基于高光谱成像技术的青花椒产地识别研究

顾佳盛[1], 刘子健[1], 周　聪[2,3], 王游游[2], 杨　健[2,3], 黄　俊[1], 王宏鹏[1*], 白瑞斌[2,3*]

(1. 浙江科技学院生物与化学工程学院, 杭州　310023; 2. 中国中医科学院中药资源中心, 道地药材品质保障与资源持续利用全国重点实验室, 北京　100700; 3. 江西省道地药材质量评价研究中心, 南昌　330000)

摘　要: 目的　基于高光谱成像技术结合机器学习建立一种青花椒产地的快速识别方法, 实现四川、贵州、云南、重庆等 10 个青花椒主要产地样品的快速无损鉴别。**方法**　本研究利用全平皿法、五点平均法和中心点法 3 种不同的兴趣区域(region of interest, ROI)提取方式, 获得平行光谱数据, 分别采用 5 种预处理方法消除数据噪声提升模型性能, 并比较了偏最小二乘判别分析(partial least square-discriminant analysis, PLS-DA)、随机森林(random forests, RF)和支持向量机(support vector machine, SVM) 3 种模型的产地识别效果。**结果**　采用全平皿法提取兴趣区域, 通过二阶导(second derivative, D2)预处理后建立的 RF 模型分类效果最佳, 训练集和测试集的准确率均可达到 100%。进一步采用连续投影算法(successive projections algorithm, SPA)选择 27 个特征波长建模, 结果表明多元散射校正(multiplicative scatter correction, MSC)-RF 模型判别效果最优, 训练集准确率为 98.8%, 测试集准确率达到 98.3%。**结论**　本研究建立的方法可实现不同青花椒主要产地样品的快速无损鉴别, 为高光谱成像技术在食品和药品领域的推广应用及专属小型化仪器装备系统的开发提供了理论依据。

关键词: 高光谱成像技术; 青花椒; 产地识别; 机器学习; 兴趣区域

Research on origin identification of *Zanthoxylum schinifolium* based on hyperspectral imaging technology

GU Jia-Cheng[1], LIU Zi-Jian[1], ZHOU Cong[2,3], WANG You-You[2], YANG Jian[2,3], HUANG Jun[1], WANG Hong-Peng[1*], BAI Rui-Bin[2,3*]

(1. *School of Biological and Chemical Engineering, Zhejiang University of Science and Technology, Hangzhou* 310023, *China*; 2. *State Key Laboratory for Quality Ensurance and Sustainable Use of Dao-di Herbs, National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing* 100700, *China*; 3. *Evaluation and Research Center of Daodi Herbs of Jiangxi Province, Nanchang* 330000, *China*)

ABSTRACT: Objective　To establish a rapid identification method of the origin of *Zanthoxylum schinifolium* based on hyperspectral imaging technology co mbined with machine learning, and realize rapid and non-destructive

基金项目: 浙江科技学院科研业务费专项资金项目(2023QN024、2023JLZD010)、中药全产业链质量技术服务平台项目(2022-230-221)、名贵中药资源可持续利用能力建设项目(2060302)、浙江省“领雁”攻关计划项目(2022C02023)

Fund: Supported by the Basic Research Special Fund Project of Zhejiang University of Science and Technology (2023QN024, 2023JLZD010), the Quality and Technical Service Platform for Traditional Chinese-Medicine Whole Industry Chain (2022-230-221), the Ability Establishment of Sustainable Use for Valuable Chinese Medicine Resources (2060302), and the “Leading Goose” Research and Development Program of Zhejiang (2022C02023)

***通信作者:** 王宏鹏, 博士, 副教授, 主要研究方向为天然产物化学。E-mail: wanghongpeng@zust.edu.cn

白瑞斌, 博士, 助理研究员, 主要研究方向为多光谱技术在中药材无损分析中的应用。E-mail: bairuibin2022@163.com

***Corresponding author:** WANG Hong-Peng, Ph.D, Associate Professor, School of Biological and Chemical Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China. E-mail: wanghongpeng@zust.edu.cn

BAI Rui-Bin, Ph.D, Assistant Professor, National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing 100700, China. E-mail: bairuibin2022@163.com

identification of 10 major origins of Sichuan, Guizhou, Yunnan and Chongqing. **Methods** In this study, 3 kinds of different region of interest (ROI) extraction methods, including full plate, five-point average and center point, were used to obtain parallel spectral data, and 5 kinds of pretreatment methods were respectively used to eliminate data noise and improve model performance. The effects of origin identification of 3 kinds of models: Partial least square-discriminant analysis (PLS-DA), random forests (RF) and support vector machine (SVM) models were compared. **Results** The full plate was used to extract regions of interest, and the RF model established after second derivative (D2) preprocessing had the best classification effect, the accuracy of both training set and test set could reach 100%. The successive projections algorithm (SPA) was further used to select 27 characteristic wavelengths for modeling. The results showed that the multiplicative scatter correction (MSC)-RF model had the best discriminating effect, the accuracy of training set was 98.8%, and the accuracy of test set was 98.3%. **Conclusion** The method established in this study can demonstrate the rapid and non-destructive identification of samples from different major producing areas of *Zanthoxylum schinifolium*, which provides a theoretical basis for the popularization and application of hyperspectral imaging technology in food and medicine fields and the development of specialized miniaturized instrument and equipment system.

