近红外高光谱的脐橙粒化检测研究

刘燕德，李茂鹏，胡 军，徐 振，崔惠桢

华东交通大学智能机电装备创新研究院，江西 南昌 330013

摘 要 脐橙粒化影响消费者食用口感，降低品质，受到广大果农和消费者的关注。脐橙粒化的检测是一项具有挑战性的任务，对品质分级具有重大意义。以不同粒化程度的赣南脐橙为研究对象，探究利用高光谱检测实现对赣南脐橙粒化程度定性判别的可行性。肉眼是无法判断脐橙粒化程度的，因此对脐橙样本做好序号标记后先测光谱再切开判断粒化程度，按照粒化程度分为无粒化(粒化面积为0%)、轻度粒化(粒化面积小于25%)、中度粒化(粒化面积25%～50%)，每类各58个脐橙样品。在这三类脐橙底部均匀取3个点，每类174个样本，共计522个样本数据用作构建原始光谱矩阵。利用近红外高光谱成像系统采集样本397.5～1 014 nm波段内的高光谱图像信息，再利用ENVI4.5软件通过选择感兴趣区域(ROI)提取样本的平均光谱信息。采用主成分分析(PCA)、连续投影算法(SPA)、无信息变量消除(UVE)三种降维方法对光谱数据进行降维处理，消除无关变量，提取有用信息。原始光谱176个波长，PCA挑选出6个主成分因子，SPA挑选17个特征波长，UVE挑选54个特征波长。以全谱数据和三种降维方法挑选出来的变量作为输入分别建立偏最小二乘判别分析(PLS-DA)和最小二乘支持向量机(LS-SVM)模型。建立的PLS-DA建模方法，PCA-PLS-DA误判率最高为25.58%，UVE-PLS-DA误判率最低为5.38%。基于RBF-Kernel和LIN-Kernel两种核函数下的LS-SVM建模方法，整体上RBF-Kernel建模效果优于LIN-Kernel，UVE波长筛选后建立的模型效果优于其他降维方法且降低了模型的误判率。基于RBF-Kernel的UVE-LS-SVM模型效果最佳，检测精度最高，分类总误判率为0.78%，达到最佳效果。该研究结果表明建立的模型能很好地对不同粒化程度的脐橙进行判别，该模型仅采用30.68%的数据，在降低光谱空间维度的同时还降低了误判率，对促进脐橙产业的品质分级发展具有一定的现实意义。

关键词 高光谱；赣南脐橙；粒化程度；无信息变量消除

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)05-1366-06

引 言

脐橙属于柑橘的优良品种，各国竞相栽培育种，种植地区几乎遍布全世界，是最受人们欢迎的水果之一，深受人们喜爱[1]。消费者大多喜欢汁水充足，酸甜可口的脐橙，但是人们也会偶尔吃到嘴里有渣的脐橙果肉，并有干瘪苦涩的口感，这种脐橙即脐橙粒化。脐橙粒化会导致脐橙失水，果肉呈颗粒状，甜度降低，粒化严重的脐橙会严重影响口感，甚至丧失食用价值，从而影响中国脐橙产业的快速发展。脐橙采摘期和贮藏期都容易发生果实粒化。粒化脐橙与无粒化脐橙混杂，在选购脐橙时难以区分以至于降低商家信誉，影响消费者的食用体验。脐橙作为我国出口量较大的水果之一，对其内部品质检测难度大的问题亟需解决。脐橙粒化传统检测主要是人工筛查，但需要丰富的经验，且存在较大的误判率，并且难以对脐橙粒化程度进行区分。探索一种无损、快速脐橙粒化程度的检测方法，可以大大降低人力物力，规范市场，同时对脐橙商品的出口、流通具有重大意义[2]。

高光谱图谱合一，包含丰富的信息，随着光谱技术的不断完善与发展，研究者们将高光谱技术广泛应用于果蔬行业的品质检测，高光谱成像技术在果蔬品质安全检测领域的应用已日臻完善且应用效果较为理想[3]。Liu等[4]探究使用近红外高光谱成像装置检测猕猴桃早期隐性损伤，采用基于核函数的偏最小二乘(kernel partial least squares，KPLS)对高光谱波段进行降维，并结合多种算法进行建模，最佳精度达到98.27%，验证了高光谱技术用于猕猴桃的早期隐性损伤检测可行性。Jan Steinbrener等[5]使用高光谱成像技术结合卷积神经网络对RGB图像数据进行训练，对颜色形状相类

