doi:10.6041/j.issn.1000-1298.2021.05.034

基于高光谱成像的油茶籽含油率检测方法

周宏平　胡逸磊　姜洪喆　许林云　王　影
（南京林业大学机械电子工程学院，南京 210037）

摘要：为了快速准确地检测油茶籽含油率、解决传统检测手段费时费力等问题，提出了一种基于高光谱成像技术的油茶籽含油率检测方法。应用光谱集Ⅰ（400～1 000 nm）和光谱集Ⅱ（900～1 700 nm）两组高光谱成像系统采集油茶籽的漫反射高光谱图像，并结合化学计量学方法建立油茶籽含油率的回归预测模型。结果显示，在不经预处理的情况下，两组光谱集数据建立的偏最小二乘回归模型精度最高：光谱集Ⅰ的预测集决定系数 R_p^2 为 0.681，均方根误差（RMSEP）为 2.89%；光谱集Ⅱ的 R_p^2 为 0.740，RMSEP 为 2.92%。通过对比 7 种不同的变量选择方法发现，两组光谱集采用遗传算法筛选特征波长后建立的 PLSR 模型精度最高：光谱集Ⅰ的 R_p^2 为 0.694，RMSEP 为 2.82%；光谱集Ⅱ的 R_p^2 为 0.779，RMSEP 为 2.54%。通过对比光谱集Ⅰ和光谱集Ⅱ的建模效果发现，使用光谱集Ⅱ建立的 PLSR 模型的性能更好，因此 900～1 700 nm 波段比 400～1 000 nm 波段更适用于油茶籽含油率的检测，进一步验证了利用高光谱成像技术实现油茶籽含油率预测值分布可视化的可行性。

关键词：油茶籽；含油率；检测；高光谱成像
中图分类号：O657.3；TS222⁺.1　　**文献标识码**：A　　**文章编号**：1000-1298(2021)05-0308-08　　**OSID**：

Detection Method of Oil Content of *Camellia oleifera* Seed Based on Hyperspectral Imaging

ZHOU Hongping　HU Yilei　JIANG Hongzhe　XU Linyun　WANG Ying
(*College of Mechanical and Electronic Engineering*, *Nanjing Forestry University*, *Nanjing* 210037, *China*)

Abstract: In order to quickly and accurately detect the oil content of *Camellia oleifera* seed and solve the time-consuming and laborious problems of traditional detection methods, a method for detecting the oil content of *Camellia oleifera* seed based on hyperspectral imagery (HSI) was proposed. Two sets of hyperspectral imaging systems, spectral set Ⅰ (400～1 000 nm) and spectral set Ⅱ (900～1 700 nm), were used to collect diffuse reflectance hyperspectral images of *Camellia oleifera* seed, and the regression prediction model of oil content of *Camellia oleifera* seed was established in combination with chemometrics. The results showed that the partial least squares regression model (PLSR) established by the two sets of spectral data without pretreatment had the highest accuracy: the determination coefficient of prediction (R_p^2) of the spectral set Ⅰ was 0.681, and the root mean square error of prediction set (RMSEP) was 2.89%; R_p^2 of spectral set Ⅱ was 0.740, and RMSEP was 2.92%. Comparing seven different variable selection methods, it was found that the two sets of spectral sets used genetic algorithm (GA) to filter the characteristic wavelength to establish the PLSR model with the highest accuracy: the spectral set Ⅰ had R_p^2 of 0.694 and RMSEP of 2.82%; the spectral set Ⅱ had R_p^2 of 0.779 and RMSEP of 2.54%. Comparing the modeling effects of spectral set Ⅰ and spectral set Ⅱ, it was found that the performance of the PLSR model established by spectral set Ⅱ was better than that of the spectral set Ⅰ, so the band of 900～1 700 nm was more suitable for the oil content detection of *Camellia oleifera* seed than the band of 400～1 000 nm. Besides, the feasibility of using HSI to visualize the distribution of the predicted value of the oil content of *Camellia oleifera* seed was further verified. This result can provide a method for the rapid detection of the oil content distribution of *Camellia oleifera* seed and the selection of high-quality its varieties.

Key words: *Camellia oleifera* seed; oil content; detection; hyperspectral imaging

收稿日期：2020－07－15　修回日期：2020－08－15
基金项目：国家重点研发计划项目(2016YFD0701501)
作者简介：周宏平(1964—)，男，教授，博士生导师，主要从事植保技术与装备、光谱检测技术研究，E-mail：hpzhou@njfu.edu.cn

0　引言

油茶属山茶科山茶属植物，为常绿小乔木或灌木[1]，是原产于我国南方的乡土树种，具有栽培历史悠久、分布区域广、栽培面积大、用途多等特点，与油橄榄、油棕、椰子并称为世界四大木本油料树种[2]。目前，我国油茶种植面积已达 437 万 hm^2，年产茶油 60 万 t，产值近千亿元[3]。油茶的主要产物是茶油，茶油含有丰富的营养成分，油酸质量分数超过 80%，不饱和脂肪酸质量分数达 90%，比橄榄油高 6 个百分点[4]。

油茶籽含油率是影响茶油产量的重要因素，也是油茶采摘时间的重要参考指标[5]。目前茶籽油常用的提取方法包括压榨法、溶剂法和水酶法[6]，这些方法存在提取率低、含有机溶剂残留、成本较高等缺点，不适合用于实验室内对单个或少许茶籽进行的含油率检测。因此，迫切需要一种能够快速、准确地检测油茶籽含油率的方法，以便于茶籽品种的鉴别和筛选。

