基于中红外光谱和化学计量学算法鉴别核桃产地及品种

何 勇，郑启帅，张 初，岑海燕*

浙江大学生物系统工程与食品科学学院，浙江 杭州 310058

摘 要 为探究中红外光谱快速检测核桃产地和品质的可行性，基于中红外光谱分析技术，并将化学计量学的算法应用于中红外光谱判别分析之中，对中国四大核桃主产区的10类主要核桃品种进行检测，取得较好效果。通过提取核桃粉末的光谱透射率，去除原始光谱首尾部分的明显噪声，对保留的700～3 450 cm^{-1}范围的光谱采用小波分析(wavelet transform，WT)算法进行去噪预处理，并采用无信息变量消除结合连续投影算法(UVE-SPA)提取光谱特征波数，采用主成分分析法(PCA)对光谱定性分析，基于反向传播神经网络(BPNN)、极限学习机(ELM)、随机森林(RF)、径向基函数神经网络(RBFNN)及偏最小二乘判别分析(PLS-DA)对全谱和特征波数建模对比。在4类不同产地核桃判别中，得到12个特征波数：803，1 355，1 418，1 541，1 580，1 727，1 747，1 868，2 338，2 462，2 824和3 166 cm^{-1}，基于特征波数分类的正确率高于全谱的分类结果，BPNN算法结合特征波数建模得到的识别正确率高达97%，RF算法分类判别效果最差，正确率仅69.70%；在10类不同品种判别中，得到10个特征波数：903，1 275，1 507，1 541，1 563，1 671，1 868，2 311，2 845和3 437 cm^{-1}，基于特征波数分类的正确率依然高于全谱的分类结果，BPNN算法结合特征波数建模得到的识别正确率高达83.3%。在特征波数通用性方面，两组特征波数范围中有2个特征波数相同：1 541和1 868 cm^{-1}，其他大多特征波数也都相近，将10类品种特征波数作为输入变量对4类不同产地的核桃进行分类，分类结果较差，因此，在10类品种监督值下选取的特征波数无法适用于4类产地的判别问题，由此推断，即使是同一原始数据，基于不同判别问题得到的特征波数在建模时通用性较差。结果表明，经UVE-SPA算法提取特征波数后，变量数可减少99%以上，有效地简化了模型，减少计算量，提高预测的稳定性；总体上，每个分类器的表现为：BPNN＞RBFNN＞ELM＞PLS-DA＞RF；基于小波变换结合特征波数选取和反向传播神经网络算法能有效地实现核桃的产地和品种识别。

关键词 光谱分析；中红外；化学计量学；核桃；分类；特征波数

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)09-2812-06

引 言

核桃仁含有丰富的蛋白质、脂肪酸和多种微量元素，对人体有益，具有重要的营养、药用和保健功能[1]。核桃内富含的多元不饱和脂肪酸和油脂，可用于调味和烹饪，从而丰富饮食结构，对心脏保护和骨质疏松有积极影响[2-3]。经常食用核桃可降低妇女患糖尿病的风险[4]，帮助人体抗衰老，防止细胞老化，还有降低胆固醇、降低血压和抑制其他慢性病的功效[5-6]。由于产地和种类不同，核桃的品质也有差别，因此根据核桃的品质进行分级具有重要意义。核桃产地和品种的鉴别主要依赖于感官体验，甚至更复杂的过程(例如化学实验分析)，这种传统方法较主观，且费时费力。所以，找到一种快速简便的判别核桃品质的方法具有重要的意义。

红外光谱，例如近红外(near infrared，NIR)光谱和中红外(mid-infrared，MIR)光谱，是一种可以检测物质中特殊分子不同吸收频率的化学分析工具。不同的分子结构可以产生不同的吸收谱带[7]。光谱技术在食品组分分析方面有着快速、简单、灵敏等优势，而中红外光谱是由分子的基频振动组成的光谱，其吸收带多而窄，吸收强度大，有显著的吸收特性，提供了更多的频率和强度信息；而且大多典型官能团的特征振动峰大多分布于中红外区；因此，中红外光谱被广

