

利用近红外高光谱图像技术快速鉴别西瓜种子品种

张 初¹, 刘 飞¹, 孔汶汶¹, 章海亮^{1,2}, 何 勇^{1*}

(1. 浙江大学生物系统工程与食品科学学院, 杭州 310058; 2. 华东交通大学机电工程学院, 南昌 330013)

摘 要: 为了研究采用近红外高光谱图像技术对西瓜种子品种快速无损鉴别的可行性, 该文采用近红外高光谱图像技术, 通过提取西瓜种子的光谱反射率, 结合 Savitzky-Golay (SG)平滑算法, 经验模态分解算法 (empirical mode decomposition, EMD) 和小波分析 (wavelet transform, WT) 对提取出的光谱数据进行去除噪声处理, 采用连续投影算法 (successive projections algorithm, SPA) 和遗传-偏最小二乘法 (genetic algorithm-partial least squares, GA-PLS) 进行特征波长选择。基于全波段光谱建立了偏最小二乘判别分析 (partial least squares-discriminant analysis, PLS-DA), 基于特征波长建立了反向传播神经网络 (back-propagation neural network, BP NN) 判别模型和极限学习机 (extreme learning machine, ELM) 判别模型。试验结果表明, 基于特征波长的 BPNN 模型和 ELM 模型的结果优于基于全部波长的 PLS-DA 模型, 基于 SG 预处理光谱提取的特征波长建立的 ELM 模型取得最优的判别效果, 建模集和预测集的判别正确率均为 100%。结果表明应用近红外高光谱成像技术对西瓜种子品种鉴别是可行的, 为西瓜种子的品种快速鉴别提供了一种新方法。

关键词: 近红外光谱, 遗传算法, 图像处理, 西瓜种子, 连续投影算法, 极限学习机

doi: 10.3969/j.issn.1002-6819.2013.20.035

中图分类号: O657.33; S651

文献标志码: A

文章编号: 1002-6819(2013)-20-0270-08

张 初, 刘 飞, 孔汶汶, 等. 利用近红外高光谱图像技术快速鉴别西瓜种子品种[J]. 农业工程学报, 2013, 29(20): 270—277.

Zhang Chu, Liu Fei, Kong Wenwen, et al. Fast identification of watermelon seed variety using near infrared hyperspectral imaging technology[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2013, 29(20): 270—277. (in Chinese with English abstract)

0 引 言

作为广受欢迎的夏季水果, 西瓜的品质直接关系到果农和消费者的切身利益。西瓜的品质包含西瓜的糖度, 风味、质地等^[1]。影响西瓜品质的因素包括西瓜品种, 种子的质量, 环境因素和营养因素等, 而在其中, 选择品种合适并且高品质的种子显得尤为重要。目前对西瓜品质的研究仅限于西瓜生长过程中的内在品质的检测, 而对西瓜种子研究较少。市场上西瓜种子的品种很多, 适应不同的地区和不同的生长环境, 且品质不一, 容易造成种子假冒的问题。对西瓜的品种进行鉴别, 有助于防止买到质次价高、假冒伪劣、不适于当地种植的西瓜种子, 保护西瓜种子生产商以及西瓜种植者的利益。

当前, 对西瓜种子的识别研究也仅仅体现在新旧种子的研究^[2], 以及对种子纯度的研究^[3]。以上研究采用观察西瓜种子外观变化、化学方法检测内部品质变化以及种子发芽试验来实现对西瓜品种的识别, 这些方法要么主观性强, 要么破坏种子、耗时费力、成本较高, 无法做到对西瓜种子品种的快速无损检测。因此, 需要研究一种快速无损检测技术对西瓜种子品种进行识别。

光谱与光谱成像技术作为近年来广泛研究的快速无损检测方式, 已经被广泛的应用于种子的鉴别以及品质分析^[4-9]。近红外光谱分析只能获得样本的光谱信息, 而无法得到样本的空间信息, 而高光谱成像技术同时获取研究对象的光谱信息和空间信息, 能够在更大范围获取样本的内外部信息, 从而更全面的对研究对象进行分析^[10-11]。高光谱成像技术可以获取种子外部形态和内部成分信息, 有助于对种子进行识别研究。目前, 尚未有采用近红外光谱技术或近红外高光谱成像技术对西瓜种子品种鉴别的研究。本文的主要目的是研究采用近红外高光谱成像技术对西瓜种子品种鉴别的可行性, 并利用化学计量学方法建立相应的鉴别模型。

收稿日期: 2013-05-05 修订日期: 2013-08-31

基金项目: 国家 863 课题资助 (2013AA102301)、十二五支撑计划项目资助 (2011BAD21B04)

作者简介: 张 初 (1987—), 男, 湖北洪湖人, 博士生, 主要从事数字农业与光谱检测研究, 杭州 浙江大学生物系统工程与食品科学学院, 310058, Email: chuzh@zju.edu.cn

*通信作者: 何 勇 (1963—), 男, 教授, 博士生导师, 主要从事数字农业, 3S 技术与农业物联网方面研究。杭州 浙江大学生物系统工程与食品科学学院, 310058。Email: yhe@zju.edu.cn

1 材料与方 法

1.1 材 料

试验用西瓜种子购买于浙江勿忘农种子公 司，西瓜种子的品种有小芳、秀芳、丽芳、浙蜜 5 号 4 个品种。将购买的种子按照品种装在直径 10 cm，厚度 1 cm 的培养皿中，将西瓜种子装满培养皿并 保持培养皿中的种子与培养皿上表面在同一水平 面上，且由上而下轻轻挤压培养皿中的种子，使种 子间空隙较小。4 个品种分别获得 32、25、32 和 32 个样本，总共 121 个样本。图 1 所示为 4 个品种 西瓜种子的近红外高光谱图像（显示为 1 000、1 200 与 1 400 nm 3 个波段下的图像合成的伪彩图）。