收稿日期：2021-04-09，**修订日期**：2021-07-21

基金项目：国家自然科学基金项目(31760344)，江西省国家科技奖后备培育项目(20192AEI91007)资助

作者简介：刘燕德，女，1967年生，华东交通大学智能机电装备创新研究院教授 e-mail：jxliuyd@163.com

似的水果和蔬菜进行分类，正确率达到 92.23%。Pu 等[6]以香蕉为研究对象，利用高光谱成像技术对香蕉成熟程度进行检测，根据特征波段(650，705 和 740 nm)建立的偏最小二乘判别分析(partial least squares discriminant analysis，PLS-DA)模型总分类准确率为 93.3%，研究表明香蕉成熟度的判别可以通过高光谱技术实现。高升等[7]以红提可溶性固形物为研究指标，同时采集红提的光谱信息和图像信息，并建立 PLSR 和最小二乘支持向量机(least squares support vector machines，LS-SVM)检测模型，提出一种图像和光谱信息结合的糖度检测模型，建立的 LS-SVM 预测集相关系数为 0.954。Zhang 等[8]利用主成分分析对含有各种缺陷的南丰蜜桔进行定性分类，分类准确率达到 96.63%。以上研究表明，高光谱成像技术可以广泛应用于水果和蔬菜的质量和品质分级检测。

以江西特产——赣南脐橙作为研究对象，探究利用高光谱技术结合多种化学计量学方法识别赣南脐橙粒化程度的可行性。通过建立 PLS-DA[9]和 LS-SVM[10]模型并评选出最优模型，结果表明高光谱检测技术可应用于脐橙粒化的检测，根据建立的脐橙粒化检测模型，可实现对脐橙品质准确分级。

1　实验部分

1.1　材料

研究对象选用江西独有的脐橙品种——赣南脐橙。赣南脐橙在中国脐橙界享有很大的知名度，畅销国内外，是深受人们喜爱的脐橙品种之一。实验样本采摘于江西省赣州市于都县的一个商业果园。样本包含三种不同粒化程度(无、轻度、中度)；重度粒化(粒化面积大于 50%)[11]样本无食用价值，且能通过肉眼轻易分辨，故不作研究。样本经采摘后均用纱布洗净擦干后储存在温度控制为 5℃的冰箱，图像和光谱信息采集之前需将脐橙样品取出放置 2～3 h，以达到正常室温。由于肉眼无法直接观察判断粒化程度，所以在采集光谱前需要给每个脐橙编号，采集完光谱信息后，再将脐橙切开，观察粒化程度；其中无粒化(粒化面积为 0%)；轻度粒化(粒化面积小于 25%)；中度粒化(粒化面积 25%～50%)。图 1为采用的三种不同粒化程度的赣南脐橙。

图 1　三种不同粒化程度的赣南脐橙

(a)：无；(b)：轻度；(c)：中度

Fig. 1　Three different granulation degrees of navel orange in southern Jiangxi

(a)：None；(b)：Mild；(c)：Moderate

最后挑选出无粒化、轻度粒化、中度粒化各 58 个脐橙样品。由于脐橙粒化均从底部产生，每个脐橙选择底部三个不同位置提取平均光谱，每类 174 个样本，共计 522 个样本。利用 Kennard-Stones 算法将样本按照 3∶1 分成训练集和测试集，其中每类训练集为 131 个样本，测试集为 43 个样本，不同粒化程度的脐橙训练集和预测集区分及样本编号如表 1 所示。

表 1　不同粒化程度脐橙的训练集和预测集区分以及样本编号

Table 1　Classification of training sets and prediction sets and sample codes of different coffee beans

Degree of granulation	Sample code	Total number of samples	Number of training set samples	Number of test set samples
None	1	174	131	43
Mild	2	174	131	43
Moderate	3	174	131	43

1.2　装置及参数

实验用双利合谱公司“GaiaSorter”盖亚高光谱成像系统如图 2 所示，该装置主要由三部分组成：成像镜头、成像光谱仪、探测器。采用短波近红外相机的光源波长范围为 397.5～1 014 nm，共计 176 个波长。在光谱采集过程中，为保证图像的清晰度，将相机曝光时间和电动平移台速度分别设置为 15 ms 和 18 mm·s^{-1}，分辨率设置为 15 ms，回程速度设置为 20 mm·s^{-1}。

