

基于近红外光谱的掺伪油茶籽油检测

郭文川^{1,2} 朱德宽¹ 张 乾¹ 杜荣宇¹

(1. 西北农林科技大学机械与电子工程学院, 陕西杨凌 712100; 2. 农业农村部农业物联网重点实验室, 陕西杨凌 712100)

摘要: 为了探索采用近红外光谱技术检测掺伪油茶籽油的潜力,以 12 个产地的玉米油、花生油、菜籽油和大豆油为掺杂油,以 5 个产地的油茶籽油为被掺杂油,制备了 455 份掺伪质量分数为 0、1%、3%、6%、10%、15% 和 20% 的掺伪油茶籽油,采集了所制备样品在 833 ~ 2 500 nm 范围内的近红外光谱。对采集的近红外光谱进行多元散射校正处理后,应用 Kennard - Stone 样本划分法按 2:1 的比例将样本划分为校正集和测试集。采用连续投影算法 (SPA)、无信息变量消除算法和竞争性自适应重加权算法提取表征掺伪油茶籽油样本的特征波长,并建立判别掺伪油茶籽油样品的支持向量机 (SVM) 和随机森林 (RF) 模型。研究结果表明,SVM 模型具有较高的灵敏度,RF 模型具有良好的特异性。基于 SPA 提取的 9 个特征波长所建立的 RF 模型的识别准确率最高,为 99.34%,对掺伪质量分数为 1% 的掺伪油茶籽油的识别准确率达到 94.74%,对掺伪质量分数为 3% 及以上的掺伪油茶籽油的识别准确率达到 100%。本研究为掺伪油茶籽油检测仪的研发提供了基础数据。

关键词: 油茶籽油; 近红外光谱; 特征波长; 支持向量机; 随机森林

中图分类号: S565; O433.4

文献标识码: A

文章编号: 1000-1298 (2020)09-0350-08

OSID:



Detection on Adulterated Oil-tea Camellia Seed Oil Based on Near-infrared Spectroscopy

GUO Wenchuan^{1,2} ZHU Dekuan¹ ZHANG Qian¹ DU Rongyu¹

(1. College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, Shaanxi 712100, China
2. Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, Yangling, Shaanxi 712100, China)

Abstract: Aiming to explore the potential of near-infrared (NIR) technology in detecting adulterated oil-tea camellia seed oil, the corn oil, peanut oil, rapeseed oil and soybean oil from 12 different production areas were used as adulteration oil, and the oil-tea camellia seed oil from five different production areas were used as adulterated oil. Totally 455 adulterated oil-tea camellia seed oil samples at the adulterated mass fractions of 0, 1%, 3%, 6%, 10%, 15% and 20% were prepared. The NIR spectra of the prepared samples were obtained at the wavelength range of 833 ~ 2 500 nm. After the collected NIR spectra were pretreated by multiple scatter correction method, the samples were divided into a calibration set and a validation set according to the ratio of 2:1 by using the Kennard - Stone sample partitioning method. Furthermore, successive projections algorithm (SPA), uninformative variable elimination and competitive adaptive reweighted sampling were used to extract the characteristic wavelengths (CWs) representing the adulterated oil-tea camellia seed oil samples from the investigated whole spectra. Then the support vector machine (SVM) and random forest (RF) classification models were established based on full spectra and extracted CWs. The results showed that the SVM model had higher true positive rates, while the RF model had better true negative rates. The established RF model based on the extracted nine CWs by using SPA had the highest recognition accuracy rate of 99.34%. Moreover, the recognition accuracy rate of the model was 94.74% for the adulterated oil-tea camellia seed oil samples whose adulterated mass fraction was 1%, and reached 100% for the adulterated oil samples whose adulterated mass fraction was equal to and greater than 3%. The research result provided basic data for the development of a portable detector for adulterated oil-tea camellia seed oil.

Key words: oil-tea camellia seed oil; near-infrared spectroscopy; characteristic wavelength; support vector machine; random forest

收稿日期: 2019 - 11 - 06 修回日期: 2019 - 12 - 25

基金项目: 国家自然科学基金面上项目 (31772065) 和国家级大学生创新创业训练计划项目 (201810712028)

作者简介: 郭文川 (1969—), 女, 教授, 博士生导师, 主要从事农产品和食品品质无损检测技术研究, E-mail: guowenchuan69@126.com

0 引言

油茶籽油富含油酸、亚油酸、亚麻酸等不饱和脂肪酸^[1],其脂肪酸组成与橄榄油相似^[2]。油茶籽油的优良品质使其市场销售价格是普通植物油的 5 ~ 10 倍。随着人们身体健康意识的增强,油茶籽油的市场需求量也逐渐增大。一些不法经营者为了谋取高额利润,向油茶籽油中掺入廉价的普通植物油,严重损害了消费者与合法经营者的正当权益。