收稿日期：2018-08-08，**修订日期**：2018-12-19

基金项目：国家“十三五”重点研发计划(2016YFD0700304)，国家重大仪器设备开发专项(2014YQ470377)，中央高校基本科研业务费专项资金项目(2017FZA5011)资助

作者简介：何 勇，1963年生，浙江大学生物系统工程与食品科学学院教授 e-mail：yhe@zju.edu.cn

*通讯联系人 e-mail：hycen@zju.edu.cn

泛应用于替代化学方法的定性和定量分析领域[8]。吴迪等[9]用中红外光谱法快速检测了奶粉中蛋白质和脂肪含量。Clegg 等[10]将中红外应用于制药，监测化学转化过程。Botelho 等[11]使用中红外光谱检测牛奶中是否有掺杂物。Vermeulen 等[7]和 Zhou 等[12]利用中红外光谱技术对酒糟的产地来源进行了分类判别以区分其品质。Kanakis 等[13]利用中红外光谱技术对薄荷的产地进行了识别，识别结果最高可达 90%。贾昌路等[14]利用近红外光谱对新疆产地不同品种的核桃进行了相对区别。不同的核桃主产区，其生长环境(温度，湿度，降雨量，光照时间等)存在很大的差异，所生产的核桃品质也会有不同，而中红外在核桃产地和品种检测方面鲜有研究。

红外光谱携带有大量信息，其中部分信息对于建立模型无相关性，甚至还会干扰建模效果，因此需要剔除无用的信息，选出相关的特征波数用于建模。本研究的主要目的是根据中红外光谱结合特征波数选择算法和化学计量算法检测核桃的产地和种类。

1　实验部分

1.1　中红外光谱的采集

从中国主要生产核桃的 4 个省份采集了 10 个核桃主产品种，每个品种重复 20 个样本。此 10 个品种的核桃肉眼区分度很小，若将其去壳后用肉眼更是无法辨识。实验前，将核桃去壳，取每个品种的核桃研磨成粉末，称量约 10 g 作为该品种的一个样本并装入干燥密封袋中。剔除 8 个无效样本，余下的 192 个样本作为实验样本，样本详细信息见表 1。所有样品放在真空干燥器中以防止吸水。实验时，将核桃粉末与溴化钾晶体按照 1∶49 的比例混合均匀，压片。红外光谱采集在 25 ℃恒温下进行。采用日本分光株式会社生产的 Jasco FTIR 4100 傅里叶变换光谱仪，测量范围为 400～4 000 cm^{-1}，分辨率为 4 cm^{-1}，每个样品扫描 32 次后取平均值作为该样品的光谱值。

表 1　样本信息

Table 1　General information of samples

Origin	Variety	Number of sample
Yunnan	No. 1：Yangbidapao	20
	No. 2：Yangbicaoguo	19
Xinjiang	No. 3：185(Hetian)	19
	No. 4：Xinfeng	20
	No. 5：Xinxin2	20
Shaanxi	No. 6：Liao4	20
	No. 7：：Xiangling	20
Hebei	No. 8：Qingxiang	16
	No. 9：Liao1	18
	No. 10：Liao8	20

1.2　特征波数选择算法

无信息变量消除结合连续投影算法(uninformative variable elimination-successive projections algorithm，UVE-SPA)是一种结合无信息变量消除法(uninformative variable elimination，UVE)和连续投影算法(successive projections algorithm，SPA)的特征变量选择算法，UVE 可去除大量的无效信息，基于 UVE 选择的变量建模可以避免模型过拟合，并提高其预测能力。SPA 主要解决共线性问题，用于选择具有最低冗余信息的波数，获得具有最小共线性的有用变量，在选取光谱特征变量中取得了广泛的应用[7, 15]。因此，结合这 2 种算法的优势，首先使用 UVE 算法，根据建模时各个变量的稳定性选取带有有效信息的变量，然后再将 UVE 选出的变量输入 SPA 算法，选择的变量数从 1 到 35 范围内变化，通过比较不同变量数对应的建模最小均方根误差(root mean square error，RMSE)，选出变量的数量。