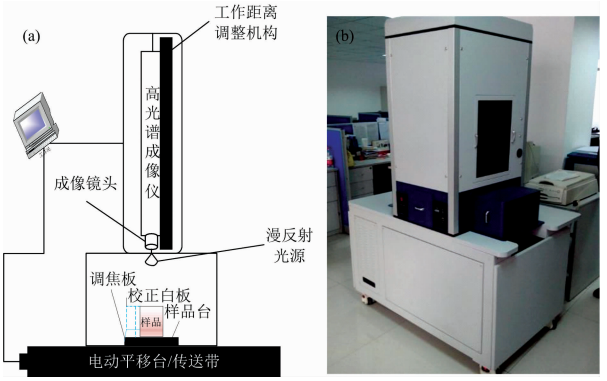

图 2 高光谱设备成像装置

(a)：原理图；(b)：实物图

Fig. 2 Hyperspectral equipment imaging device

(a)：Schematic diagram；(b)：Picture

1.3 高光谱图像采集

采用 SpecView 软件对不同粒化赣南脐橙样品进行图像采集，为得到清晰的光谱图像，系统参数设置后需对图像进行调焦处理。样品放置在电动平移台的果杯上，底部朝上且与工作平面平行，脐橙底部朝向相机镜头。在原始光谱图像采集后应利用白、暗基准进行校正，以避免各波段信息分布不均以及 CCD 相机存在暗电流效应对图像质量的影响[8]。在光谱图像校正之前需获得校正参比，获取参比主要步骤如下：首先用黑色的盖子盖住 CCD 相机镜头采集一段全黑图像，然后去掉镜头盖采集白色参比板的白色参考图像，利用该图像对样品原始图像进行校正，校正后的高光谱图像计算如式(1)所示。

$$R = \frac{I_\lambda - H_B}{B_W - H_B} \tag{1}$$

式(1)中，H_B，B_W 和 I_λ 分别为黑色参考、白色参考和原始光谱数据。待对所有光谱图像校正完毕后，利用 ENVI4.5 软件在每个脐橙底部均匀选择三个不同的位置，选择含有 1 000个像素点的感兴趣区域(region of interest，ROI)并提取平均光谱，进行后续分析处理。

1.4 模型评价标准

首先通过高光谱系统采集不同粒化程度脐橙样品的高光谱图像，并使用 ENVI4.5 软件提取光谱数据；在 Unscrambler 中采用多种预处理方法对原始光谱进行处理，并建立 PLS-DA 模型进行评价，评选最优的预处理方法。利用多种波长筛选方法对最优预处理之后的光谱数据进行筛选，建立对应的 PLS-DA 和 LS-SVM 检测模型。模型评价参数为不同粒化程度脐橙鉴别的误判率及总误判率，误判率是判错样本个数占总个数的比例，误判率越低，建模效果越好。图 3 为三种不同粒化程度脐橙鉴别流程图。

2 结果与讨论

2.1 不同粒化程度脐橙的高光谱响应特性对比与分析

高光谱的波长范围为 397.5～1 014 nm，共计 176 个波段点。图 4 为三种粒化程度脐橙的原始平均反射率光谱，每条光谱代表 1 000 个像素点的信息。从图 4 可知，三条光谱变化趋势大致相似，整体上随着粒化程度的加深，平均光谱反射率会增大，中度粒化脐橙的平均反射率最大。400～500 nm 波段反射率呈下降趋势，500 nm 处存在一个波谷，这是因为果皮对类胡萝卜素的光吸收引起的，500～820 nm 波段，平均反射率单调递增，在 820～1 014 nm 之间反射率呈下降趋势且在 980 nm 波长处有一个波谷，这可能是由于水分子中 O—H 键三级倍频的拉伸振动引起的反射率变化，此处中度粒化脐橙平均反射率＞轻度＞无粒化[12]。

