

doi:10.6041/j.issn.1000-1298.2013.09.032

基于特征光谱的草莓品种快速鉴别*

闫润^{1,2} 王新忠¹ 邱白晶¹ 史德林² 孔鹏飞¹

(1. 江苏大学现代农业装备与技术教育部重点实验室, 镇江 212013; 2. 江苏农林职业技术学院, 句容 212400)

摘要:应用近红外光谱技术对草莓品种快速鉴别进行了研究。通过比较7种预处理方法,确定SNV+MAF+2D组合预处理方法最佳。采用相关系数阈值法提取了14个与草莓品种信息相关的特征光谱。建立了基于548~562 nm范围内14个特征光谱的PLS-ANN、PLSR、PCR 3种校正模型。经预测集样品验证,主成分数为9时的PLS-ANN模型最优,其预测相关系数和预测均方根误差分别为0.967 7和0.461。结果表明,通过提取少量特征光谱建立PLS-ANN校正模型能够实现草莓品种的快速鉴别。

关键词:草莓 品种鉴别 偏最小二乘-人工神经网络 近红外光谱

中图分类号: S603.7; S668.4; O657.33 **文献标识码:** A **文章编号:** 1000-1298(2013)09-0182-05

Discrimination of Strawberries Varieties Based on Characteristic Spectrum

Yan Run^{1,2} Wang Xinzhong¹ Qiu Baijing¹ Shi Delin² Kong Pengfei¹

(1. Key Laboratory of Modern Agricultural Equipment and Technology, Ministry of Education, Jiangsu University, Zhenjiang 212013, China 2. Jiangsu Polytechnic College of Agriculture and Forestry, Jurong 212400, China)

Abstract: Near infrared spectroscopy (NIRS) technology was applied to discriminate the variety of strawberries rapidly. Compared with seven kinds of preprocessing methods, “SNV + MAF + 2D” was ascertained as the optimal scheme. Totally 14 characteristic spectrums related to strawberries information were extracted by the correlation coefficient and threshold value method. Three kinds of correction models of PLS – ANN, PLSR and PCR based on wavelength ranging from 548 nm to 562 nm were established. Verified by the prediction set samples, the optimal correction model was PLS – ANN with nine principal components. Prediction correlation coefficient and root mean square error of prediction (RMSEP) of the PLS – ANN correction model was 0.967 7 and 0.461. The results showed that a few characteristic spectrum extracted to establish PLS – ANN correction model would achieve rapid discrimination of strawberries.

Key words: Strawberries Varieties discrimination PLS – ANN Near infrared spectroscopy

引言

不同品种草莓的品质和价格差异大,但外观相似,难以鉴别。常用的感官鉴别易受主观因素影响,结果不可靠;理化实验鉴别虽可靠,但操作繁琐,破坏实验样品,成本高。

近红外光谱技术作为一种无损检测技术,以其效率高、成本低、测量方便,而被国内外学者应用于

水果品种^[1~4]鉴别。但对浆果鉴别的研究较少,已有研究主要集中在蓝莓、草莓可溶性固形物、可滴定酸含量等品质检测方面^[5~7]。在草莓品种鉴别的研究中 Sánchez 等^[8]使用 PLS – DA 等算法对 5 个商业品种草莓进行分类,识别率为 57% ~ 78%。牛晓颖等^[9]使用 BP-ANN 等算法对 3 个品种草莓进行分类,识别率达到 98.29%。但已有研究中利用全光谱定模,模型复杂,效率低。

收稿日期: 2013-03-03 修回日期: 2013-03-26
* 国家自然科学基金资助项目(31271620)、江苏省科技支撑计划资助项目(BE2011346)和江苏省农机三项工程资助项目(NJ2011-46)
作者简介: 闫润,博士生,江苏农林职业技术学院讲师,主要从事光谱应用技术与农业自动化检测研究,E-mail: yanrun123@163.com
通讯作者: 王新忠,教授,主要从事农业装备系统监测与控制研究,E-mail: wangxinzhong@gmail.com