KEY WORDS: hyperspectral imaging technology; *Zanthoxylum schinifolium*; origin identification; machine learning; region of interest

0　引　言

青花椒为芸香科花椒属植物青椒(*Zanthoxylum schinifolium*)的干燥成熟果皮, 是世界流行的大宗调味品, 也是药食同源的天然物质[1]。青花椒主要产自四川、贵州、重庆等地区, 其中四川汉源县是青花椒的著名产地[2]。研究表明, 不同地区光照、气温、降水量等环境条件不同, 导致不同地区青花椒的品质亦存在差异[3]。在市场条件下, 产地来源是青花椒品质评估的重要因素; 随着市场需求的不断增长, 青花椒的地理标志产品由于“品牌”效应的影响进一步导致了市场价格差异, 目前市场上或线上渠道销售中常出现青花椒产地冒充、以次充好的现象, 极大地扰乱了市场秩序, 损害了消费者的利益[4]。通常青花椒产地鉴别及品质评价采用气相色谱-质谱法(gas chromatography-mass spectrometry, GC-MS)[5]、高效液相色谱法(high performance liquid chromatography, HPLC)[6]等化学仪器分析方法, 这些方法虽然结果准确, 但存在检测周期长, 硬件依赖程度高, 检测效率低等问题, 无法满足市场场景下快速、无损、准确鉴别的要求。因此, 开发具有针对青花椒的快速无损检测方法具有迫切的市场需求。

高光谱成像(hyperspectral imaging, HSI)技术是一种新兴的检测技术[7], 它可以同时获取样品的光谱信息和图像信息, 将样品放置在检测平台上, 通过批量地移动扫描采集光谱数据, 且不破坏样品成本低, 具有样品无损、高通量, 效率高等特点。近年来, 使用 HSI 技术应用于农产品、食品和中药材产地鉴别已有较多报道。吴静珠等[8]采集了我国 10 个产地大米样本的高光谱图像, 基于 AlexNet 卷积神经网络, 得到的最佳性能分类模型可达 99.5%的识别准确率; 张璐等[9]基于 HSI 结合分水岭算法对聚集的酸枣仁样品进行单粒样本光谱的自动提取, 建立的最佳模型对酸枣仁产地有较强的鉴别能力; LIU 等[10]使用 HSI 系统在不破坏茶叶外观的情况下, 获取高光谱信息并建立相应的模型, 达到 97.5%的溯源准确率; 张悦等[11]使用高光谱结合化学计量学模型, 实现了不同尺度产地陈皮的准确鉴别; CAI 等[12]采用 HSI 技术提取白芍样品两侧的光谱信息, 利用具有注意力机制的卷积神经网络(convolutional neural networks, CNN)模型, 可以有效地识别白芍的地理起源。

目前, HSI 应用于青花椒产地溯源的研究报道较少[13], 区分的产地也较少, 且主要是将青花椒与另一品种红花椒进行区分, 本研究旨在探索结合 HSI 和机器学习方法对青花椒进行产地区分的可行性, 使用全平皿法(full plate, FP)、五点平均法(five-point average, FPA)和中心点法(center point, CP)等兴趣区域(region of interest, ROI)提取方法, 评估不同 ROI 提取方法的建模效果, 为食品和中药材的快速无损评价提供新的参考依据。

1　材料与方法

1.1　青花椒样品

青花椒样品均采自 2022 年 6 月至 7 月, 包括四川汉源、云南巧家、贵州关岭、重庆江津等 10 个产地, 每个产地 12 个批次, 共 120 批样品(表 1)。每个批次随机抽取 5 个样本, 共 600 个青花椒样本进行高光谱数据采集, 并通过随机划分的方法将样本按 4:1 的比例划分成训练集和测试集。

表 1 青花椒样品信息

Table 1 Informations of *Zanthoxylum schinifolium* samples

编号	采样时间	产地	编号	采样时间	产地
MC-1	2022.06	四川省雅安市汉源县	MC-6	2022.07	贵州省六盘水市盘州市
MC-2	2022.06	四川省自贡市沿滩区	MC-7	2022.07	贵州省毕节市金沙县
MC-3	2022.06	四川省凉山彝族自治州冕宁县	MC-8	2022.07	贵州省铜仁市德江县
MC-4	2022.06	云南省昭通市巧家县	MC-9	2022.06	四川省南充市嘉陵区
MC-5	2022.07	贵州省安顺市关岭县	MC-10	2022.06	重庆市江津区

1.2 高光谱成像系统及图像采集

高光谱成像系统由可见-近红外(visible and near-infrared, VNIR, 400~1000 nm, H-V16, 挪威 Norsk Elektro Optikk 公司)与短波红外(short wave infrared, SWIR, 950~2500 nm, H-S16, 挪威 Norsk Elektro Optikk 公司)两个镜头、两个卤钨灯(150 W/12 V), 探测仪, 水平移动平台, 仪器自带计算机组成。

图像采集前, 打开双侧卤钨灯, 需提前对高光谱成像系统预热 30 min, 关闭室内灯光。镜头与样品的距离为 25 cm, 平台移动速度为 1.5 mm/s, VNIR 镜头的积分时间设置为 3500 μs, 帧周期为 20000; SWIR 镜头积分时间设置为 4500 μs, 帧周期为 52142, 两个镜头的光谱分辨率为 6 nm, 扫描方式为线性扫描。将青花椒样品放在 60 mm 培养皿里均匀铺满, 每次 5 份培养皿水平摆放于移动平台上, 同时放置 Teflon 白板, 用于黑白校正。

1.3 图像校正

黑白校正是光谱数据处理中的一种常用方法, 用于消除不均匀照明以及暗电流对光谱数据的影响[14], 以黑色移动平台的反射率为 0, 白板的反射率为 1, 对光谱图像进行反射率校正, 黑白板校正计算公式(1)如下:

$$I=\frac{I_r-I_d}{I_w-I_d} \tag{1}$$

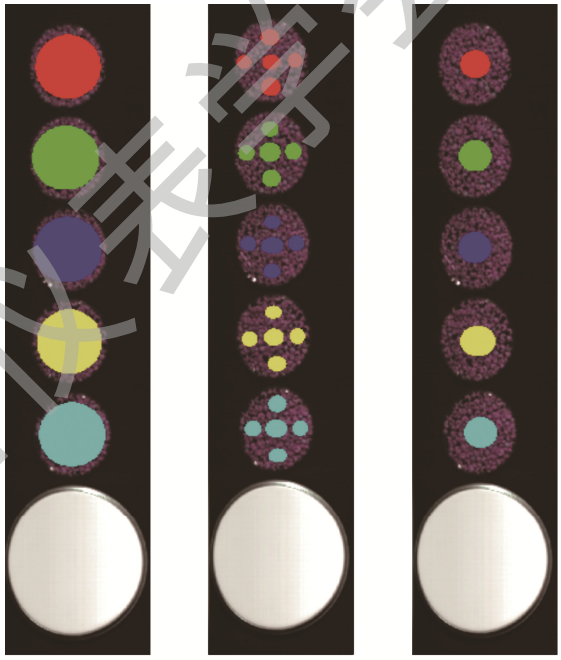
其中, I 是校正后的反射率图像; I_r是原始反射率图像; I_w是白板参考图像; I_d是黑板参考图像。

