油脂安全

基于红外光谱快速鉴别压榨油茶籽油与浸出油茶籽油的研究

王泽富[1]，吴雪辉[1,2]

（1. 华南农业大学 食品学院，广州 510642；2. 广东省油茶工程技术研究中心，广州 510642）

摘要：为规范油茶籽油市场、维护消费者权益，建立了快速、准确鉴别压榨油茶籽油和浸出油茶籽油的方法。通过傅里叶变换红外光谱仪对大量压榨油茶籽油和浸出油茶籽油样品进行扫描，提取特征波段数据，运用 Savitzky－Golay 平滑（SG）、多元散射校正（MSC）、标准正态变量变换（SNV）、一阶导数（FD）和二阶导数（SD）方法进行预处理，然后结合偏最小二乘法（PLS）、支持向量机（SVM）和 BP 人工神经网络（BPANN）建立鉴别模型。结果表明，偏最小二乘法和 BP 人工神经网络建模时，SG 平滑预处理方法最好，得到的 SG－PLS 和 SG－BPANN 两模型的验证集相关系数、验证集均方根误差、鉴别准确率分别为 0.767 9 和 0.921 2、0.322 6 和 0.205 9、88.46% 和 100%；支持向量机建模宜采用 SNV 预处理，建立的 SNV－SVM 模型验证集相关系数、验证集均方根误差和鉴别准确率分别为 0.761 4、0.882 1、88.46%。因此，红外光谱技术用于鉴别压榨油茶籽油和浸出油茶籽油是可行的。

关键词：红外光谱；油茶籽油；压榨；浸出；鉴别；偏最小二乘法；支持向量机；BP 人工神经网络

中图分类号：TS227；S794.4　　**文献标识码**：A　　**文章编号**：1003－7969（2018）11－0063－06

Rapid identification of pressed and extracted oil－tea camellia seed oils based on infrared spectroscopy

WANG Zefu[1], WU Xuehui[1,2]

（1. College of Food Science, South China Agricultural University, Guangzhou 510642, China;
2. Guangdong Engineering Research Center for Oil－Tea Camellia, Guangzhou 510642, China）

Abstract: In order to standardize the market of oil－tea camellia seed oils and safeguard the rights of consumers, a rapid and accurate method for identification of pressed and extracted oil－tea camellia seed oil was established. A large number of pressed and extracted oil－tea camellia seed oil samples were scanned by Fourier transform infrared spectroscopy to extract the characteristic band data. Savitzky－Golay smoothing (SG), multivariate scatter correction (MSC), standard normal transformation (SNV), first derivative (FD) and second derivative (SD) methods were used to preprocess, then combined with partial least squares (PLS), support vector machine (SVM) and BP artificial neural network (BPANN) to establish identification model. The results showed that when BPANN and PLS were used to establish the identification models, the results of SG were the best, and the correlation coefficient of validation (R_p), the root mean square error of validation (*RMSEP*) and the identification accuracy of the SG－PLS model and SG－BPANN model were 0.767 9 and 0.921 2, 0.322 6 and 0.205 9, 88.46% and 100% respectively. The SNV was the optimal preprocessing method for SVM modeling, and the R_p, *RMSEP* and the identification accuracy of the SNV－SVM model were 0.761 4, 0.882 1 and 88.46% respectively. Therefore, infrared spectroscopy could be applied to the identification of pressed and extracted oil－tea camellia seed oils.

收稿日期：2018－04－14；**修回日期**：2018－08－28

基金项目：林业公益性行业科研专项经费项目（201504703）；广东省林业科技创新项目（2017KJCX005）

作者简介：王泽富（1989），男，硕士研究生，研究方向为农产品加工（E-mail）1137224683@qq.com。

通信作者：吴雪辉，教授，博士（E-mail）xuehwu@scau.edu.cn。

Key words: infrared spectroscopy; oil - tea camellia seed oil; press; extraction; identification; partial least squares; support vector machine; BP artificial neural network