偏最小二乘^[10-16]与神经网络^[17-18]相结合的算法(PLS-ANN)是一种因子分析与非线性有监督训练相融合的算法,适合小样本建模,能有效处理输入与输出之间高度非线性问题,本文利用该算法对3个草莓品种进行鉴别,在对比多种预处理方法的基础上,利用相关系数阈值法提取特征光谱,进行草莓品种快速鉴别建模研究。

1 材料与方法

1.1 实验样品

草莓样品采摘自镇江市句容草莓种植园,选取本地广泛种植的3个品种:丰香、红颊、明宝。清晨采摘,挑拣着色均匀、无畸形、大小相近的3种草莓共计150个。样品采后用保温箱包装运输。保温箱底铺设1 cm厚塑料泡沫,草莓间隔轻放,避免挤伤,送回实验室。剔除有轻微损伤的样本,每种草莓各保留40个样品,共120个,逐一编号,1~40号代表丰香,种类代号为1;41~80号代表红颊,种类代号为2;81~120号代表明宝品种,种类代号为3。此处编号为后续划分校正集和预测集作准备。

1.2 实验仪器和软件

光谱测量仪器使用美国ASD (Analytical Spectral Device)公司生产的Field Spec 3型便携式光谱分析仪。光谱测量范围350~2 500 nm;350~1 000 nm波段采样间隔为1.4 nm,分辨率为3 nm;1 000~2 500 nm波段采样间隔为2 nm,分辨率为10 nm;采样频率10 Hz。光源使用光谱仪配套的石英卤素灯,色温3 200 K。分析软件使用ASD View Spec Pro系统和江苏大学研制的NIRSA软件(3.0版)。

1.3 光谱采集

为避免光源干扰,采集光谱时,将实验室门窗关闭,遮住太阳光,熄灭实验室内其余光源。调整石英卤素灯光源与水平面呈45°角。将光谱分析仪探头固定在三角架上,设定光纤视场为25°,根据被测物直径及视场角推算出探头高度,调整传感器探头位于工作台上方5 cm左右处,垂直于被测物。测量前对光谱分析仪进行白板标定。测定过程中,将样品立于自制黑色托盘上,托盘下铺设黑色台布,旋转托盘,每转120°在草莓赤道圆周处采集一次光谱,每样品共测3点,以3次采集平均值作为一个样品的原始光谱数据。在光谱采集过程中,用空调控制实验室内温度为20℃,相对湿度为70%,采集时间约3 h。

3个品种草莓近红外光谱图如图1所示。可以看出,不同品种的草莓近红外光谱图具有相似性,但仍有区别,主要在500~700 nm和900~1 100 nm波段,500~700 nm的光谱反射率急剧增加,这可能与

草莓表皮呈红色,对红光反射有关。900~1 100 nm的光谱反射率急剧下降,而该区域是C—H基团和O—H基团的吸收带,可能与草莓内部品质有关的物质对光产生了吸收。草莓品种的鉴别可优先考虑上述波段。

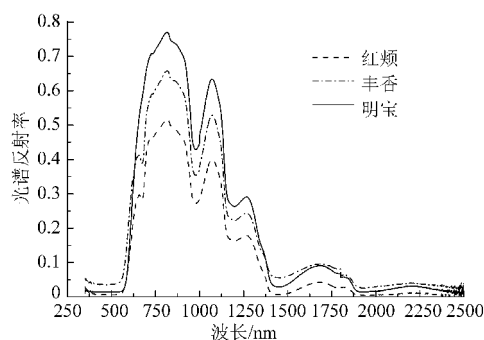


图1 3种草莓的近红外光谱图

Fig.1 Near infrared reflectance spectra of three different varieties of strawberry