1.4 ROI 的提取

ROI 的提取采用 ENVI5.3 软件中的 ROI 工具, 在每份样品上创建 ROI 区域, 并计算对应区域内所有像素的平均相对反射率作为样品的光谱数据[15]。本研究分别采用 FP、FPA 和 CP 3 种处理模式进行 ROI 提取。其中 FP 指 ROI 覆盖样品全部平皿区域; FPA 是在整个平皿区域按固定顺序选取五块大小相近的区域作为 ROI; CP 则是选取培养皿的中心区域作为 ROI, 每份样品获得 3 份平行的原始光谱数据(图 1), 用于后续的建模对比分析。

1.5 光谱预处理方法

光谱数据采集过程中, 往往会受到噪声、杂散光、基线漂移等因素带来的干扰, 影响建模效果。因此, 通常需要在建模前对原始光谱数据进行合适的预处理, 提高模型的分类效果和稳定性[16]。本研究采用多元散射校正(multiplicative scatter correction, MSC)、一阶导(derivative, D1)、二阶导(second derivative, D2)、SG 平滑(Savitzky-Golay, SG)和标准正态变量(standard normal variate, SNV) 5 种光谱预处理方法, 比较不同预处理后的分类准确率。



全平皿法 FP　五点平均法 FPA　中心点法 CP

图1 3种ROI提取方式

Fig.1 Three kinds of extraction methods of ROI

1.6 分类模型建立

1.6.1 偏最小二乘判别分析

偏最小二乘判别分析(partial least squares discriminant analysis, PLS-DA)是以偏最小二乘回归(partial least squares regression, PLSR)为基础建立的一种高维线性判别分类模型[17]。在建模过程中, 将带有标签数值的向量转换为虚拟变量矩阵, 然后利用 PLSR 模型进行预测, 将预测类值最高的归属于对应的类。在 PLS-DA 模型中, 为寻找最佳的潜变量数量, 基于训练集数据, 采用十折交叉验证的方法验证不同的潜变量, 为防止出现过拟合等情况, 潜变量的数量被限制在 12 以内。

1.6.2 随机森林

随机森林(random forests, RF)是将多棵决策树整合起

来形成森林，对所有决策树的预测结果进行多数投票的集成分类模型。随机森林中决策树数量(n_estimators)太少，易出现欠拟合；太多则模型的复杂程度越高，进而导致模型的泛化能力变差，因此有必要确定合适的决策树数量。本研究使用交叉网格搜索的方法确定决策树的数量(n_estimators，优化范围 20~150)与决策树的最大深度(max_depth，优化范围 5~30)的最优组合。此外，为了减少随机采样带来的偶然性，提高模型的泛化能力，采用 10 次十折交叉验证法获得平均准确率以评估所得到的最优组合。

1.6.3　支持向量机

支持向量机(support vector machine，SVM)基于风险最小化理论的统计学习方法，通过使用核函数，将不同类别的数据用支持向量组成的超平面分隔，并对支持向量进行优化，使超平面距离最大化[18]。本研究采用网格搜索寻优的方法在训练集数据中确定超参数核函数类型(Kernel)、多项式核函数的阶数(degree，优化范围 2~3)、惩罚项系数(C，优化范围 10^{-1}~10^{2})和核函数系数(g，优化范围 10^{-7}~10^{0})等参数。采用留一交叉验证方法对得到的最佳组合进行评估，建立最佳参数模型。

1.7　特征波长的选择

连续投影算法(successive projections algorithm，SPA)是在向量空间中使用投影操作来去除不相关变量以解决共线问题，从而选择具有最小冗余和最低共线性的最优变量[19]。本研究筛选波段数量的最小值和最大值分别设置为 10 到 50，预处理方法设置为“autoscaling”。

1.8　数据处理及评估

本研究采用软件 PyCharm (Python 3.7)对光谱数据进行预处理以及分类算法建模，PLS-DA、RF 和 SVM 通过使用 Scikit-learn 库实现。

分类准确率是评估分类模型性能的最常用指标，分类准确率是指预测正确的样本数与总样本数的比值。混淆矩阵是模型分类效果的可视化体现，能够清晰直观的呈现准确率、精确率、召回率等评估指标，具体计算如公式(2)~(4):