油茶籽油是我国特有的木本食用油脂,其脂肪酸组成与橄榄油相似,有“东方橄榄油”之称[1]。目前,油茶籽油的生产工艺主要有压榨法和溶剂浸出法。压榨法制取的油茶籽油色泽浅、风味纯正、营养成分含量高,但出油率较低;浸出法一般是对压榨后的茶籽饼进行浸出,将残留的油脂提取出来,提高出油率,但得到的油茶籽油营养成分被破坏较大,含有较多的非油脂成分,色泽深,后续精炼工艺复杂[2]。市场上压榨油茶籽油需求大于浸出油茶籽油,价格也比浸出油茶籽油高出 2 ~ 3 倍。虽然食用油产品国家标准要求在包装上标示生产工艺是“压榨”还是“浸出”法[3],但有些企业或商家为了追求高额利润,将浸出油茶籽油假冒为压榨油茶籽油,严重损害了消费者的权益。为了规范油茶籽油市场和保护消费者的利益,亟待探索出一种快速、准确的方法进行压榨油茶籽油和浸出油茶籽油的鉴别。

红外光谱技术根据分子内部原子间的相对振动和分子转动等信息来确定物质分子结构,得到样品中丰富的化学成分信息,是近年来迅速发展起来的无损检测技术,具有高灵敏度、高度计算机化等特点[4]。目前,应用近红外光谱分析技术对食用植物油掺伪鉴别有一定的研究,包括牛油果油掺伪鉴别[5],橄榄油品质分级[6]、掺伪鉴别[7]与溯源[8-9]及油茶籽油的真伪鉴别分析等方面[10-11],但未见应用红外光谱鉴别压榨油茶籽油和浸出油茶籽油的研究报道。油茶籽油生产工艺不同,营养成分组成或含量可能发生改变,导致不同基团或同一基团产生的红外光谱在吸收峰的位置和强度上有所不同,利用这种差异可以鉴别出油茶籽油的生产工艺。因此,本研究采用傅里叶变换红外光谱仪扫描大量油茶籽油加工厂采集的压榨油茶籽油和浸出油茶籽油样品,筛选两种生产工艺的油茶籽油特征指标,采用多种数据处理方法,建立鉴别模型,以期为压榨油茶籽油和浸出油茶籽油的鉴别提供一种快速、准确的方法。

1 材料与方法

1.1 样本及样本集划分

油茶籽油:86 个样本采集于广东省油茶籽油生产企业,其中压榨油茶籽油样本 46 个,浸出油茶籽油样本 40 个。

采用 SPXY(sample set portioning based on joint x - y distances)算法选取建模集和验证集样本。选择 60 个样本作为建模集,包括 27 个浸出油茶籽油和 33 个压榨油茶籽油;其余 26 个样本为验证集,包括 13 个浸出油茶籽油和 13 个压榨油茶籽油。

1.2 红外光谱采集及光谱数据预处理

采用 Nicolet iS 10 傅里叶变换红外光谱仪(赛默飞世尔科技有限公司)采集压榨油茶籽油和浸出油茶籽油红外光谱信息。光谱检测范围 4 000 ~ 400 cm^{-1},分辨率 4 cm^{-1}。每个样本重复扫描 3 次,以其平均值作为样本最终吸光度。

由于采集的红外光谱原始数据不仅包含了样本的化学信息,还包含了许多外界干扰信息,因此有必要采用合理的预处理方法消除干扰因素,以提高模型的准确性。采用 Matlab R2016b 软件对光谱数据进行预处理,包括 Savitzky - Golay 平滑(SG)、多元散射校正(MSC)、标准正态变量变换(SNV)、一阶导数(FD)和二阶导数(SD)。

1.3 模型建立与评价

采用偏最小二乘法(PLS)、支持向量机(SVM)和 BP 人工神经网络(BPANN)构建压榨油茶籽油与浸出油茶籽油鉴别模型,具体过程如图 1 所示:先将压榨油和浸出油分别进行赋值,即浸出油茶籽油样本为 0、压榨油茶籽油样本为 1,以此作为分类变量;再对原始光谱数据实施预处理,增强光谱特征,提取特征向量作为变量,构建鉴别模型,采用验证集相关系数(R_P)、验证集均方根误差(*RMSEP*)和鉴别准确率(R_r)参数来评价模型的优劣[11]。类型判别依据:分类变量预测值 y_{Pi},当 $y_{Pi} \leqslant 0.5$,则属于 0 类即浸出油茶籽油,当 $y_{Pi} > 0.5$ 则属于 1 类即压榨油茶籽油。