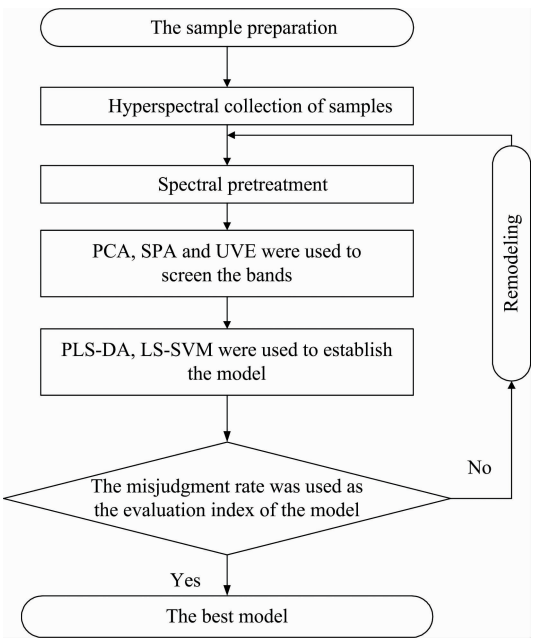

图 3 不同粒化程度脐橙模型建立流程图

Fig. 3 Modeling flow chart of detection for navel oranges with different granulation degrees

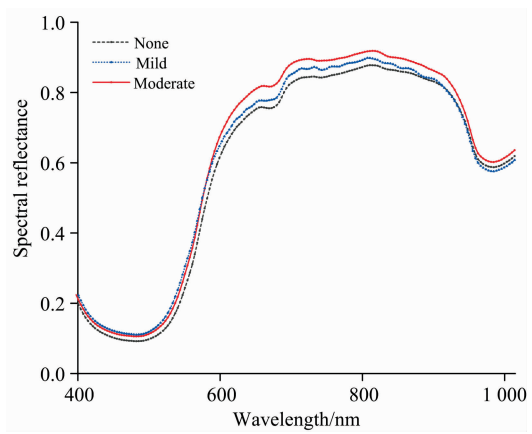

图 4 三种粒化程度脐橙的原始平均反射率光谱图

Fig. 4 Original mean reflectance spectra of navel oranges with three granulation degrees

2.2 高光谱特征变量选择

高光谱包含 176 个波长，信息较多，为了降低实验误差去除噪声，先采用多种预处理方法处理原始光谱并建立

PLS-DA 模型评价预处理方法的稳定性。多种预处理方法未能提高数据精度，降低误判率，故尝试采用降维和多种波长筛选方法降低数据维数，去除无用信息，提取有效光谱变量信息。

2.2.1　基于 PCA 的降维方法

主成分分析(principal component analysis，PCA)是常用的数据降维方法之一，变换后的变量空间是原变量空间变量的线性组合，并能代表绝大部分信息[13]。图 5 为前 20 个主成分对脐橙高光谱的累积贡献率，最大主成分因子数设置为 20，当选择 6 个主成分时，新的变量空间能代表 99%以上原始光谱的信息，故采用 6 个主成分因子作为后续建模的输入。

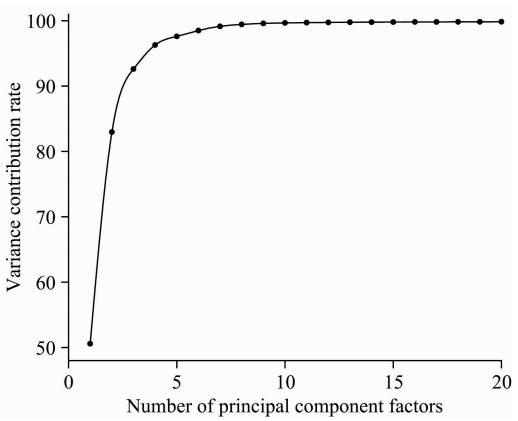

图 5　前 20 个主成分对脐橙高光谱的累积贡献率

Fig. 5　Cumulative contribution rates of the first 20 principal components of citrus hyperspectral data

2.2.2　基于 SPA 的特征波长筛选

连续投影算法(successive projections algorithm，SPA)是一种前向循环选择方法，根据计算不同样本子集的多元线性回归模型的均方根误差获得最佳样本集，通过向量投影，选择最小的冗余度和共线性的有效波长[14]。本模型在 MATLAB2014b 中建立，设定最小挑选变量数为 10 最大挑选变量数为 40。图 6 是基于 SPA 算法对样品光谱波段筛选之后的结果，其中横、纵坐标分别代表波长和平均光谱反射率。利用 SPA 算法在原始光谱 176 个变量中挑选出 17 个变量，与原始光谱相比，波长数目减少 90.34%，在一定程度上能够消除冗余信息，使模型得到简化，后续将使用经 SPA 挑选后的变量用于模型建立。

