

doi:10.6041/j.issn.1000-1298.2013.02.027

基于近红外漫反射光谱的损伤猕猴桃早期识别*

郭文川 王铭海 岳 绒

(西北农林科技大学机械与电子工程学院, 陕西杨凌 712100)

摘要: 以贮藏 1d 的碰撞损伤猕猴桃、挤压损伤猕猴桃和无损猕猴桃为对象,分别建立了猕猴桃的 Fisher 判别模型、BP 神经网络判别模型与最小二乘支持向量机(LSSVM)判别模型,综合比较了采用全光谱波长(FS)、主成分分析(PCA)提取特征变量与连续投影算法(SPA)优选特征波长作为各模型输入变量时,对各模型判别效果的影响。研究表明,SPA 优选特征波长相比于 PCA 和 FS 有较明显的优势;3 种判别模型均能基本满足实际要求,且 LSSVM 模型的识别性能最佳,其中 SPA-LSSVM 模型对预测集碰撞损伤样品、挤压损伤样品与无损样品的正确识别率分别达到 100%、95% 和 100%,总的正确识别率为 98.2%。

关键词: 猕猴桃 近红外光谱 连续投影算法 最小二乘支持向量机 识别

中图分类号: O657.33; S663.4 文献标识码: A 文章编号: 1000-1298(2013)02-0142-05

Early Recognition of Bruised Kiwifruit Based on Near Infrared Diffuse Reflectance Spectroscopy

Guo Wenchuan Wang Minghai Yue Rong

(College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, Shaanxi 712100, China)

Abstract: To detect bruised kiwifruits from intact kiwifruit early and effectively, near infrared diffused reflectance spectroscopy technology combined with Fisher discriminant function, BP neural network and least square support vector machine(LSSVM) were applied to discriminate collided kiwifruit, pressed kiwifruit and intact kiwifruit after storage of 1d, respectively. Effectiveness of the discriminant model using full spectrum(FS), feature variables based on principal component analysis(PCA) and characteristic wavelength by successive projection algorithm(SPA) was compared and evaluated. The results showed that SPA gave the best advantage compared with methods of FS and PCA. Three models all had an acceptable accuracy, especially LSSVM model had the optimal recognition performance. SPA-LSSVM had an accuracy rate of 100%, 95% and 100% for identifying collided samples, pressed fruits and intact ones respectively, and the discriminant accuracy rate for total samples was 98.2%.

Key words: Kiwifruit Near infrared spectroscopy Successive projection algorithm Least square support vector machine Recognition

引言

猕猴桃在其采摘、贮藏、包装及流通等诸多环节中常因碰撞或挤压造成果实不同程度的机械损伤,并且仅凭肉眼难以识别,若损伤果不及时剔除,随着其腐烂程度的不断加重,不仅严重影响消费者的购

买欲望,而且会感染无损猕猴桃,极大地影响猕猴桃的经济效益。因此,为了降低损伤猕猴桃造成的经济损失,采用快捷、有效的方法实现损伤猕猴桃的早期检测具有重要的意义。近年来,近红外光谱检测技术以其高效、无损、易操作等诸多优点受到广泛关注,并被广泛应用于食品、医药及化工等各个领

收稿日期: 2012-06-01 修回日期: 2012-07-16

* 国家自然科学基金资助项目(31171720)和陕西省自然科学基金资助项目(2011JM3005)

作者简介: 郭文川,教授,博士,主要从事农产品和食品品质检测技术研究,E-mail: guowenchuan69@126.com

域^[1-4]。现有应用近红外光谱检测水果品质特性的研究大多集中在定量预测各类品质指标方面^[5-8]。在定性分析方面的研究相对较少,且局限于果品的产地、品种以及病变果的识别等方面^[9-11],在损伤猕猴桃与无损猕猴桃的识别分类方面的研究尚未见报道。本文以猕猴桃为研究对象,利用无损猕猴桃和发生损伤 1 d 的碰撞损伤猕猴桃与挤压损伤猕猴桃的近红外漫反射光谱,结合化学计量学方法进行损伤猕猴桃的检测。