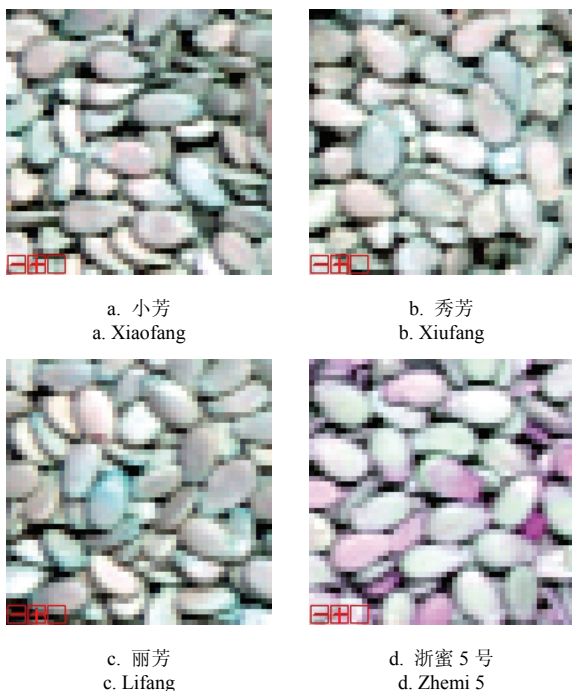


图 1 4 种西瓜种子的近红外高光谱图像 (1 000、1 200 与 1 400 nm 3 个波段下)

Fig.1 Near-infrared hyperspectral image of four cultivars of watermelon seeds (at waveband 1 000, 1 200 and 1 400 nm)

1.2 高光谱图像采集

试验采用地面高光谱成像系统实现对西瓜种 子信息采集。本试验采集系统的系统装置图如图 2 所示。

本系统主要包括 N17E-QE 成像光谱仪 (Spectral Imaging Ltd. Oulu, Finland)、C-mount 成 像镜头 OLES22(Specim, Spectral Imaging Ltd., Oulu, Finland)、Fiber-Lite DC950 线光源 (Dolan Jenner Industries Inc., USA)，IRCP0076 型电控移位平台 (Isuzu Optics Corp, 中国台湾)，计算机以及一个 暗箱组成。光谱分辨率为 5 nm。在采集高光谱图像 时，关闭暗箱以避免外部光线的干扰。

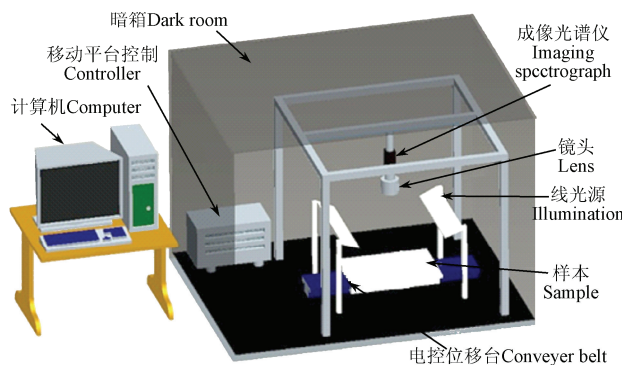


图 2 高光谱成像系统
Fig.2 Hyperspectral imaging system

在进行高光谱图像采集时，需要设置相机曝光 时间，平台移动速度以及物镜之间的距离。这 3 个 参数相互影响，图像调节的目的是使采集的图像大 小合适，清晰，不变形失真。经过反复尝试，物镜 高度设置为 31 cm，曝光时间设置为 3 000 μs，平台 移动速度设置为 30 mm/s。近红外高光谱图像分辨 率为 320×256 像素点。图像采集软件采用台湾五铃 光学公司提供的高光谱成像系统采集软件完成。图 像处理采用 ENVI4.6 软件进行处理。在进行图像处 理之前，先要对采集的光谱图像进行图像校正，图 像校正公式如下

$$R = \frac{I_{raw} - I_{dark}}{I_{white} - I_{dark}} \tag{1}$$

式中， R 是校正过的图像， I_{raw} 是原始图像， I_{white} 为白板校正图像， I_{dark} 是黑板校正图像。高光谱图 像的光谱与图像之间有着对应的关系，据此，在 西瓜种子样本上选取 15×15 像素点的感兴趣区域， 以感兴趣区域所有像素的光谱平均值作为该样本 的平均光谱。

1.3 光谱噪声去除

试验得到光谱含有由仪器和试验条件等引起 的噪声，对这些噪声的处理有助于减少噪声对光谱 分析的影响，突出光谱的有效信息。本文采用 SG 平滑算法，EMD 算法和 WT 去噪算法等对光谱进 行处理，并对 3 种去噪算法进行比较。

SG 平滑算法^[12]可以有效消减光谱数据中的随 机噪声，消噪效果受平滑点数的影响，本文中 选择 SG 二次多项式 7 点平滑对光谱数据进行处理。

EMD 算法^[13-14]依据信号自身时间尺度特征 将信号分解成一组本征模态函数 (intrinsic mode function, IMF)，通过对每一个 IMF 设置硬 (或软) 阈值门限对 IMF 进行截断处理以达到消除噪声 的目的。本文采用硬阈值法去噪，阈值的计算依 据 Donoho 等的理论^[15]。不同样本的光谱曲线计算 出的阈值不同。去噪后将所有 IMF 分量和余量求 和得

到重构后的光谱曲线。

小波分析算法^[16-17]中小波基函数和分解尺度的选择,导致不同的信号去噪效果。本文选择小波函数 Daubechies 的正交小波基 Db4 和分解尺度 7 对光谱信号去噪,采用软阈值方法去噪。

