

基于可见/近红外光谱和变量优选的 南水梨糖度在线检测

On-line detection of sugar content of Nanshui pears by Vis/NIR
spectroscopy and variable selection methods

孙 通 江水泉

SUN Tong JIANG Shui-quan

(江苏楷益智能科技有限公司, 江苏 无锡 214174)

(Jiangsu Kaiyi Intelligent Technology Co., Ltd., Wuxi, Jiangsu 214174, China)

摘要:利用可见/近红外光谱技术对南水梨糖度进行在线检测研究。南水梨样本以 0.3 m/s 速度传输,并采用 USB4000 光谱仪在 470~1 150 nm 波段范围内采集南水梨样本的光谱。然后,利用 3 种变量选择方法对波长变量进行筛选,应用偏最小二乘(PLS)方法分别建立南水梨糖度的在线预测模型,并分析预测模型性能的优劣。结果表明:可见/近红外光谱技术结合变量选择方法在线检测南水梨的糖度是可行的;竞争自适应重加权采样(CARS)方法优于无信息变量消除(UVE)及连续投影算法(SPA);CARS 方法可以有效简化预测模型并提高预测模型的性能;南水梨全光谱 PLS 及 CARS-PLS 糖度预测模型的预测集相关系数和预测均方根误差(RMSEP)分别为 0.940,0.951 和 0.467%,0.420%。

关键词:可见/近红外光谱;变量选择;糖度;南水梨

Abstract: Sugar content (SC) is one of the important internal qualities of Nanshui pears. In this research, Visible/near infrared (Vis/NIR) spectroscopy was used to detect SC of Nanshui pears on-line. Transmission speed of Nanshui pears was 0.3 m/s, and USB4000 spectrometer was used to acquire the spectra of Nanshui pear samples in the wavelength range of 470~1 150 nm. Then three variable selection methods were used to select sensitive wavelength variables, and partial least squares (PLS) was used to develop calibration models of SC for Nanshui pears, also performance of calibration models was compared. The results indicate that Vis/NIR spectroscopy combined with variable selection method is feasible for on-line detection of SC for Nanshui pears. Competitive adaptive reweighted sampling (CARS) method is superior to uninformative variable elimination (UVE) and successive projections algorithm (SPA) methods. CARS

method can simplify calibration model and improve performance of calibration model. The correlation coefficients in prediction and root mean square errors of prediction (RMSEPs) of full-PLS and CARS-PLS models of SC for Nanshui pears are 0.940,0.951 and 0.467%, 0.420%, respectively.

Keywords: visible/near infrared spectroscopy; variable selection; sugar content; nanshui pear

南水梨原产于日本,由越后与新水品种杂交而成,果实呈圆形或扁圆形,表面光亮,肉质细嫩,含糖量极高,且有特殊香味。糖度为南水梨的重要内部品质指标之一,为实现南水梨按质分级,非常有必要对其糖度进行在线快速无损检测。

可见/近红外光谱技术是一种间接分析技术,具有绿色无污染、无损及快速等特点。其原理主要是基于有机物分子中含氢基团(C—H、N—H、O—H)的倍频与合频吸收,根据基团的吸收强度可对物质进行定量检测。该技术已应用于谷物淀粉^[1-2]、肉类蛋白质^[3-4]、食用油脂脂肪酸^[5]、酒类酒精度^[6]、茶叶茶多酚^[7]、鸡蛋新鲜度^[8]及液态奶三聚氰胺^[9]等检测。对水果糖度(可溶性固形物)检测, Sun 等^[10]利用可见/近红外光谱技术对不同运动速度(0.3~0.7 m/s)下的翠冠梨进行可溶性固形物检测。结果表明,在 0.3 m/s 运动速度下,糖度预测模型性能最好,其相关系数为 0.957,预测均方根误差(RMSEP)为 0.530%。韩东海等^[11]对小型西瓜不同部位的可溶性固形物进行检测。结果表明,瓜顶和赤道部分的可溶性固形物检测结果较好,相关系数和 RMSEP 分别为 0.952,0.929 和 0.666,0.732。Li 等^[12]对不同品种的梨进行可溶性固形物检测,通过回归系数确定特征波长变量,并应用偏最小二乘回归(PLS)和最小二乘支持向量机(LS-SVM)方法建立预测模型。结果表明,LS-SVM 结合特征变量所建立的预测模型最优,其相关系数和 RMSEP 分别为 0.916 和 0.251%。欧阳爱国等^[13]为简化苹果可溶性固形物