1 材料与方法

1.1 仪器设备

采用德国 Bruker 公司的 MPA 型傅里叶变换近红外光谱仪采集光谱。光谱扫描范围: 12 000 ~ 4 000 cm^{-1} (833 ~ 2 500 nm), 共采集 2 074 个波长; 扫描次数: 32 次; 分辨率: 8 cm^{-1} ; 探头视场角: 90°; 采集条件: 室温 (23 ~ 25℃), 以仪器内置背景为参比。光谱数据由 OPUS 6.5 (Bruker Optics, Germany) 软件进行采集和导出, 分析软件采用 Unscrambler 9.8 (CAMO, Norway), SPSS 20.0 (IBM, USA) 和 Matlab 7.11 (The Math Works, USA)。

1.2 实验材料及光谱采集

供试样品为 2010 年 9 月 16 日采摘于陕西杨凌某猕猴桃园的 280 个形状规整、无损伤、无病斑、成熟良好的“西选二号”猕猴桃。随机选取 100 个样品, 用万能试验机以 10 mm/min 的速度缓慢挤压使每个果实发生 5 mm 变形。再随机选取 100 个样品依次从 0.5 m 高度自然跌落到瓷砖地面, 形成约 4 cm^2

损伤表面积的碰撞损伤果。将 100 个挤压损伤样品、100 个碰撞损伤样品和 80 个未经任何处理的无损样品标记后装于保鲜袋中置于 2℃ 下冷藏。24 h 后, 从冷藏室取出置于室温 (23 ~ 25℃) 下 5 h, 使样品温度升至室温且均匀后用于光谱采集。摄谱时光纤探头与猕猴桃果品直接接触, 在果实赤道部位选取均匀分布且无缺陷的 2 点进行测定, 并取两点平均光谱作为样品光谱, 280 个实验样品的原始光谱如图 1 所示。

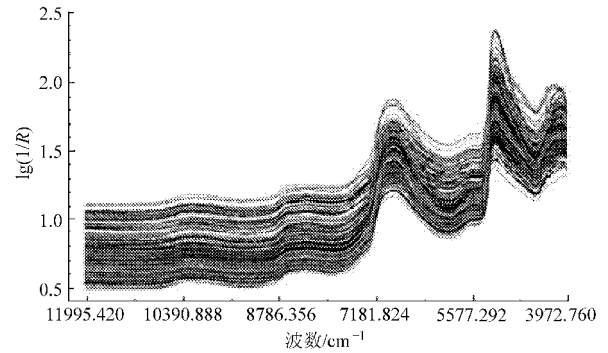


图 1 280 个猕猴桃的原始近红外光谱图

Fig. 1 Original near infrared spectra of 280 kiwifruits

1.3 数据分析与处理方法

1.3.1 Kennard - Stone 算法划分定标集

Kennard - Stone 算法是一种基于样本光谱间欧氏距离的定标集划分方法, 通过将光谱差异较大的样品放入训练集, 而将其余光谱相近的样品归入预测集, 以此充分保证定标集样品的代表性和均匀性^[12]。采用 Kennard - Stone 算法按 4:1 比例分别对 3 种实验样品进行划分, 将各自划分结果汇总后形成总体样品的训练集与预测集, 划分结果如表 1 所示。

表 1 Kennard - Stone 算法划分样本集结果

Tab. 1 Results of sample sets selected by Kennard - Stone algorithm

实验样品	碰撞损伤猕猴桃		挤压损伤猕猴桃		无损猕猴桃		总和	
	训练集	预测集	训练集	预测集	训练集	预测集	训练集	预测集
样本分布个数	80	20	80	20	64	16	224	56

1.3.2 连续投影算法

连续投影算法 (Successive projection algorithm, SPA) 作为一种前向循环的变量选择方法, 能够有效消除各波长变量之间的线性相关影响, 降低模型的复杂度, 以其简便、快速的特点在光谱信息特征变量的优选中被广泛使用^[13]。

1.3.3 Fisher 判别模型

Fisher 判别亦称典型判别, 通过寻找一个能使 N 维空间的某点降为一维数值的线性表达式, 然后应用这个表达式把 N 维空间中的已知类别总体以及未知类别归属的样本都变换为一维数据, 再根

