DOI:10.6041/j.issn.1000-1298.2012.09.030

# 基于 DPLS 和 LS-SVM 的梨品种近红外光谱识别*

刘雪梅　章海亮
(华东交通大学土木建筑学院,南昌 330013)

【摘要】 为了实现不同品种梨的快速光谱鉴别,采用主成分分析法(PCA)对光谱数据进行聚类分析,得到3种不同品种梨的特征差异,主成分分析表明,以所有建模样本主成分 PC1 和 PC2 做出的得分图,对不同种类梨具有很好的聚类作用。利用主成分分析得到的载荷图可以得到对于梨品种敏感的特征波段,用特征波段图谱作为输入建立偏最小二乘判别(DPLS)模型和最小二乘支持向量机(LS-SVM)模型。3 个品种梨各 70 个共 210 个分别建立偏最小二乘判别(DPLS)模型和最小二乘支持向量机(LS-SVM)模型。对未知的 24 个样本进行预测,LS-SVM 模型品种识别准确率达到 100%,DPLS 模型的校正及验证结果与实际分类变量的相关系数均大于 0.980,交叉验证均方根误差(RMSECV)和预测均方根误差(RMSEP)都小于 0.100,品种识别率为 100%。表明提出的方法具有很好的分类和鉴别作用,提供了梨的品种快速鉴别分析方法。

**关键词**:梨　近红外光谱　品种识别　主成分分析　偏最小二乘判别　最小二乘支持向量机
**中图分类号**:O657.33　　**文献标识码**:A　　**文章编号**:1000-1298(2012)09-0160-05

# Identification of Varieties of Pear Using Near Infrared Spectra Based on DPLS and LS-SVM Model

Liu Xuemei　Zhang Hailiang
(*School of Civil Engineering*, *East China Jiaotong University*, *Nanchang* 330013, *China*)

**Abstract**

In order to realize the rapid identification of different varieties of pears, principal component analysis (PCA) on the spectral data clustering analysis was used on three different varieties of pears to find the characteristic differences. The principal component analysis showed that the main composition PC1 and PC2 for all the modeling samples score diagrams had very good clustering effect to the different types of pears. Load diagram that got by using principal component analysis can obtain the variety sensitive characteristic wavelengths from pears, and with the characteristic band spectrum as input to build partial least-squares discriminant (DPLS) and least squares support vector machine (LS-SVM) models. Seventy pears of three varieties with 210 in total were used to build DPLS and LS-SVM models respectively. The unknown 24 samples were predicted by the models, the recognition accuracy rate of the LS-SVM model reached to 100%. The calibration and verification results of the DPLS model and the actual classification variables of the correlation coefficient was greater than 0.980. Cross validation root mean square error (RMSECV) and root mean square error of prediction (RMSEP) were less than 0.100. The varieties recognition rate was 100%. The proposed rapid identification method has good classification effects.

**Key words**　Pear, Near infrared spectral, Varieties identification, Principal component analysis, DPLS, LS-SVM

收稿日期:2011-11-27　修回日期:2012-02-15
*江西省科技支撑项目(2010BNB01200)
**作者简介**:刘雪梅,讲师,主要从事环境工程和农产品无损检测研究,E-mail:lyumu@163.com

## 引言

梨是我国主产的水果之一，梨的产后处理、品质判断及检测一直是农产品加工研究的重要课题。梨的口感、糖分含量、酸度和维生素含量等内部品质，已得到了广泛重视[1~2]。基于近红外光谱技术的梨品种区分相关研究报道较少，不同品种的梨表面色泽、皮厚以及表皮光滑程度等特征具有较大差异，这些表面光学特性和表皮微观结构的差异，为梨品种的鉴别提供了新思路。传统的品种鉴别是对其进行化学分析，不仅检测周期长，而且容易产生人为误差。所以研究一种简单、快速、非破坏性的在线梨品种鉴别技术很有必要。国内外很多学者利用光谱技术进行了品种鉴别的研究，如苹果[3]、脐橙[4]、水蜜桃[5]、杨梅[6]、羊肉产地[7]、茶叶[8]、水稻[9]、干红葡萄酒[10]、玉米[11]、黄瓜[12]和雪莲花产地[13]等。支持向量机(SVM)是一种建模方法，它通过结构风险最小化原理来提高泛化能力，较好地解决了非线性、高维数、局部极小等实际问题。LS－SVM 是对经典 SVM 的一种改进，以求解一组线性方程代替经典 SVM 中复杂的二次优化问题，降低了计算的复杂性，并且加快了计算的速度[14]。本文在主成分分析的基础上得到对于品种敏感的指纹图谱，即 600～750 nm 波长区域，把该区域的光谱反射值作为 LS－SVM 和偏最小二乘判别(DPLS)模型的输入，建立梨品种鉴别模型。

