英红九号发酵叶中茶褐素近红外定量模型的优化与验证

夏晶晶[1,2]，张敏[2]，王飞仁[1*]，王广海[1]，马成英[3]，黄涵[2]，郭嘉明[2,4*]，陈锦星[2]

（1.广东省机电职业技术学院汽车学院，广东广州 510515）（2.华南农业大学工程学院，广东广州 510642）

（3.广东省农业科学院茶叶研究所，广东广州 510640）

（4.岭南现代农业科学与技术广东省实验室茂名分中心，广东茂名 525000）

摘要：为提高近红外光谱分析方法快速实现对测定红茶在制品中的茶褐素的定量模型精度无损、快速检测，该试验利用近红外光谱技术对以英红九号发酵叶中茶褐素进行采集、提取和分析的检测为例，对其近红外定量检测模型的构建与优化进行了研究。首先，采用规范化处理（Normalize）、基线校正（Baseline）、S-G 一阶导数（Savitzky-Golay，1st S-G）、S-G 二阶导数（2nd S-G）、标准正态变量变换（Standard Normal Variate Transform，SNV）五种预处理方法对原始光谱进行预处理分析。然后，采用效果最好的一阶导数预处理方法进行波长特征提取，分别使用间隔偏最小二乘算法（Interval Partial Least Square，iPLS）、竞争自适应加权算法（Competitive Adaptive Reweighted Sampling，CARS）、变量迭代空间收缩方法（the Variable Iterative Space Shrinkage Approach，VISSA）提取波长特征。最后，使用偏最小二乘回归（Partial Least Square，PLS）预测模型进行回归建模。研究结果表明：使用一阶导数进行预处理、同时使用 CARS 方法建立的 1st-CARS-PLS 模型效果特征更显著，特征值数量为 53 个。研究表明，该试验采用的模型方法能够快速、无损地检测英红九号发酵叶中的茶褐素含量。

关键词：近红外光谱；红茶；茶褐素

文章编号：1673-9078(2023)06-313-320 DOI: 10.13982/j.mfst.1673-9078.2023.6.0761

Optimization and Verification of a Near Infrared Quantitative Model for the Theabrownin in Yinghong No.9 Fermented Leaves

XIA Jingjing[1,2], ZHANG Min[2], WANG Feiren[1*], WANG Guanghai[1], MA Chengying[3], HUANG Han[2], GUO Jiaming[2,4*], CHEN Jinxing[2]

(1.Guangdong Mechanical & Electrical Polytechnic, Automotive College, Guangzhou 510515, China)

(2.College of Engineering, South China Agricultural University, Guangzhou 510642, China)

(3.Tea Research Institute, Guangdong Academy of Agricultural Sciences, Guangzhou 510640, China)

(4.Maoming Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Maoming 525000, China)

Abstract: In order to improve the near-infrared spectroscopy analysis method for realizing quickly the use of the quantitative model for non-destructive and rapid detection of the theabrownin in black tea products, this research used near-infrared spectroscopy to collect, extract and analyze the theabrownin in the fermented leaves of Yinghong No.9 as the example. The construction and optimization of the near-infrared quantitative detection model are performed. Firstly, the original spectra were preprocessed and analyzed by five preprocessing methods: Normalization, baseline correction, S-G first derivative (Savitzky-Golay, 1st S-G), S-G second derivative (2nd S-G) and standard normal variable

引文格式：

夏晶晶,张敏,王飞仁,等.英红九号发酵叶中茶褐素近红外定量模型的优化与验证[J].现代食品科技,2023,39(6):313-320.

XIA Jingjing, ZHANG Min, WANG Feiren, et al. Optimization and verification of a near infrared quantitative model for the theabrownin in Yinghong No.9 fermented leaves [J]. Modern Food Science and Technology, 2023, 39(6): 313-320.

收稿日期：2022-06-15

基金项目：广东省乡村振兴战略专项项目（粤财农[2020]20 号）；茂名实验室自主科研项目（2021ZZ003）

作者简介：夏晶晶（1980-），男，博士研究生，副教授，研究方向：农产品无损检测，E-mail：2011010005@gdmec.edu.cn

通讯作者：郭嘉明（1987-），男，博士，副教授，研究方向：农产品无损检测，E-mail：jmguo@scau.edu.cn

transform (SNV). Then, the best 1st S-G preprocessing method was used to extract the wavelength features, using the interval partial least squares algorithm (iPLS), competitive adaptive weighting algorithm (CARS) and the variable iterative space shrinkage approach (VISSA), respectively. Finally, the partial least squares regression (PLS) prediction model was used for regression modeling. The results show that the 1st-CARS-PLS model established by using the first-order derivative for preprocessing and the CARS method has more significant effect characteristics, with the number of eigenvalues being 53. The research shows that the model method used in this experiment can rapidly and non-destructively detect the theabrownin content in the fermented leaves of Yinghong No.9.