据其间的亲疏程度判定未知类别的归属^[10]。

1.3.4 误差反向传播网络

误差反向传播 (Error back propagation, BP) 网络是通过信号正向传播和误差反向传递操作周而复始的进行, 使权值不断得到调整, 从而指导网络使输出误差减少到可接受的程度或达到预先设定的学习次数为止, 可很好地解决样品的判别分类问题。

1.3.5 最小二乘支持向量机

最小二乘支持向量机 (Least square support vector machine, LSSVM) 是基于传统 SVM 的一种新兴的非线性定量校正方法, 能有效克服经典支持向

量机模型复杂、求解速度慢的弊端,以求解一组线性方程组代替经典 SVM 中复杂的二次优化问题,从而提高建模效率,较好地解决小样本、非线性等实际问题^[14],极大改善近红外光谱的定性、定量模型的预测性能。

2 结果与分析

2.1 近红外光谱特征分析

从图 1 可以看出,3 种猕猴桃的近红外光谱之间的相似性高、重叠严重。在 $6\ 900$ 、 $5\ 200\ \text{cm}^{-1}$ 附近有明显吸收峰,但仅从光谱特征上难以区分损伤果与无损果。因此,需要基于光谱特征信息结合化学计量学方法对无损果和损伤果进行有效识别分类。

2.2 光谱主成分提取

通过对样本光谱数据进行主成分分析(PCA),可得到 3 种猕猴桃的主成分得分聚类图(图 2)。

由图 2 可以看出,碰撞损伤猕猴桃能较明显地与其他 2 种样品区分,而挤压损伤猕猴桃与无损猕

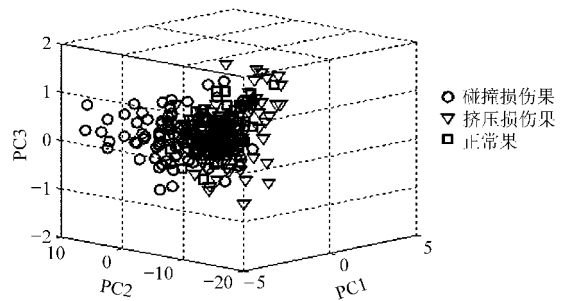


图 2 前 3 个主成分得分聚类图

Fig. 2 Principal component scores cluster plot of the first three principal components

猴桃的聚类中心相近,不易区分。经提取的前 10 个主成分对原始光谱的累积贡献率见表 2。由表 2 可知,前 3 个主成分能代表原始光谱 99% 的信息,但因损伤果与无损果间的光谱差异性较小、重叠严重,若选取过少的主成分可能丢失有效信息,影响判别效果。为充分保证模型具有较好的判别性能与较少的计算量,选取累积贡献率增加不小于 0.05% 的前 7 个主成分用于建立判别模型。

表 2 前 10 个主成分累积贡献率

Tab. 2 Cumulative variance explained for the first ten principal components

主成分	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	%
累积贡献率	70.92	97.76	99.48	99.69	99.84	99.89	99.94	99.96	99.98	99.99	

2.3 SPA 算法优选特征波长

应用 SPA 算法对贮藏期间样品集光谱数据进行特征变量选取,指定波长数范围为 3 ~ 20,根据均方根误差(RMSE)值确定选取的最佳特征光谱数量。SPA 优选过程中,样本 RMSE 随特征波长数目的变化曲线如图 3 所示,点 A 对应横坐标值表示样品集原始光谱经 SPA 优选后的最佳特征波长数目。图 3 表明,当波长数目大于 5 时, RMSE 随波长数目增加变化很小。因此,经 SPA 优选出的 5 个特征波数大致分布如图 4 所示,分别为 $10\ 826.735$ 、 $7\ 679.383$ 、 $7\ 343.820$ 、 $7\ 166.396$ 和 $5\ 311.155\ \text{cm}^{-1}$ 。