## 1　材料与方法

### 1.1　仪器设备

试验仪器为美国 ASD 公司的近红外光谱仪，其波长范围为 350～1 800 nm，光谱采样间隔 1 nm，扫描次数 15 次，分辨率为 3 nm ，探头视场角 25°，光源为12 V/45 W 卤钨灯。光谱数据分析软件为 Unscramble V9.7、Origin 7.0 和 Matlab 2010。

### 1.2　样品来源及光谱的获取

试验用梨购于江西省南昌市青云谱农产品批发市场，品种为翠冠、黄花和清香，每个品种挑选个体均匀的各 78 个，避免试验中梨个体与近红外光谱仪之间的距离剧烈变化，共计 234 个样本。全部样本随机分成建模集和预测集，建模集有 210 个样本，预测集有 24 个样本。沿梨赤道部位 120°等间隔采集 3 次光谱。

### 1.3　光谱数据预处理

为了去除高频随机噪声、基线漂移、样本不均匀、光散射等的影响，需要进行光谱预处理。本文采用 moving average 平滑法(平滑点数为 3)消除噪声后，再进行变量标准化(standard normal variate，简称 SNV) 处理。为消除光谱数据在采集时首端与末端产生的部分噪声，截取 400～1 700 nm 波段的光谱数据进行梨品种鉴别分析。

### 1.4　最小二乘支持向量机

最小二乘支持向量机(LS－SVM) 通过非线性映射函数 $\boldsymbol{\varphi}(\boldsymbol{x})$ 建立回归模型，将输入变量映射到高维特征空间，然后将优化问题改成等式约束条件。利用拉格朗日算子求解最优化问题，对各个变量求偏微分。根据 Mercer 条件，存在映射函数 $\boldsymbol{\varphi}(\boldsymbol{x})$ 和核函数 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 使得

$$\boldsymbol{\varphi}(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x}_l)=K(\boldsymbol{x}_k,\boldsymbol{x}_l)\quad(k,l=1,2,\cdots,n)\tag{1}$$

核函数为满足 Mercer 条件的任意对称函数，常用的有：线性核函数、多项式核函数、径向基函数(radial basis function，简称 RBF)、多层感知核函数等。本文采用 RBF 作为核函数，其表达式为

$$K(\boldsymbol{x}_k,\boldsymbol{x}_l)=\exp(-\|\boldsymbol{x}_k-\boldsymbol{x}_l\|^2/(2\sigma^2))\tag{2}$$

从而得到 LS－SVM 的函数估计为

$$y(\boldsymbol{x})=\sum_{k=1}^{n}\alpha_k K(\boldsymbol{x},\boldsymbol{x}_k)+b\tag{3}$$

式中　$\alpha_k$——拉格朗日算子　　$b$——偏差

LS－SVM 需要调节的参数为核参数 $\sigma^2$ 和惩罚系数 $\gamma$。惩罚系数 $\gamma$ 主要是控制对错分样本惩罚的程度，实现在错分样本的比例与算法复杂度之间的折中。RBF 核函数的核参数 $\sigma^2$ 的选择对模型准确度起到很大的作用，选的太小则会造成过学习，选的太大会造成欠学习[15]。

### 1.5　偏最小二乘判别

偏最小二乘判别(DPLS)分析法基于 PLS 回归方法，将光谱数据与分类变量进行线性回归，其判别过程为：①建立校正集样本的分类变量。②通过分类变量与光谱数据的 PLS 分析，建立分类变量和光谱数据间的 PLS 模型。③根据校正集建立的分类变量和光谱特征的 PLS 模型对预测集样品进行预测验证[4]。DPLS 方法是基于 PLS 回归的一种判别分析方法，在构造因素时考虑到了辅助矩阵以代码形式提供的类成员信息，因此具有高效的鉴别能力。具体判别标准为：计算验证集的分类变量值($Y_p$)，设 $Y_y$ 为样品预测值。①当 $|Y_p-Y_y|>0.5$，且偏差小于 0.5，判定样本属于该类。②当 $|Y_p-Y_y|<0.5$，且偏差小于 0.5，判定样本不属于该类。③当偏差大于 0.5，该判别模型不稳定。