Key words: near infrared spectroscopy; black tea; theabrownin

英红九号是国内公认的三大著名出口红茶[1]之一，该茶历史悠久，是当地市委、政府的农业重点扶持产业。英红九号产量高，色、香、味俱全，品质好，并且获得了大量的美誉[2-4]。

茶褐素（Theabrownine）的产生主要集中在制茶过程中的“发酵”阶段，该成分的含量决定了发酵过程进行的程度、影响成品茶的汤色。发酵的时间越长，茶褐素的含量随之升高。茶褐素的含量会影响成品茶汤色的明暗度、光泽度。发酵时间不足、茶褐素含量低、茶汤色则呈现红橙明亮的特征；发酵时间适中、茶褐素含量适中、汤色呈现红褐明亮的品质特征；发酵时间过度、茶褐素含量过多、汤色呈现暗红的品质特征[5]。因此，茶褐素含量可以作为英红九号发酵程度的一个指标。

茶褐素含量的标准检测方法 NY/T 3675-2020，是一种分光光度检测方法，该检测方法具有准确性良好、灵敏度高的特点，但具有操作繁琐、检测时间长且破坏样品的缺点。因此，快速、无损检测英红九号发酵叶中的茶褐素对于提高茶叶生产效率有重要帮助。近红外光谱技术是一种快速、无损的检测技术。近红外光谱技术目前已广泛应用于工业[6]、农业[7]和食品业[8]相关领域，同时，针对不同品种的红茶[9]、绿茶[10]和普洱茶[11]等也有应用。赵雅等[12]使用近红外光谱技术，基于偏最小二乘法在 1 872 nm 建立了茶多酚含量的预测模型，结果表明，此模型的相关系数达到 0.94，可以用于茶多酚含量的快速检测。石艳梅等[13]利用近红外光谱技术，采用偏最小二乘法结合多元线性回归实现了对茶叶 6 种主要成分快速检测。刘蕾等[14]基于 PLS 算法，通过不同光谱预处理方法对茶叶内含物建立校正模型。据此，近红外技术已广泛应用于食品及农产品的成分检测环节，并且针对茶叶的各种内含物的无损、快速检测进入了快速发展阶段。目前，针对茶叶在制品发酵过程中的茶褐素含量检测的研究较少，董春旺等[15]用不同的特征选择方法，对中小叶种工夫红茶的茶褐素等物质建立模型，校正集的相关系数（Correlation Coefficient of Calibration Set，Rc）为 0.952，预测集的相关系数（Correlation Coefficient of Predication Set，Rp）为 0.951，这仍然存在着模型精度不够高、稳定性不够强的问题。同时，以英红九号为研究对象的发酵叶茶褐素含量的研究尚未出现。

本文提出针对茶叶在制品发酵过程对茶褐素无损、快速检测定量模型的优化。首先，为减少试验环境、试验仪器对光谱数据带来的影响，本文采用规范化处理（Normalize）、基线校正（Baseline）、S-G 一阶导数（Savitzky-Golay，1st S-G）、S-G 二阶导数（2nd S-G）、标准正态变量变换（Standard Normal Variate Transform，SNV）五种预处理方法对原始光谱进行预处理分析。然后，为有针对性的提取光谱特征，建立简化高效的定量模型，本文采用效果最好的预处理方法进行波长特征提取，分别使用间隔偏最小二乘算法（Interval Partial Least Square，iPLS）、竞争自适应加权算法（Competitive Adaptive Reweighted Sampling，CARS）、变量迭代空间收缩方法（the Variable Iterative Space Shrinkage Approach，VISSA）提取波长特征。最后，使用偏最小二乘回归（Partial Least Square，PLS）预测模型进行回归建模。

本文结合新的试验和模型分析方法，对红茶在制品进行分析建模，验证近红外光谱技术在红茶在制品上应用的可行性，为实际生产过程中的发酵程度评估提供一种无损、快速的理论基础，提高茶叶生产效率。

1 材料与方法

1.1 试验样本

本试验材料采自广东省茶叶科学院英德-英红九号试验基地，所采用的茶青品种为英红九号。本试验样品茶青于 2020 年 6 月下旬在广东省英德市（地点）采集，采摘样本均为一芽二叶或一芽三叶。茶青采收后静置于英红九号试验基地的萎凋槽，铺放厚度约 4 cm，萎凋处理 8 h。萎凋后样本经揉捻机揉捻 40 min 后，再一次揉捻 40 min 完成本试验样品制备。样本经过萎凋、揉捻的过程后，揉捻叶放在发酵筐中，根据英红九号的常规发酵时间，调控室内温度保持在 22 ℃，共发酵 7 h。在发酵过程中，每个小时取出 5

个样本，每个样本约 100 g，共取样 5 次（第一次为发酵开始时设置为 0 点），合计 40 个发酵叶样本。取样后，将每个样本置于网袋中烘干，烘干机设置温度为 80 ℃，定时 3 h。

1.2 试验设备

本试验使用赛默飞二代傅里叶变换近红外（FT-NIR）光谱仪（Thermo Scientific Co., Waltham，MA，US）测量茶叶反射光谱，积分球的漫反射光谱范围为 12 000~3 800 cm^{-1}（833~2 630 nm）。本文将分辨率设置为 4 cm^{-1}，旋转杯的直径为 20 cm。样本扫描次数为 64（发酵叶采集次数为 64）次（可使样本在旋转杯中旋转一圈），每个样本采集三次数据取平均值。

1.3 理化测定

本文茶褐素含量的理化测定严格按照茶褐素标准检测方法 NY/T 3675-2020 进行。该方法为分光光度计测量法，试样中的茶褐素经沸水提取后，再分别用正丁醇、乙酸乙酯和碳酸氢钠溶液进行液液萃取得到测试液，于 380 nm 处测定吸光度，再根据经验系数计算茶褐素的含量。