目前,用于检测食用油品质的方法主要有气相色谱^[3]、液相色谱^[4]、低场核磁共振^[5]、激发矩阵荧光光谱^[6]、电子鼻^[7]等方法。这些方法需要使用大型昂贵的分析仪器或大量的化学试剂,因而整个检测过程费时、繁琐,且成本较高,无法满足快速鉴别掺伪油茶籽油的要求。近红外光谱分析技术是一种无损、快速、高效且无污染的现代分析技术,已广泛应用于多种领域^[8-13]。

目前,在应用近红外光谱技术检测掺伪油茶籽油方面已有一些研究报道^[14-16]。但现有研究的样本量偏少,而且大多是直接利用全光谱数据建模或将吸收峰附近的光谱作为输入变量。直接利用全光谱数据建模使得模型输入变量多、模型复杂,易出现过拟合的现象,同时无关信息的引入有可能降低模型的精度,且不利于经济实用的检测仪开发。而直接以油茶籽油光谱吸收峰的位置作为建模变量,则有可能忽略其他光谱处对建模有用的信息^[17]。为了开发便携式掺伪油茶籽油检测仪,有必要提取对掺伪油茶籽油敏感的特征波长,并分析其对建模效果的影响。本文以多个产地生产的油茶籽油、玉米油、花生油、菜籽油和大豆油为对象,制备掺伪油茶籽油,采用不同方法从全光谱数据中提取对掺伪油茶籽油敏感的特征波长,并基于全光谱数据和提取的特征波长建立识别掺伪油茶籽油的判别模型,对模型的综合性能进行分析,以期为基于多光谱技术的掺伪油茶籽油检测仪的研发提供基础数据。

1 材料与方法

1.1 试验材料

1.1.1 试验样本

试验所用油茶籽油样品共 5 个,产地分别为江西省吉安市、江西省玉山市、湖南省永州市祁阳县、浙江省杭州市和广西壮族自治区河池市巴马瑶族自治县。用作掺伪的植物油为玉米油、花生油、菜籽油和大豆油,每类植物油均来自 3 个不同产地。故用于掺伪的植物油样品共计 12 个。试验所用油均购于西安市某大型超市,所购置的油茶籽油均符合

GB/T 11765—2018。试验期间,所有样品均在保质期内。

制备样品时,向每个约 20 g 油茶籽油样品中按掺伪质量分数为 1%、3%、6%、10%、15% 和 20% 的梯度掺入 12 个用于掺伪的植物油样品,共得到 360 个掺伪油茶籽油样品。

为了增加纯油茶籽油样本量以保证后续试验所建模型具有普遍性,按质量分数 0 ~ 90% 间以 10% 为梯度将 5 个纯油茶籽油样品两两混合,共得到 95 个不同的纯油茶籽油样品。

1.1.2 试验仪器及软件

FA2004 型电子天平(上海舜宇恒平科学仪器有限公司,精度 0.1 mg);MPA 型傅里叶变换近红外光谱仪(德国 Bruker 公司,配备积分球漫反射附件,波长范围为 833 ~ 2 500 nm)。

由 MPA 型傅里叶变换近红外光谱仪自带的光谱分析软件 OPUS 6.5 (德国 Bruker 公司)采集光谱数据;由 Matlab2016a (美国 The MathWork 公司)完成光谱数据预处理、样本划分、特征波长提取和模型建立。

1.2 近红外光谱采集

利用傅里叶变换近红外光谱仪采集样本的近红外漫反射光谱。光谱扫描范围:833 ~ 2 500 nm (12 000 ~ 4 000 cm⁻¹),扫描次数:32 次;光谱分辨率:8 cm⁻¹;采集条件:室温(23 ~ 25℃),以空白比色皿为参比。采集光谱时每个样品测量 3 次,取平均值作为该样品的最终光谱。

1.3 光谱预处理

由于受到仪器自身或外界环境的干扰,所采集的近红外光谱会受到噪声、基线偏移等与建模无关信息的影响,因此在建模前需要对采集的原始光谱进行预处理。常见的光谱预处理方法有 Savitzky - Golay (S - G) 平滑、标准正态变量变换 (Standard normal variate transformation, SNV)、多元散射校正 (Multiple scatter correction, MSC)、一阶微分、二阶微分等。

S - G 平滑法是光谱分析中常用的预处理方法,它是基于最小二乘原理的移动窗口加权平均算法,能有效地提高光谱的平滑性,并降低噪声的干扰^[18]。SNV 和 MSC 法可以消除表面散射、固体颗粒大小和光程变化对近红外漫反射光谱的影响,达到去噪的效果^[19-20]。一阶微分、二阶微分等导数预处理方法能减少由系统内部引起的随机噪声,并能增强处理后信号频率的分辨率。

1.4 样品划分

基于样本光谱间欧氏距离的 Kennard - Stone

(K-S)样本划分方法,能有效地将光谱差异较大的样品选入校正集,将其余相近样品归入测试集,达到保证校正集样品具有代表性和均匀性的目的^[21]。因而 K-S 方法被广泛地用在定性研究中对样本进行划分。