2 实验结果与分析

2.1 校正集与预测集的建立

为了后续建模与测试,首先应划分样品,建立校正集和预测集。Kennard-Stone算法是一种有效的划分方法。该方法依据预先设定的数目,计算两两样品之间的距离,逐一进行筛选,直至全部样品划分结束。为达到较好预测效果,建模集与预测集样品数目常按3:1比例分配。本实验草莓样品总数120,计算得建模集样品数目90,预测集数目30。将ASD View Spec Pro输出的近红外光谱数据输入NIRSA软件,用Kennard-Stone算法,结果得到红颊、丰香、明宝3个品种的校正集样品数目分别为30、29、31;预测集样品数目分别为10、11、9。

2.2 不同品种草莓的聚类分析

主成分分析(PCA)是一种数据降维方法,在不丢失信息的前提下,将原始光谱数据转换成表征原始数据结构特征的较少数目的新变量。对3种草莓的120个样品进行主成分分析,所得聚类结果如图2所示。图2中横坐标表示每个样品第1主成分(PCA1)得分值,120个样品累积得分69.38%;纵坐标表示每个样品第2主成分(PCA2)得分值,120个样品累积得分22.16%。图2中红颊品种样本大部分能被区分,丰香和明宝虽有重叠,仍呈现较明显的条状聚类,但整体看每个品种的聚合度不高,鉴别效果不佳,故需选择其他算法定量分析。另外从图2中还看到经Kennard-Stone算法划分的草莓校正集与预测集所处位置,分布是合理的,有代表性。

2.3 光谱预处理

近红外光谱的采集易受样品状态、测量方式及

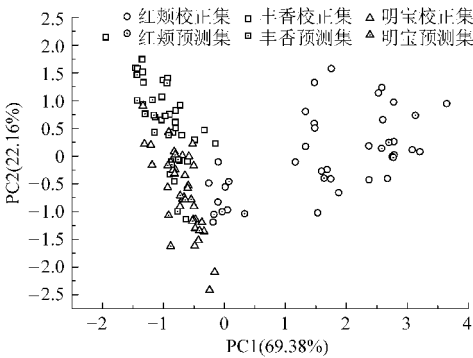


图2 全部样品聚类效果图

Fig.2 Cluster analysis of all samples

测量条件影响,如样品的颗粒不均、仪器的随机噪声、基线漂移、光散射等。有必要对草莓光谱数据进行预处理。光谱预处理方法主要有平均平滑(MAF)、卷积平滑(SGF)、一阶导数(1D)、二阶导数(2D)、多元散射校正(MSC)和标准归一化校正(SNV)等。近红外光谱往往同时被多种因素干扰,仅用单一光谱预处理方法可能得不到较好效果。本实验考虑将上述单一光谱预处理方法合理组合,通过比较,得到最佳光谱预处理方法。在 PLS-ANN 模型里对比 7 种预处理方法得到的预测结果如表 1 所示。表 1 中 R_p 为预测相关系数,表征模型的预测能力。从表 1 中看出,草莓近红外光谱经预处理后相关系数有所提高,而均方根误差则显著下降,其中 SNV + MAF + 2D 组合为最佳预处理方法。

表 1 不同预处理方法结果对比

Tab.1 Comparison of different preprocessings

预处理方法	R_p	预测均方根误差
无	0.732 9	0.643
1D	0.873 9	0.522
2D	0.881 5	0.484
MAF + 2D	0.925 5	0.462
SGF + 1D	0.931 6	0.435
SNV + MAF + 2D	0.932 5	0.385
MSC + SGF + 1D	0.901 4	0.492

2.4 特征光谱的提取

采集的草莓近红外光谱范围为 350 ~ 2 500 nm, 共计 2 151 个波长,信息有冗余,建模耗时,难以在实际生产中应用。为了提取较少数目的特征光谱,快速建立校正模型,本实验采用相关系数阈值法提取草莓的特征光谱。相关系数阈值法是一种提取特征光谱的有效方法。该方法先在全光谱范围内进行相关分析,得到每个波长与品种特征相关的相关系数,系数范围 -1 ~ 1;再设定阈值,阈值的绝对值在 0.5 ~ 1 之间。阈值取值越大,提取的特征光谱个数越少。应用相关系数阈值法提取的特征光谱如图 3