1.3　化学计量学算法

主成分分析法(principal component analysis，PCA)是一种被广泛应用到光谱数据的定性分析方法。PCA 通过线性变换将原始光谱数据投射到一些新的主成分变量(principal components，PCs)，每一个主成分都是由原始数据线性组合而成，只需要几个方差最大的主成分即可反映数据信息，大大降低了数据维度[16]。

极限学习机(extreme learning machine，ELM)是一种单隐层前馈神经网络，亦是一种快速、简单的回归和分类方法。在 ELM 算法中，只有隐含层神经元节点数需要被设置以获得独特的最佳解决方案，通过对不同神经元节点个数下的效果比较，选择出最优解。本研究隐层中的神经元以步长 1 进行寻优，设定其数量从 1 到建模集合的上限变化，根据训练误差最小的原则得出 ELM 模型中隐藏神经元的数量[17]。

随机森林(random forests，RF)是一种使用多种决策树的综合方法。RF 构造不同的决策树，决策树相互独立。为了构建随机森林，对每个决策的样本进行随机抽样。决策树节点的特征也从训练集的特征中随机选择。基于每个决策树输出分类结果。此算法训练快速并且可调，同时无需担心要像支持向量机一样调大量参数[18]。

偏最小二乘判别分析(partial least squares discrimination analysis，PLS-DA)算法基于 PLS 回归模型对目标进行判别分析[15, 19]，PLS-DA 根据代表类别的整数建立类别关于光谱的回归模型。然而，所建立模型判别样本时得出的结果带有小数位，需要对判别结果设置阈值以确定样本属于哪个类别，设置阈值为 0.5。

反向传播神经网络(back propagation neural network，BPNN)广泛应用于回归分析和判别分析[20]。利用错误反向传播修改每个训练阶段后的内部网络权重，直到训练错误或网络的训练阶段达到目标为止[21]。采用 Matlab 自带的 Neural Network Toolbox 工具箱，判别时的判别阈值与 PLS-DA 一样，设置为 0.5，设定目标偏差为 10^{-5}，学习速率为 0.6，迭代 1 000 次。

径向基函数神经网络(radical basis function neural network，RBFNN)是另一个普遍使用的人工神经网络，与 BPNN 都是非线性多层前向网络，RBFNN 通常有 3 层：输

入层，带有非线性 RBF 激活函数的隐藏层和线性输出层。网络的输出是输入和神经元参数的径向基函数的线性组合[22]。

2 结果与讨论

2.1 实验数据预处理

由于实验仪器、环境和操作等引起的系统误差，原始光谱的首尾部分有明显噪声，最终保留 700～3 450 cm^{-1} 范围的光谱，并用小波变换对光谱数据进行平滑去噪预处理。应用小波函数 Daubechies 的正交小波基 Db3 进行光谱信号去噪，其中分解尺度为 4。图 1 为随机选取某一样本处理前后的光谱，从中可以看出平滑去噪效果明显。图 2(a)和(b)分别为 4 类产地核桃和 10 类品种核桃经预处理后的中红外光谱。可以看出，各类光谱之间存在一定的差异，并且在 1 060～1 800 cm^{-1} 波数范围内有很多明显的吸收峰，主要的峰为 1 200 cm^{-1} 附近的 C—C 振动峰，1 401 cm^{-1} N═N 振动峰，1 600 cm^{-1} 附近的—$COOH^-$ 和—NH_3^+ 振动峰，1 715 cm^{-1} 处 C═O 振动峰以及 3 000～3 300 cm^{-1} 处 ═C—H 振动的大波峰[8]。