1.4 数据分析方法

1.4.1 样本划分方法

SPXY 算法（Sample Set Partitioning Based on Joint x-y Distance，SPXY）是一种将 X 变量和 Y 变量同时考虑在内的样本划分方法[16]。本文运用 SPXY 算法在发酵叶样本中选出 28 个作为校正集，12 个作为预测集。

其距离计算公式为：

$$d_y(p,q)=\sqrt{(y_p-y_q)^2},\ p,q\in[1,n] \quad (1)$$

$$d_{xy}(p,q)=\frac{d_x(p,q)}{max_{p,q\in[1,n]}d_x(p,q)}+\frac{d_y(p,q)}{max_{p,q\in[1,n]}d_y(p,q)},\ p,q\in[1,n] \quad (2)$$

式中：

$d_y(p,q)$——点 p 与点 q 之间浓度的距离；

y_p 与 y_q——分别为 p 与 q 的浓度位置。

为了确保样本在光谱空间具有同样的权重，将得到距离 $d_{xy}(p,q)$，这个为样本在两个空间中的标准化距离公式。

1.4.2 光谱预处理方法

为了消除噪声干扰和基线漂移对模型性能的影响，一般在建模之前对光谱数据进行预处理。目前比较常用的近红外光谱预处理方法有规范化处理（Normalize）、基线校正（Baseline）、S-G 一阶导数（Savitzky-Golay，1st S-G）、S-G 二阶导数（2nd S-G）、标准正态变量变换（Standard Normal Variate Transform，SNV），能有效提高光谱的平滑性，减少高频噪声干扰。规范化处理（Normalize）将采集回来的数据标准化、基线校正（Baseline）消除设备带来的检测响应，SNV 主要是减少固体颗粒物大小不均和物体表面散射以及光程变换对光谱数据的影响，从而达到去除噪声的目的，S-G 导数方法用来消除背景干扰和基线校正，提高分辨率和灵敏度。

1.4.3 特征选择方法

1.4.3.1 竞争自适应重加权算法（Competitive Adaptive Reweighted Sampling，CARS）

CARS 算法是由 Li 等[17]提出的一种新型波长选择方法，其主要是基于“适者生存”的达尔文进化法则建立的。该算法通过随机抽样，并基于 PLS 算法得出权重，再通过随机加权抽样试验去将波长点进行替换、淘汰从而获得一个 RMSECV 值最小的特征子集作为最优子集。刘雪松等[18]使用三种不同的算法对黄岑中的黄岑苷成分进行特征提取，并建立了 PLS 预测模型，其中 CARS 算法的效果最好，其校正相关系数从 0.92 提升到 0.98。唐海涛等[19]利用不同类型的土壤光谱数据和有机质数据经过 CARS 算法挑选特征后建立随机森林模型。

1.4.3.2 区间偏最小二乘法（Interval Partial Least Square，iPLS）

iPLS 算法是由 Norgaard 等[20]提出的一种波长区间筛选方法。这一种方法的主要作用是提供不同细分光谱的相关信息全图，并且找出重要的光谱区域，消除无关信息区域的干扰。iPLS 方法主要基于均方根误差（Root Mean Squared Error of Cross-Validation，RMSECV）进行挑选，以及其他参数如决定系数（Squared Correlation Coeffcient，R2），和斜率、偏移量等进行评估。Yang 等[21]采用 PLS、iPLS、SiPLS 等不同的回归分析方法，采用多种预处理方法，其中，使用 iPLS 算法获得的模型的相关系数从 0.83 到 0.92。

1.4.3.3 变量迭代空间收缩方法（the Variable Iterative Space Shrinkage Approach，VISSA）

变量迭代空间收缩方法[22]是一种将集群模型和加权二进制矩阵抽样结合的一种变量挑选方法。

（1）创建一个 K*P 的二进制矩阵，以获得 K 个子数据建立 K 个子模型。并且计算每个子模型的 RMSECV 值，选取 RMSECV 值最低的一些子模型，并计算所有变量的权重。最后记录这些被选取的最佳子模型的平均 RMSECV。

（2）根据步骤（1）的权重去更新二进制矩阵，并创建新的子模型，再计算最佳子模型的 RMSECV 均值。

（3）比较（1、2）内的平均 RMSECV 值，取平均 RMSECV 值更小的模型，直到选的 RMSECV 值没有进一步的改善为止，得到一个变量集。

重复（1~3），得到新的变量集，将新的变量集与前一个变量集的 RMSECV 进行比较，如果 RMSECV 值更小，迭代出最小的 RMSECV 值。

1.4.4 偏最小二乘回归算法（Partial Least Square，PLS）

由 GELADI，P 提出的 PLS 算法是结合多元线性回归（MLR）、主成分回归（PCR）和典型相关回归（CCA）的特性。先使用 PCR 消除无用的噪声信息，再结合 MLR 建立回归模型。光谱矩阵和浓度矩阵分解的同时，将浓度矩阵的信息引入到光谱矩阵分解过程中，在每计算一个新主成分前，将光谱矩阵的得分与浓度矩阵的得分进行交换，使得到的光谱矩阵主成分直接与浓度进行关联，既克服了 PCR 只对光谱矩阵进行分解的缺点，又可以与 MLR 的特性相结合。

