文章编号:1001-9014(2006)06-0417-04

基于主成分和多类判别分析的可见-红外光谱水蜜桃品种鉴别新方法

李晓丽[1], 胡兴越[2], 何 勇[1]
(1. 浙江大学 生物系统工程与食品科学学院,浙江 杭州 310029;
2. 浙江大学邵逸夫医院,浙江 杭州 310016)

摘要:提出了一种用可见-近红外漫反射光谱技术快速鉴别水蜜桃品种的新方法.应用可见-近红外光谱仪测定三个品种水蜜桃的光谱曲线,再用主成分分析法对不同品种样本进行聚类分析,获取了水蜜桃可见-近红外光谱的特征信息,同时结合多类判别分析技术建立水蜜桃品种鉴别的模型.对经过预处理的光谱数据进行主成分分析,分析表明,以样本在第一主成分和第二主成分上的得分做出的二维散点图,对不同种类水蜜桃具有很好的聚类,能定性区分不同种类水蜜桃;经过主成分分析得到的前8个主成分的累积可信度已达94.38%,说明这8个变量能够代表绝大部分原始光谱的信息.从75个样本中随机抽取60个样本用于建立8个主成分变量的多类判别分析品种鉴别模型,余下的15个样本用于验证,准确率为100%.说明本文提出的方法具有明显的分类和鉴别作用.

关 键 词:可见-近红外光谱;水蜜桃;主成分分析;多类判别分析;鉴别
中图分类号:S123 **文献标识码**:A

NEW APPROACH OF DISCRIMINATION OF VARIETIES OF JUICY PEACH BY NEAR INFRARED SPECTRA BASED ON PCA AND MDA MODEL

LI Xiao-Li[1], HU Xing-Yue[2], HE Yong[1]
(1. College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310029, China;
2. Sir Run Run Shaw Hospital, Zhejiang University, Hanzhou 310026, China)

Abstract: A new method for discrimination of varieties of juicy peach by means of visible-near infrared spectroscopy (NIRS) was developed. First, the spectral curves of three varieties juicy peaches were measured by spectrometer; the pretreated spectra data of juicy peach were analyzed through principal component analysis (PCA). Then the diagnostic information from PCA was used as inputs of multiple discriminant analysis (MDA) for pattern recognition. The 2-dimontional plot was drawn with first and second principal components, which indicated that it was a good clustering analysis for classification varieties of juicy peach. The result of the analysis suggested that the reliabilities of first 8 principal components were more than 94.38%. 60 samples from three varieties selected randomly. Then they were used to build discriminating model. 15 unknown samples were validated by this model. The recognition rate is 100%. This model is reliable and practicable. So this study can offer a new approach to the fast discrimination of varieties of juicy peach.

Key words: visible-near infrared spectra; juicy peach; principal component analysis (PCA); multiple discriminant analysis (MDA); discrimination

引言

桃原产我国,栽培历史悠久,种质资源丰富,居核果类果树首位.水蜜桃(Prunus persica)是我国广泛种植的桃,它以果肉柔嫩、甜蜜多汁、香气浓郁、风味独特而享誉大江南北,但是我国目前栽培的水蜜桃品种较多,品种间差异较大、良莠不齐,需要研究一种简单、快速、无损的水蜜桃品种鉴别方法,不仅

收稿日期:2005-12-19,**修回日期**:2006-05-17 **Received date**: 2005-12-19, **revised date**: 2006-05-17
基金项目:国家自然科学基金(30671213)、高等学校优秀青年教师教学科研奖励计划(02411)、高等学校博士学科点专项科研基金(20040335034)和浙江省重大科技攻关(2005C12029)资助项目
作者简介:李晓丽(1982-),女,四川广安人,浙江大学生物系统工程与食品科学学院博士生,主要研究方向为应用机器视觉技术进行农产品品质分析.

具有极大的经济价值，而且对于桃产业的健康持续发展具有重要的意义. 本文以光谱技术为基础研究了水蜜桃的品种快速无损鉴别方法.