高光谱成像是一种集光谱和图像于一体的技术，图像中的每个像素点都包含特定位置的光谱信息，与传统近红外光谱相比，其优势之一是可实现被测物各组分分布情况的可视化[7]。文献[8]利用近红外光谱建立了橄榄果肉中油含量的偏最小二乘回归模型（Partial least squares regression，PLSR），模型的校正集相关系数 R_c 为 0.848，交叉验证均方根误差（Root mean square error of cross validation，RMSECV）为 0.901。文献[9]采集了不同成熟度油棕的高光谱信息，发现 750 nm 是适用于棕油含量定量检测的最佳波长。文献[10]利用两组光谱集（光谱集 Ⅰ：400 ~ 1 000 nm；光谱集 Ⅱ：1 000 ~ 2 500 nm），结合化学计量方法检测不同品种花生的含油量，建立的 PLSR 模型的预测集决定系数分别为 0.696 和 0.923。文献[11]研究发现，使用高光谱数据建立的径向基神经网络模型能较好地预测油茶籽脂肪酸的成分含量。文献[12]利用 30 份油茶籽的近红外光谱数据建立了含油率的 PLSR 模型，其校正集相关系数为 0.93。这些研究结果为使用光谱分析技术进行油料作物含油率检测奠定了良好的理论基础。

目前，国内外有关采用高光谱成像技术检测油茶籽含油率的研究尚未见报道。本文旨在探讨波长在 400 ~ 1 000 nm 和 900 ~ 1 700 nm 范围内的高光谱相机检测油茶籽含油率的可行性，并对比两个波段的检测效果，以期开发一种快速检测油茶籽含油率的方法，为油茶籽优质育种与品质快检分选提供理论依据与技术基础。

1　材料与方法

1.1　实验材料

油茶籽样品来自安徽省芜湖市无为县联合行政村联合农业发展有限公司的油茶林，于 2019 年 10 月 10 日和 10 月 14 日采摘油茶鲜果共 109 个，当天带回实验室进行果高、果径、质量等形态参数的测量，并置于冰箱 4℃ 环境下保存，于第二天取出、去壳，采集油茶鲜籽的高光谱图像，干燥后采集含油率数据。

1.2　高光谱成像系统

高光谱成像系统采用南京林业大学生物质包装无损检测实验室搭建的高光谱成像无损检测平台，主要包括两台光谱仪（GaiaField - V10E - AZ4 型，400 ~ 1 000 nm（光谱集 Ⅰ）；GaiaField - N17E 型，900 ~ 1 700 nm（光谱集 Ⅱ））、两台探测器（sCMOS 相机、InGaAs 相机）、一条白色食品级传送带（HSIA - CSD800 型）、一套由 12 只 50 W 的卤素灯和漫反射穹顶组成的照明系统以及一台计算机。其中成像光谱仪的光谱分辨率分别为 2.8 nm 和 5 nm，被测物品置于传送带上的载物台，步进电机驱动传送带使被测物品运动，暗箱用于屏蔽外界杂散光对数据采集的干扰。

1.3　高光谱图像采集和校正

高光谱图像数据获取基于计算机上的 SpecView 软件，将高光谱仪器预热 30 min 后进行油茶籽图像采集。为了避免获取的图像失真，经过多次预实验确定最佳的数据采集参数如下：光谱仪 Ⅰ 的曝光物距为 300 mm，曝光时间为 1.2 ms，电控位移台扫描速度是 0.601 4 nm/s，扫描线实际长度是 200 mm，图像分辨率是 800 像素 × 664 像素；光谱仪 Ⅱ 的曝光物距为 300 mm，曝光时间为 7.5 ms，电控位移台扫描速度是 2.256 8 nm/s，扫描线实际长度是 200 mm，图像分辨率是 640 像素 × 542 像素；每次采集同一个油茶果中 3 粒油茶籽的高光谱图像，并取其平均值作为样本的最终光谱数据。

由于高光谱图像采集过程中存在暗电流的影响，而且不同波段下成像系统光源的强度分布也不均匀，从而导致获取的高光谱图像中含有较大的噪声。因此要对其进行黑白校正以消除暗电流的影响，校正方法为[13-14]

$$R_e = \frac{R_0 - D}{W - D} \tag{1}$$

式中　R_e——校正后的漫反射光谱图像数据

R_0——样本原始的漫反射光谱图像数据

D——暗图像数据

W——白板的漫反射图像数据

1.4 含油率测定

油茶籽含油率采用 NAI - ZFCDY - 6Z 型脂肪测定仪(上海那艾精密仪器有限公司)按照 GB 5009.6—2016 规定的方法测定。首先采用 BSM -220.4 型分析天平(上海卓精电子科技有限公司)称取油茶仁 2.000 g,磨碎后移入滤纸筒内,并将滤纸筒放入索氏抽提器的抽提筒内,连接已干燥至质量恒定的接收瓶,由抽提器冷凝管上端加入60 mL无水乙醚至瓶内容积的 2/3 处,于水浴上加热,使无水乙醚不断回流抽提 6 h。然后取下接收瓶,回收无水乙醚,待接收瓶内溶剂剩余 1 ~ 2 mL 时在水浴上蒸干,再于 100℃干燥箱内干燥 1 h,放干燥器内冷却 0.5 h 后称量。最后按照文献[15]的方法计算油茶籽含油率,公式为

$$R_{ratio} = \frac{M_{oil}}{M_{seed}} \times 100\% \tag{2}$$

式中 M_{oil}——出油质量

M_{seed}——茶仁质量

1.5 变量选择方法

高光谱数据量大且数据之间的共线性严重,影响模型的运算速度[16]。因此采用以下 7 种方法提取有效信息变量并进行对比,从而得到最优的变量选择方法:

(1)连续投影算法(Succesive projections algorithm,SPA)是一种前向选择算法,通过在光谱中寻找最低限度冗余光谱信息变量集,使得变量之间的共线性最小化[17]。该方法要预先设置选择的变量数范围,最终选择的变量数在该范围内对应最低的均方根误差(Root mean square error,RMSE)。本研究中选择的最佳变量数范围为 5 ~ 30,光谱集Ⅰ筛选了 8 个特征波长,光谱集Ⅱ筛选了 11 个特征波长。