所示。图 3a 为阈值大于等于 0.7 时提取的特征光谱(设定阈值大于等于 0.7 的方法称为阈值法 1),得到图中被命名为“ I ”、“ II ”、“ III ”、“ IV ”的 4 个光谱波段:528 ~ 530 nm、32 ~ 533 nm、537 ~ 566 nm 和 585 ~ 597 nm,共计 44 个特征波长点。图 3b 为阈值大于等于 0.8 时提取的特征光谱(设定阈值大于等于 0.8 的方法称为阈值法 2),满足条件的有 528 ~ 529 nm、548 ~ 562 nm、598 ~ 602 nm 波段。但是 528 ~ 529 nm 和 598 ~ 602 nm 波段的相关系数在阈值临界点附近,并考虑到对校正模型误差的影响,实取在图中被命名为“ I ”的 548 ~ 562 nm 光谱波段的 14 个特征波长点。

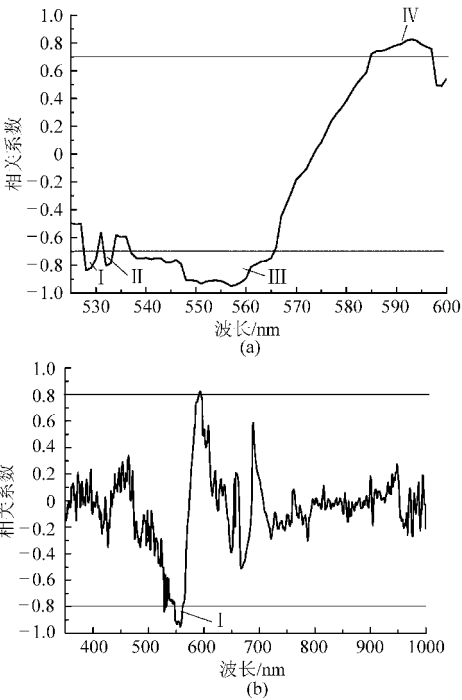


图 3 阈值法提取的特征光谱

Fig.3 Characteristic spectra extracted by threshold value method

为验证图 3a 阈值大于等于 0.7 和图 3b 阈值大于等于 0.8 特征光谱的有效性,将图 3a、图 3b 中提取的特征光谱分别输入 PLS-ANN 模型,预测结果如表 2 所示。从表 2 中看出,图 3b 中的 14 个特征光谱的预测效果最好。预测集草莓样品种类的真实值与预测值如表 3 所示。从表 3 中看出仅将一个红颊草莓样品误判为明宝,识别率为 96.67%。

表 2 特征光谱建模效果的对比

Tab.2 Comparison of models with different characteristic spectrum

方法	波长个数	阈值	R_p	预测均方根误差
全波长法	2 151	0	0.932 5	0.385
阈值法 1	44	0.7	0.942 9	0.542
阈值法 2	14	0.8	0.967 7	0.461

表 3 应用 PLS – ANN 模型的 30 个预测样品的预测结果

Tab.3 Prediction results of 30 samples with

PLS – ANN model

预测样本 编号	真实值	预测值	预测样本 编号	真实值	预测值
4	1	1.038 9	67	2	2.192 3
6	1	0.929 1	69	2	2.231 9
8	1	1.013 3	72	2	2.027 2
10	1	1.026 6	74	2	2.045 2
13	1	0.980 1	76	2	2.159 3
15	1	1.063 0	78	2	2.242 8
23	1	1.238 5	84	3	3.017 2
25	1	1.166 5	91	3	3.051 7
27	1	0.983 0	94	3	2.996 7
30	1	0.914 2	96	3	3.091 4
32	1	1.031 0	98	3	3.086 6
51	2	2.113 2	106	3	2.832 6
59	2	2.259 7	108	3	3.012 4
61	2	1.895 4	115	3	2.927 6
63	2	2.851 2	117	3	2.989 6