1.4 特征波长选择

光谱信息之间存在大量的冗余和共线性信息特征,对光谱有效信息的提取产生了较大的干扰,且大量光谱数据造成模型复杂、计算量大的问题。本文采用连续投影算法 (successive projections algorithm, SPA) 和遗传-偏最小二乘法 (genetic algorithm-partial least squares, GA-PLS) 进行特征波长的选择,以减少信息冗余和共线性的影响,简化模型,减少计算量。

SPA 是一种特征变量前向选择算法,在光谱特征波长中取得了广泛的应用^[18-19]。本文采用 SPA 算法对不同去噪处理的光谱进行特征波长选择。进行 SPA 计算时,以建模集样本的光谱数据和类别赋值为输入,设置选择特征波长数的范围为 5~30。

GA-PLS 算法是一种结合遗传算法 (GA) 和 PLS 算法的特征变量选择算法^[20-21]。运行 GA-PLS 算法的时候,种群大小设置为 30,交叉概率设置为 0.5,变异概率为 0.01,迭代次数为 100 次,通过运行 GA-PLS 算法 100 次,以选择频率最高的波长为特征波长。

1.5 判别分析方法

本文基于全谱建立 PLS-DA 判别分析模型,分别基于特征波长建立 BPNN 以及 ELM 判别分析模型。PLS-DA 算法是基于 PLS 回归模型建立的判别分析算法^[22-23],通过建立光谱数据与类别特征之间的回归模型,进行判别分析。回归模型得到的样本的预测值不是整数,需要设置阈值以判断样本的归属。本文中阈值设置为 0.5^[23],即预测值与实际值之差的绝对值小于 0.5,则判别正确,反之,则判别错误。

BP 神经网络在回归分析和判别分析中具有广泛的应用^[24-26]。本文中,BP 神经网络为单隐含层的 3 层神经网络,其中输入层神经元个数由输入变量的个数决定,而输出层则只有 1 个神经元,隐含层的神经元则经过不断尝试确定。BP 神经网络采用 Matlab 自带的 Neural Network Toolbox 工具箱,其中学习速率为 0.6,迭代次数为 1 000,目标偏差为 10^{-5} ,其他设置为工具箱默认设置,隐含层激活函数为 tansig,输出层激活函数为 purelin。本文中,BP 神经网络判别时的判别阈值与 PLS-DA 一样,设置为 0.5。

ELM 算法^[27-29]中,只需设置隐含层神经元节点

数,通过对不同神经元节点个数下的效果比较,便可获得最优解。本文中隐含层神经元个数从 1 到 83 以步长 1 进行寻优,以最小训练误差下的神经元个数为 ELM 模型隐含层神经元个数。

2 结果与分析

2.1 西瓜种子的原始光谱曲线

本试验采集的是 874~1 734 nm 波长范围的 256 个波段的近红外光谱数据,原始光谱图如图 3 所示。

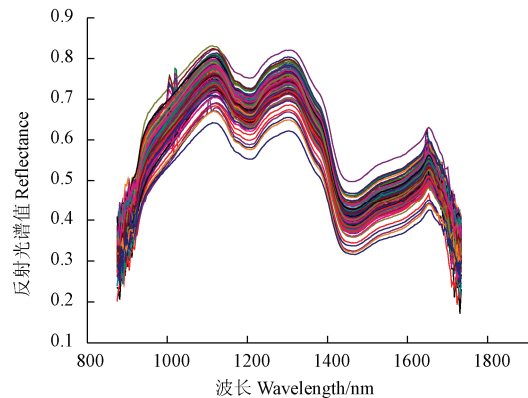


图 3 全部西瓜种子样本原始反射光谱图

Fig.3 Raw reflectance spectra of all watermelon seeds samples

可以看到,在光谱前端和后端有较明显的噪声,因此切去前端和后端噪声明显的部分,采用波段 51 到波段 230 (1 042~1 646 nm) 之间的光谱进行分析。同时,在图中可以看出,4 个品种的西瓜种子的光谱曲线没有十分明显的差异。按照 Kennard-Stone 算法^[30]将样本分成建模集和预测集,其中建模集 83 个样本,预测集 38 个样本。4 个品种西瓜种子分别赋值为 1、2、3、4 (表 1),不同品种西瓜种子建模集和预测集样本的划分如表 1 所示。

表 1 不同品种西瓜种子类别赋值与建模集和预测集样本划分

Table 1 Class assignment and division of calibration set and prediction set of different cultivars of watermelon seeds

	小芳 Xiaofang	秀芳 Xiufang	丽芳 Lifang	浙蜜 5 号 Zhemi 5
类别赋值 Assigned class	1	2	3	4
建模集 Calibration set	22	17	22	22
预测集 Prediction set	10	8	10	10

2.2 主成分分析定性分析

对西瓜种子光谱数据进行主成分分析 (principal component analysis, PCA) 可知,第一主成分 (principle component 1, PC1) 的贡献率为 93.38%, PC2 的贡献率为 5.55%, PC1 和 PC2 累计贡献率为