2 结果与分析

2.1 茶褐素理化测定结果

表 1 40 个英红九号茶褐素样本实测值

Table 1 Measured values of 40 samples of Yinghong No.9 tea theabrownine

样本编号	茶褐素含量/%	样本编号	茶褐素含量/%
1	3.999	21	6.437
2	4.105	22	6.204
3	4.074	23	5.96
4	3.827	24	5.97
5	3.949	25	6.024
6	4.771	26	6.559
7	4.904	27	6.295
8	5.056	28	6.118
9	4.495	29	6.118
10	4.818	30	6.129
11	5.634	31	6.956
12	5.459	32	6.582
13	5.248	33	6.381
14	5.386	34	6.288
15	5.219	35	6.46
16	6.189	36	7.272
17	5.745	37	6.775
18	5.694	38	6.61
19	5.911	39	6.617
20	5.784	40	7.066

采用分光光度法对 40 个发酵叶样本中茶褐素的含量进行检测，试验的结果如下图 1 所示，经过样本总量的平均计算，发酵叶样本中的茶褐素含量在 3.83%~7.27%之间。经过对茶褐素的平均含量绘图、茶褐素含量实测值（如表 1 所示），以及统计性分析（如表 2 所示）可以得知，随着发酵的时间增加，茶褐素的最小值在未进行发酵时（设置为 0 点），而最大值在发酵 7 h 后，符合英红九号制茶过程中的发酵规律。

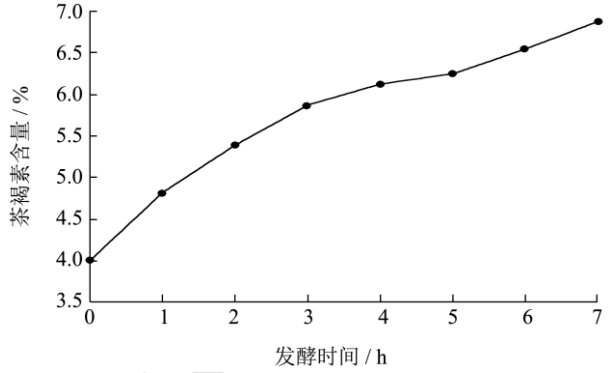


图 1 英红九号发酵叶置于发酵槽中经过 7 h 的茶褐素含量

Fig.1 The contents of theofulin in yinghong No.9 fermented leaves were placed in the fermentation tank for 7 hours

表 2 英红九号茶褐素实测值的描述性统计

Table 2 Descriptive statistics of the measured value of Yinghong No.9 tea theabrownine

内含物	最大值/%	最小值/%	均值/%	标准偏差
茶褐素	7.27	3.83	5.73	0.92

2.2 样品近红外光谱数据分析

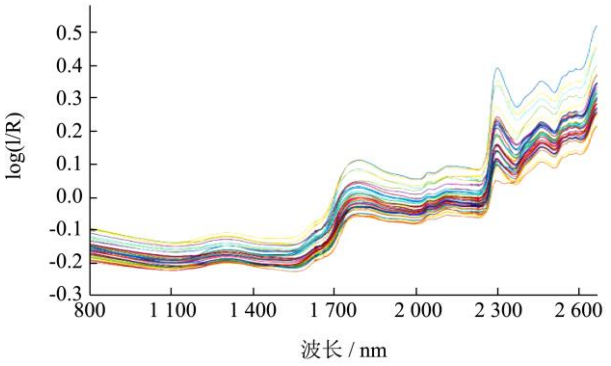


图 2 发酵叶样本的近红外光谱数据

Fig.2 Near infrared original spectrum of fermented leaves

为减少试验误差，提高近红外光谱数据的精度，本文将设置光谱仪扫描次数为 64（发酵叶采集次数为 64）次（可使样本在旋转杯中旋转一圈），每个样本采集三次数据取平均值。40 个样本的原始光谱如图 2 所示，整体而言，近红外光谱的变化趋势基本一致，说明本试验的操作可靠，但存在光谱信息冗余、噪音多的问题，需要进行进一步的预处理。试验表明，发酵

叶样本的近红外光谱吸收峰主要集中在 1 700~2 000 nm、2 300~2 600 nm 之间，且样本的走势统一，光谱的采集结果可靠。

2.3 茶褐素定量模型建立

2.3.1 近红外光谱数据预处理

为了进一步消除近红外光谱中噪声及环境因素对试验精度的影响，分别采用规范化处理（Normalize）、基线校正（Baseline）、S-G 一阶导数（Savitzky-Golay，1st S-G）、S-G 二阶导数（2nd S-G）、标准正态变量变换（Standard Normal Variate Transform，SNV）五种预处理方法对原始光谱进行预处理。同时，偏最小二乘回归（Partial Least Squares，PLS）模型是近红外光谱中最广泛使用的建模方法之一，通过不同预处理方法后所建的 PLS 模型精度如表 3 所示。经过一阶导数预处理的茶褐素预测集精度最高和稳定性最高，其训练集决定系数 R_c^2 为 0.96、交叉验证集决定系数 R_{cv}^2 为 0.89、预测集决定系数 R_p^2 为 0.94，训练集均方根误差（Root Mean Square Error of Calibration，RMSEC）为 0.18、交叉训练集均方根误差（Root Mean Square Error of Cross Validation，RMSECV）为 0.31、预测集均方根误差（Root Mean Square Error of Prediction，RMSEP）为 0.24。