(2)竞争性自适应重加权算法(Competitive adaptive reweighted sampling, CARS)是一种以回归系数作为变量重要性指标的变量选择方法。该方法利用自适应重加权采样技术和指数衰减函数优选出每次循环所构建的 PLSR 模型中回归系数绝对值大的变量点,并将交互验证选出 N 个 PLSR 子集模型中 RMSECV 最小的子集定义为最优变量子集[18]。本研究中将蒙特卡洛采样次数设置为 2 000,每次运行程序选择的校正集和预测集样本数比例为 2∶1。光谱集Ⅰ筛选出 14 个特征波长,光谱集Ⅱ筛选出 16 个特征波长。

(3)粒子群优化算法(Particle swarm optimization,PSO)是一种源于对鸟群捕食行为研究的进化计算技术。在 PSO 中,每个优化问题的潜在解可看作高维空间上的一个粒子,所有粒子都拥有速度以及由目标函数决定的适应值,粒子们通过追随当前的最优粒子在解空间中搜索[19]。本研究中粒子种群大小设置为 20,迭代次数为 1 000,以 $F = R^2$作为适应度函数(其中 R^2 表示决定系数)。光谱集Ⅰ筛选出 18 个特征波长,光谱集Ⅱ筛选出 18 个特征波长。

(4)蚁群优化(Ant colony optimization,ACO)算法是模拟蚂蚁的合作和适应机制等自然行为的一种正反馈式算法。每个蚂蚁在其所经过的路径上会遗留一种叫做信息素的挥发性物质,蚂蚁通过信息素及其强度的反馈机制选择路径,所有蚂蚁找到的特定路径便是解决目标问题的最优方案[20]。本研究中设置蚁群大小为 30,光谱窗口为 1,迭代次数为 100,适应度函数 $F = (1 + Q_{RMSECV})/R^2$(其中 Q_{RMSECV} 表示交叉验证均方根误差)。光谱集Ⅰ筛选出 18 个特征波长,光谱集Ⅱ筛选出 18 个特征波长。

(5)模拟退火(Simulated annealing,SA)算法是一种基于固体物理退火原理而研发的随机全局优化算法,SA 在解决组合优化问题时先从某一模拟较高初温开始,随着温度参数的不断下降,结合 Metropolis 标准在解空间中随机寻找目标函数的全局最优解[21]。本研究设置初始温度 $T_0 = 50$℃,第 k 个温度控制参数值 $T_k = 0.96T_0$,终止温度 $T_f = 0$℃,第 k 个马尔可夫链的长度 $L_k = 50$,迭代次数为 100。光谱集Ⅰ筛选出 30 个特征波长,光谱集Ⅱ筛选出 26 个特征波长。

(6)区间随机蛙跳(Interval random frog,iRF)算法是基于随机蛙跳算法提出的一种波长间隔选择方法。基本思想是将整个光谱按照特定宽度划分成若干子区间,通过每个区间光谱点的绝对回归系数总和来评估区间,找到最佳区间组合[22-23]。本研究参数设置如下:移动窗口大小 $\omega = 3$,初始化子集变量个数为 5,最大主成分数为 10,迭代次数为 500。光谱集Ⅰ筛选出 10 个特征波长,光谱集Ⅱ筛选出 10 个特征波长。

(7)遗传算法(Genetic algorithm, GA)是模拟达尔文生物进化论的自然选择和遗传学机理的生物进化过程的计算模型,是一种通过模拟自然进化过程搜索最优解的方法[24]。本研究设置进化代数为 150,算法运行次数 30,种群大小为 64,初始时平均 5 个波长构成一个染色体,染色体个数为 20,变异概率为 1%。光谱集Ⅰ筛选出 26 个特征波长,光谱集Ⅱ筛选出 28 个特征波长。

1.6　建模方法及模型评价

本文应用 PLSR 建立油茶籽含油率的检测模型。PLSR 是光谱分析中应用最广泛的化学计量方法，该方法同时将自变量和因变量数据矩阵进行分解，把因变量信息引入到自变量数据的分解过程中，使得自变量主成分直接与样品被测组分含量相关联[25]。在 PLSR 中，确定潜变量数是保证模型性能的关键，本研究采用交叉验证法（Cross validation，CV）通过最小 RMSECV 确定最优的潜变量数。

本文采用决定系数 R^2 和 RMSE 作为评价 PLSR 模型的指标，R^2 越高、RMSE 越低说明 PLSR 模型的预测性能越好。数据处理软件包括 ENVI 5.1、Matlab 2014a。

1.7　含油率分布可视化

使用单一的化学计量方法很难测量样品每个部分的化学成分，而高光谱成像的优势在于可以通过校准模型对高光谱图像中样品的每个像素点的化学成分进行预测，从而得到整个样品的理化成分含量分布图[26-27]。预测值的准确性主要依赖于校准模型的性能，也可通过选择特征波长减少数据冗余，获得更好的模型结果。

2　结果与分析

2.1　原始光谱

由于使用 3 颗完整的油茶籽作为 1 个样品进行含油率测定，所以选取 3 颗油茶籽的表面作为每个样品的感兴趣区域（Region of interest，ROI）。光谱集Ⅰ和光谱集Ⅱ分别在 936.2 nm 和 1 133.9 nm 处设置反射率阈值为 0.3，可有效提取油茶籽的高光谱信息[28]。此外光谱集Ⅰ在 336.2～416.6 nm 和 994.9～1 092.5 nm 范围内存在较大的噪声，光谱集Ⅱ在874.0～1 038.3 nm 和 1 564.9～1 731.0 nm 范围内存在明显的噪声，所以分别取 416.6～994.9 nm 和1 038.3～1 564.9 nm 作为两组光谱集的有效波段，原始光谱如图 1 所示。