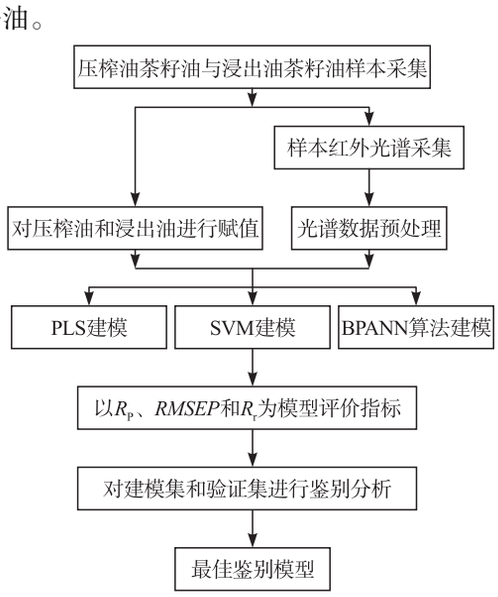


图 1 模型的建立与评价

2　结果与分析

2.1　压榨油茶籽油与浸出油茶籽油特征光谱分析

图 2 为压榨油茶籽油和浸出油茶籽油样本的红外光谱图。从图 2 可以看出，压榨油茶籽油和浸出油茶籽油样本的红外光谱在 4 000 ~ 400 cm^{-1} 范围内差别微小，均在 3 007、2 924、2 852、1 747、1 462、1 377、1 163、723 cm^{-1} 有吸收峰，肉眼难以区分两者的差别。虽然两者的特征吸收峰一样，但是吸光度及变化趋势却有所差异，因此需要采用化学计量学方法进行分类鉴别。

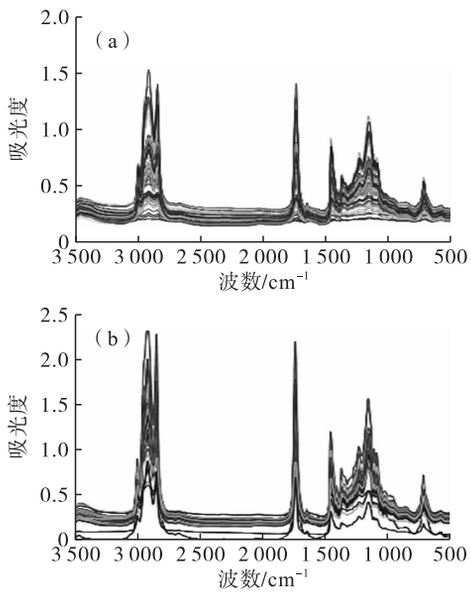


图 2　浸出油茶籽油(a)和压榨油茶籽油(b)的红外光谱图

2.2　压榨油茶籽油与浸出油茶籽油 PLS 鉴别模型的建立及预测

2.2.1　主成分数的确定

采用交叉验证法确定回归模型中最佳主成分数，建立 PLS 模型，以交叉验证均方根误差(*RMSECV*)和相关系数(R_{CV})确定 PLS 模型的最佳主成分数。研究不同主成分数对应的 R_{CV} 和 *RMSECV*，其变化曲线如图 3 所示。由图 3 可知，*RMSECV* 曲线随主成分数的增加呈递减后又上升趋势，R_{CV} 曲线随主成分数的增加呈递增后又下降趋势。综合考虑，确定回归模型中的最佳主成分数为 4。

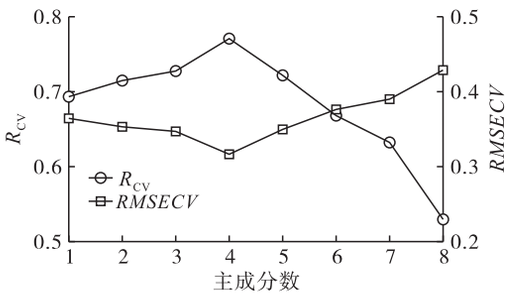


图 3　主成分数与相关系数和交叉验证均方根误差关系

2.2.2　不同预处理方法的模型结果

选取较佳主成分数进行 PLS 建模，5 种预处理方法对应的模型结果如表 2 所示。由表 2 可知，经 SG、MSC 和 FD 预处理后，相比原始光谱模型的 R_P 分别上升了 0.142 2、0.009 7、0.002 3，然而经 SNV 和 SD 预处理，相比原始光谱其模型 R_P 则是分别下降了 0.088 4、0.009 7。这很可能是因为 SG 和 MSC 通过消除样本散射从而滤掉了一些噪声，而经过 FD、SD 和 SNV 虽然消除了噪声但同时也滤掉了一部分有用信息。在上述 5 种预处理方法中，SG 平滑预处理后建立的模型验证集 R_P 最大，且验证集 *RMSEP* 最小，因此 SG 平滑联合偏最小二乘法(SG - PLS)建模效果较优。