导数预处理可以消除光谱数据背景漂移造成的影响，一阶导数可以消除背景的常数平移[23]。本文使用了一阶导数的光谱如图 3 所示，光谱数据消除了背景造成的常数偏移，减轻了仪器带来的试验影响。综合比较建模结果后，采用经过一阶导数后的预处理茶褐素光谱数据进行后续研究。

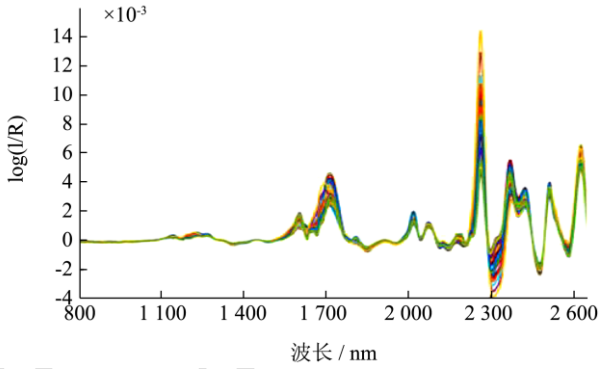


图 3 茶褐素光谱经过一阶求导后的光谱数据

Fig.3 Spectral data of thearubin spectrum after Savitzky-Golay

表 3 茶褐素不同预处理后的 PLS 模型精度

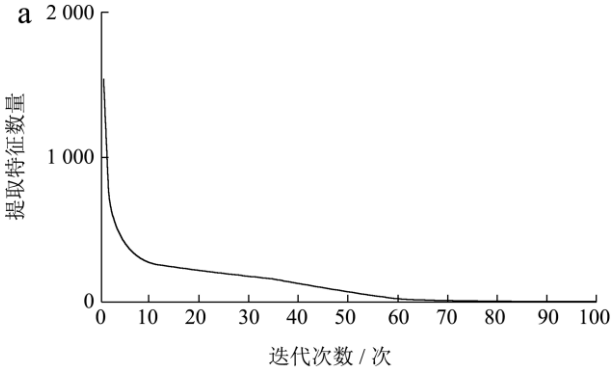
Table 3 Accuracy of PLS model after different pretreatment of theabrownine

物质	预处理方法	R_c^2	R_{cv}^2	R_p^2	RMSEC	RMSECV	RMSEP
茶褐素	Normalize	0.91	0.85	0.93	0.28	0.35	0.25
	Baseline	0.90	0.70	0.87	0.29	0.51	0.35
	1st, Savitzky-Golay	0.96	0.89	0.94	0.18	0.31	0.24
	2nd, Savitzky-Golay	0.95	0.79	0.82	0.21	0.43	0.38
	SNV	0.95	0.88	0.93	0.21	0.31	0.26

2.3.2 茶褐素光谱特征提取

2.3.2.1 基于竞争自适应加权算法（CARS）的茶褐素特征提取

本研究采用 CARS 算法对英红九号的茶褐素含量进行变量筛选，设定最大迭代次数为 100 以进行相关物质的特征选择。图 4 显示了 CARS 算法对茶褐素特征选择的变化过程。其中，图 4a 显示了在进行 100 次迭代的过程中，随着迭代的次数增加，被选择的波长特征数量逐渐减少，下降的趋势由快变慢，并迭代到一定的程度时特征数量趋于平缓。图 4b 呈现的是在迭代过程中 RMSECV 值的变化，选取迭代过程中 RMSECV 值最小的变量组是本算法的核心。在图中可以看出，RMSECV 值有缓慢下降至最低点然后再逐渐上升的趋势，这是因为在迭代过程中，与茶褐素无关的变量不断被剔除，而到了最低点后，与茶褐素相关的变量也有被剔除的趋势。而图 4c 呈现的是变量的回归系数路径，在三张图中分别可以看出有些回归系数的绝对值不断变大，而有些变量回归系数却不断变小。因此，可以得出采用 CARS 算法在在迭代次数为 51 时 RMSECV 值最小，被选择的特征变量数为 54 个。



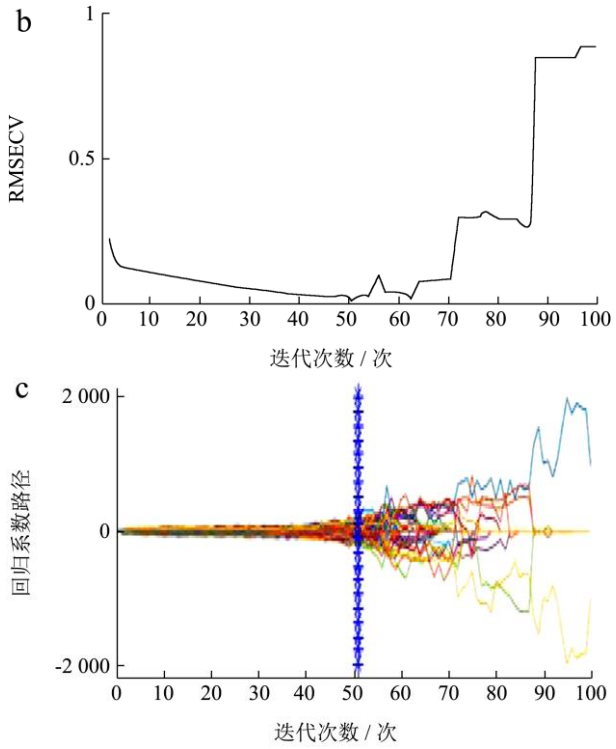


图 4 基于竞争自适应加权算法的茶褐素三种数据的变化过程

Fig.4 The change process of three kinds of theofronin data based on competitive adaptive weighting algorithm