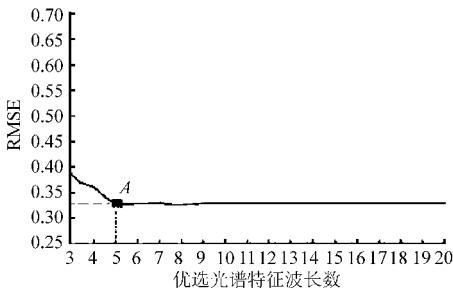


图 3 SPA 特征波长优选过程中 RMSE 的变化

Fig. 3 Changing curves of RMSE during selecting optimal characteristic wavelength

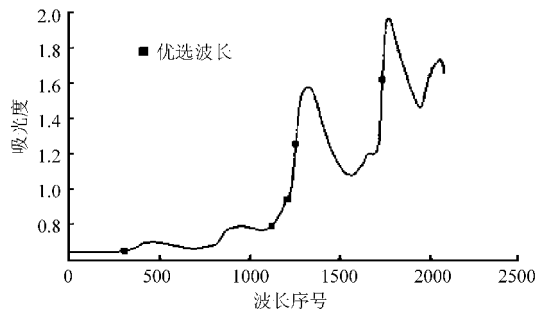


图 4 SPA 算法优选的波长

Fig. 4 Selected wavelengths by SPA algorithm

2.4 判别模型的建立与预测分析

2.4.1 Fisher 判别模型

用数字 1、2 和 3 分别代表碰撞损伤猕猴桃、挤压损伤猕猴桃与无损猕猴桃,因全光谱(Full spectrum, FS)下 2 074 个波长的 Fisher 函数的建立计算量很大,模型建立用时太长,没有任何实际意义。所以,利用 SPSS 软件建立校正集样品的 PCA - Fisher 模型与 SPA - Fisher 模型,所建各 Fisher 判别函数系数及常量如表 3 所示。再将预测集样品各特征变量分别代入 3 个类别函数 F_1 、 F_2 和 F_3 进行计算,根据比较结果将样品归入计算结果最大的对应类别中。

表3 Fisher 判别函数参数值

Tab.3 Parameter values of Fisher discriminant function

判别函数系数	PCA - Fisher			SPA - Fisher		
	碰撞损伤猕猴桃 (F_1)	挤压损伤猕猴桃 (F_2)	无损猕猴桃 (F_3)	碰撞损伤猕猴桃 (F_1)	挤压损伤猕猴桃 (F_2)	无损猕猴桃 (F_3)
变量1	0.634	-0.242	-0.476	5 801.724	7 600.033	8 578.030
变量2	-1.025	0.637	0.497	-4 773.792	-6 379.211	-7 170.527
变量3	-0.013	-0.223	0.127	-1 846.455	-2 434.330	-2 779.815
变量4	-10.645	4.440	8.395	912.847	1 182.565	1 299.436
变量5	6.132	-2.463	-6.107	338.915	429.839	479.185
变量6	-10.870	2.963	10.128			
变量7	4.080	-2.786	-1.372			
常量	-2.962	-1.573	-2.142	-273.491	-309.708	-346.683

2.4.2 BP 神经网络模型

将 224 个训练集样本全光谱下 2 074 个波长、PCA 提取主成分得分变量与 SPA 算法优选特征波长分别作为网络输入变量建立 3 层 BP 神经网络模型,其中输入层传递函数选取为 tansig,隐含层传递函数为 logsig,输出层传递函数为 trainscg,目标均方差设置为 0.01,网络输出 3 类猕猴桃,网络其他各参数设置如表 4 所示。充分训练后的网络将用于预测集样品的识别。

表4 BP 网络的最优网络结构与参数

Tab.4 Optimal parameters and structure of BP - ANN

参数	FS - BP	PCA - BP	SPA - BP
输入层	2074	7	5
隐含层	20	10	7
输出层	3	3	3
学习率	0.3	0.2	0.3
训练次数	103	64	83

2.4.3 LSSVM 模型

采用径向基核函数(RBF),其中参数回归误差权重 γ 与 RBF 核函数的核参数 δ^2 的选取通过二次多维非线性无约束寻优法与交互验证对训练集进行训练实现。LSSVM 建立时 tunelssvm 函数中交叉次数(L-fold)设置为 20,各模型参数 γ 和 δ^2 的选取结果如表 5 所示。