## 2　试验结果与分析

### 2.1　梨样本的近红外漫反射光谱

3 种梨的典型近红外光谱曲线如图 1 所示。从

图中可以看出,不同品种梨的光谱曲线具有一定的特征性和指纹性,这一差异为梨的不同品种鉴别奠定了数学基础。应用 Matlab 软件,把同一个梨 3 个不同部位的光谱曲线做平均处理,形成反射率矩阵,用主成分分析法对其聚类。前 3 个主成分 PC1、PC2、PC3 的特征值及累计可信度分别为 66%、93%、97%。由于前 2 个主成分的累计可信度已达 93%,所以仅用前 2 个主成分就可以表示原始近红外光谱的主要信息。

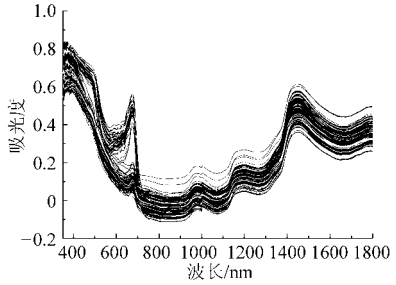


图 1 3 种梨近红外光谱曲线

Fig. 1 Near infrared reflectance spectra of three different variety pears

## 2.2 最佳定标模型的确定

主成分分析(PCA)方法是经典的特征抽取和降维技术之一,它可以在不具备任何相关知识背景的情况下对未知样品进行类别归属的判别。图 2 表示 210 个建模样本的主成分 PC1、PC2 得分图。图 2 中 3 种梨明显分成 3 类,说明主成分 PC1、PC2 对 3 种梨有较好的聚类作用。因此,本研究基于主成分 PC1、PC2 对 210 个样本的得分图进行聚类分析。从图 2 中可以看出,品种为翠冠的 70 个样本聚合度较好而且均位于 $Y$ 轴的右方,其他 140 个样本大部分位于 $Y$ 轴的左方。品种为黄花的 70 个样本聚合度较好,紧密的分布在图 2 的第 2、3 象限。品种为清香的梨样本聚合度不如另外两种,分布较为分散,但是除了有 3 个样本位于第 1 象限,其他全部分布在第 2、3 象限之内,而另 2 个品种均在 2 个象限内。

## 2.3 基于指纹图谱建立 LS-SVM 品种预测模型

全波段从 350 ~ 1 800 nm 共有 1 451 个点,但是,采用全光谱计算时,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缺乏相关关系。为此在主成分分析的基础上,选取对梨品种敏感的波段作为输入建立 LS-SVM 品种预测模型。图 3 表示主成分 PC1 和主成分 PC2 在整个波长范围的载荷图。从图 3 中可以看出主成分与全波长变量的相关程度。从图 3a 可以看出主成分 PC1 与波长 600 ~ 650 nm 的相关性较大,即主成分与波长 600 ~ 650 nm 范围的反射值的相关性较强;从图 3b 可以看出主成分 PC2 与波长 600 ~ 750 nm 范围的反射值相关性较强。又知主成分 PC1、PC2 的累积可信度已达到 93%,即主成分 PC1、PC2 几乎能完全解释所有原变量,而且从图 2 可以看出主成分 PC1、PC2 对 3 类梨有较好的聚类作用,所以与它们紧密相关的 600 ~ 650 nm 和 600 ~ 750 nm 是对梨品种极为敏感的特征波长。所以从 350 ~ 1 800 nm 范围的 1 451 个采样点中选出波长在 600 ~ 750 nm 范围的共计 151 个采样点的反射值作为 LS-SVM 的输入变量建立鉴别模型,核参数 $\sigma^2$ 和惩罚系数 $\gamma$ 取值都为 0.001。对未知的 24 个样本进行预测,预测准确率为 100%(表 1)。