基金项目:无锡市农业科技支撑项目(编号:CLE02N1507)

作者简介:孙通,男,博士。

通讯作者:江水泉(1972—),男,江苏楷益智能科技有限公司高级工程师,硕士。E-mail:jsq55@163.com。

收稿日期:2015—12—07

预测模型,采用多种方法进行波长变量筛选,获得36个特征波长变量,并采用特征变量建立预测模型。结果表明,该预测模型的相关系数和RMSEP分别为0.90和0.70%。此外,还有其他学者^[14-18]也对水果糖度进行检测研究,但均未在线检测条件下对水果糖度检测进行波长变量筛选,而静态检测条件下的波长变量筛选结果不适用于在线检测;且尚未有南水梨糖度在线检测的研究报道。

本研究拟利用可见/近红外光谱技术对南水梨糖度进行在线检测研究,采用3种变量选择方法筛选与南水梨糖度相关的波长变量,应用PLS建立南水梨糖度(SC)在线预测模型,并比较不同变量选择方法的优劣。

1 材料与方法

1.1 试验材料

南水梨:购自南昌水果批发市场,样本数量为199个。试验前,对样本进行简单处理包括表皮清理和样本编号。样本处理后,按照2:1比率并根据随机原则将样本分配到校正集和预测集。因此,校正集和预测集样本分为133个和66个。

1.2 试验装置及光谱采集

试验装置由光谱仪、光源、光纤、托盘及电脑等组成,具体见图1。光谱仪器为微型USB4000光谱仪(美国Oceanoptics公司),配备3648像素CCD检测器。光源为2盏功率150W的卤钨灯,两卤钨灯光源中心与水果赤道处于同一水平面,且对称分布于水果两侧。光纤芯径为100μm,与USB4000光谱仪连接。水果位于托盘上,托盘由电机带动,传输速度为0.3m/s。

样本光谱采集前,先获取暗场及参比的光谱。对于南水梨样本,将其以任意方位置于托盘上,在上述传输速度下采集其光谱。样本光谱的积分时间为30ms,暗场和参比的积分时间为50ms。试验所用参比为聚四氟乙烯球,直径为80mm。

1.3 糖度含量测定

样本光谱采集后,按照GB12295—90测定南水梨的真实糖度值。首先,将去皮后的南水梨样本果肉部分放入榨汁机,用中速滤纸过滤得澄清果汁。然后取部分澄清摇匀后果汁,滴于PR-101α型数字折光仪(日本Atago公司)的测量窗口,测定南水梨样本的真实糖度值。

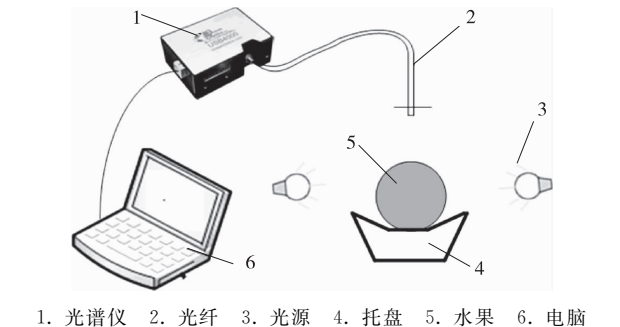


图1 可见/近红外在线检测系统示意图
Figure 1 Schematic diagram of Vis/NIR online detection system

1.4 数据处理与分析

采用竞争自适应重加权采样(CARS)、无信息变量消除(UVE)及连续投影算法(SPA)对样本光谱波长变量进行选择,筛选与南水梨糖度相关的重要变量,对筛选出的重要变量应用PLS方法建立南水梨糖度在线预测模型。

对于CARS变量选择,其蒙特卡罗(MC)采样次数为50次,最大因子数由MC交叉验证确定。对于选择的一系列波长变量子集,采用PLS交互验证建模获得结果,最优的波长变量子集则根据交互验证均方根误差(RMSECV)值最小原则确定。对于UVE变量选择,样本光谱矩阵中加入1554个随机噪声变量,变量筛选的阈值为随机噪声变量稳定性最大绝对值的99%。对于SPA变量选择,其最小和最大的可选波长变量数为1和50,并采用交叉验证建模确定最优波长变量数及组合。CARS、UVE及SPA变量选择方法的具体原理及步骤详见文献^[19-21]。