现代可见-近红外光谱分析技术，可充分利用全谱段或多波长下的光谱数据进行定性或定量分析. 由于可见-近红外光谱技术分析具有速度快、效率高、成本低、测试重现性好、测量方便等特点，已被越来越多地应用于食品工业、石油化工、制药工业等领域. 有学者研究利用可见-近红外光谱技术区别咖啡品种[1]，道地山药[2]、苹果品种[3]、检测皮棉杂质[4]等. 但是可见-近红外波谱段具有信息量大、有噪声干扰、波谱重叠等特征，如果直接运用原始光谱数据进行判别分析往往导致模型精度低、稳定性差. 我们应用主成分分析(PCA)结合多类判别分析方法来挖掘光谱中的有用信息，实现水蜜桃品种的快速区别.

主成分分析在不丢失主要光谱信息的前提下，选择为数较少的新变量来代替原来的变量，解决了由于谱带的重叠而无法分析的困难. 多类判别分析法是一种集"有效特征选择与状态识别"功能于一体的统计分析法，我们将两者有机地结合起来建立水蜜桃不同品种的可见-近红外光谱鉴别模型.

1 材料与方法

1.1 仪器设备

实验使用美国 ASD(Analytical Spectral Device)公司的 Handheld FieldSpec 光谱仪，其光谱采样间隔(波段宽)1.5nm，测定范围 325 ~ 1075nm，分辨率 3.5 nm. 光谱数据以 ASCII 码形式导出进行处理，分析软件为 ASD ViewSpec Pro, Unscramble 和 DPS.

1.2 样品来源及光谱的获取

从超市买来 3 种水蜜桃，蜜露水蜜桃(产地浙江省奉化)、大白桃水蜜桃(产地浙江省金华)、红仙玖水蜜桃(山东省)各 25 个，选择大小均匀的个体，避免试验中水蜜桃个体与可见-近红外光谱仪的距离剧烈变化，共计 75 个样本. 光谱仪置于水蜜桃的上方，探头视场角为 20°，对每一个水蜜桃扫描 30 次，从水蜜桃的赤道部位等距的依次取 3 个部位，3 个桃赤道部相差约 120°.

1.3 光谱数据预处理

为了去除来自高频随机噪音、基线漂移、样本不均匀、光散射等影响，需要进行光谱预处理. 采用 Move average 平滑法，选用平滑点数为 9，此时能很好滤除各种因素产生的高频噪音，再进行二阶求导处理，消除基线漂移. 光谱曲线在首端和末端有较大噪音，为了减少噪音，提高精度，去掉前 75nm 和后 75nm 波段，选用 400 ~ 1000nm 光谱范围进行研究[5,6].

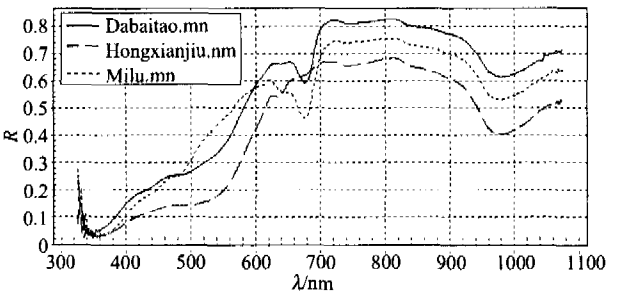

图 1 3 类水蜜桃的典型可见-近红外光谱反射图

Fig. 1 Near infrared reflectance spectra of three different variety juicy peaches

1.4 多类判别分析法

判别分析主要是判别未知样本应划归哪一个已知总体. 在判别未知样本前，首先要研究不同总体的性质和特征，根据已知总体的多种观测指标建立判别函数，并以它作为样本划归某一总体的依据. 最后将待判别样品的有关数值代入判别函数，即能判别该样本的类别.

2 试验结果与分析

2.1 水蜜桃品种模型的建立

3 个品种水蜜桃典型可见-近红外反射光谱曲线如图 1 所示. 横坐标为波长范围 325 ~ 1075nm，纵坐标为光谱漫反射率. 从图 1 中可以看出，3 种水蜜桃光谱范围内有很大的差异，但是存在较大的基线漂移，所以运用二阶求导消除基线漂移，如图 2 所示. 图 2 中不同品种水蜜桃在 600 ~ 700nm 光谱范围内的谱线差异更明显. 所以，不同品种水蜜桃的光谱图有明显区别，并具有一定的特征性和指纹性，这一差异为水蜜桃的不同品种鉴别奠定了数学基础[7].