2.2.3　基于 UVE 的特征波长筛选

无信息变量消除(uninformative variable elimination，UVE)方法是建立在偏最小二乘法(PLS)回归系数基础上的波长筛选方法，采用回归系数来衡量波段的显著性[15]。图 7 为样品光谱经过 UVE 波长筛选的结果，平行于 X 轴的两条平行线表示的是 UVE 波长筛选的阈值，以光谱矩阵处稳定性最大值的 99%作为阈值分隔线，值为±35.998 3，稳定性在阈值分隔线内的变量将会被剔除，取阈值分隔线之外的变量作为输入变量。176 个光谱变量经过 UVE 筛选后剩下 55 个变量，约占原始数据的 30.68%，使模型得到简化。

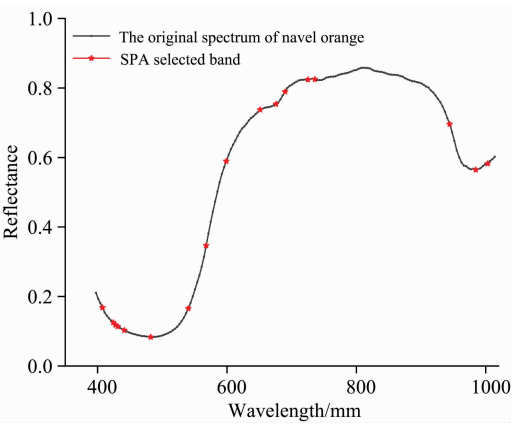

图 6　SPA 波长变量选择结果

Fig. 6　SPA wavelength variable selection results

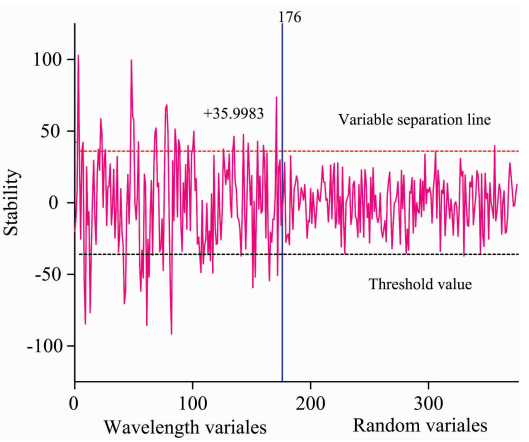

图 7　UVE 变量筛选稳定性结果图

Fig. 7　UVE variable screening stability result graph

2.3　不同粒化程度脐橙的高光谱 PLS-DA 模型建立

建模之前采用 PCA 降维方法和 SPA、UVE 算法对全光谱进行波长筛选，然后建立 PLS-DA 模型并比较建模效果。如表 2 为不同波长变量选择方法的 PLS 模型结果对比，经 UVE 算法筛选波长后，建立的 PLS-DA 模型效果罪佳，建模集的相关系数 R_p 和均方根误差 RMSEP 分别为 0.895 和 0.261，主成分个数为 7，误判率最低为 5.38%。

2.4　不同粒化程度脐橙的高光谱 LS-SVM 模型建立

LS-SVM 常采用径向基核函数(RBF-Kernel)和线性核函数(Lin-Kernel)两种核函数建立模型，γ 和 σ^2 是 LS-SVM 模型评价的重要参数。如表 3 为不同降维方法分别基于两种核函数建立的 LS-SVM 模型结果对比，结果表明：Lin 核函数的效果低于 RBF 核函数，经过 UVE 波长筛选后，两种核函数的误判率均为最低，预测集误判率分别为 0.78%和 1.55%，说明 LS-SVM 模型精度优于 PLS-DA 模型，建立的较优的 UVE-LS-SVM 模型。因此，使用 UVE 波长筛选方法可以很好地实现对不同粒化程度赣南脐橙的定性判别。

表 2 基于不同降维方法的 PLS-DA 模型的比较

Table 2 Comparison of PLS-DA models based on different dimension reduction methods