南水梨糖度在线预测模型性能由相关系数(r)、校正均方根误差(RMSEC)、RMSECV及RMSEP评价。

2 结果与讨论

2.1 样本糖度统计

表1为南水梨校正集和预测集样本的糖度平均值、范围及标准偏差。由表1可知,校正集和预测集样本的糖度平均值相差0.27%,标准偏差相差0.07%,且校正集样本的糖度值范围略大于预测集样本。由此可知,校正集及预测集样本的分配较为合理,校正集样本所建立的预测模型能较好地适用于预测集样本。

2.2 变量选择

在变量选择前,对样本光谱波段进行初步筛选。由图2可知,在600~900nm波段范围内,样本光谱有较为明显的波峰和波谷,可能包含较多有用的光谱信息;而在600nm以下及900nm以上噪声较大,光谱信噪比低。为排除光谱噪声

表1 南水梨校正集和预测集样本的糖度平均值、范围及标准偏差
Table 1 Means, ranges, and S. D. s of SC of nanshui pears in the calibration and prediction sets

数据集	样本量	平均值/%	标准偏差/%	范围/%
校正集	133	11.02	1.30	7.80~13.50
预测集	66	11.29	1.37	7.80~13.30

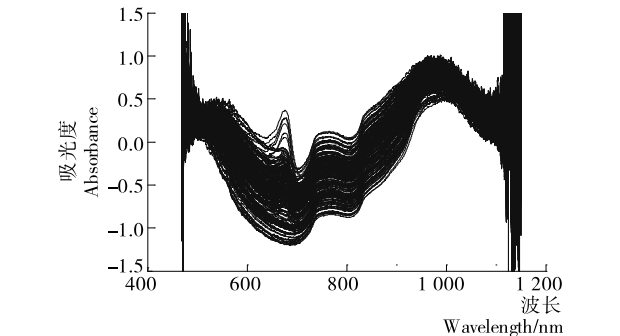


图2 所有南水梨样本的可见/近红外光谱
Figure 2 Vis/NIR spectra of all Nanshui pear samples

的影响,选择 600~900 nm 波段用于后续 的变量选择分析,该波段范围内共有 1 554 个波长变量。

2.2.1 CARS 利用 CARS 方法对校正集样本光谱进行波长变量筛选。图 3 为南水梨糖度的 CARS 变量选择分析结果。由图 3(a)可知,随着采样次数的增加,被选择的波长变量数先迅速下降而后缓慢减少,体现了 CARS 方法的粗选和精选两个过程。由图 3(b)可知,采样次数为 1~20 时, RMSECV 值随采样次数增加而缓慢下降,表明无信息波长变量或冗余波长变量正逐步被剔除;采样次数为 21~50 时, RMSECV 值随采样次数增加而逐渐上升,表明有用信息波长变量开始被剔除。图 3(c)为被选择波长变量的回归系数的变化趋势。根据 RMSECV 值最小原则,确定最优的波长变量子集。图 3(c)中“*”所标注的采样次数(20 次采样)对应的 RMSECV 值最小,为 CARS 方法的最佳结果。经 CARS 变量选择后,共有 118 个有用信息波长变量被保留。

2.2.2 UVE 利用 UVE 方法对校正集样本光谱进行波长变量筛选,其结果见图 4。图 4 中,1 554 个随机噪声变量位

于右侧,而 1 554 个真实波长变量位于左侧,竖直线为两者的分界线。两水平虚线为 UVE 分析中变量稳定性的上下阈值,该阈值为随机噪声变量稳定性最大绝对值的 99%,并根据此阈值对波长变量进行筛选。若波长变量的稳定性值处于两阈值之外,则被作为有用信息波长变量而保留;反之,则被作为无信息波长变量而剔除。经 UVE 分析后,共有 910 个波长变量被保留。