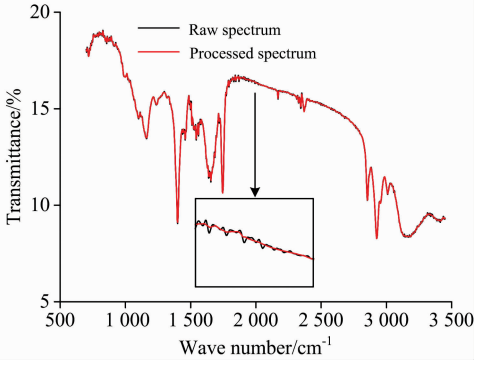

图 1 随机选取样本的原始光谱和小波变换处理后的光谱

Fig. 1 Raw mid-infrared spectrum and mid-infrared spectrum preprocessed by wavelet transform of a randomly selected sample

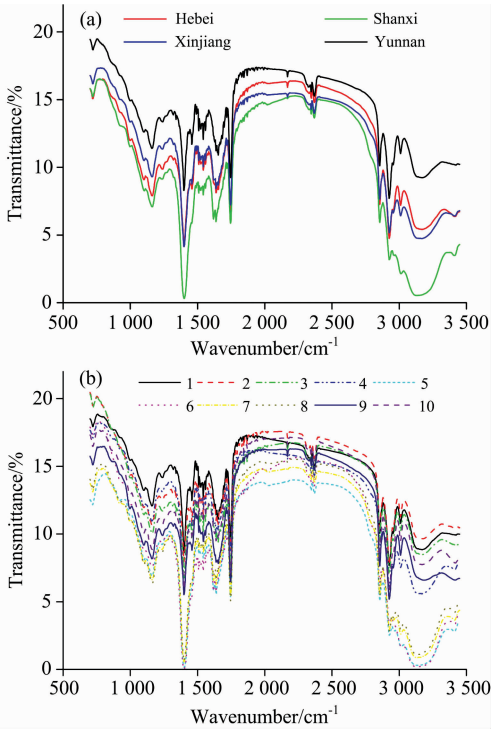

图 2 4 类产地预处理后的平均光谱(a)和 10 类品种预处理后的平均光谱(b)

Fig. 2 Average spectra after preprocessing of four origins (a) and average spectra after preprocessing of ten varieties (b)

2.2 不同产地分类

利用 PCA 对进行平滑预处理后的光谱数据分析，结果如图 3 所示。从中可以发现：PC1，PC2 和 PC3 分别解释了 86.49%，6.77%和 4.1%方差，前 3 个主成分的解释方差可达到 97.36%；云南产地与陕西产地的核桃区分明显，新疆产地的核桃与其他 3 类产地的核桃略有重合，河北产地的核桃与其他产地的核桃区分度不大。

对 4 类产地数据进行特征波数选取时，UVE 算法提取变量数为 586，SPA 建模使用变量数为 12(803，1 355，1 418，1 541，1 580，1 727，1 747，1 868，2 338，2 462，2 824和 3 166 cm^{-1})时对应的均方根误差最小为 0.341 1，图 4 为最终选出的特征波数结果。