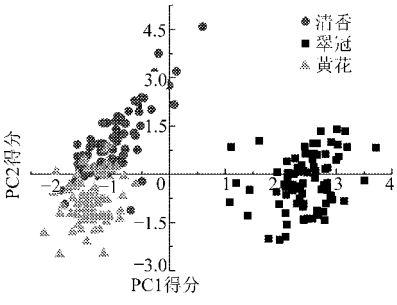


图 2 梨样品 PC 得分图

Fig. 2 PCA scores plots (PC1 and PC2) for pear samples

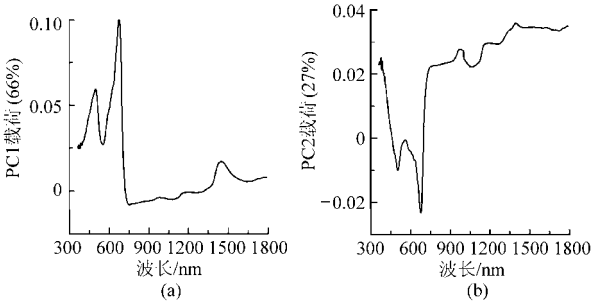


图 3 PC1 和 PC2 在全波段载荷图

Fig. 3 Loading plot of PC1 and PC2 across the entire spectral region

(a) PC1 (b) PC2

## 2.4 DPLS 判别模型的建立与验证

DPLS 方法是基于 PLS 方法建立的样本分类变量与 NIR 光谱特征间的回归模型。因此,首先需要按照样本实际类别特征,赋予校正集样本的分类变量值(表 1)。然后,利用 PLS 回归方法对校正集样本的 NIR 光谱与样本对应的分类变量进行回归分析,并建立 NIR 光谱特征与分类变量间的 PLS 模型。利用近红外光谱结合 DPLS 方法建立的判别模型,校正结果如表 2 和图 4 所示。

从表 2 可知,通过 DPLS 回归分析建立的梨品种和分类变量间的相关性较好,分类变量的实测值和模型的预测值的相关系数均大于 0.980,校正样品的识别率均为 100%,说明模型拟合较好。

**表 1　LS－SVM 模型对未知样品预测结果**

**Tab. 1　Prediction result of unknown samples with LS－SVM model**

| 预测样本序号 | 真实值 | 预测值 | 预测样本序号 | 真实值 | 预测值 | 预测样本序号 | 真实值 | 预测值 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0.996 2 | 9 | 2 | 1.931 3 | 17 | 3 | 3.016 0 |
| 2 | 1 | 1.002 1 | 10 | 2 | 1.958 8 | 18 | 3 | 2.993 5 |
| 3 | 1 | 1.002 5 | 11 | 2 | 1.912 6 | 19 | 3 | 3.086 4 |
| 4 | 1 | 1.003 5 | 12 | 2 | 1.925 0 | 20 | 3 | 2.997 0 |
| 5 | 1 | 1.039 3 | 13 | 2 | 1.934 4 | 21 | 3 | 3.075 5 |
| 6 | 1 | 0.997 1 | 14 | 2 | 1.935 0 | 22 | 3 | 2.962 3 |
| 7 | 1 | 1.000 3 | 15 | 2 | 2.101 8 | 23 | 3 | 3.024 9 |
| 8 | 1 | 1.120 4 | 16 | 2 | 2.111 1 | 24 | 3 | 3.137 4 |

注：真实值 1 代表翠冠，2 代表清香，3 代表黄花。

**表 2　DPLS 模型的校正和验证结果**

**Tab. 2　Calibration and validation results of DPLS models**

| 梨品种 | 光谱预处理方法 | 因子数 | 校正集 | | | 验证集 | | |
|---|---|---|---|---|---|---|---|---|
| | | | $r_c$ | RMSECV | 识别率/% | $r_v$ | RMSEP | 识别率/% |
| 翠冠 | | 15 | 0.999 | 0.021 | 100 | 0.999 | 0.023 | 100 |
| 黄花 | SG 平滑＋变量标准化 | 16 | 0.989 | 0.081 | 100 | 0.982 | 0.085 | 100 |
| 清香 | | 14 | 0.988 | 0.082 | 100 | 0.984 | 0.092 | 100 |