不同品种水蜜桃的光谱曲线差异，不仅跟它们的表皮特征有关而且跟它们的内部品质如含糖量、酸度等密切相关. 水蜜桃含糖量(可溶性固性物)采用 WYT-4 型手持糖度折光计(泉州中友光学仪器有

表 1 3 种水蜜桃的含糖量和酸度

Table 1 Sugar content and acidity content of three varieties of juicy peaches

Characteristic Parameter	Sugar content			Acidity content		
	Mean	Range	Std. dev.	Mean	Range	Std. dev.
Milu	11.62	8.8-13.9	1.345	4.31	3.6-5.0	0.407
Hongxianjiu	8.80	6.0-11.0	1.223	4.41	4.1-4.9	0.185
Dabaitao	10.35	7.5-13.0	1.424	4.64	4.4-4.8	0.130

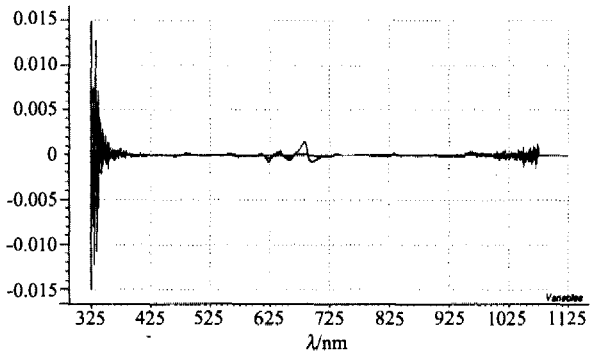

图 2　预处理后的 3 个种类水蜜桃的典型可见-近红外光谱反射图

Fig. 2　Near infrared reflectance spectra of three different varieties juicy peache after pretreatment

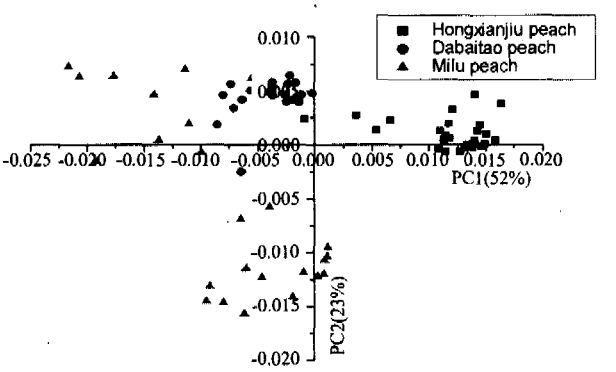

图 3　75 个样本的主成分 1 和主成分 2 的得分图

Fig. 3　Scores plot obtained from the PCA (PC1 vs PC2) of 75 samples

限公司,精度 ±1%)进行测量,水蜜桃酸度采用台式酸度离子计 pH213(上海精密仪器仪表有限公司,精度 ±0.01/ ±0.002pH)进行测量.3 个品种水蜜桃含糖量和酸度如表 1 所列.

2.2　应用主成分分析对不同品种水蜜桃进行聚类

主成分分析不仅能够降低数据维数[8],而且能够通过样本在各因子空间的得分确定所属的类别,所以新变量能够更加形象地表征原样本的品质差异、品种区别等.光谱数据经预处理并选择光谱范围后,对其做主成分分析.以样本在第一主成分和第二主成分上的得分作图,结果见图 3.

图 3 为主成分 1、2 所作的二维散点得分图,图中横坐标表示每个样本的第一主成分得分值,纵坐标表示每个样本的第二主成分得分值.图 2 中蜜露水蜜桃、大白桃水蜜桃、红仙玖水蜜桃明显的分成 3 类,说明主成分 1、2 对 3 种水蜜桃有较好的聚类作用.从图 3 中可以看出,大白桃的 25 个样本聚合度很好,紧密地分布在图 3 中坐标系的第二象限附近;红仙玖的 25 个样本与其它 2 个品种的样本分界很清楚,它们都位于图 2 中坐标系的第一象限附近即坐标系中纵坐标的右边,而其它 2 个品种的样本大都位于坐标系中纵坐标的左边.蜜露水蜜桃的 25 个样本的聚合度没有前 2 个品种好,它们分布在图 3 坐标系中的第三、四象限,但是没有跟另 2 个品种混合起来,它们之间的分界线清楚.分析表明主成分分析对 3 种水蜜桃有一定的聚类作用,能定性区别不同品种水蜜桃.