可见和近红外波段最主要的吸收带是由于强泛音和含氢键 O—H（来自水）、C—H（来自脂肪和油）、N—H（来自蛋白质）的组合吸收而产生，油茶籽的光谱反射率曲线在光谱集Ⅰ和光谱集Ⅱ上存在一定的差异（图 1），这些差异可能与油茶籽的质量属性、表面结构不均匀性以及表面无规律散射有关[29]。图 1a 中，930 nm 附近的吸收峰与 C—H 伸展的第三泛音有关[25]。图 1b 中，在 1 400 nm 附近有较强的水吸收峰，在 1 220 nm 处的吸收峰与油含量相关，是由脂肪组分中甲基或亚甲基基团的 C—H 伸展的第一和第二泛音所致[30]。

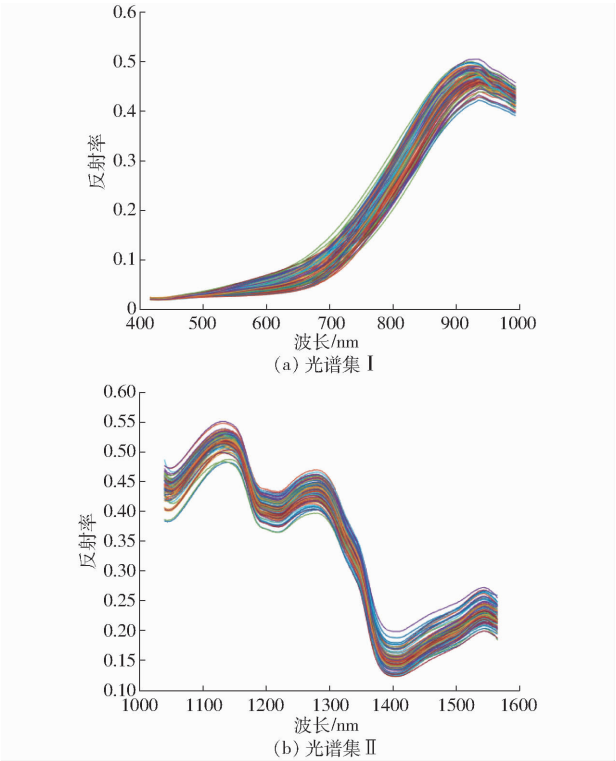

图 1　油茶籽原始反射光谱曲线

Fig. 1　Original reflection spectrum curves of *Camellia oleifera* seed

2.2　油茶籽含油率统计

本研究采用蒙特卡罗交叉验证（Monte Carlo cross-validation，MCCV）识别并剔除了 7 个奇异样本。首先通过对所有油茶籽数据作交互检验，确定最优主成分数，然后利用 MCCV 每次随机选取 67% 的样本建立 PLSR 模型，剩余 33% 的样本用于预测。经过 5 000 次蒙特卡罗采样后，计算每个样本预测残差的均值和标准偏差，将具有较高的均值和标准偏差的样本定义为奇异样本并从总样本中剔除[31-32]。表 1 是剩余的 102 份油茶籽样品含油率的统计，使用 SPXY 算法[33]将总样本按照 2∶1 划分为校正集和预测集。油茶籽含油率在 19.17% ～45.12% 之间，说明样本之间的差异性较大，有利于建立稳定的校准模型。

表 1　102 份油茶籽样品含油率

Tab. 1　Oil content of 102 *Camellia oleifera* seeds

	数量	含油率/%			
		最小值	最大值	平均值	标准差
校正集	68	19.17	45.12	33.04	5.85
预测集	34	20.33	43.13	33.11	5.12
总样本	102	19.17	45.12	33.07	5.59

2.3　全光谱 PLSR 建模

原始光谱数据中，存在多种因素对数据的准确性造成影响，如采集过程中光源强度分布不均匀、摄

像头暗电流的存在以及油茶籽自身形状的不规则性等,因此在建模前通过预处理方法消除目标信息和噪声干扰。分别使用 MSC(多元散射校正)、SNV(标准正态变换)、SG(Savitzky - Golay 卷积平滑)、Normalize(归一化)、Detrend(去趋势)共 5 种方法对原始光谱进行预处理,并建立油茶籽含油率的 PLSR 预测模型,建模结果如表 2、3 所示。

表 2 光谱集 Ⅰ PLSR 模型预测结果

Tab. 2 PLSR model predicted results of spectrum set Ⅰ

预处理方法	潜变量数	R_c^2	RMSEC/%	R_p^2	RMSEP/%	RMSECV/%
无	9	0.683	3.27	0.681	2.89	4.41
MSC	10	0.715	3.10	0.560	3.51	4.48
SNV	9	0.716	3.10	0.529	3.64	4.13
SG	12	0.730	3.02	0.567	3.52	4.56
Normalize	11	0.760	2.85	0.573	3.53	4.40
Detrend	10	0.754	2.88	0.650	3.09	4.04

表 3 光谱集 Ⅱ PLSR 模型预测结果

Tab. 3 PLSR model predicted results of spectrum set Ⅱ

预处理方法	潜变量数	R_c^2	RMSEC/%	R_p^2	RMSEP/%	RMSECV/%
无	8	0.813	2.44	0.740	2.92	3.24
MSC	8	0.778	2.66	0.718	2.95	3.38
SNV	9	0.798	2.54	0.694	3.03	3.22
SG	9	0.721	2.99	0.691	3.04	3.95
Normalize	10	0.814	2.44	0.645	3.38	3.24
Detrend	9	0.751	2.83	0.599	3.58	3.64