表 2　不同预处理方法的 PLS 模型结果

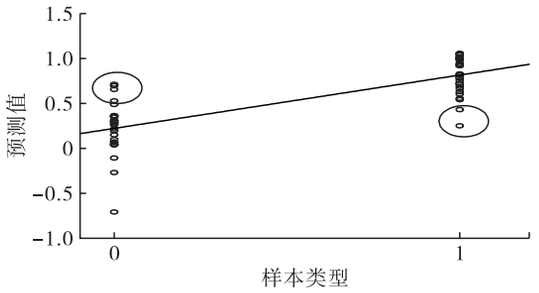
预处理方法	建模集		验证集	
	相关系数 R_{CV}	均方根误差 *RMSECV*	相关系数 R_P	均方根误差 *RMSEP*
原始光谱	0.762 4	0.321 9	0.625 7	0.390 0
Savitzky - Golay 平滑(SG)	0.771 0	0.276 7	0.767 9	0.322 6
多元散射校正(MSC)	0.679 0	0.365 2	0.635 4	0.386 9
标准正态变量变换(SNV)	0.698 4	0.356 0	0.537 3	0.431 9
一阶导数(FD)	0.755 8	0.325 7	0.628 0	0.391 3
二阶导数(SD)	0.755 1	0.326 1	0.616 0	0.393 8

2.2.3　SG - PLS 模型对压榨油茶籽油与浸出油茶籽油的鉴别分析

应用 SG - PLS 模型对建模集和验证集样本进行鉴别分析，其结果如图 4、图 5 所示。

从图 4、图 5 可以看出，该模型对建模集和验证集大部分样本鉴别结果准确，只有少数样本出现误判，统计分析结果如表 3 所示。

由表 3 可知，该模型对建模集和验证集的误判个数分别为 6、3。建模集样本鉴别过程中，将 4 个浸出油样本(预测值分别为 0.711 2、0.695 1、0.526 5、0.654 3)误判为压榨油样本，2 个压榨油样本(预测值为 0.429 1、0.248 6)误判为浸出油样本。验证集样本鉴别过程中，将 1 个浸出油样本(预测值为 0.616 5)误判为压榨油样本，2 个压榨油样本(预测值为 0.3374、0.405 5)误判为浸出油样本。SG - PLS 模型对建模集和验证集的鉴别准确率分别为 90%、88.46%。



注:图中圈内的点表示鉴别错误的样本。下同。

图 4 SG – PLS 对建模集样本鉴别结果

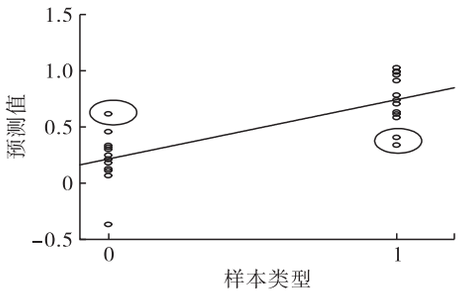


图 5 SG – PLS 对验证集样本鉴别结果

表 3 SG – PLS 模型对压榨油与浸出油鉴别统计结果

样本类型	实际类别	预测结果		鉴别准确率/%	相关系数 *R*	均方根误差 *RMSE*
		0	1			
建模集	0	23	4	90	0.771 0	0.276 7
	1	2	31			
验证集	0	12	1	88.46	0.767 9	0.322 6
	1	2	11			