2.3　基于多类判别分析建立品种鉴别模型

光谱波段从 400 ~ 1000nm 共有 600 个点,但是采用 600 个点计算时,计算量大,而且有些区域样品的光谱信息很弱,与样品的组成或性质间缺乏相关关系.所以通过主成分分析,选取对于水蜜桃品种敏感的新变量来建立品种鉴别模型.可大大简化模型,提高计算速度和分类准确度.表 2 所列表示前 8 个主成分对原始变量的解释程度.

如表 2 可知,前 8 个主成分能够解释原始波长变量的 94.38%,说明前 8 个主成分可代表原可见-近红外光谱的主要信息[6].

以 1、2、3 符号分别代表蜜露水蜜桃、大白桃水蜜桃和红仙玖水蜜桃的品种,从 75 个样品中随机抽取 60 个作为训练集,其余部分作为测试集,进行判别分析.光谱数据的前 8 个主成分,用于多类判别分析法,建立多类判别分析品种鉴别模型.以下是 3 类水蜜桃的判别方程:

蜜露水蜜桃:

$$\begin{aligned} Y_1(x) = & -30.7339 + 3851.213x_1 + 1700.889x_2 \\ & - 898.221x_3 + 3767.616x_4 + 7080.834x_5 \\ & + 6138.231x_6 + 3100.931x_7 + 1291.52x_8; \end{aligned}$$

大白桃水蜜桃:

$$\begin{aligned} Y_2(x) = & -7.0525 - 1437.6x_1 - 38.8619x_2 \\ & - 210.5343x_3 - 1919.43x_4 - 3655.57x_5 \\ & - 1766.59x_6 - 2079.52x_7 - 1773.83x_8; \end{aligned}$$

红仙玖水蜜桃:

$$\begin{aligned} Y_3(x) = & -9.8321 - 2041.95x_1 - 1329.99x_2 \\ & + 541.1353x_3 - 1286.32x_4 - 2829.92x_5 \\ & - 3084x_6 - 707.631x_7 + 382.1324x_8. \end{aligned}$$

将已知 60 个样本代入已建立的判别方程,按各

表 2　前 8 个主成分及其累积贡献率

Table 2　8 Principal components and reliabilities

Principal component	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Accumulative reliabilities	53.62%	76.63%	84.12%	87.95%	90.27%	92.17%	93.40%	94.38%

表3 多类判别分析的回代验证结果

Table 3 Resubstitution rate using multiple discriminant analysis

Actual Varieties	Varieties discriminated by model			Total	Correct rate
	Milu	Dabaitao	Hongxianjiu		
Milu	200	0	0	20	100%
Dabaitao	0	20	0	20	100%
Hongxianjiu	0	3	17	20	85%

母体的后验概率重新归类. 其判别结果见表3所列. 从表中可以看出,所建立的判别模型对蜜露水蜜桃、大白桃水蜜桃和红仙玖水蜜桃的判识效果均相当显著,正确率分别为100%、100%和85%. 从总体上看,在将已知样本代入判别函数进行回判时,其总的回判率达95%,说明所建立的判别模型是可靠的.

为了检验所建立的判别分析模型的可靠程度,对未参与建立模型的15个样本进行预测验证. 15个样本依次是蜜露水蜜桃5个,大白桃水蜜桃5个,红仙玖水蜜桃5个. 将15个样本的相关数据分别代入3个判别函数分中,得到表4. 从表4中我们可以看到15个未参加建模的样本,前5个被判别为"1"品种,第6个到第10个被判别为"2"品种,最后5个被判别为"3"品种. 模型的判别品种与实际品种完全一致,除了第10号样本的后验概率是99.99%,其它样本的后验概率都是100%,该模型对于未知样本预测正确率达到100%.