98.93%，PC1 和 PC2 能够解释绝大部分的变量。

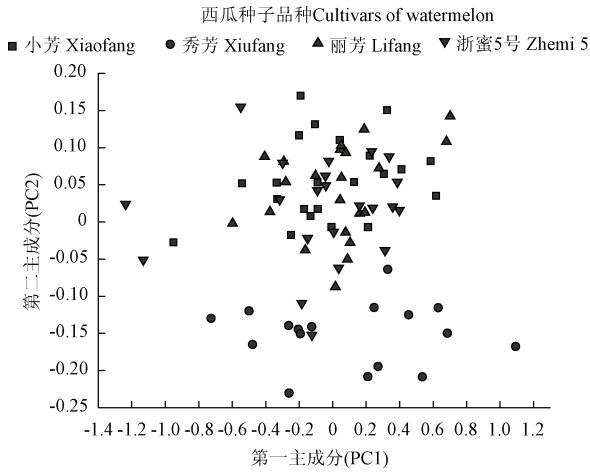


图 4 第一主成分 (PC1) 和第二主成分 (PC2) 的得分分布图

Fig.4 Scores scatter plot of PC1 and PC2

由图 4 可知，各个西瓜种子品种之间较难直接区分出来，得分图中均有重合的地方。因此需要对光谱数据进行进一步的分析处理，以鉴别不同品种的西瓜种子。

2.3 光谱噪声去除

以小芳品种中的 1 号样本的光谱曲线为例，EMD 分解的结果如图 5 所示。其中 *IMF1*、*IMF2*、*IMF3* 为信号分解后的本征模态函数 (intrinsic mode function) 分量，*r* 为光谱信号的余量。*IMF1* 表现为含噪声的高频分量。对 *IMF1* 计算得出阈值为 0.0014，对 *IMF1* 进行硬阈值去噪处理之后，将所有 *IMF* 分量和 *r* 求和得到重构后的光谱曲线。EMD 分解采用的是极值对称延拓法抑制边缘效应。

小芳品种中的 1 号样本的原始光谱曲线以及分别经 SG7 点平滑去噪、EMD 去噪以及 WT 去噪后的光谱曲线如图 6 所示。

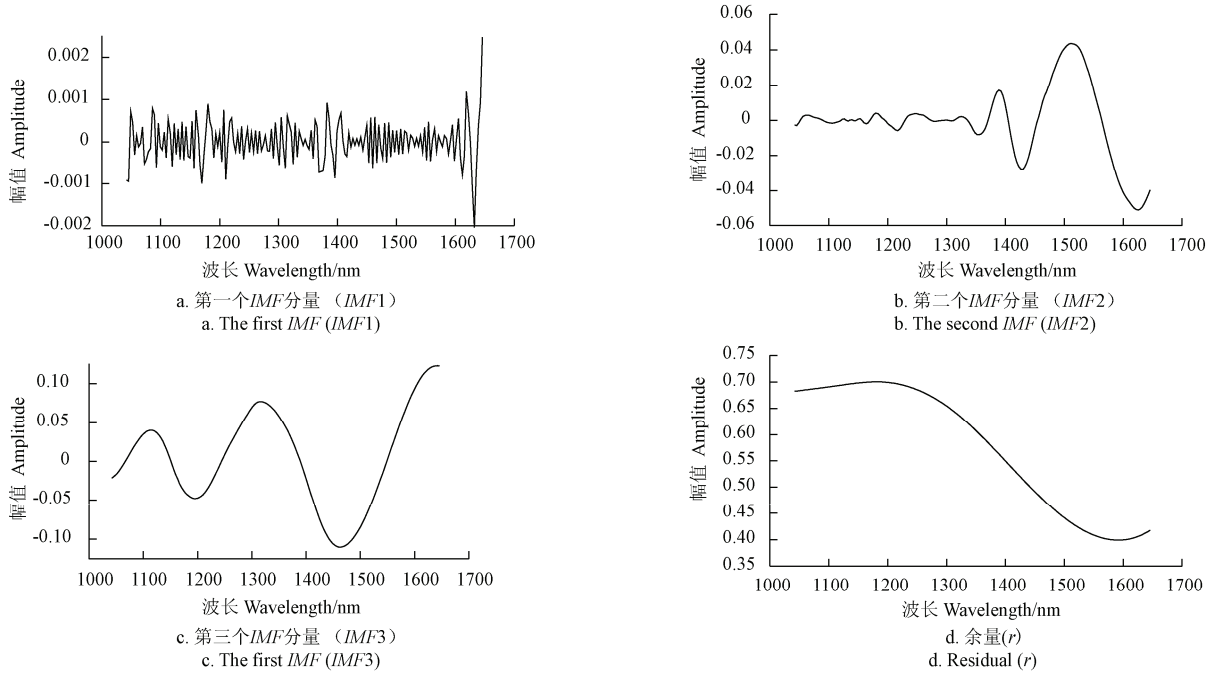
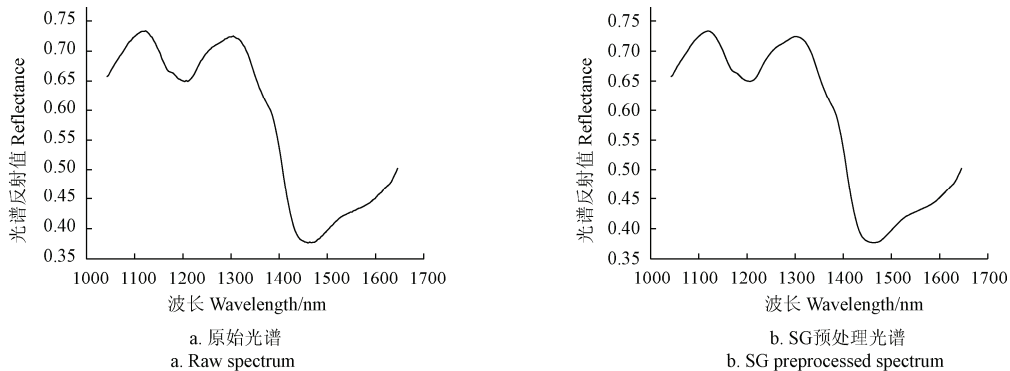


图 5 EMD 分解后的信号(小芳品种中的 1 号样本)