$$\text{分类准确率}/\%=\frac{TP+TN}{TP+TN+FP+FN}\times 100\% \tag{2}$$

$$\text{精确率}/\%=\frac{TP}{TP+FP}\times 100\% \tag{3}$$

$$\text{召回率}/\%=\frac{TP}{TP+FN}\times 100\% \tag{4}$$

其中，TP 为真阳性的样本个数；TN 为真阴性的样本个数；FP 为假阳性的样本个数；FN 为假阴性的样本个数。

2　结果与分析

2.1　原始光谱曲线分析

青花椒样品的光谱如图 2 所示，不同产地青花椒光谱的变化趋势是相似的，说明其主要化学成分的物质组成是

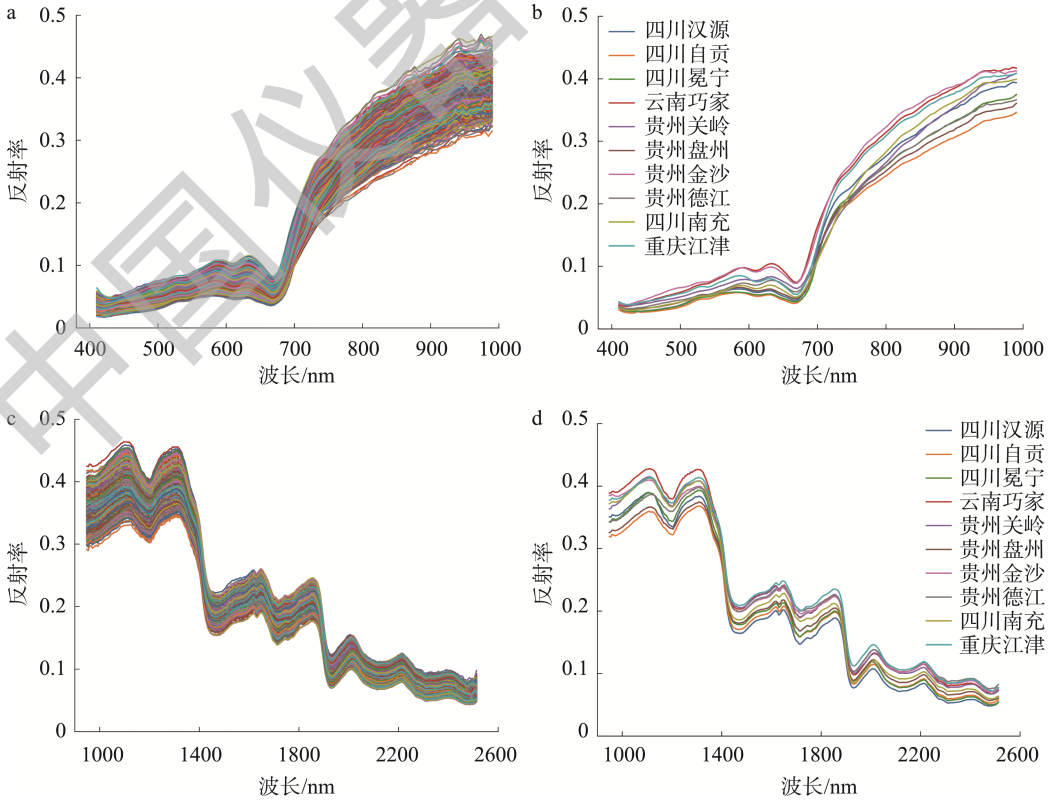


图2　不同产地青花椒在VNIR波段(a、b)和SWIR波段(c、d)的原始光谱曲线及平均光谱曲线图

Fig.2　Original and average spectral curves in the VNIR (a, b) and SWIR (c, d) bands of *Zanthoxylum schinifolium* from different regions

一致的，但在通过求均值获得的平均光谱中，600~700 nm 的波谷处，云南巧家和贵州金沙这两个产地样本的平均反射率要高于其他产地；1000~1300 nm 处，云南巧家样本的平均反射率同样最高，而四川自贡样本的平均反射率则略低于其他产地；1500~2000 nm 之间，重庆江津样本的平均反射率高于其他产地，而四川汉源样本的平均反射率则相对较低，说明不同产地青花椒的成分含量有较明显的产地差异[20]。

对不同产地青花椒的反射率数据进行降维处理，绘制的主成分分析(principal component analysis, PCA)得分图(图 3)，保留了前两个主成分，能够解释 89%以上的信息变量。云南巧家和四川自贡在 PC1 轴上相距较远，说明两个产地间的差异会比较大，这与获得的平均光谱图中两个产地的平均反射率相差较大的情况相一致。四川汉源、自贡、冕宁和南充 4 个产地的青花椒样本之间互有重叠，这与 4 个产地样本的地理位置相近、化学成分含量相似有关[21]；而贵州德江的样本大多表现出较为显著的分离趋势，说明贵州德江青花椒的成分含量与其他产地的样本之间存在较大的差异。PCA 结果表明，不同产地青花椒的化学成分具有一定的差异，表现为 PCA 得分图中四川汉源、贵州德江等产地样本的分离趋势；但云南巧家、贵州金沙、重庆江津等产地差异较小，PCA 得分图中样本并未表现出明显的分离趋势。因此，需要进一步采用 PLS-DA, RF 和 SVM 等模式识别方法构建青花椒的产地判别模型。

2.2 提取方式的选择

为了保证结果的可对比性，使用 FP、FPA、CP 3 种提取方式对原始数据进行提取，并结合 PLS-DA，RF 和 SVM 3 种模型进行分析，结果如表 2 所示。在 VNIR 波段、SWIR 波段和全波段范围内，大多数情况下 FP 方式的模型分类结果较好，分类准确率高于另外两种方式。在全波段范围内，FP 方式的 3 种模型测试集准确率都能达到 90%以上，原始数据下获得了较好的分类效果，其中最优的 SVM 模型分类结果，训练集准确率为 97.9%，测试集准确率为 95.0%。CP 方式的分类准确率较差，测试集准确率仅有 60%~80%左右。

由以上结果可知，FP 的分类准确率较优，分析认为是提取 ROI 时 FP 选取的兴趣区域范围大于另外两种提取方式，能够覆盖较多的信息变量，同时 FPA 的结果要优于 CP，因为 FPA 选到的范围多于 CP。虽然 FPA 和 CP 不易选到样本间的空隙范围，但可能还是选到的信息变量不够多，导致判别模型的分类准确率不如 FP 的分类准确率，由此可以得出，进行 ROI 兴趣区域提取时选择较大的范围才有更好的效果。

2.3 预处理方法和分类模型的选择

使用最优提取方式 FP 的光谱数据，基于不同的预处理方法，对比模型的判别效果，由表 3 可知，全波段相比于 VNIR 和 SWIR 波段的分类结果，准确率获得了一定的提升，PLS-DA 的分类准确率与 VNIR 波段(89.6%, 83.3%)，SWIR 波段(93.3%, 91.7%)相比，全波段能够达到 96.3%和 95.0%，同样 RF 的分类准确率与 VNIR 波段(96.9%, 93.3%)，SWIR 波段(96.5%, 87.5%)相比，全波段准确率能够达到 99.2%和 94.2%。