1.5 特征波长选择

1.5.1 连续投影算法

基于向量投影分析的连续投影 (Successive projections algorithm, SPA) 算法能够在光谱矩阵中充分寻找含有最低限度冗余信息的变量组,使变量之间的共线性达到最小。同时能极大地减少建模所用特征波长的数量,提高建模的速度和效率^[22]。

对于定性分析问题,文献[23]提出提取的波长数量可以由 G 值确定。 G 值定义为

$$G = \frac{1}{n} \sum_{i=1}^n g_i \tag{1}$$

其中
$$g_i = \frac{r^2(x_i - \mu_{I_i})}{\min_{I_i \neq J_i} r^2(x_i - \mu_{J_i})} \tag{2}$$

式中 n ——总样品数
 x_i ——SPA 选定波长下第 i 个样品的吸光度
 μ_{I_i} ——第 i 个样品所属类别的平均吸光度
 μ_{J_i} ——第 i 个样品对应错误类别的平均吸光度
 $r(x_i - \mu_{I_i})$ —— x_i 与 μ_{I_i} 马氏距离
 $I_i、J_i$ ——第 i 个样品对应的正确、错误类别
 g_i 应该尽可能小,即 x_i 应该靠近其真实类别样本的中心,并且与错误类别样本中心相距较远。

1.5.2 无信息变量消除算法

无信息变量消除 (Uninformative variable elimination, UVE) 算法通过引入一定数目的随机变量到光谱矩阵中,建立偏最小二乘回归 (Partial least squares regression, PLS) 交互验证模型,根据各波长稳定性指数,即回归系数向量的均值与标准偏差的商的绝对值,决定光谱变量是否被选取^[24]。

1.5.3 竞争性自适应重加权算法

竞争性自适应重加权 (Competitive adaptive reweighted sampling, CARS) 算法是将每个波长作为一个单独的个体,利用自适应加权采样技术筛选出 PLS 模型中回归系数绝对值大的波长点,淘汰回归系数绝对值小的波长点,利用交互验证选出均方根误差 (Root mean square error of cross validation, RMSECV) 最低的波长变量子集作为优选波长变量子集^[25]。

1.6 建模方法及评价指标

1.6.1 支持向量机

支持向量机 (Support vector machine, SVM) 是

一种以结构风险最小化为基础思想的有监督学习模式识别算法^[26]。该算法将原始数据映射到高维空间以构建最优的分类超平面,然后假设分类器误差与平行超平面间的距离成反比关系,从而解决常规空间里数据间线性不可分的问题^[27]。SVM 在解决小样本、非线性和高维模式识别问题中表现出许多特有的优势,并在很大程度上克服了“维数灾难”和“过学习”等问题。

1.6.2 随机森林

随机森林 (Random forest, RF) 是一种用于分类和回归的机器学习方法。该方法组合多个决策树算法对相同现象产生重复的预测结果。RF 算法对每棵决策树进行自助法重采样,使误差估计的计算能够基于袋外样本数据。该算法的优点体现在对数据集中的噪声有较强的鲁棒性,不需要另外预留部分数据做交叉验证^[28]。

1.6.3 评价指标

本研究将识别准确率、灵敏度、特异性作为油茶籽油判别模型的评价指标。识别准确率为纯油茶籽油样品和掺伪油茶籽油样品被正确判别的百分比;灵敏度是指纯油茶籽油样品被正确判别为纯油茶籽油的百分比;特异性是指掺伪油茶籽油样品被正确判别为掺伪油茶籽油的百分比。

2 结果与讨论

2.1 光谱分析

图 1 所示是本研究中所用的 455 个纯油茶籽油和掺伪油茶籽油样品的原始吸收光谱。由图 1 可知,在 833 ~ 2 500 nm 的近红外光谱范围内,油茶籽油主要有 4 个吸收峰。1 200 nm 左右的吸收峰为—C—H 伸缩振动的二级倍频;1 400 nm 左右为—C—H 一级倍频及二级倍频的合频吸收;1 700 ~ 1 800 nm 之间为—CH₂、—CH₃、—CH = CH 中—C—H 伸缩振动的一级倍频;2 150 nm 左右为—C—H 伸缩振动的合频吸收^[29-30]。从波形上看,纯油茶籽油与掺伪油茶籽油的光谱曲线重叠严重、

