了进一步探索本研究中校正模型的预测性能, 将其与相同波长组合下的 SVR 模型进行了比较, 发现线性 PLSR 和非线性 SVR 算法分别建立的校正模型均具有较好的预测性能, 但 SVR 的预测结

果略低于 PLSR。研究表明,近红外光谱结合机器学习算法是一种快速、无损检测蓝莓营养物质含量的方法,波长选择技术可以提高预测模型对蓝莓中 SSC 和 V_C 的预测性能。机器学习与波长筛选算法相结合便于建立无损检测通用模型,能够有效预测蓝莓中多种营养物质,为其质量评价及监测提供了一种有效的途径。

本研究在蓝莓贮藏温度 10℃ 条件下,共采集 300 组样本的光谱数据、SSC 和 V_C 含量,为近红外光谱快速检测蓝莓 SSC 和 V_C 含量提供了理论依据和数据支持。为进一步提高模型预测精度,增强模型的泛化性能,在今后的研究中将增加样本量,扩大蓝莓种类和贮藏温度的采集范围,对数据处理方法及预测模型进行更深入的研究。

参考文献

- [1] 张丽娟,夏其乐,陈剑兵,等.近红外光谱的三种蓝莓果渣花色苷含量测定[J].光谱学与光谱分析,2020,40(7):2246-2252. [ZHANG L J, XIA Q L, CHEN J B, et al. Determination of anthocyanins in three blueberry pomace by near-infrared spectroscopy[J]. Spectroscopy and Spectral Analysis, 2020, 40(7): 2246-2252.]
- [2] 褚小立,陈瀑,李敬岩,等.近红外光谱分析技术的最新进展与展望[J].分析测试学报,2020,39(10):1181-1188. [ZHU X L, CHEN P, LI J Y, et al. The latest progress and prospect of near-infrared spectroscopy analysis technology[J]. Journal of Analysis and Testing, 2020, 39(10): 1181-1188.]
- [3] 王健伟,陶飞.近红外光谱技术在农产品检测中的应用研究进展[J].安徽农学通报,2021,27(17):155-158. [WANG J W, TAO F, et al. Research progress on the application of near-infrared spectroscopy in the detection of agricultural products[J]. Anhui Agricultural Science Bulletin, 2021, 27(17): 155-158.]
- [4] LAN W J, BAETEN V, JAILLAIS B, et al. Comparison of near-infrared, mid-infrared, Raman spectroscopy and near-infrared hyperspectral imaging to determine chemical, structural and rheological properties of apple purees[J]. Journal of Food Engineering, 2022: 111002.
- [5] 陈蓓,郑恩让,郭拓.多种光谱变量筛选算法在红参提取近红外建模中的应用[J].光谱学与光谱分析,2021,41(8):2443-2449. [CHEN B, ZHENG E R, GUO T. Application of multiple spectral variable screening algorithms in near-infrared modeling of red ginseng extraction[J]. Spectroscopy and Spectral Analysis, 2021, 41(8): 2443-2449.]
- [6] AGULHEIRO SANTOS A C, RICARDO RODRIGUES S, LARANJO M, et al. Non-destructive prediction of total soluble solids in strawberry using near infrared spectroscopy[J]. Journal of the Science of Food and Agriculture, 2022, 102(11): 4866-4872.
- [7] 姜妍,蒙何,赵熠荣,等.青饲大豆主要品质参数的近红外光谱快速分析[J].光谱学与光谱分析,2022,42(3):919-923. [JIAN Y, MENG H, ZHAO Y R, et al. Rapid analysis of main quality parameters of green forage soybean by near-infrared spectroscopy[J]. Spectroscopy and Spectral Analysis, 2022, 42(3): 919-923.]
- [8] SYLVIA T, MARIANNE G B, KAREN K P, et al. Predicting pear (cv. Clara Frijs) dry matter and soluble solids content with near infrared spectroscopy[J]. LWT-Food Science and Technology, 2014, 59(2): 1107-1113.
- [9] 曹建康,姜微波,赵玉梅.果蔬采后生理生化实验指导[M]. 北京:中国轻工业出版社,2007:50-51. [CAO J K, JIANG W B, ZHAO Y M. Postharvest physiological and biochemical experiment guidance for fruits and vegetables[M]. Beijing: China Light Industry Press, 2007: 50-51.]
- [10] 延芸芸.食物中维生素C含量的测定[J].大众标准化,2015(11):67-69. [YAN Y Y. Determination of vitamin C content in food[J]. Mass Standardization, 2015(11): 67-69.]
- [11] 黄锦佳,吴岚芳.热带园艺产品采后实验原理与技术[M].北京:中国林业出版社,2012:172-174. [HUANG J J, WU L F. The principle and technology of postharvest experiment of tropical horticultural products[M]. Beijing: China Forestry Press: 172-174.]
- [12] 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. [LIU C L, HU Y J, WU S N, et al. Research on the method of removing singular samples from near-infrared spectroscopy[J]. Journal of Food Science and Technology, 2014, 32(5): 74-79.]
- [13] CHE W K, SUN L J, ZHANG Q, et al. Application of visible/near-infrared spectroscopy in the prediction of azodicarbonamide in wheat flour[J]. Journal of Food Science, 2017, 82(10): 2516-2525.
- [14] 王世芳,韩平,崔广禄,等.SPXY算法的西瓜可溶性固形物近红外光谱检测[J].光谱学与光谱分析,2019,39(3):738-742. [WANG S F, HAN P, CUI G L, et al. Near-infrared spectral detection of watermelon soluble solids by SPXY algorithm[J]. Spectroscopy and Spectral Analysis, 2019, 39(3): 738-742.]
- [15] 王赋腾,孙晓荣,刘翠玲,等.光谱预处理对便携式近红外光谱仪快速检测小麦粉灰分含量的影响[J].食品工业科技,2017,38(10):58-66. [WANG F T, SUN X R, LIU C L, et al. Influence of spectral pretreatment on rapid detection of ash content in wheat flour by portable near-infrared spectrometer[J]. Food Industry Science and Technology, 2017, 38(10): 58-66.]
- [16] 曹念念,刘强,彭菁,等.基于近红外光谱技术的黄桃脆片可溶性固形物和硬度定量检测方法[J].食品与机械,2021,37(3):51-57. [CAO N N, LIU Q, PENG J, et al. Quantitative detection method for soluble solids and hardness of yellow peach chips based on near-infrared spectroscopy[J]. Food and Machinery, 2021, 37(3): 51-57.]
- [17] HEIL K, SCHMIDHALTER U. An evaluation of different nir-spectral pre-treatments to derive the soil parameters C and N of a humus-clay-rich soil[J]. Sensors, 2021, 21(4): 1423-1423.
- [18] XU H R, QI B, SUN T, et al. Variable selection in visible and near-infrared spectra: Application to on-line determination of sugar content in pears[J]. Journal of Food Engineering, 2012, 109(1): 142-147.
- [19] 李浩光,于云华,逢燕,等.基于遗传算法的近红外光谱定性分析特征波长提取方法研究[J].光谱学与光谱分析,2021,41(8):2437-2442. [LI H G, YU Y H, PANG Y, et al. Research on extraction method of characteristic wavelength for qualitative analysis of near-infrared spectrum based on genetic algorithm[J]. Spectroscopy and Spectral Analysis, 2021, 41(8): 2437-2442.]
- [20] 张立欣,杨翠芳,陈杰,等.基于变量优选的苹果糖分含量近红外光谱检测[J].食品与机械,2021,37(10):112-118. [ZHANG L X, YANG C F, CHEN J, et al. Detection of sugar content in apples by near-infrared spectroscopy based on variable optimization[J]. Food and Machinery, 2021, 37(10): 112-118.]
- [21] 陈蓓,郑恩让,郭拓.多种光谱变量筛选算法在红参提取近红外建模中的应用[J].光谱学与光谱分析,2021,41(8):2443-2449. [CHEN B, ZHEN E R, GUO T. Application of multiple spectral variable screening algorithms in near-infrared modeling of