在构建的 3 种分类模型中，SVM 在 MSC、SNV 预处理后，测试集准确率有较大提高，全波段的测试集准确率均达到 99.2%，并且 3 个波段范围内的预测最优结果都是结合 SNV 预处理，说明 SNV 可以作为 SVM 模型的最优预处理方法。结果表明 RF 模型适合处理青花椒的光谱数据，获得的效果最好，分类准确率基本能达到 90%及以上，在全波段范围，除 SG 平滑外，其余 4 种预处理方法准确率均有提升，特别是 D1 和 D2 预处理，分类准确率均达到 98%以上，其中 D2-RF 训练集和测试集准确率均达到 100%。因此，SNV-SVM、D2-RF 能够较好的实现青花椒的产地识别分类。

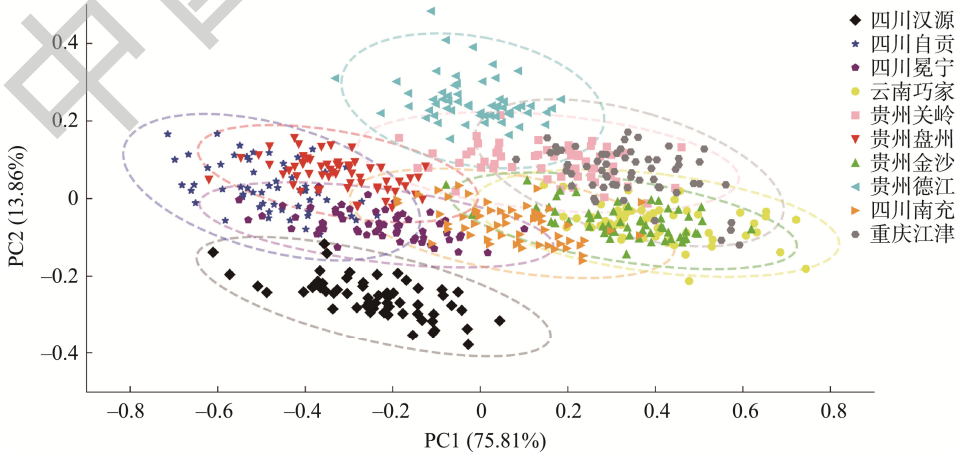


图3 不同产地青花椒PCA得分图

Fig.3 PCA score plot of *Zanthoxylum schinifolium* from different regions

表 2　样品不同兴趣区域选取和判别模型分类准确率结果(%)

Table 2　Selection of different regions of interest of samples and results of classification accuracy of discriminative model (%)

提取方式	分类模型	VNIR 波段		SWIR 波段		全波段	
		训练集	测试集	训练集	测试集	训练集	测试集
FP	PLS-DA	89.6	83.3	93.3	91.7	96.3	95.0
	RF	96.9	93.3	96.5	87.5	99.2	94.2
	SVM	88.1	87.5	93.5	89.2	97.9	95.0
FPA	PLS-DA	87.7	88.3	89.8	90.0	94.8	92.5
	RF	90.0	83.5	90.6	86.7	95.2	87.2
	SVM	81.3	70.8	87.1	85.0	94.8	90.8
CP	PLS-DA	81.0	81.7	86.7	82.5	87.9	87.5
	RF	80.4	68.5	86.9	75.6	90.4	75.2
	SVM	66.7	60.0	80.6	80.8	90.6	83.3

表 3　样品预处理方法和判别模型分类准确率结果(%)

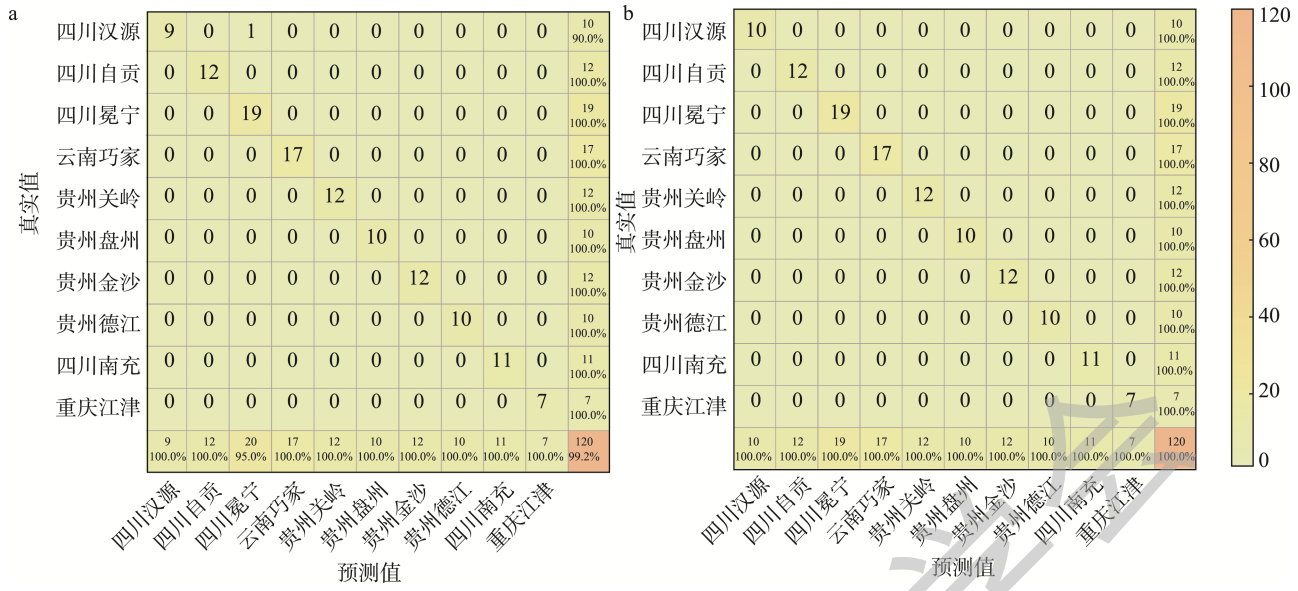
Table 3　Results of sample pretreatment method and classification accuracy of discriminative model (%)