Fig.5 Decomposition signal of EMD (sample No.1 of cultivar Xiaofang)



a. 原始光谱
a. Raw spectrum

b. SG 预处理光谱
b. SG preprocessed spectrum

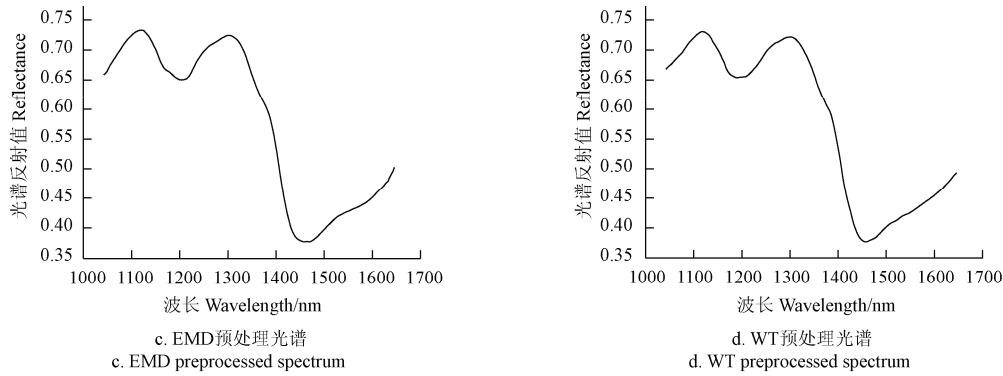


图 6 小芳品种中 1 号样本的原始光谱曲线、SG 预处理光谱曲线、EMD 预处理光谱曲线和 WT 预处理光谱曲线
Fig.6 Raw spectrum, SG preprocessed spectrum, EMD preprocessed spectrum and WT preprocessed spectrum of sample No.1 of cultivar Xiaofang

2.4 基于全波段光谱的 PLS-DA 判别模型

对经过 3 种去噪方法处理的光谱进行 PLS-DA 判别分析的结果如表 2 所示。

表 2 PLS-DA 判别分析结果
Table 2 Discriminant results of PLS-DA models

预处理 Pretreatments	主成分 PCs	建模集 Calibration set		预测集 Prediction set	
		识别数 Correct number	识别率 Identification rate/%	识别数 Correct number	识别率 Identification rate/%
SG	11	74	89.16	30	78.95
EMD	11	76	91.57	30	78.95
WT	11	66	79.52	23	60.53

注: SG (Savitzky-golay)为 SG 平滑; EMD(empirical mode decomposition)为经验模态分解; WT(wavelet transform)为小波分析, 下同。

由表 2 可知, 基于 EMD 去噪后的光谱识别效果最好, 建模集识别正确率和预测集识别正确率均最高。基于 EMD 去噪后的光谱取得了最佳的 PLS-DA 判别分析结果, 但是预测集的识别正确率低于 80%。小波分析虽然有效去除了光谱中的部分噪声, 但也消除了部分的有效信息, 导致 PLS-DA 判别效果最低。

表 4 BP 神经网络判别分析结果

Table 4 Discriminant results of BPNN models

预处理 Pretreatment	节点数 Node number	SPA				GA-PLS			
		建模集 Calibration set		预测集 Prediction set		建模集 Calibration set		预测集 Prediction set	
		识别数 Correct number	识别率 Identification rate/%	识别数 Correct number	识别率 Identification rate/%	识别数 Correct number	识别率 Identification rate/%	识别数 Correct number	识别率 Identification rate/%
SG	8	72	86.75	32	84.21	77	92.77	33	86.84
EMD	6	63	75.90	29	76.32	74	89.16	33	86.84
WT	6	71	85.54	30	78.95	71	85.54	30	78.95

从表 4 可知, 采用 GA-PLS 提取的基于 SG 预处理光谱的特征波长建立的 BP 神经网络识别模型的识别效果最好, 建模集识别正确率为 92.77%, 预测集识别正确率为 86.84%。采用 SPA 选择的基于 EMD 预处理光谱的特征波长取得了最差的识别效果, 建模集和预测集的识别正确率均低于 80%。比

2.5 特征波长选择

本文以建模集样本的光谱数据和类别赋值为输入, 分别基于 SPA 和 GA-PLS 算法选择特征波长。选出的特征波长的个数如表 3 所示。从表 3 可知, GA-PLS 算法所选择的特征波长的个数要多于 SPA 算法, 基于 SG 预处理后的光谱挑选的特征波长个数最多, 而基于 WT 预处理后的光谱选择的特征波长最少。

表 3 SPA 算法和 GA-PLS 算法选择的特征波长个数

Table 3 Number of wavelengths selected by SPA and GA-PLS

特征波长选择方法 Methods	SG	EMD	WT
SPA	24	16	15
GA-PLS	38	33	32

注: SPA(successive projections algorithm)为连续投影算法; GA-PLS (genetic algorithm-partial least squares)为遗传-偏最小二乘法, 下同。

2.6 基于特征波长的 BP 神经网络模型

基于特征波长的 BP 神经网络模型的计算结果如表 4 所示。

较分别基于 SPA 和 GA-PLS 算法提取的特征波长建立的 BP 神经网络模型可知, 基于 GA-PLS 提取特征波长的 BP 神经网络的整体识别效果要优于基于 SPA 提取特征波长的 BP 神经网络, 且基于 WT 预处理的光谱在 2 种情况下取得了同样的识别正确率。单独比较不同预处理方法对 BP 神经网络模型

识别效果影响可知, 基于 SG 预处理的光谱取得了最佳的识别效果, 而基于 EMD 预处理光谱和 WT 预处理光谱的识别效果因为特征波长选择方法的不同而不同。

2.7 基于特征波长的 ELM 判别模型

ELM 模型最优的隐含层神经元个数和最优的判别效果见表 5。从表 5 的结果可知, 采用 SPA 提取的基于 SG 预处理光谱和 WT 预处理光谱的特征波长建立的 ELM 模型以及采用 GA-PLS 提取的基