- red ginseng extraction[J]. *Spectroscopy and Spectral Analysis*, 2021, 41(8): 2443–2449.]
- [22] MISHRA P, HERRMANN I, ANGILERI M. Improved prediction of potassium and nitrogen in dried bell pepper leaves with visible and near-infrared spectroscopy utilising wavelength selection techniques[J]. *Talanta*, 2021, 225: 121871–121971.]
- [23] 撒继铭, 江河, 谢凯文, 等. 基于随机蛙跳算法的特征波长优选[J]. *光学学报*, 2021, 41(15): 235–243. [SA J M, JIANG H, XIE K W, et al. Characteristic wavelength optimization based on random frog leaping algorithm[J]. *Acta Optics Sinica*, 2021, 41(15): 235–243.]
- [24] BAI X L, ZHANG L, KANG C Y, et al. Near-infrared spectroscopy and machine learning-based technique to predict quality-related parameters in instant tea[J]. *Scientific Reports*, 2022, 12(1): 3833–3833.]
- [25] RODRIGUEZ JUAN D, PEREZ A, LOZANO J A. Sensitivity analysis of k-fold cross validation in prediction error estimation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, 32(3): 569–575.]
- [26] LIU R, QI S Y, LU J, et al. Measurement of soluble solids content of three fruit species using universal near infrared spectroscopy models[J]. *Journal of Near Infrared Spectroscopy*, 2015, 23(5): 301–309.]
- [27] DANIEL D S C, NELSON F O M, MURILO S F, et al. Development of predictive models for quality and maturation stage attributes of wine grapes using vis-NIR reflectance spectroscopy[J]. *Postharvest Biology and Technology*, 2019, 150: 166–178.]
- [28] 郭美丽, 覃锡忠, 贾振红, 等. 基于改进的网格搜索 SVR 的话务预测模型[J]. *计算机工程与科学*, 2014, 36(4): 707–712.]
- [GUO M L, QIN X Z, JIA Z H, et al. Traffic prediction model based on improved grid search SVR[J]. *Computer Engineering and Science*, 2014, 36(4): 707–712.]
- [29] KOVALENKO IGOR V, RIPPKE GLEN R, HURBURGH CHARLES R. Determination of amino acid composition of soybeans (*Glycine max*) by near-infrared spectroscopy[J]. *Journal of Agricultural and Food Chemistry*, 2006, 54(10): 3458–3491.]
- [30] 杜聪, 邵建华, 杨薇, 等. 网格搜索法优化的支持向量机室内可见光定位[J]. *激光杂志*, 2021, 42(3): 104–109. [DU C, SHAO J H, YANG W, et al. Indoor visible light localization with support vector machine optimized by grid search method[J]. *Laser Journal*, 2021, 42(3): 104–109.]
- [31] 罗小燕, 陈慧明, 卢小江, 等. 基于网格搜索与交叉验证的 SVM 磨机负荷预测[J]. *中国测试*, 2017, 43(1): 132–135. [LUO X Y, CHEN H M, LU X J, et al. SVM mill load prediction based on grid search and cross-validation[J]. *China Test*, 2017, 43(1): 132–135.]