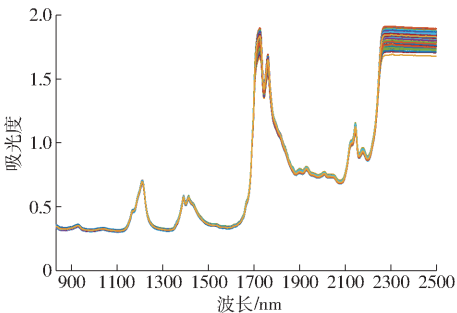


图 1 所有样品的原始近红外光谱

Fig. 1 Original near-infrared spectra of all samples

相似度很高,很难直接从中分辨出掺伪油茶籽油。因此本研究采用化学计量学方法对光谱进行预处理,并结合机器学习方法建立定性判别掺伪油茶籽油的模型。由于在光谱的起始和末尾部分存在较大的噪声,因此将1 100 ~ 2 200 nm(1 178 个中心波长)作为有效光谱区域。

2.2 光谱预处理和样本划分

为了减少无关因素对建模效果的影响,分别用 MSC、SNV、S-G 平滑、一阶微分和二阶微分共 5 种预处理方法对光谱进行预处理,然后使用 SVM 建立油茶籽油掺伪判别模型。以径向基函数作为 SVM 模型的核函数,并根据十折交叉验证和网格搜索法选择各模型交叉验证识别准确率最高时对应的惩罚因子和核参数作为建模参数。具体参数见表 1。

表 1 不同预处理方法下 SVM 建模的参数
Tab.1 Determined parameters of SVM models by different pretreatment methods

预处理方法	惩罚因子 C	核参数 γ
无	21 544. 36	0. 129 155 0
MSC	21 544. 36	0. 215 443 5
SNV	7 742. 636	0. 215 443 5
S-G 平滑	21 544. 36	0. 215 443 5
一阶微分	7 742. 636	464. 159 0
二阶微分	7 742. 636	464. 159 0

对比不同预处理方法下校正识别准确率和交叉验证识别准确率,确定最优的光谱预处理方法。不同预处理方法的建模结果如表 2 所示。由于 MSC 预处理方法使得建立的 SVM 模型具有最高的交叉验证识别准确率,因此在后续的研究中只对 MSC 预处理后的光谱进行分析。

表 2 不同预处理方法下 SVM 建模的预测结果
Tab.2 Prediction results of SVM models by different pretreatment methods

预处理方法	校正识别准确率	交叉验证识别准确率	%
无	99. 34	93. 62	
MSC	98. 90	96. 70	
SNV	98. 90	95. 16	
S-G 平滑	98. 46	94. 94	
一阶微分	99. 34	95. 60	
二阶微分	99. 34	95. 16	

对经 MSC 处理后的光谱采用 K-S 法按 2:1 的比例分别对 360 个掺伪油茶籽油和 95 个纯油茶籽油进行样本划分,得到校正集样品 304 个(240 个掺伪油茶籽油和 64 个纯油茶籽油样品),测试集样品 151 个(120 个掺伪油茶籽油样品和 31 个纯油茶籽油样品)。

2.3 特征波长选择

2.3.1 SPA 算法

将该算法提取的最小波长数设定为 1,最大数设定为 30,计算不同特征波长数下的 G ,结果如图 2 所示。由图 2 可见,当特征波长数小于 9 时,随着波长数的增加, G 迅速减小,但当波长数量大于 9 时, G 逐渐增大。以 G 最小处的波长数作为最佳波长数。因此本文以 9 个特征波长作为输入的特征变量,该 9 个特征波长分别是 1 163. 64、1 235. 16、1 382. 72、1 419. 82、1 458. 97、1 633. 64、1 733. 02、1 756. 50、1 896. 56 nm。

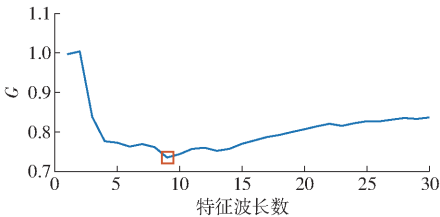


图 2 不同特征波长数的 G 变化曲线
Fig. 2 Calculated G values at different numbers of characteristic wavelengths

2.3.2 UVE 算法

基于 UVE 算法的掺伪油茶籽油特征波长选取结果如图 3 所示。其中,竖线左侧为全光谱 1 178 个波长的稳定性指数分布曲线,右侧为相同数量的随机变量稳定性指数分布曲线。以随机变量稳定性指数最大绝对值的 99% 作为变量筛选的阈值,即稳定性指数在两条水平虚线以外的特征波长被选中。利用 UVE 算法共选择出 207 个特征波长,其分布如图 4 所示。

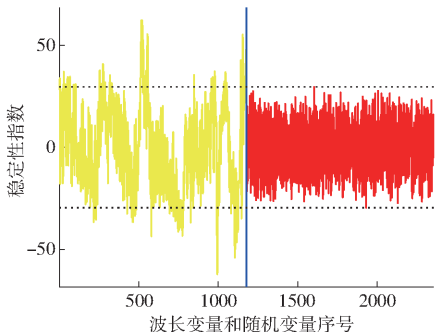


图 3 各波长变量和随机变量下的稳定性指数
Fig. 3 Stability indices at different wavelength variables and random variables