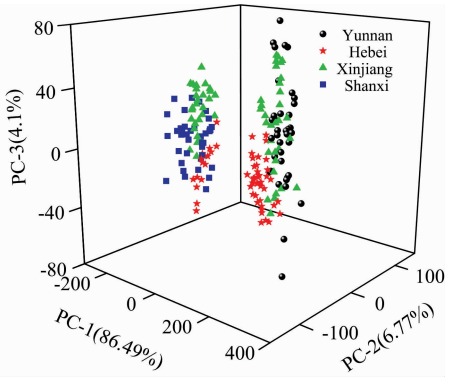

图 3 不同产地得分图

Fig. 3 Score plot for different origins

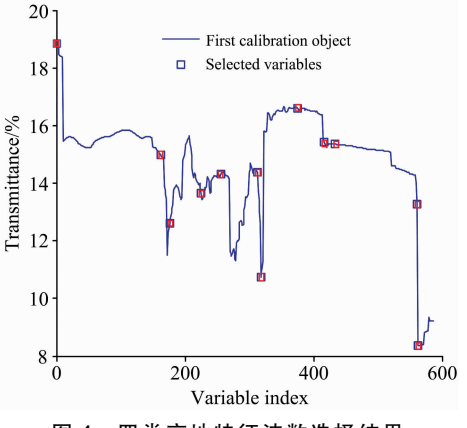

图 4 四类产地特征波数选择结果

Fig. 4 Result of characteristic variables selection of four origins

如表 2 所示，将全光谱和选取的 4 类产地特征波数作为输入量对几种分类器进行比较，校正集样本数量∶预测集样本数量＝2∶1。通常 4 类不同产地的分类难度并不大，平均正确率能达到 80%以上。从表 2 可以看出，基于特征波数分类的正确率高于全谱的分类结果。BPNN 取得了最好的效果，正确率达到 97.0%。RF 分类判别效果最差，最高正确率仅 69.70%，且对全谱和特征波数的判别没有明显差别，说明可能 RF 不适合作为核桃产地的分类器。PLS-DA 分类器对输入的特征波数表现良好，然而对全谱的判别结果较差，分析认为该分类器对噪声较为敏感。RBFNN 的识别效果较为良好，且模型比较稳定，受输入变量变化的影响较小。ELM 虽建模效果比较稳定，但是判别结果一般。

表 2　不同产地分类结果

Table 2　Classification results of different origins

Variables input	Chemometrics	Parameter	Calibration set accuracy/%					Prediction set accuracy/%				
			Yunnan	Xinjiang	Shanxi	Hebei	Overall	Yunnan	Xinjiang	Shanxi	Hebei	Overall
Full spectrum	ELM	62	73.1	79.5	96.2	54.3	74.6	84.6	70.0	100	52.6	74.2
	RF	40	100	100	100	100	100	61.5	65.0	57.1	89.5	69.7
	RBFNN	66	100	100	100	100	100	69.2	100	64.3	94.7	84.9
	PLS-DA	12	96.2	97.5	100	100	98.4	69.2	50.0	71.4	84.2	68.2
Characteristic wavenumbers	ELM	40	80.8	89.7	88.5	88.6	87.3	92.3	85.0	64.3	84.2	81.8
	RF	30	100	100	100	100	100	46.2	80.0	50.0	89.5	69.7
	RBFNN	4	100	100	100	100	100	92.3	90.0	71.4	94.7	87.9
	PLS-DA	8	92.3	97.4	92.3	97.1	95.2	84.6	95.0	92.9	100	93.9
	BPNN	8	100	100	100	100	100	100	100	85.7	100	97.0

注：Parameter 含义：ELM：学习机的节点数；RF：随机森林数；RBFNN：隐含层节点数；PLS-DA：潜在变量数；BPNN：隐藏层中的神经元数目，下同

2.3　不同品种分类

在比较 4 类不同产地判别分析后，尝试对 10 类不同品种的核桃进行分类。同样用 PCA 对 10 类数据进行分析，结果如图 5 所示。不同产地之间区分明显，而同一产地不同品种之间区分度不高，每个品种都与其他几类品种有重叠，与图 3 相同，同一产地的核桃聚集较明显，而来自新疆产地的第 3，4 和 5 个品种较分散，且与其他品种混淆，尤其第 3 类品种分散明显。

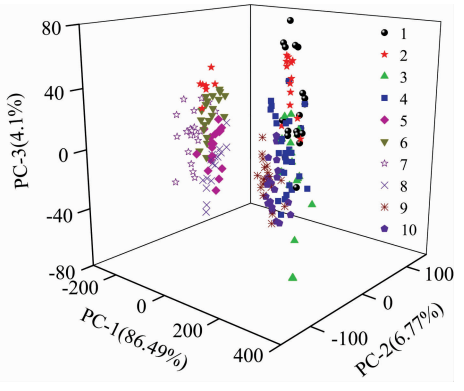

图 5　不同品种得分图

Fig. 5　Score plot of different varieties

对 10 类品种数据进行特征波数选取时，UVE 算法提取变量数为 1 316，SPA 建模使用变量数为 10(903，1 275，1 507，1 541，1 563，1 671，1 868，2 311，2 845 和3 437 cm^{-1})时对应的均方根误差最小为 0.777 2，图 6 为最终选出的特征波数结果。4 类产地识别和 10 类品种识别选取的特征波数中，有 2 个特征波数相同：1 541 和 1 868 cm^{-1}，其他大多特征波数也都相近。由于核桃主要成分是脂肪和蛋白质，查阅文献[8]可知，选出的特征波数大多落在脂肪和蛋白质的特征峰范围内。