表5 LSSVM 模型参数

Tab.5 Parameters of LSSVM models

参数	FS - LSSVM	PCA - LSSVM	SPA - LSSVM
γ	58 506.72	1 949.37	4.23
δ^2	42 686.94	107.10	12.21

2.4.4 各判别模型预测结果分析

根据以上建立的各种模型分别对预测集样品进行判别分析,各模型判别结果见表 6。

表6 Fisher、BP 和 LSSVM 判别模型对预测集样品识别结果

Tab.6 Recognition results using discriminant model of Fisher, BP and LSSVM

判别模型	碰撞损伤猕猴桃		挤压损伤猕猴桃		无损猕猴桃		总正确 判别率/%
	误判数/个	正确判别率/%	误判数/个	正确判别率/%	误判数/个	正确判别率/%	
PCA - Fisher	1	95	4	80	5	68.8	82.1
SPA - Fisher	1	95	4	80	4	75.0	83.9
FS - BP	2	90	3	85	2	87.5	87.5
PCA - BP	2	90	1	95	2	87.5	91.1
SPA - BP	0	100	5	75	1	93.8	89.3
FS - LSSVM	0	100	1	95	1	93.8	96.4
PCA - LSSVM	3	75	2	90	2	87.5	87.5
SPA - LSSVM	0	100	1	95	0	100	98.2

注:表中正确判别率为判别正确的样品个数除以总样品个数再乘以 100%。

从表 6 可以看出, Fisher 判别分析中 SPA - Fisher 模型对预测集样品的总正确判别率为 83.9%,略好于 PCA - Fisher 模型的 82.1%。BP 神经网络判别分析中 PCA - BP 模型的判别性能最优,总正确判别率达到 91.1%。LSSVM 模型判别分析

中,FS - LSSVM 和 SPA - LSSVM 模型对预测集样品的总正确判别率分别达到 96.4% 和 98.2%,均显著高于 PCA - LSSVM,具有很好的判别性能,且因 SPA - LSSVM 模型的输入变量少、建模效率高,所以在实际应用中 SPA - LSSVM 模型要优于 FS - LSSVM。

总体而言,3种模型输入变量选取方法中,SPA算法能够很好地提取特征变量,并减少计算量;3种判别模型均可用于识别早期损伤猕猴桃与无损猕猴桃,但SPA-LSSVM判别效果最好,PCA-Fisher模型识别效果最差。

3 结束语

分别建立了猕猴桃样品的Fisher判别模型、BP神经网络判别模型与LSSVM判别模型,并综合比较

了全光谱波长、PCA提取变量和SPA优选特征波长参与建模下各模型的判别效果。研究表明,SPA优选方法能够有效地减少模型输入变量、提高建模效率并满足预测要求;3种不同判别模型均可满足识别早期损伤猕猴桃的实际需要,而LSSVM模型的识别性能最好,其中SPA-LSSVM模型对预测集碰撞损伤、挤压损伤与无损猕猴桃的正确识别率分别达到100%、95%和100%,总正确识别率为98.2%,为无损、准确识别早期损伤猕猴桃提供了技术参考。