2.3.3 CARS 算法

图 5 为应用 CARS 算法筛选特征波长过程中特征波长数、RMSECV 以及回归系数随运行次数的变化图。由图 5 可见,当运行次数从 1 次增加到 28 次,特征波长数从迅速下降到缓慢下降,RMSECV 逐步降低,表明在 1 ~ 28 次筛选过程中淘汰了较多无关变量,模型精度逐步提高。当运行次数大于 28 次

时,随着运行次数的增加,RMSECV 缓慢或快速增大,模型精度下降。在运行次数为 28 次时, RMSECV 降到最低,此时有 35 个波长变量被保留下来。因此以该 35 个波长为特征波长,其分布如图 6 所示。

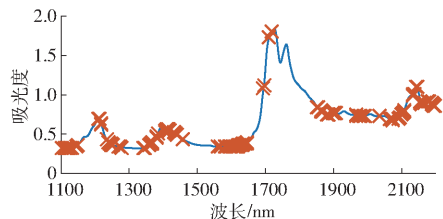


图 4 基于 UVE 算法筛选的特征波长

Fig. 4 Selected characteristic wavelengths by using UVE algorithm

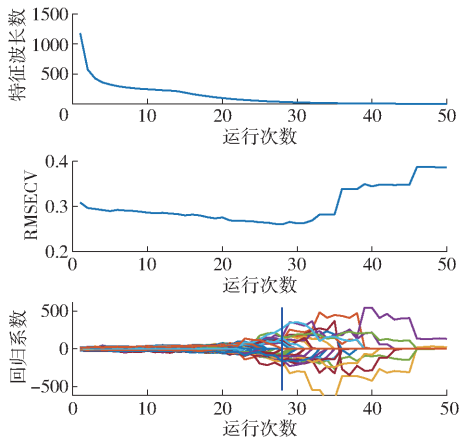


图 5 基于 CARS 算法筛选特征波长的过程

Fig. 5 Process of selecting characteristic wavelengths by using CARS algorithm

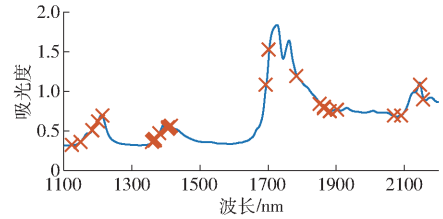


图 6 基于 CARS 算法筛选的特征波长

Fig. 6 Selected characteristic wavelengths by using CARS algorithm

从利用 SPA、UVE 和 CARS 算法提取的特征波长看,特征波长数明显少于全光谱(1 100 ~ 2 200 nm)下的 1 178 个波长,分别仅是全光谱下波长数的 0.936 9%、17.57% 和 2.971%。这说明提取特征波长对于简化模型、提高运算速度有很重要的作用。此外,从提取的特征波长看,除了吸收峰或吸收峰附近的波长外,吸收峰之间的一些波长也是对掺伪油茶籽油敏感的特征波长。

2.4 建模结果及建模方法比较

2.4.1 建模参数的选择和设定

以径向基函数作为 SVM 模型的核函数。根据

十折交叉验证和网格搜索法选取惩罚因子 C 和核参数 γ 。建立 RF 模型时,以不同特征波长提取方法下预测准确率最高时的决策树个数作为 RF 模型的决策树数量。确定的参数见表 3。

表 3 不同特征波长提取方法确定的 SVM 和 RF 模型的参数

Tab. 3 Determined parameters of SVM and RF models by different methods of characteristic wavelength selection

特征波长 提取方法	SVM		RF 决策树
	惩罚因子 C	核参数 γ	数量
FS(全光谱)	27 825.61	0.215 443 5	300
UVE	21 544.36	12.915 50	250
CARS	3 593.814	77.426 36	95
SPA	166.810 1	744.263 6	100

2.4.2 SVM 模型建模结果

以不同方法所提取的特征波长作为建模输入变量时,所建立的 SVM 模型对校正集和测试集的纯油茶籽油和掺伪油茶籽油的判别结果见表 4。由表 4 可以看出,基于 SPA、UVE 和 CARS 所提取的特征波长建立的 SVM 模型(SPA-SVM、UVE-SVM 和 CARS-SVM)对测试集的识别准确率分别为 96.03%、96.69% 和 96.69%,均高于基于全光谱建立的 SVM 模型(FS-SVM)对测试集 94.70% 的识别准确率。在灵敏度方面,SPA-SVM 的灵敏度与 FS-SVM 相同,均为 96.77%,UVE-SVM 和 CARS-SVM 的灵敏度最高,达到 100%,说明 UVE 和 CARS 提高了 SVM 模型对纯油茶籽油样本的识别能力。在特异性方面,SPA-SVM、UVE-SVM 和 CARS-SVM 的特异性均为 95.83%,高于 FS-SVM 的 94.17%,说明 3 种特征波长提取方法均提高了 SVM 模型对掺伪油茶籽油样本的识别能力。