从表 2、3 中可以看出:对于光谱集 Ⅰ,使用原始光谱及 5 种预处理方法建立的 PLSR 模型的校正集决定系数 R_c^2 在 0.68 ~ 0.76 之间,其中 Normalize - PLSR 的 R_c^2 最高,为 0.760,且其校正集均方根误差(Root mean square error of calibration,RMSEC)最低,为 2.85%。但原始光谱的 PLSR 模型的预测集决定系数 R_p^2 最高,为 0.681,且相应的预测集均方根误差(Root mean square error of prediction,RMSEP)相对于其他预处理方法低,几种预处理方式建模后的 RMSECV 均在 4.0% 以上。对于光谱集 Ⅱ,使用原始光谱及 5 种预处理方法建立的 PLSR 模型的 R_c^2 在 0.72 ~ 0.82 之间,其中 Normalize - PLSR 的 R_c^2 最高,为 0.814,且其 RMSEC 在 5 种预处理方法中最低,为 2.44%。但原始光谱的 PLSR 模型的 R_p^2 最高,为 0.740,且相应的 RMSEP 也相对于其他预处理方法低,几种预处理方式建模后的 RMSECV 在 3.2% ~ 4.0% 之间。

比较各种预处理数据建立的 PLSR 模型,发现光谱集 Ⅰ 和光谱集 Ⅱ 均使用全波长原始数据建立的 PLSR 模型精度较高,并且光谱集 Ⅱ 的 PLSR 模型性能明显优于光谱集 Ⅰ 的模型性能。

2.4 特征波长选择

图 2 所示为 7 种方法对两组光谱集筛选出的特征波长的分布情况。对于光谱集 Ⅰ,使用 SPA 选择的波长数量最少,使用 SA 选择的波长数量最多。使用 SPA 和 iRF 选择的波长分布在特定的区域,使用其他方法选择的波长相对分散不连续。进行变量筛选后,光谱集 Ⅰ 变量数量减少了 83.5% ~ 93.0%。对于光谱集 Ⅱ,使用 iRF 选择的波长数量最少,使用 GA 选择的波长数量最多。使用 7 种方法选择的波长均相对分散不连续。进行变量筛选后,光谱集 Ⅱ 变量数量减少了 91.1% ~ 96.5%。可以发现不同方法选择的特征波长数量不同,因此确定最优变量选择方法显得非常必要。

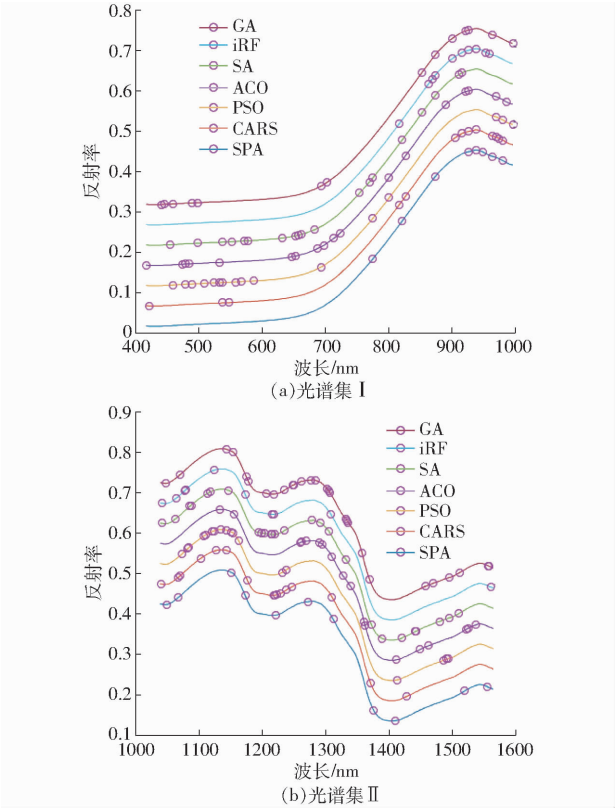

图 2 不同方法选择的特征波长分布

Fig. 2 Distributions of characteristic wavelengths selected by different methods

2.5 特征波长 PLSR 建模

不同变量选择方法筛选出的特征波长 PLSR 建模结果如表 4、5 所示。对于光谱集 Ⅰ,7 种变量选择方法建立的 PLSR 模型的预测性能差别较大,R_p^2 在 0.43 ~ 0.70 之间,但总体来说相对于全光谱建模的精度有所提升。其中 PSO - PLSR 模型性能最差:R_c^2 = 0.570,RMSEC 为 3.81%,R_p^2 = 0.435,RMSEP 为 3.86%,RMSECV 为 4.92%,原因是所选的特征波长在较短的区域内是连续的,而一些

有效的波长却被剔除。与全光谱的 PLSR 模型相比，PSO－PLSR 模型的 R_p^2 降低了 36.123%，RMSEP 升高了 33.564%。GA－PLSR 模型性能最好，如图 3a 所示：R_c^2 = 0.734，RMSEC 为 3.00%，R_p^2 = 0.694，RMSEP 为 2.82%，RMSECV 为 3.63%。与全光谱的 PLSR 模型相比，经 GA 降维后的 PLSR 模型的 R_p^2 升高了 1.873%，RMSEP 降低了 2.422%。iRF－PLSR 模型的校正集结果较好，而预测集结果较差，CARS－PLSR 模型的结果与全光谱建模相当。

表 4　光谱集Ⅰ特征波长 PLSR 模型预测结果

Tab. 4　PLSR model predicted results based on characteristic wavelengths of spectrum set Ⅰ

降维方法	波段数量	潜变量数	R_c^2	RMSEC/%	R_p^2	RMSEP/%	RMSECV/%
SPA	8	7	0.679	3.28	0.553	3.48	3.80
CARS	13	7	0.721	3.07	0.673	2.95	3.62
PSO	16	9	0.570	3.81	0.435	3.86	4.92
ACO	18	9	0.600	3.67	0.535	3.45	4.91
SA	19	12	0.564	3.84	0.462	3.82	5.43
iRF	10	7	0.687	3.25	0.466	4.06	3.90
GA	14	11	0.734	3.00	0.694	2.82	3.63