2.3 压榨油茶籽油与浸出油茶籽油 SVM 鉴别模型的建立与预测

2.3.1 SVM 参数寻优

网格全局寻优算法是将参数的寻优范围划分为网格形式并遍历网格内的所有参数点去搜寻最优值。网格寻优算法的精确度与参数寻优范围及所设步长有关,可通过扩大寻优范围或减小步长来提高精确度[12]。当平均误差 *MSE* 最小时,其对应的惩罚系数(C)和松弛系数(g)便是最优值。

2.3.2 不同预处理方法的模型结果

利用网格全局寻优算法分别优化惩罚系数 C 和松弛系数 g 并建立分类模型,5 种预处理方法对应的模型预测结果如表 4 所示。从表 4 可知,预处理后,R_P 相比于原光谱都有不同程度的上升,在上述 5 种预处理方法中,SNV 的 R_P 最大,因此 SNV 联合支持向量机(SNV – SVM)建模效果最优。

表 4 不同预处理方法的 SVM 模型结果

预处理方法	支持向量机数 *nSV*	参数		建模集		验证集	
		惩罚系数 *C*	松弛系数 *g*	R_{CV}	*RMSECV*	R_P	*RMSEP*
原始光谱	45	8	0.250 0	0.983 5	0.090 7	0.493 3	0.861 5
Savitzky – Golay 平滑(SG)	42	8	0.031 2	0.852 2	0.260 8	0.596 4	0.489 3
多元散射校正(MSC)	56	0.5	0.015 6	0.904 1	0.216 1	0.626 2	0.520 6
标准状态变量变换(SNV)	60	1	0.500 0	0.993 3	0.057 9	0.761 4	0.882 1
一阶导数(FD)	34	1	2	0.762 0	0.328 3	0.601 6	0.409 3
二阶导数(SD)	34	2	8	0.788 3	0.308 5	0.603 3	0.415 4

2.3.3 SNV – SVM 模型对压榨油茶籽油与浸出油茶籽油的鉴别分析

用 SNV – SVM 模型对建模集和验证集样本进行鉴别分析,其结果如图 6、图 7 所示。

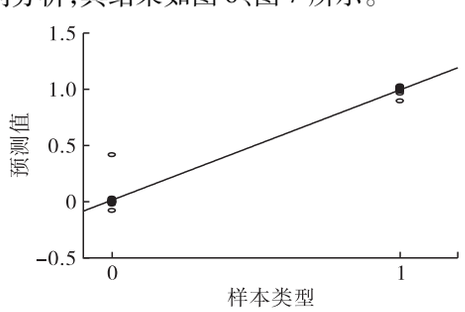


图 6 SNV – SVM 对建模集样本鉴别结果

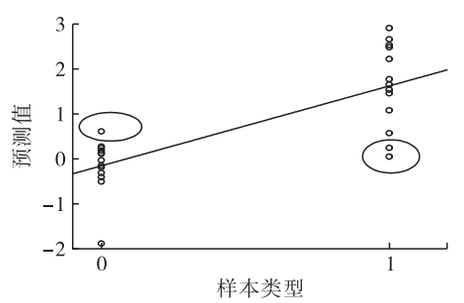


图 7 SNV – SVM 对验证集样本鉴别结果

从图 6、图 7 可以看出,SNV – SVM 模型对建模集样本类型鉴别完全正确,验证集样本有个别误判,具体结果如表 5 所示。

从表 5 可知,模型对建模集样本鉴别准确率为

100%。对验证集误判个数为 3,其中,将 1 个浸出油样本(预测值为 0.603 9)误判为压榨油样本,将 2 个压榨油样本(预测值分别为 0.042 7、0.239 2)误判为浸出油样本。SNV – SVM 模型对建模集和验证集的鉴别准确率分别为 100%、88.46%。

表 5　SNV – SVM 模型对压榨油与浸出油鉴别统计结果

样本类型	实际类别	预测结果		鉴别准确率/%	相关系数 *R*	均方根误差 *RMSE*
		0	1			
建模集	0	27	0	100	0.993 3	0.057 9
	1	0	33			
验证集	0	12	1	88.46	0.761 4	0.882 1
	1	2	11			