所建立的 SVM 模型对不同掺伪质量分数油茶籽油的识别准确率如表 5 所示。表 5 说明 FS-SVM、UVE-SVM 和 CARS-SVM 对掺伪质量分数为 3% 以上的油茶籽油样本的识别准确率达到 100%,UVE-SVM 和 CARS-SVM 对掺伪质量分数为 1% 的油茶籽油样本的识别准确率为 73.68%,高于 FS-SVM 的 63.16%,说明 UVE 和 CARS 提高了 SVM 模型对掺伪质量分数为 1% 的油茶籽油样本的识别能力。而 SPA 方法虽然使 SVM 模型对掺伪质量分数为 1% 的油茶籽油样本的识别准确率提高到 78.95%,但对掺伪质量分数为 3% 的油茶籽油样本的识别准确率却下降到 96.43%,说明 SPA 方法在减少波长输入的同时也删去了对 SVM 建模有用的信息。

表 4 不同特征波长提取方法下 SVM 和 RF 模型对纯油菜籽油和掺伪油茶籽油的识别结果

Tab.4 Identification results of SVM and RF models for pure and adulterated oil-tea camellia seed oil by using different characteristic wavelength selecting methods							
建模方法	特征波长 提取方法	校正集			测试集		
		灵敏度	特异性	识别准确率	灵敏度	特异性	识别准确率
SVM	FS	96.88	97.50	97.37	96.77	94.17	94.70
	SPA	95.31	99.17	98.36	96.77	95.83	96.03
	UVE	100	100	100	100	95.83	96.69
	CARS	98.44	98.75	98.68	100	95.83	96.69
RF	FS	100	100	100	90.32	98.33	96.69
	SPA	100	100	100	100	99.17	99.34
	UVE	100	100	100	93.55	98.33	97.35
	CARS	100	100	100	93.55	99.17	98.01

表 5 不同模型对不同掺伪质量分数油茶籽油的识别准确率

Tab.5 Identification accuracy of different models for adulterated oil-tea camellia seed oil at different mass fractions							
建模方法	特征波长 提取方法	掺伪质量分数/%					
		1	3	6	10	15	20
SVM	FS	63.16	100	100	100	100	100
	SPA	78.95	96.43	100	100	100	100
	UVE	73.68	100	100	100	100	100
	CARS	73.68	100	100	100	100	100
RF	FS	94.74	96.43	100	100	100	100
	SPA	94.74	100	100	100	100	100
	UVE	94.74	96.43	100	100	100	100
	CARS	94.74	100	100	100	100	100

2.4.3 RF 模型建模结果

不同特征波长提取方法下,基于所提取的特征变量建立的 RF 模型对校正集和测试集中纯油茶籽油和掺伪油茶籽油的判别结果见表 4。由表 4 可见,基于 SPA、UVE 和 CARS 所提取的特征波长建立的 RF 模型(SPA-RF、UVE-RF 和 CARS-RF)的识别准确率分别为 99.34%、97.35% 和 98.01%,均高于基于全光谱所建立的 RF 模型(FS-RF)。在灵敏度上,SPA-RF、UVE-RF 和 CARS-RF 的灵敏度分别为 100%、93.55% 和 93.55%,高于 FS-RF 的 90.32%,说明 3 种特征波长提取方法均提高了 RF 模型对纯油茶籽油的识别能力。在特异性上,UVE-RF 与 FS-RF 的特异性相等,为 98.33%,SPA-RF 和 CARS-RF 的特异性均为 99.17%,说明利用 SPA 和 CARS 特征波长提取方法提高了 RF 模型对掺伪油茶籽油的识别能力。

所建立的 RF 模型对不同掺伪质量分数油茶籽油的识别准确率见表 5。由表 5 可见,不管是基于全光谱,还是基于不同方法所提取的特征波长,所建立的 RF 模型对掺伪质量分数为 1% 的掺伪油茶籽油的识别准确率均为 94.74%。当掺伪质量分数为 3% 时,SPA-RF 和 CARS-RF 的识别准确

率为 100%,高于 FS-RF 和 UVE-RF 96.43% 的识别准确率。结果表明,全光谱中含有对 RF 模型冗余的信息,而 SPA 和 CARS 方法比 UVE 方法能有效地从全光谱中提取出对掺伪油茶籽油敏感的特征波长。

2.4.4 结果比较

当对 SVM 和 RF 模型性能进行比较时,由表 4 可知,除 SPA 外,测试集其他特征波长提取方法下所建 SVM 模型的灵敏度均高于 RF 模型的灵敏度,说明 SVM 模型对纯油茶籽油的识别能力更强。而 RF 模型的特异性均高于 SVM 模型的特异性,说明 RF 模型对掺伪油茶籽油的识别能力更强。

由表 5 可知,RF 模型对掺伪质量分数为 1% 的掺伪油茶籽油的识别准确率(94.74%)明显高于 SVM 模型的最高识别准确率 78.95%。而当掺伪质量分数在 3% 以上时,两种模型的判别能力基本相当。