注：$r_c$为校正集相关系数；$r_v$为验证集相关系数；RMSECV 为交叉验证均方根误差；RMSEP 为预测均方根误差。

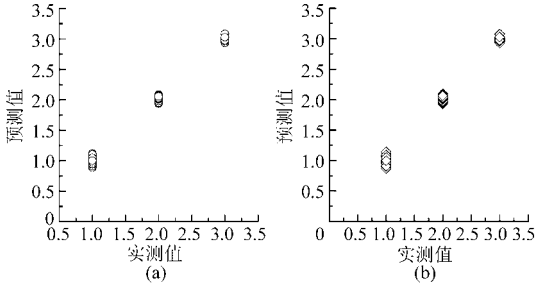


图 4　DPLS 校正集和预测集的 DPLS 模型回归图

Fig. 4　Calibration set and prediction set of DPLS regression model diagram

（a）校正集　（b）预测集

图 4a 所示为所有校正集样本（包括翠冠、黄花和清香 3 个梨品种）分类变量的 DPLS 预测值与实测值的回归图。从图 4b 中可知，DPLS 模型都能清楚地将 3 类梨样本区分开，即分散在实测值等于 1 的线上的梨品种样本点和实测值等于 2 和 3 的其他 2 个品种的梨样本明显分开，说明 DPLS 模型具有很高的可靠性，可以用于检验和判别新的样本。

## 3　结束语

应用主成分分析结合偏最小二乘判别和最小二乘支持向量机建立了梨品种鉴别的模型，两种模型的预测效果都较好，识别率达到 100%。说明运用近红外光谱技术可以快速、准确、无损地对梨品种进行鉴别，模型用于鉴别梨品种是成功的。通过主成分分析得到与梨品种密切相关的特征波段为 600～750 nm，并且用特征波段建立偏最小二乘判别和最小二乘支持向量机品种鉴别模型效果较好，说明该波段是品种鉴别的指纹波段，可以进一步研究确定特征波长，为开发简单、低成本、高精度的仪器奠定了基础。

## 参　考　文　献

1　刘燕德，孙旭东. 近红外漫反射光谱检测梨内部指标可溶性固性物的研究[J]. 光谱学与光谱分析，2008，28(4)：797～800.

Liu Yande, Sun Xudong. Research on detecting internal quality SSC of pear using NIR diffuse reflection spectroscopy[J]. Spectroscopy and Spectral Analysis, 2008, 28(4): 797～800. (in Chinese)

2　刘燕德，应义斌. 傅里叶近红外光谱的雪青梨酸度偏最小二乘法定量分析[J]. 光谱学与光谱分析，2006，26(8)：1 454～1 456.

Liu Yande, Ying Yibing. The pear acidity quantified analysis using PLS methods and Fourier transform near infrared spectroscopy[J]. Spectroscopy and Spectral Analysis, 2006, 26(8): 1 454～1 456. (in Chinese)

3　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[J]. 光谱学与光谱分析，

2006,26(5): 850 ~ 853.
He Yong, Li Xiaoli, Shao Yongni. Discrimination of varieties of apple using near infrared spectra based on principal component analysis and artificial neural network model[J]. Spectroscopy and Spectral Analysis, 2006, 26(5): 850 ~ 853. (in Chinese)

4 郝勇,孙旭东,高荣杰,等. 基于可见/近红外光谱与 SIMCA 和 PLS - DA 的脐橙品种识别[J],农业工程学报,2010,26(12):373 ~ 377.
Hao Yong, Sun Xudong, Gao Rongjie, et al. Application of visible and near infrared spectroscopy to identification of navel orange varieties using SIMCA and PLS - DA methods[J]. Transactions of the CSAE, 2010, 26(12): 373 ~ 377. (in Chinese)

5 李晓丽,胡兴越,何勇. 基于主成分和多类判别分析的可见近红外光谱水蜜桃品种鉴别新方法[J]. 红外与毫米波学报,2006, 25(6): 417 ~ 420.
Li Xiaoli, Hu Xingyue, He Yong. New approach of discrimination of varieties of juicy peach by near infrared spectra based on PCA and MDA model[J]. Journal of Infrared Millimeter Waves, 2006, 25(6): 417 ~ 420. (in Chinese)

6 何勇,李晓丽. 近红外光谱杨梅品种鉴别方法的研究[J]. 红外与毫米波学报,2006,25(3): 192 ~ 194.
He Yong, Li Xiaoli. Discrimination of varieties of waxberry using near infrared spectra[J]. Journal of Infrared Millimeter Waves, 2006, 25(3): 192 ~ 194. (in Chinese)