分类模型	预处理方法	VNIR 波段		SWIR 波段		全波段	
		训练集	测试集	训练集	测试集	训练集	测试集
PLS-DA	原始数据	89.6	83.3	90.8	87.5	96.3	95.0
	MSC	90.2	92.5	91.9	90.8	97.5	95.8
	D1	89.6	85.0	94.6	91.7	98.3	97.5
	D2	88.8	88.3	82.9	70.8	93.1	90.0
	SG 平滑	87.9	83.3	90.2	87.5	96.5	95.0
	SNV	90.4	93.3	91.7	91.7	97.7	95.8
RF	原始数据	96.9	93.3	96.5	87.5	99.2	94.2
	MSC	96.5	90.8	99.0	96.7	98.3	97.5
	D1	99.6	98.5	99.4	95.8	99.8	98.9
	D2	100.0	99.4	99.6	97.5	100.0	100.0
	SG 平滑	95.0	92.5	96.7	88.3	99.0	94.2
	SNV	96.7	90.8	99.2	97.5	98.8	97.5
SVM	原始数据	88.1	87.5	93.5	89.2	97.9	95.0
	MSC	91.7	88.3	95.0	93.3	97.5	99.2
	D1	98.3	98.3	97.1	94.2	96.3	94.2
	D2	98.3	98.3	97.7	92.5	95.8	91.7
	SG 平滑	86.7	84.2	93.1	89.2	96.9	95.0
	SNV	99.4	99.2	99.6	97.5	99.6	99.2

2.4　模型分类性能评估

混淆矩阵是衡量模型分类准确率最直观、可视化效果最清晰的一种方法。混淆矩阵中的列代表真实类别, 行代表预测类别, 矩阵的对角线单元格代表每一类别预测正确的样本数量。最右列显示每一类别的真实样本数量以及该类别的召回率, 最底行显示被预测为该类别的样本数量以及该类别的精确率。

SNV-SVM 模型测试集结果的混淆矩阵见图 4a, 模型分类准确率为 99.2%, 同时四川汉源的召回率为 90.0%, 四川冕宁的精确率为 95.0%, 结果显示一份四川汉源的样品被错误预测成四川冕宁的样品。图 4b 是 D2-RF 模型结果的混淆矩阵, 测试集分类准确率为 100.2%, 产地分类均正确, 各个产地的召回率和精确率均达到 100.0%, 优于 SNV-SVM 模型的分类结果, 说明 D2-RF 模型可以达到更好的分类效果, 适用于青花椒样品的产地识别。



注: a: SNV-SVM模型; b: D2-RF模型。

图4 分类结果的混淆矩阵

Fig.4 Confusion matrix for classification results

2.5 特征波长的选择

对特征波长曲线的不同峰值进行关联分析(图 5), 800~900 nm 的波长与青花椒挥发油中 N—H、C—H 和 O—H 化学键震动相关[22]。980 nm 和 1430 nm 附近的吸收峰主要反映了水分含量[23]。1300 nm 附近的峰主要与青花椒中酰胺B类物质的第一个泛音有关, 而1500 nm附近的谷是由蛋白质相关的 N—H 拉伸振动的第一泛音引起的[24]。

基于全波段-RF 模型, 经 5 种预处理方法处理后, 用 SPA 选取特征波长建模, 分类准确率如表 4 所示, MSC-SPA-RF 的分类结果最优, 仅选取 27 个波段, 测试集准确率有 98.3%, 与MSC-RF (97.5%)相比, 提升了 0.8%, 除此之外SG 平滑-SPA-RF (97.5%)与 SG 平滑-RF (94.2%)相比, 测试集准确率提升了 3.3%, 可以得出特征波长的选取与全波段分类效果基本一致, 分类准确率均达到 94%以上且有所提高, 分析认为特征波长的选择减少了全波段里无关变量的干扰, 保留了更多有效信息的同时, 提升了模型的预测效果[25]。

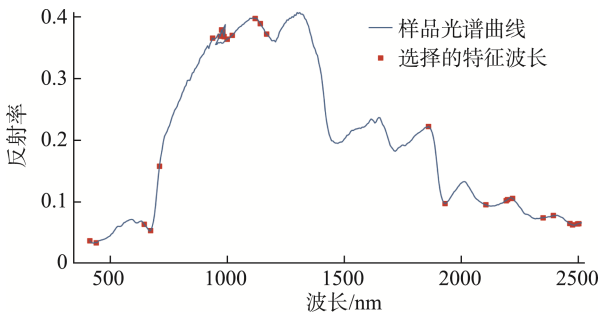


图5 MSC-SPA的特征波长选择

Fig.5 Feature wavelength selection for MSC-SPA

表 4 基于 SPA 特征波段的模型预测结果(%)

Table 4 Model prediction results based on SPA feature bands (%)

特征波长算法	预处理方法	特征波长数量	RF	
			训练集	测试集
SPA	原始数据	27	99.8	95.8
	MSC	27	98.8	98.3
	D1	23	99.4	95.0
	D2	36	100.0	97.5
	SG 平滑	33	99.6	97.5
	SNV	27	99.2	97.5

3 讨论与结论

原始光谱数据经由 MSC、SNV 预处理后, 多数情况下的分类准确率都有较大提升, 两种预处理方法有相似之处, 它们消除了青花椒样品在摆放时颗粒分布不均匀和颗粒大小不同而引起的散射效应, 增强了光谱信息之间的相关性, 因此在分类模型中获得的准确率也相似[26]。全波段的分类准确率与 VNIR 和 SWIR 波段内的准确率相比, 也获得了一定的提高, 这是因为全波段结合了 VNIR 可见光和 SWIR 近红外两个镜头下不同类型的波长信息, 相比于单一镜头, 全波段的波长信息更为全面和丰富, 能够进一步提高建模的可靠性和准确性[27]。