在所建立的 8 种掺伪油茶籽油判别模型中,测试集 SPA-RF 的识别准确率、灵敏度和特异性均最高,分别为 99.34%、100% 和 99.17%(表 4),且该模型对掺伪质量分数为 1% 和 3% 及以上的掺伪油茶籽油的识别准确率分别为 94.74% 和 100%(表 5)。此外,该模型的输入波长数只有 9 个,对于简化模型,降低运算时间和便携式掺伪油茶籽油检测仪的开发具有重要的意义。

3 结论

(1)研究了 SPA、UVE 和 CARS 3 种特征波长选择方法下 SVM 和 RF 模型对不同掺伪质量分数(0~20%)油茶籽油的识别能力。结果表明,这 3 种特征波长提取方法均可提高所建模型对油茶籽油的识别准确率、灵敏度和特异性,SVM 模型具有较高的灵敏度,而 RF 模型具有良好的特异性。

(2)在所建立的 8 个模型中,SPA-RF 模型具有最佳识别能力,其识别准确率、灵敏度、特异性

分别为 99.34%、100% 和 99.17%，对掺伪质量分数为 1% 的掺伪油茶籽油识别准确率达 94.74%，对掺伪质量分数为 3% 及以上的掺伪油茶籽油的

识别准确率达到 100%。本研究为基于多光谱技术开发便携式掺伪油茶籽油检测仪提供了基础数据。

参 考 文 献

- [1] SHI T, ZHU M T, CHEN Y, et al. ^1H NMR combined with chemometrics for the rapid detection of adulteration in camellia oils [J]. Food Chemistry, 2018, 242:308–315.
- [2] MA J, YE H, RUI Y, et al. Fatty acid composition of camellia oleifera oil [J]. Journal für Verbraucherschutz und Lebensmittelsicherheit, 2010, 6(1):9–12.
- [3] MERCHAK N, RIZK T, SILVESTRE V, et al. Olive oil characterization and classification by ^{13}C NMR with a polarization transfer technique: a comparison with gas chromatography and ^1H NMR [J]. Food Chemistry, 2018, 245:717–723.
- [4] BAJOUB A, PACCHIAROTTA T, ELENA H F, et al. Comparing two metabolic profiling approaches (liquid chromatography and gas chromatography coupled to mass spectrometry) for extra-virgin olive oil phenolic compounds analysis: a botanical classification perspective [J]. Journal of Chromatography A, 2016, 1428:267–279.
- [5] ZHU W, WANG X, CHEN L. Rapid detection of peanut oil adulteration using low-field nuclear magnetic resonance and chemometrics [J]. Food Chemistry, 2017, 216:268–274.
- [6] 刘燕德, 万常斓. 芝麻油掺伪的近红外透射光谱检测技术 [J/OL]. 农业机械学报, 2012, 43(7):136–140.
LIU Yande, WAN Changlan. Analysis of sesame oil adulteration using near infrared transmission spectroscopy [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2012, 43(7):136–140. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20120725&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2012.07.025. (in Chinese)
- [7] HAI Z, WANG J. Detection of adulteration in camellia seed oil and sesame oil using an electronic nose [J]. European Journal of Lipid Science and Technology, 2006, 108(2):116–124.
- [8] 郭文川, 王铭海, 岳绒. 基于近红外漫反射光谱的损伤猕猴桃早期识别 [J/OL]. 农业机械学报, 2013, 44(2):142–146.
GUO Wenchuan, WANG Minghai, YUE Rong. Early recognition of bruised kiwifruit based on near infrared diffuse reflectance spectroscopy [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2013, 44(2):142–146. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20130227&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2013.02.027. (in Chinese)
- [9] 郭文川, 刘大洋. 猕猴桃膨大果的近红外漫反射光谱无损识别 [J/OL]. 农业机械学报, 2014, 45(9):230–235.
GUO Wenchuan, LIU Dayang. Identification of expanded kiwifruits by near-infrared diffused spectroscopy [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2014, 45(9):230–235. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20140937&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2014.09.037. (in Chinese)
- [10] 王转卫, 迟茜, 郭文川, 等. 基于近红外光谱技术的发育后期苹果内部品质检测 [J/OL]. 农业机械学报, 2018, 49(5):348–354.
WANG Zhuanwei, CHI Qian, GUO Wenchuan, et al. Internal quality detection of apples during late developmental period based on near-infrared spectral technology [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(5):348–354. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20180541&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2018.05.041. (in Chinese)
- [11] 张德荣, 方慧, 何勇. 可见/近红外光谱图像在作物病害检测中的应用 [J]. 光谱学与光谱分析, 2019, 39(6):1748–1756.
ZHANG Derong, FANG Hui, HE Yong. Research of crop disease based on visible/near infrared spectral image technology: a review [J]. Spectroscopy and Spectral Analysis, 2019, 39(6):1748–1756. (in Chinese)
- [12] 李真, 周立红, 叶正良, 等. 近红外光谱分析技术在药物质量分析中的应用进展 [J]. 药物评价研究, 2016, 39(4):686–692.
LI Zhen, ZHOU Lihong, YE Zhengliang, et al. Research progress on application of near-infrared reflectance spectroscopy in pharmaceutical analysis [J]. Drug Evaluation Research, 2016, 39(4):686–692. (in Chinese)
- [13] 陈全胜, 王名星, 郭志明, 等. 鸡肉中假单胞菌的近红外光谱快速识别 [J/OL]. 农业机械学报, 2017, 48(8):328–334.
CHEN Quansheng, WANG Mingxing, GUO Zhiming, et al. Rapid identification of *Pseudomonas* spp. in chicken by near-infrared spectroscopy [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(8):328–334. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20170839&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2017.08.039. (in Chinese)
- [14] 张菊华, 朱向荣, 李高阳, 等. 近红外光谱法结合化学计量学方法用于茶油真伪鉴别分析 [J]. 分析化学, 2011, 39(5):748–752.
ZHANG Juhua, ZHU Xiangrong, LI Gaoyang, et al. Identification of camellia oils by near infrared spectroscopy combined with chemometrics [J]. Chinese Journal of Analytical Chemistry, 2011, 39(5):748–752. (in Chinese)