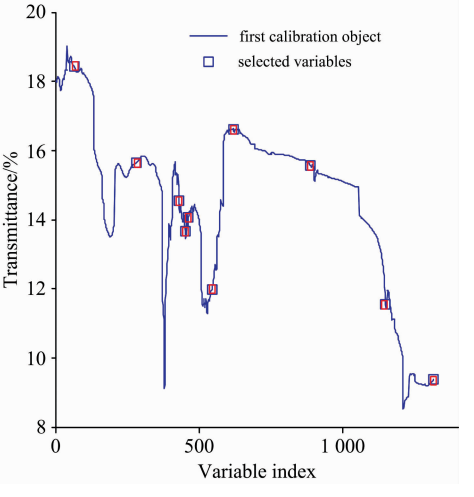

图 6　十类品种特征波数选择结果

Fig. 6　Result of characteristic variables selection of ten varieties

将校正集样本数量与预测集样本数量的比例依然设为 2∶1，利用分类器分类结果如表 3 所示。随着判别种类增多，同产地不同品种之间的区别度较小，分类难度也随之加大，只有 BPNN 的判别正确率可以达到 80%以上，RBFNN 的判别效果次于 BPNN，最高可达 68.2%，RBFNN 表现依然比较稳定，这说明 RBFNN 判别方法对输入变量的变化不敏感。ELM 与 RBFNN 效果相当，PLS-DA 与 RF 选择效果最

差。整体上，UVE-SPA 特征波数建模效果好于全光谱。

2.4 基于 10 类品种特征波数的四类产地判别

在对 4 类不同产地以及 10 类不同品种进行判别分析后，为比较因监督值不同选取的特征波数对建模效果的影响，尝试将 10 类品种特征波数作为输入变量对 4 类不同产地的核桃进行分类，判别结果如表 4 所示。

通过比较表 2 和表 4，可以明显看出，基于 4 类产地监督值选取的特征波数的判别结果总体优于基于 10 类品种监督值选取的特征波数。结合前 2 个判别部分可知，虽然在选择特征波数时监督值类别为 10，但其所选特征波数并不能应用于监督值类别为 4 的分类判别。

表 3 不同种类分类结果

Table 3 Classification results of varieties

Variables input	ELM		RF		RBFNN		PLS-DA		BPNN	
	Cal/%	Pre/%	Cal/%	Pre/%	Cal/%	Pre/%	Cal/%	Pre/%	Cal/%	Pre/%
Full spectrum	77.0	66.6	100	54.6	100	68.2	67.5	42.4	—	—
Characteristic wavenumbers	89.7	66.7	100	48.5	100	60	62.7	51.5	97.6	83.3

表 4 不同输入变量产地判别结果

Table 4 Classification results of different input variables

Classifier	Parameter	Calibration set accuracy/%					Prediction set accuracy/%				
		Yunnan	Xinjiang	Shanxi	Hebei	Overall	Yunnan	Xinjiang	Shanxi	Hebei	Overall
UVE-SPA-ELM	56	88.5	97.4	96.2	100	96.0	61.5	90.0	71.4	94.7	81.8→
UVE-SPA-RF	88	100	100	100	100	100	58.9	75.0	50.0	89.5	69.7→
UVE-SPA-RBFNN	70	100	100	100	100	100	76.9	35.0	28.6	79.0	54.6↓
UVE-SPA-PLS-DA	6	69.2	84.6	96.2	82.9	84.1	53.9	90.0	100	89.5	84.9↓
UVE-SPA-BPNN	8	92.2	97.4	92.3	100	96.0	100	100	93.3	94.7	97.0→