表 5　光谱集Ⅱ特征波长 PLSR 模型预测结果

Tab. 5　PLSR model predicted results based on characteristic wavelengths of spectrum set Ⅱ

降维方法	波段数量	潜变量数	R_c^2	RMSEC/%	R_p^2	RMSEP/%	RMSECV/%
SPA	11	8	0.831	2.32	0.741	2.79	2.81
CARS	16	7	0.813	2.44	0.779	2.68	2.99
PSO	18	10	0.771	2.71	0.582	3.90	3.76
ACO	18	6	0.568	3.71	0.442	4.16	4.59
SA	26	9	0.785	2.62	0.705	3.05	3.50
iRF	10	8	0.713	3.03	0.647	3.62	4.17
GA	28	8	0.852	2.18	0.779	2.54	2.63

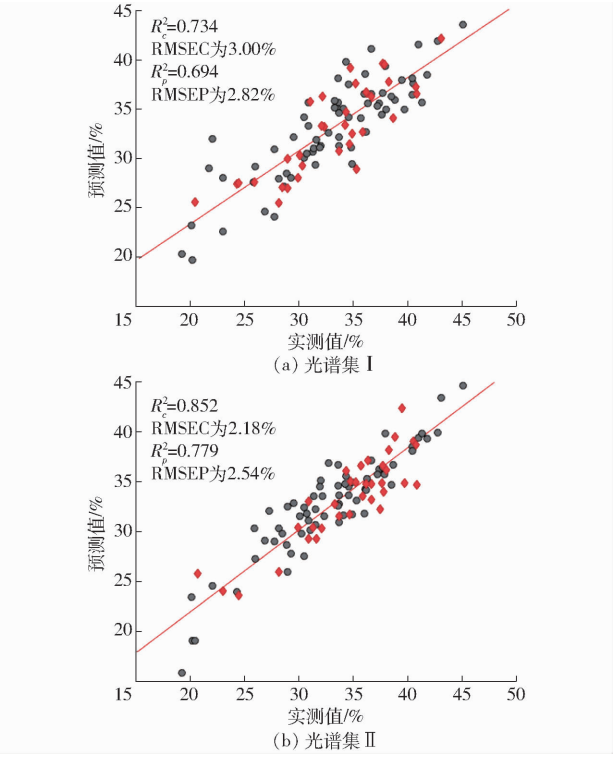

图 3　GA－PLSR 模型预测结果

Fig. 3　GA－PLSR model predicted results

对于光谱集Ⅱ，7 种变量选择方法建立的 PLSR 模型的预测性能差别较大，R_p^2 在 0.44～0.78 之间，但总体来说相对于全光谱建模的精度有所提升。其中 ACO－PLSR 模型性能最差：R_c^2 = 0.568，RMSEC 为 3.71%，R_p^2 = 0.442，RMSEP 为 4.16%，RMSECV 为 4.59%，原因是某些波长的共线性程度较高。与全光谱的 PLSR 模型相比，ACO－PLSR 模型的 R_p^2 降低了 40.270%，RMSEP 升高了 42.466%。GA－PLSR 模型性能最好，如图 3b 所示：R_c^2 = 0.852，RMSEC 为 2.18%，R_p^2 = 0.779，RMSEP 为 2.54%，RMSECV 为 2.63%。与全光谱的 PLSR 模型相比，GA－PLSR 模型的 R_p^2 升高了 5.006%，RMSEP 降低了 13.014%。PSO－PLSR 模型的校正集结果较好，而预测集结果较差，CARS－PLSR 模型的结果与 GA－PLSR 相当。

对比光谱集Ⅰ和光谱集Ⅱ经变量选择后建立的 PLSR 模型对油茶籽含油率的预测能力发现，使用光谱集Ⅱ建立的模型性能更好，因为油茶籽在 900～1 700 nm 范围内呈现的光谱特征峰多于 400～1 000 nm 内的特征峰。

2.6　油茶籽含油率预测值的可视化

由于光谱集Ⅱ的 PLSR 模型效果明显优于光

谱集Ⅰ,因此使用光谱集Ⅱ的最优校正模型(GA-PLSR)计算高光谱图像中油茶籽每个像素点的含油率,再使用伪彩色图像处理方法生成含油率的分布图,如图 4 所示。图中可以明显地观察到油茶籽含油率由小到大变化,并且具有相似光谱特征的像素点对应的含油率预测值近似,在图像中以相似的颜色显示,但该图中最小预测值和最大预测值都超出了校正集的参考值范围,说明存在预测误差。此外,每个油茶籽四周边缘处颜色鲜艳,对应较高的含油率,推测是因为边缘处只有油茶籽壳而没有茶仁,从而导致较高的预测误差。

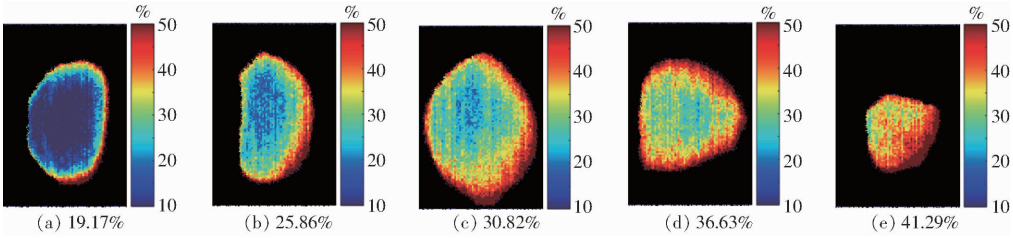

图 4 油茶籽含油率分布可视化结果

Fig. 4 Visualizations of oil content distribution of *Camellia oleifera* seed

该结果证明了利用高光谱成像技术实现油茶籽含油率含量分布可视化的可行性,对不同油茶籽含油率的快速无损评估具有重要意义。

3 结论

(1)通过对比不同的预处理方式发现,光谱集Ⅰ的原始光谱的 PLSR 模型精度最高:R_p^2 为 0.681、RMSEP 为 2.89%,光谱集Ⅱ同样是原始光谱的 PLSR 模型精度最高:R_p^2 为 0.740、RMSEP 为 2.92%。