参 考 文 献

- Gardner C M, Tan H, Hull E L, et al. Detection of lipid core coronary plaques in autopsy specimens with a novel catheter-based near-infrared spectroscopy system[J]. *JACC: Cardiovascular Imaging*, 2008, 1(5): 638~648.
- Gomez A H, He Y, Pereira A G. Non-destructive measurement of acidity, soluble solids and firmness of Satsuma mandarin using Vis/NIR-spectroscopy techniques[J]. *Journal of Food Engineering*, 2006, 77(2):313~319.
- 李文龙, 邢丽红, 薛东升, 等. 一种基于近红外光谱技术的熊胆粉鉴别方法[J]. *光谱学与光谱分析*, 2011, 31(3):673~676.
Li Wenlong, Xing Lihong, Xue Dongsheng, et al. An authentication method of bear bile powder based on the near infrared spectroscopy[J]. *Spectroscopy and Spectral Analysis*, 2011, 31(3):673~676. (in Chinese)
- 周子立, 张瑜, 何勇, 等. 基于近红外光谱技术的大米品种快速鉴别方法[J]. *农业工程学报*, 2009, 25(8):131~135.
Zhou Zili, Zhang Yu, He Yong, et al. Method for rapid discrimination of varieties of rice using visible NIR spectroscopy[J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2009, 25(8):131~135. (in Chinese)
- Camps C, Christen D. Non-destructive assessment of apricot fruit quality by portable visible-near infrared spectroscopy[J]. *LWT - Food Science and Technology*, 2009, 42(6):1125~1131.
- 刘卉, 郭文川, 岳绒. 猕猴桃硬度近红外漫反射光谱无损检测[J]. *农业机械学报*, 2011, 42(3):145~149.
Liu Hui, Guo Wenchuan, Yue Rong. Non-destructive detection of kiwifruit firmness based on near-infrared diffused spectroscopy[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2011, 42(3):145~149. (in Chinese)
- 刘燕德, 罗吉, 陈兴苗. 可见/近红外光谱的南丰蜜桔可溶性固形物含量定量分析[J]. *红外与毫米波学报*, 2008, 27(2):119~122.
Liu Yande, Luo Ji, Chen Xingmiao. Analysis of soluble solid content in Nanfeng manarain fruit with visible near infrared spectroscopy[J]. *Journal of Infrared and Millimeter Waves*, 2008, 27(2):119~122. (in Chinese)
- 王加华, 韩东海. 基于遗传算法的苹果糖度近红外光谱分析[J]. *光谱学与光谱分析*, 2008, 28(10):2308~2311.
Wang Jiahua, Han Donghai. Analysis of near infrared spectra of apple SSC by genetic algorithm optimization[J]. *Spectroscopy and Spectral Analysis*, 2008, 28(10):2308~2311. (in Chinese)
- 李桂峰, 赵国建, 王向东, 等. 苹果质地品质近红外无损检测和指纹分析[J]. *农业工程学报*, 2008, 24(6):169~173.
- 李顺峰, 张丽华, 刘兴华, 等. 基于主成分分析的苹果霉心病近红外漫反射光谱判别[J]. *农业机械学报*, 2011, 42(10):158~161.
Li Shunfeng, Zhang Lihua, Liu Xinghua, et al. Discriminant analysis of apple moldy core using near infrared diffuse reflectance spectroscopy based on principal component analysis[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2011, 42(10):158~161. (in Chinese)
- 马本学, 饶秀勤, 应义斌, 等. 基于近红外漫反射光谱的香梨类别定性分析[J]. *光谱学与光谱分析*, 2009, 29(12):3288~3290.
Ma Benxue, Rao Xiuqin, Ying Yibin, et al. Qualitative analysis of fragrant pear class based on near infrared diffuse reflectance spectroscopy[J]. *Spectroscopy and Spectral Analysis*, 2009, 29(12):3288~3290. (in Chinese)
- 李华, 王菊香, 邢志娜, 等. 改进的K/S算法对近红外光谱模型传递影响的研究[J]. *光谱学与光谱分析*, 2011, 31(2):362~365.
Li Hua, Wang Juxiang, Xing Zhina, et al. Influence of improved Kennard/Stone algorithm on the calibration transfer in near-infrared spectroscopy[J]. *Spectroscopy and Spectral Analysis*, 2011, 31(2):362~365. (in Chinese)
- 洪涯, 洪添胜, 代芬, 等. 连续投影算法在砂糖橘总酸无损检测中的应用[J]. *农业工程学报*, 2010, 26(增刊2):380~384.
- 吴桂芳, 何勇. 应用可见/近红外光谱进行纺织纤维鉴别的研究[J]. *光谱学与光谱分析*, 2010, 30(2):331~335.
Wu Guifang, He Yong. Identification of varieties of textile fibers by using Vis/NIR infrared spectroscopy technique[J]. *Spectroscopy and Spectral Analysis*, 2010, 30(2):331~335. (in Chinese)