2.5 校正模型的比较

上述光谱预处理及特征光谱提取均用 PLS – ANN 模型实现。为探索进一步提高草莓品种识别率,本实验尝试采用 PLS – ANN、偏最小二乘回归 (PLSR) 及主成分回归 (PCR) 3 种模型建模,预测结果如表 4 所示。从表 4 中看出,主成分数为 9 时,PLS – ANN 模型最优。此时设定的 PLS – ANN 主成分数为 9,输入层节点数为 9,隐层节点数为 14,训

练误差 0.001,动量因子 0.95。

表 4 建模方法比较

Tab.4 Comparison of different models

建模方法	主成分数	R_p	预测均方根误差
PLS – ANN	8	0.934 7	0.711
PLS – ANN	9	0.967 7	0.461
PLS – ANN	10	0.901 6	0.826
PLSR	3	0.882 4	0.513
PLSR	4	0.849 3	0.549
PCR	2	0.888 0	0.505
PCR	8	0.872 2	0.523

3 结论

- (1)通过 7 种预处理方法对比,确定了最佳预处理方法为 SNV + MAF + 2D 组合。
- (2)采用相关系数阈值法可以从全波长范围有效提取出草莓光谱 548 ~ 562 nm 波段的 14 个特征光谱,所建 PLS – ANN 模型识别率为 96.67%。
- (3)建立了基于 14 个特征光谱的 PLS – ANN、PLSR、PCR 3 种校正模型,通过预测集验证,确定当主成分数为 9 时的 PLS – ANN 模型最优,其预测集相关系数和预测均方根误差分别为 0.967 7 和 0.461。
- (4)上述结果表明通过提取少量特征光谱建立 PLS – ANN 校正模型,能实现草莓品种的快速鉴别。

参 考 文 献

1 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):3 288 ~ 3 290.
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):3 288 ~ 3 290. (in Chinese)

2 曹芳,吴迪,何勇,等. 基于可见-近红外反射光谱技术的葡萄品种鉴别方法的研究[J]. 光学学报,2009,29(2):537 ~ 540.
Cao Fang,Wu Di,He Yong, et al. Variety discrimination of grapes based on visible-near reflection infrared spectroscopy[J]. Acta Optica Sinica,2009,29(2):537 ~ 540. (in Chinese)

3 He Y, Li X, Shao Y. Quantitative analysis of the varieties of apple using NIR by PCA and BP model[J]. Advances in Artificial Intelligence Lecture Notes in Computer Science,2005,3 809:1 053 ~ 1 056.

4 Xie L,Ying Y,Ying T. Classification of tomatoes with different genotypes by visible and short-wave near-infrared spectroscopy with least-squares support vector machines and other chemometrics[J]. Journal of Food Engineering,2009,94(1):34 ~ 39.

5 金长江. 基于近红外光谱与机器视觉技术的浆果品质检测研究[D]. 哈尔滨:东北农业大学,2011.

6 王姗姗,孙红男,陈健,等. 基于近红外光谱技术检测蓝莓可溶性固形物含量[J]. 食品工业科技,2012,33(1):340 ~ 341.
Wang Shanshan, Sun Hongnan, Chen Jian, et al. Determination of soluble solids content of blueberries with near infrared spectroscopy[J]. Science and Technology of Food Industry,2012,33(1):340 ~ 341. (in Chinese)

7 Sinelli N, Spinardi A, Di Egidio V, et al. Evaluation of quality and nutraceutical content of blueberries (*Vaccinium corymbosum* L.) by near and mid-infrared spectroscopy[J]. Postharvest Biology and Technology,2008,50(1):31 ~ 36.