于 SG 预处理光谱的特征波长建立的 ELM 模型的建模集和预测集均取得了 100% 的识别正确率, 采用 GA-PLS 提取基于 EMD 预处理光谱的特征波长建立的 ELM 模型取得了最差的识别效果, 但建模集和预测集识别正确率仍高于 90%。基于 SPA 的 ELM 模型的识别效果要优于基于 GA-PLS 的 ELM 模型。对比不同预处理方法可以看出, 基于 SG 预处理光谱提取的特征波长建立的 ELM 模型取得了最佳的识别效果。

表 5 ELM 判别分析结果
Table 5 Discriminant results of ELM models

预处理 Pretreatments	节点数 Node number	SPA				GA-PLS				
		建模集 Calibration set		预测集 Prediction set		建模集 Calibration set		预测集 Prediction set		
		识别数 Correct number	识别率 Identification rate/%	识别数 Correct number	识别率 Identification rate/%	识别数 Correct number	识别率 Identification rate/%	识别数 Correct number	识别率 Identification rate/%	
SG	22	83	100	38	100	18	83	100	38	100
EMD	20	83	100	36	94.74	26	83	100	35	92.11
WT	27	83	100	38	100	19	81	97.59	38	100

2.8 PLS-DA 模型, BP 神经网络模型和 ELM 模型 的比较

综合对比 PLS-DA 模型, BP 神经网络模型和 ELM 模型的识别效果可知, 除了采用基于 SPA 选择的 EMD 预处理光谱的特征波长建立的 BP 神经网络模型外, PLS-DA 模型的识别效果最差, ELM 模型的识别效果最优。在所有模型中, 基于 SG 预处理的光谱建立的模型取得了最优的效果。基于 EMD 预处理光谱建立的所有模型中, 仅 PLS-DA 模型效果在所有的 PLS-DA 模型中较优, 而 BP 神经网络模型和 ELM 模型的效果都较差。所有的 ELM 模型都获得了超过 90% 的识别正确率, 基于 SG 预处理光谱的 ELM 模型的建模集和预测集识别正确率均为 100%。ELM 取得最优的效果可能是因为 ELM 继承了单隐含层前馈神经网络的优点, 且 ELM 只含有 1 个隐含层, 在计算过程中只需要选择隐含层激励函数和隐含层节点数而不需要调整输入权值和隐元的偏置, 训练过程简单, 能获得全局最优解且泛化能力强。试验结果表明, 基于 SG 平滑算法和 ELM 算法能有效的实现西瓜种子的品种识别。

3 结论与讨论

采用高光谱成像技术鉴别西瓜种子的品种, 分别采用 SG 平滑算法、EMD 算法和 WT 算法对提取出的光谱数据进行去噪处理, 同时采用 SPA 算法和 GA-PLS 算法对去噪后的光谱提取特征波长, 并分别基于全谱建立 PLS-DA 判别模型和基于特征波长建立 BP 神经网络模型和 ELM 模型, 取得了比较好的识别效果。基于 SG 平滑算法预处理的光谱提取

出的特征波长建立的 ELM 模型取得了最佳效果, 建模集和预测集识别正确率均达到 100%。在今后的研究中将考虑研究尽可能多的西瓜品种, 建立更为稳健以及适用范围更广的西瓜种子品种判别模型, 并研究高光谱成像技术在西瓜品质检测中的应用的可行性及效果。

[参 考 文 献]

- [1] 张帆, 宫国义, 王倩, 等. 西瓜品质构成分析[J]. 果树学报, 2006, 23(2): 266—269.
Zhang Fan, Gong Guoyi, Wang Qian, et al. Analysis of watermelon quality structure[J]. Journal of Fruit Science, 2006, 23(2): 266—269. (in Chinese with English abstract)
- [2] 杨宝成. 新旧西瓜种子鉴别方法[J]. 农业知识 (瓜果菜), 2011(2): 29.
- [3] 周先林, 樊丹, 张淑霞, 等. 西瓜杂交种子纯度鉴定方法研究进展[J]. 中国瓜菜, 2011, 24(5): 46—49.
Zhou Xianlin, Fan Dan, Zhang Shuxia, et al. Advances of watermelon hybrid purity test[J]. China Cucurbits and Vegetables, 2011, 24(5): 46—49. (in Chinese with English abstract)
- [4] 朱大洲, 王坤, 周光华, 等. 单粒大豆的近红外光谱特征及品种鉴别研究[J]. 光谱学与光谱分析, 2010, 30(12): 3217—3221.
Zhu Dazhou, Wang Kun, Zhou Guanghua. The NIR spectra based variety discrimination for single soybean seed[J]. Spectroscopy and Spectral Analysis, 2010, 30(12): 3217—3221. (in Chinese with English abstract)
- [5] 刘天玲, 苏琪雅, 孙群, 等. 基于 NIR 分析和模式识别技术的玉米种子识别系统[J]. 光谱学与光谱分析, 2012, 32(6): 1550—1553.
Liu Tianling, Su Qiya, Sun Qun, et al. Recognition of corn seeds based on pattern recognition and near infrared spectroscopy technology[J]. Spectroscopy and Spectral Analysis, 2012, 32(6): 1550—1553. (in Chinese with English abstract)