收集到的不同产地青花椒样品之间, 无论是形状、大小还是颜色鲜艳程度等外观都很相似, 用目视的方法难以快速区分各个产地[28], 而 VNIR 可见光镜头下往往能够捕捉到样品颜色的差异, 在 600~680 nm 处的叶绿素吸收波谷, 能够反映样品的颜色特征[29], 云南巧家和贵州金沙在此波

谷处平均反射率要高于其他产地, 表明这两个产地的青花椒颜色相对较深。从特征波长曲线可知, 在 600~680 nm 处存在 643.4 nm、670.5 nm 两个特征波段, 包含着与样品颜色有关的有效信息, 在数据分类中起着重要作用[30], 同时在可见光镜头下的特征波长范围大都集中在 935~990 nm 处, 说明这个区间的波长含有较多信息变量, 可以利用拥有较多有效信息的特征波长区间, 为简化识别青花椒光谱特征波长的小型仪器提供了设计思路。

本研究基于高光谱成像技术结合机器学习算法对 10 个产地的青花椒样品进行快速鉴别。构建 PLS-DA、SVM 和 RF 3 种模型用于青花椒样品的高光谱数据进行建模分析, 使用不同的 ROI 提取方式进行对比, FP 方式选取到的 ROI 范围较大, 训练集和测试集准确率均达到 90%以上, 在 3 种模型上的分类效果要优于 FPA 和 CP 法。以 FP 作为最优 ROI 提取方法, 分别使用 5 种不同的方法对光谱数据预处理后, 模型的准确率都得到进一步的提升, 其中 D2-RF 的模型分类效果最优, 训练集和测试集准确率均达到 100%, 此外, 通过 SPA 选取得到信息量大的特征波长, 获得了与全波段相似的分类效果, 与采用全波段建立的模型相比, 大大地降低了模型的复杂度。总体来说, 本研究实现了不同产地青花椒的快速无损判别, 为保护青花椒地理标志产品的市场提供了一种新的识别方法。

参考文献

[1] 国家药典委员会. 中华人民共和国药典[M]. 北京: 中国医药科技出版社, 2020.
National Pharmacopoeia Commission. Pharmacopoeia of the People's Republic of China [M]. Beijing: China Medical Science and Technology Press, 2020.

[2] 贾利蓉, 赵志峰, 雷绍荣, 等. 汉源青花椒挥发油的成分分析[J]. 食品与机械, 2008, 24(3): 105–108.
JIA LY, ZHAO ZF, LEI SR, *et al*. Analysis of chemical components of volatile oil from Hanyuan *Zanthoxylum schinifolium* Sieb. et Zucc [J]. Food Mach, 2008, 24(3): 105–108.

[3] 赵静珂, 李鑫, 黄登艳, 等. 不同青花椒品种挥发油成分的对比分析[J]. 生物资源, 2021, 43(5): 467–473.
ZHAO JK, LI X, HUANG DY, *et al*. Analysis of volatile components of essential oil in *Zanthoxylum armatum* DC. cultivars [J]. Biotic Res, 2021, 43(5): 467–473.

[4] 吴习宇, 祝诗平, 黄华, 等. 近红外光谱技术鉴别花椒产地[J]. 光谱学与光谱分析, 2018, 38(1): 68–72.
WU XY, ZHU SP, HUANG H, *et al*. Near infrared spectroscopy for determination of the geographical origin of Huajiao [J]. Spectrosc Spect Anal, 2018, 38(1): 68–72.

[5] 李亚南, 崔传坚, 陈江琳, 等. 基于挥发性风味物质的花椒产地溯源技术研究[J]. 食品工业科技, 2022, 43(2): 293–303.
LI YN, CUI CJ, CHEN JL, *et al*. Tracing the geographical origin of *Zanthoxylum bungeanum* by volatile compounds [J]. Sci Technol Food Ind, 2022, 43(2): 293–303.

[6] 课净璇. 基于化学指纹图谱的花椒产地鉴别与应用[D]. 雅安: 四川农业大学, 2018.
KE JX. Identification and application of chemical fingerprints of Chinese prickly ash (*Zanthoxylum*) [D]. Ya'an: Sichuan Agricultural University, 2018.

[7] HUANG HP, HU XJ, TIAN JP, *et al*. Rapid and nondestructive prediction of amylose and amylopectin contents in sorghum based on hyperspectral imaging [J]. Food Chem, 2021, 359: 129954.

[8] 吴静珠, 李晓琪, 林珑, 等. 基于 AlexNet 卷积神经网络的大米产地高光谱快速判别[J]. 中国食品学报, 2022, 22(1): 282–288.
WU JZ, LI XQ, LIN L, *et al*. Fast hyperspectral discrimination of rice origin based on Alexnet convolutional neural network [J]. J Chin Inst Food Sci Technol, 2022, 22(1): 282–288.

[9] 张璐, 茹晨雷, 殷文俊, 等. 基于近红外高光谱成像结合分水岭算法鉴别酸枣仁药材的产地[J]. 药物分析杂志, 2021, 41(4): 726–734.
ZHANG L, RU CL, YIN WJ, *et al*. Identification of *Ziziphi spinosae* Semen from different habitats based on near-infrared hyperspectral imaging technology and watershed algorithm [J]. Chin J Pharm Anal, 2021, 41(4): 726–734.

[10] LIU Y, HUANG JL, LI MH, *et al*. Rapid identification of the green tea geographical origin and processing month based on near-infrared hyperspectral imaging combined with chemometrics [J]. Spectrochim Acta A, 2022, 267: 120537.

[11] 张悦, 王游游, 张婷, 等. 高光谱结合图分割算法快速鉴别不同尺度产地陈皮[J]. 化学试剂, 2023, 45(1): 136–143.
ZHANG Y, WANG YY, ZHANG T, *et al*. Identification of *Citri reticulatae pericarpium* form different scales geographical origin by hyperspectral imaging combined with image segmentation algorithm [J]. Chem Reag, 2023, 45(1): 136–143.