2.4　压榨油茶籽油与浸出油茶籽油 BPANN 鉴别模型的建立与预测

2.4.1　输入层的确定

由于采集的红外光谱数据维数高,如果将其直接作为人工神经网络的输入变量,则会导致输入层数过多,会使模型复杂而且泛化能力低[13]。同时,数据中一些无关信息也会降低模型质量。因此,采用主成分分析法(PCA)对光谱数据进行降维处理以剔除冗余信息,来提高建模速度和质量[14]。根据主成分累积贡献率确定主成分数,主成分累积贡献率见图 8。从图 8 可知,前 6 个主成分累积贡献率到达 99.33%,说明前 6 个主成分包含了原始 901 个波数中 99.33% 的信息。因此,选择最佳的主成分数 6 作为输入层数。

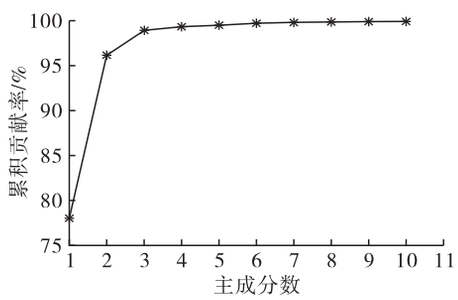


图 8　主成分的累积贡献率

2.4.2　不同预处理方法的模型结果

采用 3 层 BP 人工神经网络,输入层为 2.4.1 优选的主成分,选用 tansig 和 pureline 作为隐含层和输入层的转移函数,训练函数选用 trainlm。人工神经网络的相应参数设置:最大训练步数 1 000,目标误差 0.000 1,速率 0.05。不同预处理方法模型结果如表 6 所示。

表 6　不同预处理方法 BPANN 模型结果

预处理方法	建模集		验证集	
	相关系数 R_{CV}	均方根误差 *RMSECV*	相关系数 R_P	均方根误差 *RMSEP*
原始光谱	0.855 2	0.257 2	0.688 0	0.389 7
Savitzky – Golay 平滑(SG)	0.939 4	0.157 9	0.921 2	0.205 9
多元散射校正(MSC)	0.827 8	0.293 0	0.714 0	0.389 7
标准正态变量变换(SNV)	0.877 3	0.240 4	0.608 2	0.459 5
一阶导数(FD)	0.891 1	0.247 9	0.640 7	0.492 1
二阶导数(SD)	0.884 8	0.177 9	0.705 0	0.333 9

由表 6 可知,不同的预处理方法建模集的相关系数分布范围为 0.827 8 ~ 0.939 4,验证集的相关系数分布范围为 0.608 2 ~ 0.921 2,其中 SG 平滑数据建模效果最优,其验证集 R_P、*RMSEP* 分别为 0.921 2、0.205 9。因此,SG 平滑和主成分分析联合 BP 人工神经网络(SG – BPANN)建模效果最优。

2.4.3　SG – BPANN 模型对压榨油茶籽油与浸出油茶籽油的鉴别分析

应用 SG – BPANN 模型对建模集和验证集样本进行鉴别分析,其结果如图 9、图 10 所示。

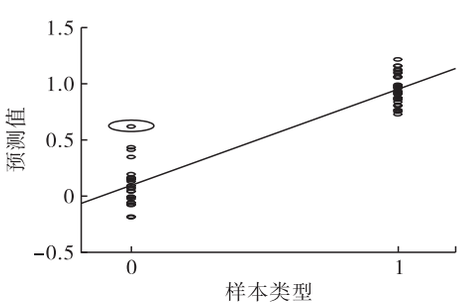


图 9　SG – BPANN 对建模集样本鉴别结果

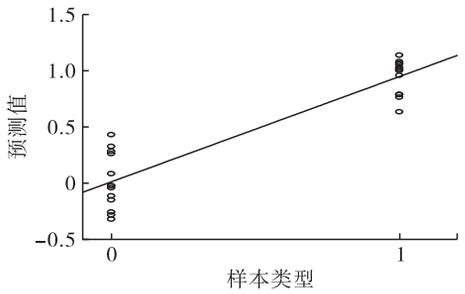


图 10 SG－BPANN 对验证集样本鉴别结果

从图 9、图 10 可以看出，该模型对压榨油茶籽油鉴别效果很好，没有出现错误。统计建模集、验证集鉴别正确和错误的个数，具体结果如表 7 所示。

从表 7 可知，模型对建模集中的 1 个浸出油样本（预测值为 0.618 5）误判为压榨油样本，验证集样本全部鉴别正确。SG－BPANN 模型对建模集和验证集的鉴别准确率分别为 98.33%、100%，该模型鉴别效果很好。