- [6] 王庆, 薛卫青, 马晗煦, 等. 近红外光谱技术定量分析玉米杂交种纯度[J]. 农业工程学报, 2012, 28(增刊2): 259—264.
Wang Qi, Xue Weiqing, Ma Hanxu, et al. Quantitative analysis of seed purity for maize using near infrared spectroscopy [J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2012, 28(Supp.2): 259—264. (in Chinese with English abstract)
- [7] 黄敏, 朱晓, 朱启兵, 等. 基于高光谱图像的玉米种子特征提取与识别[J]. 光子学报, 2012, 41(7): 868—873.
Huang Min, Zhu Xiao, Zhu Qibing, et al. Morphological characteristics of maize seed extraction and identification based on the hyperspectral image [J]. Acta Photonica Sinica, 2012, 41(7): 868—873. (in Chinese with English abstract)
- [8] 邹伟, 方慧, 刘飞, 等. 基于高光谱图像技术的油菜籽品种鉴别方法研究[J]. 浙江大学学报: 农业与生命科学版, 2011, 37(2): 175—180.
Zou Wei, Fang Hui, Liu Fei, et al. Identification of rapeseed varieties based on hyperspectral imagery [J]. Journal of Zhejiang University: Agriculture and Life Sciences, 2011, 37(2): 175—180. (in Chinese with English abstract)
- [9] Zhang X L, Liu F, He Y, et al. Application of hyperspectral imaging and chemometric calibrations for variety discrimination of maize seeds[J]. Sensors, 2012, 12(12): 17234—17246.
- [10] 周竹, 李小昱, 陶海龙, 等. 基于高光谱成像技术的马铃薯外部缺陷检测[J]. 农业工程学报, 2012, 28(21): 221—228.
Zhou Zhu, Li Xiaoyu, Tao Hailong, et al. Detection of potato external defects based on hyperspectral imaging technology[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2012, 28(21): 221—228. (in Chinese with English abstract)
- [11] Huang M, Wan X M, Zhang M, et al. Detection of insect-damaged vegetable soybeans using hyperspectral transmittance image[J]. Journal of Food Engineering, 2013, 116(1): 45—49.
- [12] 褚小立, 袁洪福, 陆婉珍, 等. 近红外分析中光谱预处理及波长选择方法进展与应用[J]. 化学进展, 2004, 16(4): 528—542.
Chu Xiaoli, Yuan Hongfu, Lu Wanzhen, et al. Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique[J]. Progress in Chemistry, 2004, 16(4): 528—542. (in Chinese with English abstract)
- [13] 张志勇, 李刚, 林凌, 等. EMD 和 SPA 算法在光谱法检测面粉过氧化苯甲酰添加量中的应用[J]. 光谱学与光谱分析, 2012, 32(10): 2815—2819.
Zhang Zhiyong, Li Gang, Lin Ling, et al. Application of EMD and SPA algorithm in the detection of benzoyl peroxide addition in flour by spectroscopy[J]. Spectroscopy and Spectral Analysis, 2012, 32(10): 2815—2819. (in Chinese with English abstract)
- [14] 蔡剑华, 王先春, 胡惟文, 等. 基于 EMD 的土壤有机质含量近红外光谱检测[J]. 农业机械学报, 2010, 41(9): 182—186.
Cai Jianhua, Wang Xianchun, Hu Weiwen, et al. Near-infrared spectrum detection of soil organic matter content based on empirical mode decomposition[J]. Transactions of the Chinese society for agriculture machinery, 2010, 41(9): 182—186. (in Chinese with English abstract)
- [15] Trad D O, Travassos J M. Wavelet filtering of magnetotelluric data[J]. Geophysics, 2000, 65(2): 482—491.
- [16] 吴桂芳, 何勇. 小波阈值降噪模型在红外光谱信号处理中的应用研究[J]. 光谱学与光谱分析, 2009, 29(12): 3246—3249.
Wu Guifang, He Yong. Application of wavelet threshold denoising model to infrared spectral signal processing[J]. Spectroscopy and Spectral Analysis, 2009, 29(12): 3246—3249. (in Chinese with English abstract)
- [17] 梁亮, 杨敏华, 臧卓, 等. 基于小波去噪与 SVR 的小麦冠层含氮率高光谱测定[J]. 农业工程学报, 2010, 26(12): 248—253.
Liang Liang, Yang Minhua, Zang Zhuo, et al. Determination of wheat canopy nitrogen content ratio by hyperspectral technology based on wavelet denoising and support vector regression[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2010, 26(12): 248—253. (in Chinese with English abstract)
- [18] Liu F, He Y. Application of successive projections algorithm for variable selection to determine organic acids of plum vinegar [J]. Food Chemistry, 2009, 115(4): 1430—1436.
- [19] Liu F, Jiang Y, He Y. Variable selection in visible/near infrared spectra for linear and nonlinear calibrations: a case study to determine soluble solids content of beer[J]. Anal Chim Acta, 2009, 635(1): 45—52.
- [20] Leardi R. Application of genetic algorithm-PLS for feature selection in spectral data sets[J]. Journal of Chemometrics, 2000, 14(5/6): 643—655.
- [21] Leardi R, Gonzalez A L. Genetic algorithms applied to feature selection in PLS regression: how and when to use them[J]. Chemometrics and Intelligent Laboratory Systems, 1998, 41(2): 195—207.
- [22] Luna A S, da Silva A P, Pinho J S, et al. Rapid characterization of transgenic and non-transgenic soybean oils by chemometric methods using NIR spectroscopy[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2013, 100: 115—119.
- [23] Canaza-Cayo A W, Cozzolino D, Alomar D, et al. A feasibility study of the classification of Alpaca (Lama pacos) wool samples from different ages, sex and color by means of visible and near infrared reflectance spectroscopy[J]. Computers and Electronics in Agriculture, 2012, 88: 141—147.
- [24] 刘飞, 王莉, 何勇, 等. 应用可见/近红外光谱进行黄酒品种的判别[J]. 光谱学与光谱分析, 2008, 28(3): 586—589.
Liu Fei, Wang Li, He Yong, et al. Discrimination of varieties of Yellow wines using Vis/NIR spectroscopy[J]. Spectroscopy and Spectral Analysis, 2008, 28(3): 586—589. (in Chinese with English abstract)
- [25] Gong A P, Qiu Z J, He Y, et al. A non-destructive method for quantification the irradiation doses of irradiated sucrose using Vis/NIR spectroscopy[J]. Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy, 2012, 99: 7—11.
- [26] 夏天, 吴文斌, 周清波, 等. 冬小麦叶面积指数高光谱遥感反演方法对比[J]. 农业工程学报, 2013, 29(3): 139—147.
Xia Tian, Wu Wenbin, Zhou Qingbo, et al. Comparison of two inversion methods for winter wheat leaf area index based on hyperspectral remote sensing[J]. Transactions of