Model	Variable selection methods	Number of variable	PCs	R_c	RMSEC	R_p	RMSEP	Error rate of prediction set/%
PLS-DA	Original data	176	12	0.910	0.210	0.890	0.281	7.55
	PCA	6	5	0.708	0.442	0.659	0.474	25.58
	SPA	17	7	0.832	0.330	0.827	0.338	15.55
	UVE	54	7	0.912	0.244	0.895	0.261	5.38

表 3 不同降维方法与 LS-SVM 方法建立的模型性能比较

Table 3 Comparison of model performance between different dimension reduction methods and LS-SVM method

Methods	No. of variable	RBF-Kernel			LIN-Kernel		
		γ, σ^2	Error rate of training set/%	Error rate of test set/%	γ	Error rate of training set/%	Error rate of test set/%
Full spectrum	176	1.796×10^4, 672.223	1.27	4.65	1.568	2.29	4.65
PCA	6	6.781, 0.735	1.78	1.55	1.039	5.09	17.05
SPA	17	1.362×10^4, 122.775	0.76	2.33	1.111	0.76	4.65
UVE	54	1.802×10^4, 500.116	0%	0.78%	1.667	0.25	1.55

如图 8 为基于 RBF-Kernel 的 UVE-LS-SVM 模型预测集结果图，从图中可以看出，轻度粒化脐橙样本仅有一例误判为无粒化，其他样品均准确判别，综合 PLS-DA 和 LS-SVM 两种建模方法，LS-SVM 建模效果整体优于 PLS-DA。经 UVE 变量筛选后的 LS-SVM 模型误判率最低为 0.78%，因此，UVE-LS-SVM 模型更适合对不同粒化程度脐橙进行高精度判别。

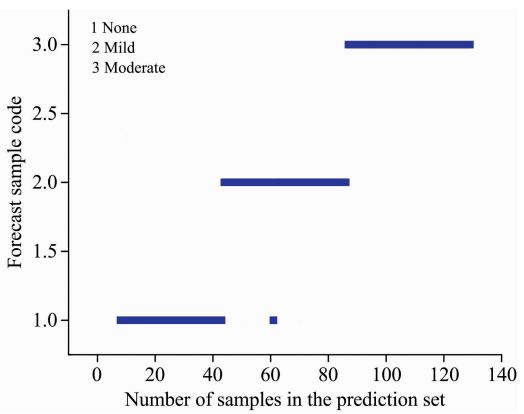

图 8 基于 RBF-Kernel 的 UVE-LS-SVM 模型预测集结果

Fig. 8 Prediction results of UVE-LS-SVM model based on RBF-Kernel

3 结 论

以赣南脐橙为研究对象，验证利用高光谱检测技术对不同粒化程度脐橙定性判别的可行性。分别利用 PCA 降维方法和 SPA、UVE 两种波长筛选方法缩小矩阵空间，在一定程度上降低数据空间维数。利用 PLS-DA 和 LS-SVM 与不同波长筛选方法结合建立定性判别模型，结果表明 LS-SVM 建模效果整体优于 PLS-DA，RBF-Kernel 核函数效果优于 LIN-Kernel 核函数，经 UVE 波长筛选后的变量结合 LS-SVM 中 RBF-Kernel 建模效果最佳，UVE-LS-SVM 误判率达到最低为 0.78%。本实验为脐橙粒化提供了一种快速无损的光谱检测方法，对柑橘品质分级具有重要意义，也为粒化水果检测提供借鉴。

References

[1] YAO Shi-xiang, LI Qiu-yu, CAO Qi, et al(姚世响，李秋雨，曹 琦，等). Food and Fermentation Industries(食品与发酵工业), 2020, 46(18): 259.

[2] Jie Dengfei, Wu Shuang, Wang Ping, et al. Food Analytical Methods, 2021, 14(2): 280.

[3] Jiang H, Wang W, Zhuang H, et al. Food Analytical Methods, 2019, 12(10): 2205.

[4] Liu Y, Yang Z, Cao J, et al. Detection of Invisible Damage of Kiwi Fruit Based on Hyperspectral Technique, International Conference on Brain Inspired Cognitive Systems, 2019: 373.

[5] Steinbrener J, Posch K, Leitner R, et al. Computers and Electronics in Agriculture, 2019, 162: 364.

[6] Pu Y, Sun D, Buccheri M, et al. Food Analytical Methods, 2019, 12(8): 1693.