2.2.3 SPA 利用 SPA 方法对校正集样本光谱进行波长变量筛选。图 5 为南水梨糖度的 SPA 变量选择分析结果。由图 5 可知,当被选择的波长变量数由 1 增加到 2、由 4 增加到 5 及由 7 增加到 10 时,均方根误差(RMSE)迅速下降,表明此类波长变量为南水梨糖度相关的重要波长变量。当被选择的波长变量数由 10 增加到 17 时, RMSE 继续缓慢下降,此类波长变量均为有用信息变量;而后随着波长变量数的继续增加, RMSE 反而逐渐上升,表明模型中已引入无信息的噪声或冗余波长变量,致使模型的 RMSE 不降反升。根据 RMSE 最小原则,最终选择 17 个波长变量为最优组合,其波长变量分别为 600. 2, 610. 0, 612. 3, 615. 7, 619. 2, 632. 7, 688. 2, 697. 1, 730. 4, 753. 9, 780. 3, 807. 7, 878. 5, 881. 6, 884. 1, 897. 6 nm。

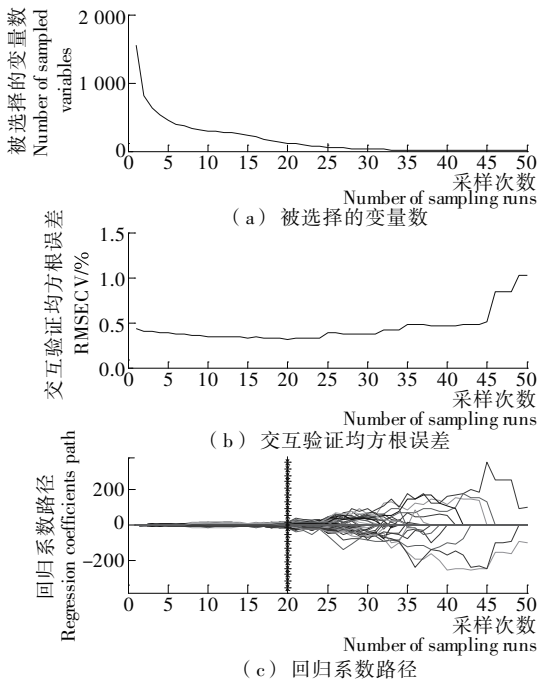


图 3 南水梨糖度的 CARS 变量选择分析结果

Figure 3 Results of CARS method of SC for Nanshui pears

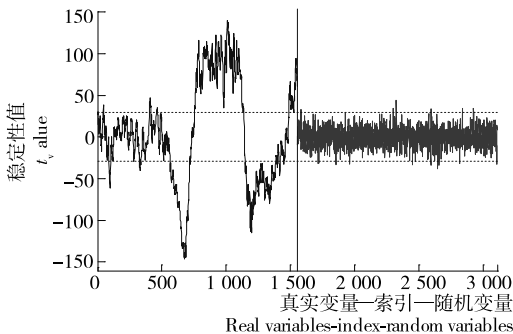


图 4 南水梨糖度的 UVE 变量选择分析结果

Figure 4 Results of UVE method of SC for Nanshui pears

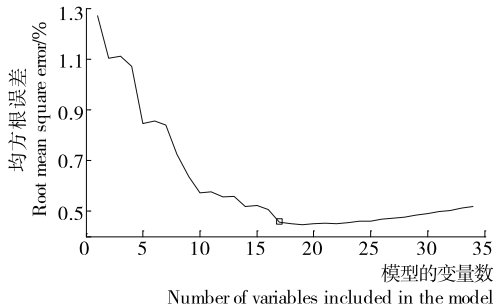


图 5 南水梨糖度的 SPA 变量选择分析结果

Figure 5 Results of SPA method of SC for Nanshui pears

2.3 预测模型建立

根据 CARS、UVE 及 SPA 方法的变量选择结果,应用 PLS 方法分别建立南水梨糖度在线预测模型,并对预测集样本进行预测。表 2 为不同变量选择方法的南水梨糖度 PLS 模型结果。由表 2 可知, CARS—PLS 模型的性能优于全光谱 PLS 模型,预测集相关系数由 0. 940 上升为 0. 951,而 RMSEP 却由 0. 467% 下降为 0. 420%。此外, CARS—PLS 模型所用的波长变量数为 118 个,仅占全光谱波长变量数的 7. 6%。由此表明, CARS 方法可以简化预测模型,并提高预测模型的预测精度。对于 UVE—PLS 及 SPA—PLS 模型,其模型性能劣于全光谱 PLS 模型,预测集相关系数下降而 RMSEP 上升,但建模所用的波长变量数大大减少;特别是 SPA—PLS 模型,其建模波长变量数仅占全光谱的 1. 1%。对比 CARS—PLS、UVE—PLS 及 SPA—PLS 模型性能可知, CARS—PLS 模型性能最优,表明 CARS 方法优于 UVE 及 SPA 方法。图 6 为 CARS—PLS 糖度模型对预测集样本的预测结果。