注：a 为提取特征数量的变化；b 为 RMSECV 变化；c 为回归系数路径变化。

2.3.2.2 基于区间偏最小二乘法（iPLS）的茶褐素特征提取

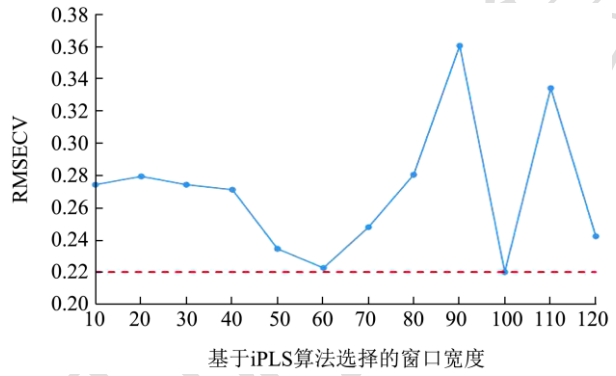


图 5 随着窗口宽度变化的 RMSECV 值

Fig.5 RMSECV value that varies with window width

iPLS 方法主要基于 RMSECV 进行挑选，以及其他参数如决定系数，和斜率、偏移量等进行评估。首先使用移动窗口的方法确定最佳窗口宽度：

（1）确定起始窗口宽度和移动步长。本文设置步长为 10，起始宽度为 20。

（2）对每个窗口的数据进行 PLS 建模。并得出这步骤建立的模型的最小交互验证的均方根误差 RMSECV，并保存该数据。

（3）重复（1）、（2）步，直到窗口宽度为 120。模型运行到宽度为 120 即结束，运行的结果如图 5 所示。RMSECV 值会随着选取不同的波长的而改变大小。其中，当窗口宽度为 100 时，RMSECV 值为 0.22，特征数量为 198 个波长点，综合比较后，得出宽度保持在 100 左右更能突出算法的优势。

本文确定最佳窗口宽度为 100，特征数量为 198 个波长点。经过 PLS 建模后，通过对比 RMSECV 最小值来确定最佳的窗口宽度。本文将以全光谱建模的 RMSECV 值作为阈值，再选择 RMSECV 值低于其阈值的区间作为特征区间。

2.3.2.3 基于空间迭代收缩选变量方法（VISSA）的茶褐素特征提取

变量迭代空间收缩方法是一种将集群模型和加权二进制矩阵抽样结合的一种变量挑选方法，通过若干次的迭代将变量挑选出来，且每次迭代得到的更小的 RMSECV 值所对应的变量集为被选的变量集，一直到 RMSECV 值没有变化为止。茶褐素的 RMSECV 值变化值如图 6 所示，茶褐素的 1 557 个特征经过 73 次迭代，减少到 344 个特征，在迭代过程中，RMSECV 值先快速下降，剔除了与茶褐素无关的特征，到了第 59 次迭代时，RMSECV 值就维持在一个水平上，为 0.20。这是因为随着迭代过程的推进，所选的特征基本稳定在一个范围内，尽管在每次迭代中的权重仍然发生改变，但不会改变最后选取变量集的结果，所以到了迭代后期 RMSECV 值不会再发生改变，说明迭代的效果可靠性。

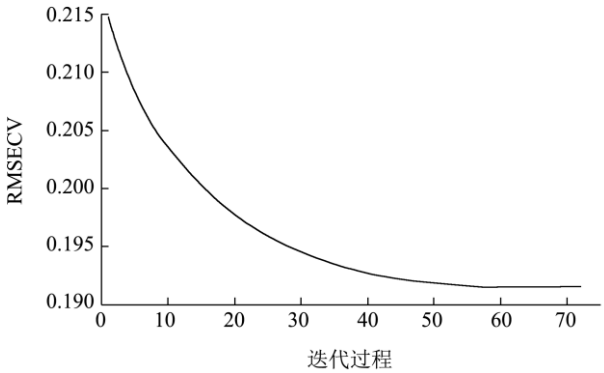


图 6 基于变量迭代空间收缩方法的迭代过程的 RMSECV 值变化

Fig.6 Variation of RMSECV value in the iterative space shrinkage approac based

2.4 茶褐素定量模型验证

茶褐素的近红外光谱经过经过一阶导数预处理后，对使用 CARS、iPLS、VISSA 三种方法挑选波长特征后的数据集进行 PLS 线性回归建模，建模的结果如表 4 所示。

表 4 三种特征选择方法建立的 PLS 茶褐素模型结果

Table 4 Results of PLS theofloocin model established by three feature selection methods