- the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2013, 29(3): 139–147. (in Chinese with English abstract)
- [27] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: Theory and applications[J]. *Neurocomputing*, 2006, 70(1/2/3): 489–501.
- [28] Jiang H, Zhu W X. Determination of pear internal quality attributes by fourier transform near infrared (FT-NIR) spectroscopy and multivariate analysis[J]. *Food Analytical Methods*, 2013, 6(2): 569–577.
- [29] 李若诚, 许文方, 华英杰, 等. 基于近红外光谱和极限学习机的普洱茶中游离氨基酸总量检测[J]. *长春工业大学学报: 自然科学版*, 2012, 33(3): 269–273.
- Li Ruocheng, Xu Wenfang, Hua Yingjie, et al. Free amino acid content detection in Purer tea based on near infrared spectroscopy and Extreme Learning Machine[J]. *Journal of Changchun University of Technology: Natural Science Edition*, 2012, 33(3): 269–273. (in Chinese with English abstract)
- [30] 唐玉莲, 梁逸曾, 范伟, 等. 应用近红外光谱快速鉴别不同年龄段人食用的奶粉品种[J]. *红外*, 2010, 31(1): 30–35.
- Tang Yulian, Liang Yizeng, Fan Wei, et al. Fast discrimination of varieties of different age rank milk powder using near infrared spectroscopy[J]. *Infrared*, 2010, 31(1): 30–35. (in Chinese with English abstract)

Fast identification of watermelon seed variety using near infrared hyperspectral imaging technology

Zhang Chu¹, Liu Fei¹, Kong Wenwen¹, Zhang Hailiang^{1,2}, He Yong^{1*}

(1. School of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China;

2. School of Mechatronics Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract: Watermelon seed variety selection plays a vital role in watermelon planting, and the variety of watermelon seeds directly affect the yield and quality of watermelons. In this study, we aimed to identify the cultivars of watermelon seeds by using a novel, rapid, non-invasive, and low cost technique named hyperspectral imaging. 121 samples of four different cultivars of watermelon seeds were investigated, and a near-infrared hyperspectral imaging system (874-1734 nm with 256 bands) was established to acquire the hyperspectral images of the samples. A region of interest (ROI) with 15×15 pixels of the hyperspectral image of each sample was defined, and the average reflectance spectrum of the ROI were extracted. To remove the absolute noises of the spectra, only the spectral range 1 042-1 646 nm was used for analysis, and to reduce the noises existed in spectral range 1 042-1 646 nm, the extracted 121 reflectance spectra were preprocessed by Savitzky-Golay smoothing (SG), Empirical Mode Decomposition (EMD), and Wavelet Transform (WT) methods. The preprocessed spectra were then used to select sensitive wavelengths by Successive Projections Algorithm (SPA) and Genetic Algorithm-partial least squares (GA-PLS) methods. Different numbers of sensitive wavelengths were selected by different variable selection methods with different preprocessing methods. 24, 16, and 15 sensitive wavelengths were selected by SPA with spectra preprocessed by SG, EMD, and WT, respectively. Moreover, 38, 33, and 32 sensitive wavelengths were selected by GA-PLS with spectra preprocessed by SG, EMD, and WT, respectively. Partial least squares – discriminant analysis (PLS-DA) was used to build discriminant models with the full spectra, and back-propagation neural network (BPNN) and extreme learning machine (ELM) were applied to build discriminant models with the selected wavelength variables. A PLS-DA model with spectra preprocessed by EMD obtained the best identification rate among all PLS-DA models, with an identification rate of 91.57% in the calibration set and 78.95% in the prediction set. SPA-BPNN models showed relatively worse results than GA-PLS-BPNN models with the same spectral preprocessing methods. The SG-GA-PLS-BPNN model obtained the best performance among all BPNN models, with an identification rate of 92.77% in the calibration set and 86.84% in the prediction set. Compared with the PLS-DA models and the BPNN models, ELM models obtained the best results. All ELM models obtained an identification rate over 90% in the calibration set and the prediction set, and the SG-SPA-ELM model, SG-GA-PLS-ELM model, and WT-SPA-ELM model obtained the identification rate of 100% of calibration and prediction. The overall results showed that BPNN and ELM models performed better than PLS-DA models, and the ELM models with the selected wavelengths based on SG preprocessed spectra obtained the best results, with 100% classification accuracy for both the calibration set and the prediction set. The SG preprocessing method showed the best performance in all PLS-DA, BPNN, and ELM models. The results indicated that it was feasible to use near-infrared hyperspectral imaging to identify the watermelon seed varieties, and near-infrared hyperspectral imaging provided an alternate way of rapid identification of watermelon seed variety. ELM, as a single hidden layer feed-forward network, was an effective classification method in watermelon seed cultivar identification. Moreover, the results in this paper showed the great potential of hyperspectral imaging in the seed industry for on-line identification of seed cultivars and detection of the seed quality parameters.

Key words: near infrared spectroscopy, genetic algorithm, imaging processing, watermelon seed, successive projections algorithm, extreme learning machine

(责任编辑: 刘丽英)