3 结 论

基于中红外技术鉴别核桃产地和品种，采用小波变换算法对提取出的光谱数据进行平滑去噪处理，UVE-SPA 算法提取特征波数后建模，在特征波数仅 10 和 12 个的情况下，变量数由原来的 2 853 减少了 99.6%，由模型结果显示基于特征波数建模的判别效果好于基于全谱的识别效果，表明中红外光谱结合特征波数选择方法可有效地简化模型，减少计算量。采用 UVE-SPA 算法对去噪后的光谱提取特征波数，应用化学计量学算法建模判别，结果显示，BPNN 算法的分类判别效果最优，10 类品种判别正确率可达 83.3%，四类产地判别正确率可达 97%，而对于同一原始数据，在 10 类品种监督值下选取的特征波数无法适用于四类产地的判别问题，由此推断，即使是同一原始数据，不同类别的判别效果也不一定好，在今后的研究中可以就同一物质光谱特征波数的建模共享性问题做出更深入的研究。

基于中红外光谱和化学计量学算法对中国四大核桃主产区的 10 类核桃进行了光谱检测识别。综合实验结果表明，基于小波变换结合特征波数选取和 BPNN 算法能有效地实现对核桃的产地和品种识别。后期研究可将尽可能多的核桃品种作为研究对象，探索核桃专属特征波数，建立更稳健、适用范围更广的核桃判别模型。

References

[1] Rajaram S, Haddad E H, MejiaA, et al. American Journal of Clinical Nutrition, 2009, 89(5): 1657S.

[2] Papoutsi Z, KassiE, Chinou I, et al. British Journal of Nutrition, 2008, 99(4): 715.

[3] Sze-Tao KW C, Sathe S K. Journal of the Science of Food and Agriculture, 2000, 80(9): 1393.

[4] Pan A, Sun Q, Manson J A E, et al. Journal of Nutrition, 2013, 143(4): 512.

[5] Regueiro J, Sánchezgonzález C, Vallverdúqueralt A, et al. Food Chemistry, 2014, 152: 340.

[6] Banel D K, Hu F B. American Journal of Clinical Nutrition, 2009, 90(1): 56.

[7] Vermeulen P, Pierna J A F, Abbas O, et al. Food Chemistry, 2015, 189: 19.

[8] HE Yong, LIU Fei, LI Xiao-li, et al(何 勇, 刘 飞, 李晓丽, 等). Spectroscopy and Imaging Technology in Ariculture(光谱及成像技术在农业中的应用). Beijing: Science Press(北京: 科学出版社), 2016. 27.

[9] WU Di, HE Yong, FENG Shui-juan, et al(吴 迪, 何 勇, 冯水娟, 等). Journal of Infrared and Millimeter Waves(红外与毫米波学报), 2008, 27(3): 180.

[10] Clegg I M, Daly AM, Donnelly C, et al. Applied Spectroscopy, 2012, 66(5): 574.

[11] Botelho B G, Reis N, Oliveira L S, et al. Food Chemistry, 2015, 181: 31.

[12] Zhou X F, Yang Z L, Haughey S A, et al. Food Chemistry, 2014, 189: 13.
[13] Kanakis, C D, Petrakis E A, Kimbaris A C, et al. Phytochemical Analysis, 2012, 23(1): 34.
[14] JIA Chang-lu, GAO Shan, ZHANG Hong, et al(贾昌禄, 高 山, 张 宏, 等). Hubei Agricultural Sciences(湖北农业科学), 2016, 55(10): 2560.
[15] Luna A S, Da S A, Pinho J S, et al. Spectrochimica Acta Part A Molecular & Biomolecular Spectroscopy, 2013, 100(12): 115.
[16] Dong W, Ni Y, Kokot S. Journal of Agricultural and Food Chemistry, 2013, 61(3): 540.
[17] Ding S F, Zhao H, Zhang Y N, et al. Artificial Intelligence Review, 2015, 44(1): 103.
[18] Zhang C, Ma Y. Random Forests//Ensemble Machine Learning. Boston, MA, USA: Springer, 2012. 157.
[19] Wang L, Lee F S C, Wang X, et al. Food Chemistry, 2006, 95(3): 529.
[20] Gong A P, Qiu Z J, He Y, et al. Spectrochim Acta A Mol. Biomol. Spectrosc., 2012, 99(99C): 7.
[21] Wythoff B J. Chemometrics and Intelligent Laboratory Systems, 1993, 18(2): 115.
[22] Mohammadi R, Ghomi S M T F, Zeinali F. Engineering Applications of Artificial Intelligence, 2014, 36: 204.