物质	方法	R_c^2	R_{cv}^2	R_p^2	RMSEC	RMSECV	RMSEP	特征数
茶褐素	1st-PLS	0.96	0.89	0.94	0.18	0.31	0.24	1 557
	1st-CARS-PLS	0.99	0.96	0.97	0.07	0.18	0.16	53
	1st-iPLS	0.99	0.96	0.94	0.09	0.19	0.22	198
	1st-VISSA-PLS	0.96	0.93	0.96	0.18	0.26	0.17	340

建模结果如图 7、图 8 所示，横坐标为茶褐素实测值，纵总坐标为茶褐素预测值，虚线为斜率为 1 参考线，当图中圆点越接近虚线，代表表示模型对茶褐素实测值与茶褐素预测值比例越接近 1 实测值，表明模型预测精度越准确高。茶褐素的波长特征变量经过 CARS、iPLS、VISSA 三种特征挑选算法后，分别得到了 53、198、340 个波长特征，仅使用一阶导数进预处理的建模结果为：R_c^2 为 0.96、R_{cv}^2 为 0.89、R_p^2 为 0.94，RMSEC 为 0.18、RMSECV 为 0.18、RMSEP 为 0.24。1st-CARS-PLS 结果为 R_c^2 为 0.99、R_{cv}^2 为 0.96、R_p^2 为 0.97，RMSEC 为 0.07、RMSECV 为 0.18、RMSEP 为 0.18。1st-iPLS 模型结果为 R_c^2 为 0.99、R_{cv}^2 为 0.96、R_p^2 为 0.94，RMSEC 为 0.09、RMSECV 为 0.19、RMSEP 为 0.22。1st-VISSA-PLS 模型结果为为 R_c^2 为 0.96、R_{cv}^2 为 0.93、R_p^2 为 0.96，RMSEC 为 0.18、RMSECV 为 0.26、RMSEP 为 0.17。

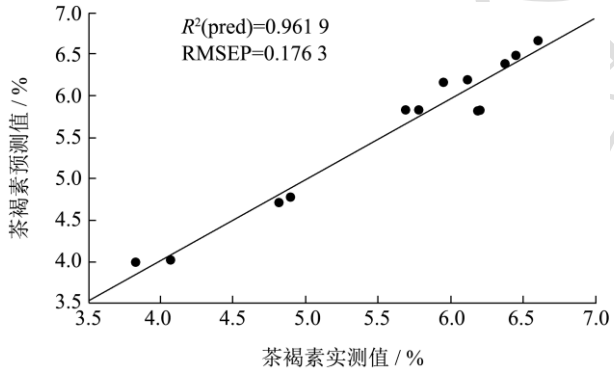


图 7 使用原始光谱建立的茶褐素 PLS 线性预测模型

Fig.7 The linear prediction model of theofupupin PLS was established by using the original spectrum

由建模结果分析可知，CARS、iPLS、VISSA 三种特征挑选方法对提升茶褐素的近红外光谱预测模型有提升作用，并且能将茶褐素的光谱信息进行了筛选，简化了模型，模型的稳定性随着 RMSE 值的降低也有了一定的提升。而 1st-CARS-PLS 模型效果更佳，使特征值从 1 557 个降低到 53 个，其校正集决定系数 R_c^2 从 0.95 提升到 0.99，交叉验证集决定系数 R_{cv}^2 从 0.91 提升到 0.96，预测集回归系数也从 0.96 提升到 0.97。前言提及，董春旺[15]用不同的特征选择方法，对茶褐素等物质建立模型，校正集的相关系数（Correlation Coefficient of Calibration Set，Rc）为 0.952，预测集的相关系数（Correlation Coefficient of Predication Set，Rp）为 0.951，另外，RMSECV 为 0.344，RMSEP 为 0.248。而本文建立的 1st-CARS-PLS 模型的结果为 R_c^2 为 0.99、R_{cv}^2 为 0.96、R_p^2 为 0.97，模型效果更好，另外，RMSEC 为 0.07，RMSECV 为 0.18，RMSEP 为 0.16，模型稳定性更高。

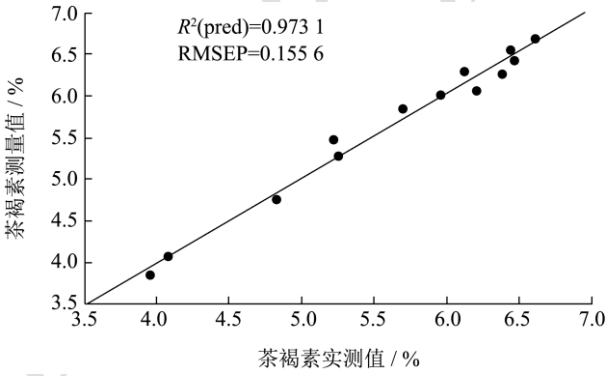


图 8 基于 CARS 算法挑选特征建立的茶褐素 PLS 线性预测模型

Fig.8 A linear prediction model of theofupupin PLS was established based on the selection features of CARS algorithm

3 结论

本文采用红外光谱技术对英红九号发酵叶的茶褐素进行无损、快速检测进行了研究，并对该物质的含量进行了定量分析，并结合最优预处理方法以及提取特征波长方法对模型进行了优化。

通过使用规范化处理（Normalize）、基线校正（Baseline）、S-G 一阶导数（Savitzky-Golay，1st S-G）、S-G 二阶导数（2nd S-G）、标准正态变量变换（Standard Normal Variate Transform，SNV）五种预处理方法建模对比，通过初步建模分析，结果表明使用一阶导数进行预处理效果最好。