7 张宁,张德权,李淑荣,等. 近红外光谱结合 SIMCA 法溯源羊肉产地的初步研究[J]. 农业工程学报,2008,24(12):309 ~ 312.
Zhang Ning, Zhang Dequan, Li Shurong, et al. Preliminary study on origin traceability of mutton by near infrared reflectance spectroscopy coupled with SIMCA method [J]. Transactions of the CSAE, 2008, 24(12): 309 ~ 312. (in Chinese)

8 李晓丽,何勇,裘正军. 一种可见近红外光谱快速鉴别茶叶品种的新方法[J]. 光谱学与光谱分析,2007,27(2):279 ~ 282.
Li Xiaoli, He Yong, Qiu Zhengjun. A new method to fast discrimination of tea varieties using visible/near infrared spectroscopy[J]. Spectroscopy and Spectral Analysis, 2007, 27(2):279 ~ 282. (in Chinese)

9 李晓丽,唐明月,何勇,等. 基于可见/近红外光谱的水稻品种快速鉴别研究[J]. 光谱学与光谱分析, 2008,28(3):578 ~ 581.
Li Xiaoli, Tang Mingyue, He Yong, et al. Discrimination of varieties of paddy based on Vis/NIR spectroscopy combined with chemometrics spectroscopy and spectral[J]. Spectroscopy and Spectral Analysis, 2008, 28(3):578 ~ 581. (in Chinese)

10 吴桂芳,蒋益虹,王艳艳,等. 基于独立主成分和 BP 神经网络的干红葡萄酒品种的鉴别[J]. 光谱学与光谱分析,2009,29(5):1 268 ~ 1 271.
Wu Guifang, Jiang Yihong, Wang Yanyan, et al. Discrimination of varieties of dry red wines based on independent component analysis and BP neural network[J]. Spectroscopy and Spectral Analysis, 2009, 29(5):1 268 ~ 1 271. (in Chinese)

11 杨蜀秦,宁纪锋,何东健. BP 人工神经网络识别玉米品种的研究[J]. 西北农林科技大学学报: 自然科学版, 2004, 32 (增刊 1): 162 ~ 164.
Yang Shuqin, Ning Jifeng, He Dongjian. Identification of corn breeds by BP neural network[J]. Journal of Northwest A&F University: Natural Science Edition, 2004, 32 (Supp. 1):162 ~ 164. (in Chinese)

12 袁挺,纪超,陈英,等. 基于光谱成像技术的温室黄瓜识别方法[J]. 农业机械学报,2011,42(11):172 ~ 176.
Yuan Ting, Ji Chao, Chen Ying, et al. Greenhouse cucumber recognition based on spectral imaging technology [J]. Transactions of the Chinese Society for Agricultural Machinery, 2011,42(11):172 ~ 176. (in Chinese)

13 赵杰文,蒋培,陈全胜. 雪莲花产地鉴别的近红外光谱分析方法[J]. 农业机械学报,2010,41(8):111 ~ 114.
Zhao Jiewen, Jiang Pei, Chen Quansheng. Discrimination of snow lotus from different geographical origins by near infrared spectroscopy [J]. Transactions of the Chinese Society for Agricultural Machinery, 2010,41(8):111 ~ 114. (in Chinese)

14 王莉,何勇,刘飞,等. 应用光谱技术和支持向量机分析方法快速检测啤酒糖度和 pH 值[J]. 红外与毫米波学报,2008,27(1):51 ~ 55.
Wang Li, He Yong, Liu Fei, et al. Rapid detection of sugar content and pH in beer by using spectroscopy techniqque combined with support vector machines [J]. Journal of Infrared Millimeter Waves, 2008, 27(1):51 ~ 55. (in Chinese)

15 刘飞,王莉,何勇. 应用有效波长进行奶茶品种鉴别的研究[J]. 浙江大学学报:工学版,2010,44(3): 619 ~ 624.
Liu Fei, Wang Li, He Yong. Application of effective wavelengths for variety identification of instant milk tea[J]. Journal of Zhejiang University: Engineering Science, 2010, 44(3): 619 ~ 624. (in Chinese)