表4 多类判别分析模型的验证结果

Table 4 Predicting rate of unknown test samples using multiple discriminant analysis

No.	Discriminated by model	Posterior Probability	No.	Discriminated by model	Posterior Probability
(1)	1	100%	(9)	2	100%
(2)	1	100%	(10)	2	99.99%
(3)	1	100%	(11)	3	100%
(4)	1	100%	(12)	3	100%
(5)	1	100%	(13)	3	100%
(6)	2	100%	(14)	3	100%
(7)	2	100%	(15)	3	100%
(8)	2	100%			

Note: (1) ~ (5), Milu juicy peach; (6) ~ (10), Dabaitao juicy peach; (11) ~ (15) Hongxianjiu juicy peach; 1-Milu juicy peach; 2-Milu juicy peach; 3-Hongxianjiu juicy peach.

3 结语

用主成分分析方法结合多类判别分析建立了水蜜桃品种鉴别的模型,该模型性能稳定,预测未知样本识别率达到100%. 说明运用可见-近红外光谱技术可以快速、准确、无损的对水蜜桃品种进行鉴别. 我们用于水蜜桃品种分析的可见-近红外光谱在325~1075nm,说明该波长范围是对水蜜桃品种敏感的特征波段. 提出的主成分分析结合多类判别分析法,特别适用于处理光谱分析中的大量数据,不仅能够从大量光谱信息中提取有用信息,降低数据维数,而且能够运用已知样本的性质、特征建立品种识别模型,定性判别未知样本的品种. 本文提出的方法为其他果品的品种识别分析提供了一种新的途径.

REFERENCES

[1] Esteban-Diez I, Gonzalez-Saiz J M, Pizarro C. An evaluation of orthogonal signal correction methods for the characterisation of arabica and robusta coffee varieties by NIRS [J]. *Analytica. Chimica. Acta.*, 2004, **514**(1):57—67.

[2] SUN Su-Qin, TANG Jun-Ming, YUAN Zi-Min, *et al.* Discrimination of trueborn tuber dioscoreae by fingerprint infrared spectra and principle component analysis [J]. *Spectroscopy and Spectral Analysis* (孙素琴,汤俊明,袁子民,等. 道地山药红外指纹图谱和聚类分析的鉴别研究. **光谱学与光谱分析**),2003,**23**(2):258—260.

[3] HE Yong, LI Xiao-Li, SHAO Yong-Ni. Discrimination of varieties of apple using near infrared spectra based on principal component analysis and artificial neural network model [J]. *Spectroscopy and Spectral Analysis* (何勇,李晓丽,邵咏妮. 基于主成分分析和神经网络的近红外光谱苹果品种鉴别方法研究. **光谱学与光谱分析**), 2006,**26**(5):850—853.

[4] JIA Kong-Yao, DING Tian-Huai. Novel method of detecting foreign fibers in lint by fiber's infrared absorption characteristic [J]. *J. Infrared Millim. Waves* (郑东耀,丁天槐. 利用纤维红外吸收特性的皮棉杂质检测新方法. **红外与毫米波学报**),2005,**24**(2):147—150.

[5] YIN Qiu, SU Xiao-Zhou, XU Zhao-An, *et al.* Analysis on the ultra-spectral characteristics of water environmental parameters about lake[J]. *J. Infrared Millim. Waves* (尹球,疏小舟,徐兆安,等. 湖泊水环境指标的超光谱响应特征分析. **红外与毫米波学报**), 2004,**23**(6): 427—435.

[6] HE Yong, FENG Shui-Juan, DENG Xun-Fei, *et al.* Study on lossless discrimination of varieties of yogurt using the Visible/NIR-spectroscopy [J]. *Food Research International*, 2006,**39**(6):645—650.

[7] WANG Hai-Shui, WANG Dong-Mei, XI Shi-Xuan. The application of near infrared spectroscopy for qualitative and quantitative analysis [J]. *Analysis and Testing Technology Intruments* (王海水,汪冬梅,席时权. 近红外光谱在品质分析和定量分析中的应用. **分析测试技术与仪器**), 2002,**8**(3):136—138.

[8] LI Zhi-Yong, KUANG Gang-Yao, YU Wen-Xian, *et al.* Algorithm on small target detection base on principal component of hyperspectral imagery [J]. *J. Infrared Millim. Waves* (李智勇,匡纲要,郁文贤,等. 基于高光谱图像主成分分量的小目标检测算法研究. **红外与毫米波学报**), 2004,**23**(4):286—290.