通过对比使用 CARS、iPLS、VISSA 方法挑选特征后建模，结果表明建立的 1st-CARS-PLS 模型效果最好，使特征值从 1,557 个降低到 53 个，其校正集决定系数 R_c^2 从 0.96 提升到 0.99，交叉验证集决定系数 R_{cv}^2 从 0.89 提升到 0.96，预测集回归系数 R_p^2 也从 0.94 提升到 0.97。

以上结果表明，样本茶褐素的实际含量与预测值

有准确、可靠的关联性，可以在一定程度上对英红九号发酵叶的茶褐素含量进行预测，对目前比较空白的茶褐素无损检测、以及制茶发酵过程的把控提供了理论基础。

参考文献

[1] 凌彩金,范杰文,王秋霜,等.英红九号红茶生化成分指纹图谱初探[J].广东农业科学,2018,45(12):95-100.

[2] PAN Shunshun, DENG Xuming, SUN Shili, et al. Black tea affects obesity by reducing nutrient intake and activating AMP-activated protein kinase in mice [J]. Molecular Biology Reports, 2018, 45(5): 689-697.

[3] QI Dandan, LI Junxing, QIAO Xiaoyan, et al. Non-targeted metabolomic analysis based on ultra-high-performance liquid chromatography quadrupole time-of-flight tandem mass spectrometry reveals the effects of grafting on non-volatile metabolites in fresh tea leaves (*Camellia sinensis* L.) [J]. Journal of Agricultural and Food Chemistry, 2019, 67(23): 6672-6682.

[4] ZHANG Wenji, CAO Junxi, LI Zhigang, et al. HS-SPME and GC/MS volatile component analysis of Yinghong No. 9 dark tea during the pile fermentation process [J]. Food Chemistry, 2021, 357: 129654.

[5] 马婉君,马士成,刘春梅,等.六堡茶的化学成分及生物活性研究进展[J].茶叶科学,2020,40(3):289-304.

[6] 崔帅,冯凌.基于近红外光谱技术的工业循环冷却水氨氮含量检测方法[J].广东化工,2021,48(8):290-291.

[7] 张晋,曹晓宁,田翔,等.近红外光谱法快速检测藜麦蛋白含量[J].安徽农业科学,2021,49(9):175-179.

[8] 李佳佳,洪慧龙,万明月,等.基于近红外光谱的大豆茎秆化学组分含量检测模型构建与应用[J].中国农业科学,2021,54(5):887-900.

[9] 卢莉,程曦,张渤,等.小种红茶茶多酚和咖啡碱近红外定量分析模型的建立[J].茶叶科学,2020,40(5):689-695.

[10] 单瑞峰,甄书仙,陈瑶,等.基于近红外光谱技术的日照绿茶茶鲜叶模型的优化[J].分析科学学报,2018,34(1):80-84.

[11] 宁井铭,宛晓春,张正竹,等.近红外光谱技术结合人工神经网络判别普洱茶发酵程度[J].农业工程学报,2013,29(11):255-260.

[12] 赵雅,王博思,赵明富.基于光谱技术的茶叶品质参数茶多酚含量快速检测方法研究[J].半导体光电,2018,39(4):591-594.

[13] 石艳梅,伍庆,谭高好,等.近红外光谱法对黔茶 6 种主要成分的快速测定[J].贵州科学,2016,34(4):74-76.

[14] 刘蕾,罗文文,龚淑英,等.采用近红外光谱技术定量分析绿茶中的主要呈味物质[J].中国食品学报,2008,8(6):109-115.

[15] 董春旺,梁高震,安霆,等.红茶感官品质及成分近红外光谱快速检测模型建立[J].农业工程学报,2018,34(24):306-313.

[16] CHEN Weihao, CHEN Huazhou, FENG Quanxi, et al. A hybrid optimization method for sample partitioning in near-infrared analysis [J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2021, 248: 119182.

[17] LI Hongdong, LIANG Yizeng, XU Qingsong, et al. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration [J]. Analytica Chimica Acta, 2009, 648(1): 77-84.

[18] 刘雪松,张丝雨,赵曼茜,等.近红外光谱结合不同变量筛选方法用于黄芩提取过程中黄芩苷含量预测[J].药学学报,2019,54(1):138-143.

[19] 唐海涛,孟祥添,苏循新,等.基于 CARS 算法的不同类型土壤有机质高光谱预测[J].农业工程学报,2021,37(2):105-113.

[20] Norgaard L, Saudland A, Wagner J, et al. Interval partial least-squares regression (ipls): a comparative chemometric study with an example from near-infrared spectroscopy [J]. Applied Spectroscopy, 2000, 54(3): 413-419.

[21] YANG Zhenfa, XIAO Hang, ZHANG Lei, et al. Fast determination of oxides content in cement raw meal using NIR-spectroscopy and backward interval PLS with genetic algorithm [J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2019, 223: 117-127.

[22] 邓百川,云永欢,梁逸曾.空间迭代收缩选变量方法[C]//.中国化学会第 29 届学术年会摘要集-第 19 分会:化学信息学与化学计量学.2014:48.

[23] 张勇,王振华,赵蔚.红外快速检测技术在国道 242 项目中的应用研究[J].交通节能与环保,2022,18(1):132-138.