表 7 SG－BPANN 模型对压榨油与浸出油鉴别统计结果

样本类型	实际类别	预测结果		鉴别准确率/%	相关系数 *R*	均方根误差 *RMSE*
		0	1			
建模集	0	26	1	98.33	0.939 4	0.157 9
	1	0	33			
验证集	0	13	0	100	0.921 2	0.205 9
	1	0	13			

3 结 论

根据压榨油茶籽油与浸出油茶籽油在 3 007、2 924、2 852、1 747、1 462、1 377、1 163、723 cm^{-1}处的红外光谱吸收特征，分析了 SG、MSC、SNV、FD、SD 5 种不同预处理方式对 PLS、SVM 和 BPANN 3 种模型的预测精确度，优选出 SG－PLS、SNV－SVM、SG－BPANN 3种能快速鉴别压榨油茶籽油和浸出油茶籽油的模型，对验证集的鉴别准确率分别为 88.46%、88.46% 和 100%，其中 SG－BPANN 模型的验证集相关系数 R_P 最大、验证集均方根误差 *RMSEP*最小和鉴别准确率最高，对压榨油茶籽油和浸出油茶籽油的鉴别效果最好。研究结果为压榨油茶籽油和浸出油茶籽油的鉴别提供了一种快递、准确的方法，也为拓展红外光谱的应用提供了科学依据。

参考文献：

[1] 吴雪辉，黄永芳，谢治芳. 油茶籽油的保健功能作用及开发前景[J]. 食品科技，2005(8)：94－96.

[2] 刘肖丽，吴雪辉. 不同提取方法对油茶籽油品质的影响[J]. 食品工业科技，2012，33(24)：307－310.

[3] 国家卫生和计划生育委员会. 油茶籽油：GB/T 11765—2003 [S]. 北京：中国标准出版社，2003.

[4] DAI F，BERGHOLT M S，BENIAMIN A J，et al. Rapid identification of potato cultivars using NIR－excited fluorescence and Raman spectroscopy[J]. Spectrosc Spect Anal，2014，34(3)：677－680.

[5] QUINONES－LSLAS N，MEZA－MÁRQUEZ O G，OSORIO－REVILLA G，et al. Detection of adulterants in avocado oil by mid－FTIR spectroscopy and multivariate analysis[J]. Food Res Int，2013，51(1)：148－154.

[6] WOODCOCK T，DOWNEY G，O'DONNELL C P. Confirmation of declared provenance of european extra virgin olive oil samples by NIR spectroscopy[J]. J Agric Food Chem，2008，56(23)：11520－11525.

[7] INAREJOS－GARCÍA A M，GÓMEZ－ALONSO S，FREGAPANE G，et al. Evaluation of minor components，sensory characteristics and quality of virgin olive oil by near infrared (NIR) spectroscopy[J]. Food Res Int，2013，50(1)：250－258.

[8] BINETTI G，DEL C L，RAGONE R，et al. Cultivar classification of Apulian olive oils：use of artificial neural networks for comparing NMR，NIR and merceological data [J]. Food Chem，2017，219：131－138.

[9] MOSSOBA M M，AZIZIAN H，FARDIN－KIA A R，et al. First application of newly developed FT－NIR spectroscopic methodology to predict authenticity of extra virgin olive oil retail products in the USA[J]. Lipids，2017，52(5)：443－455.

[10] 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油茶籽油真伪鉴别分析[J]. 分析化学，2011，39(5)：748－752.

[11] 文韬，郑立章，龚中良，等. 基于近红外光谱技术的油茶籽油原产地快速鉴别[J]. 农业工程学报，2016，32(16)：293－299.

[12] 孟滔，周新志，雷印杰. 基于自适应遗传算法的 SVM 参数优化[J]. 计算机测量与控制，2016，24(9)：215－217.

[13] 刘宇佳，贺丽苹，张泳，等. 近红外光谱－人工神经网络的模型优化用于银耳产地识别研究[J]. 食品工业科技，2016，37(3)：261－264，269.

[14] 李仲，刘明地，吉守祥. 基于枸杞红外光谱人工神经网络的产地鉴别[J]. 光谱学与光谱分析，2016，36(3)：720－723.