(2)通过对比 SPA、CARS、PSO、ACO、SA、iRF、GA 等 7 种不同的特征波长筛选方法发现:对于光谱集Ⅰ,GA-PLSR 模型的精度最高,R_p^2 为 0.694,RMSEP 为 2.82%;对于光谱集Ⅱ,GA-PLSR 模型精度最高,R_p^2 为 0.779,RMSEP 为 2.54%。并且两组光谱集的 GA-PLSR 模型精度均比原始光谱的 PLSR 模型精度略有提升。

(3)通过对比光谱集Ⅰ和光谱集Ⅱ的建模效果发现,使用光谱集Ⅱ的原始数据或降维后的数据建立的回归模型的精度均比光谱集Ⅰ高,因此 900~1 700 nm 波段更适用于油茶籽含油率的快速无损检测。

参 考 文 献

[1] 邹锡兰, 吴尚清. 国务院批准颁布《全国油茶产业发展规划(2009—2020 年)》 千亿油茶产业待破局[J]. 中国经济周刊, 2009(47):45-46.

[2] 庄瑞林. 中国油茶[M]. 2 版. 北京:中国林业出版社, 2008.

[3] 焦玉海, 欧日明. 加大政策支持力度推进油茶产业健康发展[J]. 经济林研究, 2018,36(3):191.
JIAO Yuhai, OU Riming. Increase policy support to promote the healthy development of the *Camellia oleifera* industry[J]. Nonwood Forest Research, 2018,36(3):191. (in Chinese)

[4] 贾治邦, 封加平, 邓三龙. 在创新中推进油茶产业高质量发展——来自大三湘的调研报告[J]. 中国林业产业, 2019(3):6-15.
JIA Zhibang, FENG Jiaping, DENG Sanlong. Promoting the high-quality development of oil tea industry in innovation—a research report from Dasanxiang[J]. China Forestry Industry, 2019(3):6-15. (in Chinese)

[5] 尹维万. 油茶收摘要适时[J]. 湖南林业, 2003(8):23.
YIN Weiwan. *Camellia oleifera* abel need be harvested at the right time[J]. Hunan Forestry Science Technology, 2003(8):23. (in Chinese)

[6] 李宁. 不同方法提取茶籽油的工艺对比研究[J]. 粮食与食品工业, 2013,20(1):11-13.
LI Ning. Comparative study on technology of tea-seed oil extraction in different ways[J]. Cereal & Food Industry, 2013, 20(1):11-13. (in Chinese)

[7] GOMEZ-SANCHIS J, GOMEZ-CHOVA L, AlEIXOS N, et al. Hyperspectral system for early detection of rottenness caused by *Penicillium digitatum* in mandarins[J]. Journal of Food Engineering, 2008,89(1):80-86.

[8] CAYUELA J A, CAMINO M D C P. Prediction of quality of intact olives by near infrared spectroscopy[J]. European Journal of Lipid Science & Technology, 2010,112(11):1209-1217.

[9] JUNKWON P, TAKIGAWA T, OKAMOTO H, et al. Hyperspectral imaging for nondestructive determination of internal qualities for oil palm (*Elaeis guineensis* Jacq. var. tenera)[J]. Agricultural Information Research, 2009,18(3):130-141.

[10] JIN H, MA Y, LI L, et al. Rapid and non-destructive determination of oil content of peanut (*Arachis hypogaea* L.) using hyperspectral imaging analysis[J]. Food Analytical Methods, 2016,9(7):2060-2067.

[11] 蒋蘋, 罗亚辉, 胡文武, 等. 基于高光谱的油茶籽内部品质检测最优预测模型研究[J]. 农机化研究, 2015,37(7):56-60.
JIANG Pin, LUO Yahui, HU Wenwu, et al. Research on optimal predicting model for the detection of internal quality by hyperspectral technology[J]. Journal of Agricultural Mechanization Research, 2015,37(7):56-60. (in Chinese)

[12] 原姣姣, 王成章, 陈虹霞, 等. 近红外漫反射光谱法测定油茶籽含油量的研究[J]. 林产化学与工业, 2011,31(3):28-32.

YUAN Jiaojiao, WANG Chengzhang, CHEN Hongxia, et al. Determination of oil content of *Camellia oleifera* seeds by near infrared reflectance spectroscopy[J]. Chemistry and Industry of Forest Products, 2011,31(3):28 – 32. (in Chinese)

[13] 郑涛, 刘宁, 孙红, 等. 基于高光谱成像的马铃薯叶片叶绿素分布可视化研究[J/OL]. 农业机械学报, 2017,48(增刊):153 – 159.
ZHENG Tao, LIU Ning, SUN Hong, et al. Visualization of chlorophyll distribution of potato leaves based on hyperspectral imaging technology[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2017,48(Supp.):153 – 159. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no = 2017S025&flag = 1. DOI:10.6041/j.issn.1000-1298.2017.S0.025. (in Chinese)

[14] 李红,张凯,陈超,等. 基于高光谱成像技术的生菜冠层含水率检测[J/OL]. 农业机械学报,2021,52(2):211 – 217,274.
LI Hong, ZHANG Kai, CHEN Chao, et al. Detection of moisture content in lettuce canopy based on hyperspectral imaging technique[J/OL]. Transactions of the Chinese Society for Agricultural Machinery,2021,52(2):211 – 217,274. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no = 20210219&flag = 1. DOI:10.6041/j.issn.1000-1298.2021.02.019. (in Chiense)

[15] 黄佳聪, 阚欢, 万晓军, 等. 腾冲红花油茶果实成熟度及堆沤处理对油产量及其品质的影响[J]. 林业科学研究, 2012,25(5):612 – 615.
HUANG Jiacong, KAN Huan, WAN Xiaojun, et al. Effects of fruit maturity and compost on oil yield and quality of camellia reticulate[J]. Forest Research, 2012,25(5):612 – 615. (in Chinese)

[16] WU D, SUN D W. Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: a review—Part I: fundamentals[J]. Innovative Food Science & Emerging Technologies, 2013,19(2):1 – 14.