Identification of Walnut Origins and Varieties with Mid-Infrared Spectroscopy Analysis Technique

HE Yong, ZHENG Qi-shuai, ZHANG Chu, CEN Hai-yan*
College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China

Abstract To explore the feasibility of rapid detection of the origin and quality of walnut by using mid-infrared spectroscopy, mid-infrared spectroscopy and chemometrics algorithms were used to classify walnuts of ten varieties from four major origins and finally good results were achieved. After extracting the transmittance spectra of walnut powder, the apparent noise was removed in the head and the tail of the original spectrum, and the remaining spectrum of 700～3 450 cm^{-1} was denoised by wavelet transform (WT) algorithm. The spectral characteristic wavenumber was extracted by uninformative variable elimination combined withsuccessive projections algorithm (UVE-SPA). Qualitative analysis of the spectrum was performed by principal component analysis (PCA). Back propagation neural network (BPNN), extreme learning machine (ELM), random forests (RF), radial basis function neural network (RBFNN) and partial least squares discrimination analysis (PLS-DA) were used for modeling based on the full spectrum and characteristic wavenumbers. For the discrimination of four different origins, 12 characteristic wavenumbers were selected: 803, 1 355, 1 418, 1 541, 1 580, 1 727, 1 747, 1 868, 2 338, 2 462, 2 824, and 3 166 cm^{-1}, the discrimination accuracy of characteristic wavenumbers was much higher than that of full spectrum, and the accuracy of BPNN algorithm combined with characteristic wavenumbers reached 97%. The result of RF algorithm was the worst, and the accuracy was only 69.70%. For the discrimination of ten varieties, 10 characteristic wavenumbers were selected: 903, 1 275, 1 507, 1 541, 1 563, 1 671, 1 868, 2 311, 2 845, 3 437 cm^{-1}, the discrimination accuracy of characteristic wavenumbers was still much higher than that of full spectrum. The accuracy of BPNN algorithm combined with characteristic wavenumbersreached 83.3%. In terms of the versatility of characteristic wavenumbers, there were two same characteristic wavenumbers in the two sets of characteristic wavenumbers: 1 541 and 1 868 cm^{-1}, and most of the other characteristic wavenumbers were similar. The spectra based on characteristic wavenumbers of 10 varieties were used as input variables to discriminate walnuts' origins, and the result was poor. Therefore, the characteristic wavenumbers selected under the supervisory value of 10 varieties could not be applied to discriminate 4 types of producing origins. Even with the same original data, characteristic wavenumbers selected based on different discriminant problems were less versatile in modeling. After extracting the characteristic wavenumbers by UVE-SPA algorithm, the discrimination results showed that the number of variables can be reduced by more than 99%, which effectively simplified the model, reduced the amount of calculation, and improved the stability of prediction. In general, the performance of each classifier is: BPNN>RBFNN>ELM>PLS-DA>RF. The experimental results showed that the identification of walnut origins and varieties can be realized effectively based on wavelet transform, characteristic wavenumber selection and back propagation neural network algorithm.

Keywords Spectral analysis; Mid-infrared; Chemometrics; Walnut; Classification; Characteristic wavenumber

* Corresponding author
(Received Aug. 8, 2018; accepted Dec. 19, 2018)