[7] GAO Sheng, WANG Qiao-hua(高 升，王巧华). Chinese Journal of Luminescence(发光学报), 2019, 40(12): 1574.

[8] Zhang Hailiang, Zhang Shuai, Dong Wentao, et al. Infrared Physics and Technology, 2020, 108: 103341.

[9] Henseler J, Hubona G S, Ray P A, et al. Industrial Management and Data Systems, 2016, 116(1): 2.

[10] Deng W, Yao R, Zhao H, et al. Soft Computing, 2019, 23(7): 2445.
[11] WANG Miao, ZHANG Jing, HE Yan, et al(王　淼, 张　晶, 贺　妍, 等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2016, 32(7): 290.
[12] Yang Yichao, Sun Dawen, Wang Nannan. Computers and Electronics in Agriculture, 2015, 113: 203.
[13] Qin B, Li Z, Luo Z, et al. Optical and Quantum Electronics, 2017, 49(7): 244. 1.
[14] Tang R, Chen X, Li C, et al. Applied Spectroscopy, 2018, 72(5): 740.
[15] YU Hui-chun, FU Xiao-ya, YIN Yong, et al(于慧春, 付晓雅, 殷　勇, 等). Journal of Nuclear Agricultural Sciences(核农学报), 2020, 34(3): 582.

Detection of Citrus Granulation Based on Near-Infrared Hyperspectral Data

LIU Yan-de, LI Mao-peng, HU Jun, XU Zhen, CUI Hui-zhen
School of Mechanical and Electrical Engineering, East China Jiaotong University, Nanchang　330013, China

Abstract The granulation of navel orange affects consumers' taste and reduces its quality. It has attracted the attention of fruit farmers and consumers. The detection of navel orange granulation is challenging and has great significance for quality classification. In this paper, the different granulation degrees of Gannan navel oranges are used as the research object to explore the qualitative determination of the granulation degree of Gannan navel oranges by using hyperspectral detection technology. Since the degree of granulation of navel oranges cannot be judged by the naked eye, the samples of navel oranges are marked with serial numbers, and then the spectrum is measured. Finally, the samples were cut to determine the degree of granulation. According to the degree of granulation, it is classified as non-granulation (the granulation area is 0%); light granulation (granulation area less than 25%); and medium granulation (granulation area 25%～50%). Take 3 points uniformly at the bottom of these three types of navel oranges, each with 174 samples, and a total of 522 sample data are used as the rows for constructing the spectral matrix. The near-infrared hyperspectral imaging system was used to collect the hyperspectral image information of the sample in the 397.5～1 014 nm band and then use the ENVI 4.5 software was used to extract the average spectral information the sample by selecting the Region of Interest (ROI). Three dimensionality reduction methods: Principal Component Analysis (PCA), Successive Projections Algorithm (SPA), and Uninformative Variable Elimination (UVE) are used to reduce the dimensionality of the spectral data to eliminate irrelevant variables and extract useful information. The original spectrum has 176 wavelengths. PCA selects 6 principal component factors. SPA selects 17 characteristic wavelengths, and UVE selects 54 characteristic wavelengths. The full spectrum data and the variables selected by the three-dimensionality reduction methods are used as input to establish Partial Least Squares Discriminant Analysis (PLS-DA) and Least Squares Support Vector Machines (LS-SVM) model. In the established PLS-DA modeling method, the highest false positive rate of PCA-PLS-DA is 25.58%, and the lowest false-positive rate of UVE-PLS-DA is 5.38%. The LS-SVM modeling method is based on the two kernel functions of RBF-Kernel and LIN-Kernel, and the effect of RBF-Kernel modeling is better than that of LIN-Kernel generally. And the model established after UVE wavelength screening is better than other dimensionality reduction methods, which reduces the model's false positive rate. The UVE-LS-SVM model based on RBF-Kernel has the best effect and the highest detection accuracy, and the total misjudgment rate of classification is 0.78%, achieves the best results. This study shows that the established model can distinguish navel oranges with different granulation degrees. The model reduces the spectral dimension while also reducing the misjudgment rate with only 30.68% of the data, which is useful for promoting the quality of the navel orange industry with certain practical significance.

Keywords Hyperspectral; Gannan navel orange; Granulation degree; Uninformative Variable Elimination

(Received Apr. 9, 2021; accepted Jul. 21, 2021)