8 Sánchez M,de la Haba M J,Benítez-López M, et al. Non-destructive characterization and quality control of intact strawberries based on NIR spectral data[J]. Journal of Food Engineering,2012,110(1):102 ~ 108.

9 牛晓颖,邵利敏,赵志磊,等. 基于 BP-ANN 的草莓品种近红外光谱无损鉴别方法研究[J]. 光谱学与光谱分析,2012,32(8):2095 ~ 2099.
Niu Xiaoying,Shao Limin,Zhao Zhilei, et al. Nondestructive discrimination of strawberry varieties by NIR and BP-ANN[J].

- Spectroscopy and Spectral Analysis,2012,32(8):2095~2099. (in Chinese)
- 10 张淑娟,张海红,赵艳茹,等. 鲜枣可溶性固形物可见/近红外光谱检测建模方法比较[J]. 农业机械学报,2012,43(3):108~112.
Zhang Shujuan,Zhang Haihong,Zhao Yanru, et al. Comparison of modeling methods of fresh jujube soluble solids measurement by NIR spectroscopy[J]. Transactions of the Chinese Society for Agricultural Machinery,2012,43(3):108~112. (in Chinese)
- 11 Mendoza F,Lu R,Ariana D, et al. Integrated spectral and image analysis of hyperspectral scattering data for prediction of apple fruit firmness and soluble solids content[J]. Postharvest Biology and Technology,2011,62(2):149~160.
- 12 Meza-Márquez O G,Gallardo-Velázquez T,Osorio-Revilla G. Application of mid-infrared spectroscopy with multivariate analysis and soft independent modeling of class analogies (SIMCA) for the detection of adulterants in minced beef[J]. Meat Science,2010,86(2):511~519.
- 13 邹小波,陈正伟,石吉勇,等. 基于近红外高光谱图像的黄瓜叶片色素含量快速检测[J]. 农业机械学报,2012,43(5):152~156.
Zou Xiaobo,Chen Zhengwei,Shi Jiyong, et al. Rapid detection of cucumber leaves pigments based on near infrared hyper-spectral image technology[J]. Transactions of the Chinese Society for Agricultural Machinery,2012,43(5):152~156. (in Chinese)
- 14 刘燕德, 万常澜. 芝麻油掺伪的近红外透射光谱检测技术[J]. 农业机械学报,2012,43(7):136~140.
Liu Yande,Wan Changlan. Analysis of sesame oil adulteration using near infrared transmission spectroscopy[J]. Transactions of the Chinese Society for Agricultural Machinery,2012,43(7):136~140. (in Chinese)
- 15 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)
- 16 石吉勇,邹小波,赵杰文,等. 黄瓜叶片叶绿素含量近红外光谱无损检测[J]. 农业机械学报,2011,42(5):178~182.
Shi Jiyong,Zou Xiaobo,Zhao Jiewen, et al. NIR spectra in non-invasive measurement of cucumber leaf chlorophylls content[J]. Transactions of the Chinese Society for Agricultural Machinery,2011,42(5):178~182. (in Chinese)
- 17 郭文川,王铭海,岳绒. 基于近红外漫反射光谱的损伤猕猴桃早期识别[J]. 农业机械学报,2013,44(2):142~146.
Guo Wenchuan,Wang Minghai,Yue Rong. Early recognition of bruised kiwifruit based on near infrared diffuse reflectance spectroscopy[J]. Transactions of the Chinese Society for Agricultural Machinery,2013,44(2):142~146. (in Chinese)
- 18 冯雷,陈双双,冯斌,等. 大豆豆荚炭疽病严重度的光谱检测[J]. 农业机械学报,2012,43(8):175~179.
Feng Lei, Chen Shuangshuang, Feng Bin, et al. Spectral detection on disease severity of soybean pod anthracnose [J]. Transactions of the Chinese Society for Agricultural Machinery,2012,43(8):175~179. (in Chinese)