[12] CAI ZY, HUANG ZH, HE MY, *et al*. Identification of geographical origins of *Radix paeoniae* Alba using hyperspectral imaging with deep learning-based fusion approaches [J]. Food Chem, 2023, 422: 136169.

[13] 胡航伟, 朱玲, 陈拾旸, 等. 花椒产地溯源技术研究进展[J]. 食品安全质量检测学报, 2023, 14(13): 110–116.
HU HW, ZHU L, CHEN SY, *et al*. Research progress on traceability technologies of *Zanthoxylum bungeanum* Maxim. origins [J]. J Food Saf Qual, 2013, 14(13): 110–116.

[14] SHAO YY, SHI YK, QIN YD, *et al*. A new quantitative index for the assessment of tomato quality using Vis-NIR hyperspectral imaging [J]. Food Chem, 2022, 386: 132864.

[15] 周聪, 王慧, 杨健, 等. 基于高光谱成像技术的中药栀子产地识别[J]. 中国中药杂志, 2022, 47(22): 6027–6033.
ZHOU C, WANG H, YANG J, *et al*. Origin identification of *Gardeniae fructus* based on hyperspectral imaging technology [J]. China J Chin Mater Med, 2022, 47(22): 6027–6033.

[16] 第五鹏瑶, 卞希慧, 王姿方, 等. 光谱预处理方法选择研究[J]. 光谱学与光谱分析, 2019, 39(9): 2800–2806.
DIWU PY, BIAN XH, WANG ZF, *et al*. Study on the selection of spectral preprocessing methods [J]. Spectrosc Spect Anal, 2019, 39(9): 2800–2806.

[17] BAI ZZ, HU XJ, TIAN JP, *et al*. Rapid and nondestructive detection of sorghum adulteration using optimization algorithms and hyperspectral imaging [J]. Food Chem, 2020, 331: 127290.

[18] LV YP, LV WB, HAN KX, *et al*. Determination of wheat kernels damaged by *Fusarium* head blight using monochromatic images of effective

wavelengths from hyperspectral imaging coupled with an architecture self-search deep network [J]. Food Control, 2022, 135: 108819.

[19] WANG YY, YANG J, YU S, *et al*. Prediction of chemical indicators for quality of *Zanthoxylum* spices from multi-regions using hyperspectral imaging combined with chemometrics [J]. Front Sustain Food Syst, 2022, 6: 1036892.

[20] PAN SW, ZHANG X, XU WB, *et al*. Rapid on-site identification of geographical origin and storage age of tangerine peel by near-infrared spectroscopy [J]. Spectrochim Acta A, 2022, 271: 120936.

[21] LONG WJ, ZHANG Q, WANG SR, *et al*. Fast and non-destructive discriminating the geographical origin of Hangbaiju by hyperspectral imaging combined with chemometrics [J]. Spectrochim Acta A, 2023, 284: 121786.

[22] KE JX, RAO LB, ZHOU LM, *et al*. Non-destructive determination of volatile oil and moisture content and discrimination of geographical origins of *Zanthoxylum bungeanum* Maxim. by hyperspectral imaging [J]. Infrared Phys Technol, 2020, 105: 103185.

[23] HUANG HP, HU XJ, TIAN JP, *et al*. Rapid and nondestructive determination of sorghum purity combined with deep forest and near-infrared hyperspectral imaging [J]. Food Chem, 2022, 377: 131981.

[24] WU N, JIANG HB, BAO YD, *et al*. Practicability investigation of using near-infrared hyperspectral imaging to detect rice kernels infected with rice false smut in different conditions [J]. Sens Actuators B-Chem, 2020, 308: 127696.

[25] 程介虹, 陈争光. 基于高光谱数据的乳香产地快速鉴别[J]. 黑龙江八一农垦大学学报, 2021, 33(4): 93–98.
CHENG JH, CHEN ZG. Rapid identification of Frankincense origin based on hyperspectral data [J]. J Heilongjiang Bayi Agric Univ, 2021, 33(4): 93–98.

[26] HU Y, KANG ZL. The rapid non-destructive detection of adulteration and its degree of Tieguanyin by fluorescence hyperspectral technology [J]. Molecules, 2022, 27(4): 1196.

[27] 殷文俊, 茹晨雷, 郑洁, 等. 基于高光谱成像技术融合光谱和图像特征鉴别不同产地的甘草[J]. 中国中药杂志, 2021, 46(4): 923–930.
YIN WJ, RU CL, ZHENG J, *et al*. Fusion of spectrum and image features to identify *Glycyrrhizae radix et Rhizoma* from different origins based on hyperspectral imaging technology [J]. China J Chin Mater Med, 2021, 46(4): 920–930.

[28] FAN LH, HUANG YL, ZHAO R, *et al*. Geographical-origin discrimination and volatile oil quantitative analysis of *Zanthoxylum bungeanum* Maxim. with a portable near-infrared spectrometer [J]. Anal Method, 2019, 11(41): 5301–5310.

[29] YANG L, GAO HQ, MENG LW, *et al*. Nondestructive measurement of pectin polysaccharides using hyperspectral imaging in mulberry fruit [J]. Food Chem, 2021, 334: 127614.

[30] SUN Y, LI YH, PAN LQ, *et al*. Authentication of the geographic origin of Yangshan region peaches based on hyperspectral imaging [J]. Postharvest Biol Technol, 2021, 171: 111320.

(责任编辑：韩晓红　郑　丽)

作者简介

顾佳盛，硕士研究生，主要研究方向为光谱数据分析。
E-mail: 222103855041@zust.edu.cn

王宏鹏，博士，副教授，主要研究方向为天然产物化学。
E-mail: wanghongpeng@zust.edu.cn

白瑞斌，博士，助理研究员，主要研究方向为多光谱技术在中药材无损分析中的应用。
E-mail: bairuibin2022@163.com