[15] 原姣姣,王成章,陈虹霞. 近红外透射光谱分析油茶籽油掺入豆油的研究[J]. 中国粮油学报,2012,27(3):110-114.
YUAN Jiaojiao, WANG Chengzhang, CHEN Hongxia. Study on the quantitative analysis of camellia oil adulterated with soybean oil by near infrared transmittance spectroscopy (NITS)[J]. Journal of the Chinese Cereals and Oils Association, 2012,27(3):110-114. (in Chinese)

[16] 莫欣欣,周莹,孙通,等. 可见/近红外光谱的油茶籽油三元体系掺假检测模型优化[J]. 光谱学与光谱分析,2016,36(12):3881-3884.
MO Xinxin, ZHOU Ying, SUN Tong, et al. Model optimization of ternary system adulteration detection in camellia oil based on visible/near infrared spectroscopy[J]. Spectroscopy and Spectral Analysis, 2016,36(12):3881-3884. (in Chinese)

[17] 孙通,吴宜青,李晓珍,等. 基于近红外光谱和子窗口重排分析的山茶油掺假检测[J]. 光学学报,2015,35(6):350-357.
SUN Tong, WU Yiqing, LI Xiaozhen, et al. Discrimination of camellia oil adulteration by NIR spectra and subwindow permutation analysis[J]. Acta Optica Sinica, 2015,35(6):350-357. (in Chinese)

[18] CHEN H, PAN T, CHEN J, et al. Waveband selection for NIR spectroscopy analysis of soil organic matter based on SG smoothing and MWPLS methods[J]. Chemometrics and Intelligent Laboratory Systems, 2011, 107(1):139-146.

[19] CHEN Q S, JIANG P, ZHAO J W. Measurement of total flavone content in snow lotus (*Saussurea involucrate*) using near infrared spectroscopy combined with interval PLS and genetic algorithm[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2010, 76(1):50-55.

[20] GELADI P D, MACDOUGALL D B, MARTENS H. Linearization and scatter-correction for near-infrared reflectance spectra of meat[J]. Applied Spectroscopy, 1985, 39(3):491-500.

[21] WU W, WALCZAK B, MASSART D L, et al. Artificial neural networks in classification of NIR spectral data: design of the training set[J]. Chemometrics and Intelligent Laboratory Systems, 1996, 33(1): 35-46.

[22] SOARES S F C, GOMES A A, ARAUJO M C U, et al. The successive projections algorithm[J]. TrAC Trends in Analytical Chemistry, 2013, 42: 84-98.

[23] PONTES M J C, GALVAO R K H, ARAUJO M C U, et al. The successive projections algorithm for spectral variable selection in classification problems[J]. Chemometrics & Intelligent Laboratory Systems, 2005, 78(1):11-18.

[24] CENTNER V, MASSART D L, DE NOORD O E, et al. Elimination of uninformative variables for multivariate calibration[J]. Analytical Chemistry, 1996, 68(21): 3851-3858.

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

[26] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[27] SUYKENS J A K, VANDEWALLE J. Least squares support vector machine classifiers[J]. Neural Processing Letters, 1999, 9(3): 293-300.

[28] BREIMAN L. Random forests[J]. Machine Learning, 2001,45(1): 5-32.

[29] 褚小立. 近红外光谱分析技术实用手册[M]. 北京:机械工业出版社,2016:1-13.

[30] 张小超,吴静珠,徐云. 近红外光谱分析技术及其在现代农业中的应用[M]. 北京:电子工业出版社,2012.