[17] ARAUJO M C U, SALDANHA T C B, GALVO R K H, et al. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis[J]. Chemometrics & Intelligent Laboratory Systems, 2001,57(2):65 – 73.

[18] LI H, LIANG Y, XU Q, et al. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration[J]. Analytica Chimica Acta, 2009,648(1):77 – 84.

[19] ZAVALA A E M, DIHARCE E R V, AGUIRRE A H. Particle evolutionary swarm for design reliability optimization: evolutionary multi-criterion optimization[C]// Proceedings of Third International Conference, EMO 2005, Guanajuato, 2005.

[20] STUTZLE M D T. Ant colony optimization[M]. Chicago:Bradford Company, 2004.

[21] BANGERT P. Optimization: simulated annealing[M]. Berlin:Springer Berlin Heidelberg, 2012.

[22] YUN Y H, LI H D, LESLIE R E W, et al. An efficient method of wavelength interval selection based on random frog for multivariate spectral calibration[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2013,111:31 – 36.

[23] 龙燕, 连雅茹, 马敏娟, 等. 基于高光谱技术和改进型区间随机蛙跳算法的番茄硬度检测[J]. 农业工程学报, 2019,35(13):270 – 276.
LONG Yan, LIAN Yaru, MA Minjuan, et al. Detection of tomato hardness based on hyperspectral technology and modified interval random frog algorithm[J]. Transactions of the CSAE, 2019,35(13):270 – 276. (in Chinese)

[24] LEARDI R, GONZALEZ A L E. Genetic algorithms applied to feature selection in PLS regression: how and when to use them[J]. Chemometrics & Intelligent Laboratory Systems, 1998,41(2):195 – 207.

[25] WOLD S, SJOSTROM M, ERIKSSON L. PLS-regression: a basic tool of chemometrics[J]. Chemometrics & Intelligent Laboratory Systems, 2001,58(2):109 – 130.

[26] ELMASRY G, SUN D W, ALLEN P. Near-infrared hyperspectral imaging for predicting colour, pH and tenderness of fresh beef[J]. Journal of Food Engineering, 2012,110(1):127 – 140.

[27] 邵园园,王永贤,玄冠涛,等. 基于高光谱成像的肥城桃品质可视化分析与成熟度检测[J/OL]. 农业机械学报,2020,51(8):344 – 350.
SHAO Yuanyuan,WANG Yongxian,XUAN Guantao,et al. Visual detection of SSC and firmness and maturity prediction for Feicheng peach by using hyperspectral imaging[J/OL]. Transactions of the Chinese Society for Agricultural Machinery,2020,51(8):344 – 350. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no = 20200838&flag = 1. DOI:10.6041/j.issn.1000-1298.2020.08.038. (in Chiense)

[28] 王伟,姜洪喆,贾贝贝,等. 基于高光谱成像的生鲜鸡肉糜中大豆蛋白含量检测[J/OL]. 农业机械学报,2019,50(12):357 – 364.
WANG Wei,JIANG Hongzhe,JIA Beibei,et al. Detection of soybean protein content in fresh minced chicken meat using hyperspectral imaging[J/OL]. Transactions of the Chinese Society for Agricultural Machinery,2019,50(12):357 – 364. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no = 20191241&flag = 1. DOI:10.6041/j.issn.1000-1298.2019.12.041. (in Chiense)

[29] WU D, SUN D W, HE Y. Application of long-wave near infrared hyperspectral imaging for measurement of color distribution in salmon fillet[J]. Innovative Food Science & Emerging Technologies, 2012,16:361 – 372.

[30] HOURANT P, BAETEN V, MORALES M T, et al. Oil and fat classification by selected bands of near-infrared spectroscopy[J]. Applied Spectroscopy, 2000,54(8):1168 – 1174.

[31] XU Q, LIANG Y. Monte Carlo cross validation[J]. Chemometrics and Intelligent Laboratory Systems, 2001,56(1):1 – 11.

[32] 李洪东, 曹东升, 许青松, 等. 蒙特卡罗交互检验在检测近红外数据奇异样本中的应用[C]// 全国第二届近红外光谱学术会议, 长沙, 2008.
LI Hongdong, CAO Dongsheng, XU Qingsong, et al. Outlier detection of near-infrared spectroscopy using Monte Carlo method[C]// The Second Asian NIR Symposium, Changsha, 2008. (in Chinese)

[33] 展晓日, 朱向荣, 史新元, 等. SPXY 样本划分法及蒙特卡罗交叉验证结合近红外光谱用于橘叶中橙皮苷的含量测定[J]. 光谱学与光谱分析, 2009,29(4):964 – 968.
ZHAN Xiaori, ZHU Xiangrong, SHI Xinyuan, et al. Determination of hesperidin in tangerine leaf by near-infrared spectroscopy with SPXY algorithm for sample subset partitioning and Monte Carlo cross validation[J]. Spectroscopy and Spectral, 2009,29(4):964 – 968. (in Chinese)