表 2 不同变量选择方法的南水梨糖度 PLS 模型结果

Table 2 PLS models results of SC for Nanshui pears with different variable selection methods

方法	变量数	校正集			预测集	
		r_c	RMSEC%	RMSECV%	r_p	RMSEP%
无	1 554	0.957	0.373	0.442	0.940	0.467
CARS	118	0.970	0.312	0.338	0.951	0.420
UVE	910	0.951	0.399	0.451	0.936	0.494
SPA	17	0.951	0.400	0.476	0.894	0.611

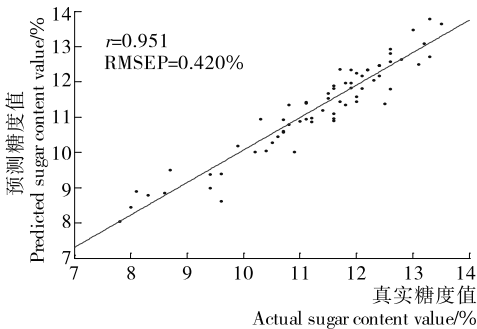


图 6 CARS—PLS 糖度模型对预测集样本的预测结果

Figure 6 Predicted results of SC in prediction samples by CARS—PLS model

3 结论

本研究利用可见/近红外光谱技术结合变量选择方法对南水梨糖度进行在线检测研究。结果表明,CARS 方法优于 UVE 及 SPA 方法,该方法可以有效筛选重要波长变量,简化预测模型及提高模型的预测精度。南水梨 CARS—PLS 糖度模型的预测集相关系数和 RMSEP 分别为 0.951 和 0.420%。由此可知,可见/近红外光谱技术结合变量选择方法可以用于南水梨糖度的在线检测。下一步将扩大南水梨样本来源,探讨样本产地及采摘年份对糖度预测结果的影响,以期解决实际生产线中南水梨糖度检测的稳定性问题。

参考文献

[1] Li Jun-hui, Danao MGC, Chen Shih-fang, et al. Prediction of starch content and ethanol yields of sorghum grain using near infrared spectroscopy[J]. Journal of Near Infrared Spectroscopy, 2015, 23(2): 85-92.

[2] 冯光, 杨亚春, 宋丰顺, 等. 近红外反射技术建立合肥地区精米直链淀粉含量测定模型[J]. 核农学报, 2012, 26(6): 942-946.

[3] Huang Lin, Zhao Jie-wen, Zhang Yan-hua, et al. Simultaneous multi-component analysis of pork meat during bacterial spoiling process by FT—NIR evaluated with a non-linear algorithm[J]. Analytical Methods, 2012, 4(11): 3 816-3 823.

[4] 黄伟, 杨秀娟, 曹志勇, 等. 近红外反射光谱快速检测滇南小耳猪肉中水分、粗脂肪及粗蛋白含量的研究[J]. 中国畜牧杂志, 2015, 51(7): 73-77.

[5] Wu Di, Chen Xiao-jing, Shi Pin-yan, et al. Determination of alpha-linolenic acid and linoleic acid in edible oils using near-infrared spectroscopy improved by wavelet transform and uninfor-

ative variable elimination[J]. Analytica Chimica Acta, 2009, 634 (2): 166-171.

[6] Ye Meng-qi, Gao Zheng-peng, Zhao Li, et al. Rapid detection of volatile compounds in apple wines using FT—NIR spectroscopy [J]. Food Chemistry, 2016, 190: 701-708.

[7] 张海东, 李贵荣, 李若诚, 等. 近红外光谱结合极限学习机和 GA—PLS 算法检测普洱茶茶多酚含量[J]. 激光与光电子学进展, 2013, 50(4): 180-186.

[8] 赵杰文, 毕夏坤, 林颢, 等. 鸡蛋新鲜度的可见—近红外透射光谱快速识别[J]. 激光与光电子学进展, 2013, 50(5): 209-216.

[9] 程文宇, 管晓, 刘静. 近红外光谱技术检测液态奶中微量三聚氰胺的可行性研究[J]. 食品与机械, 2015, 31(1): 71-74, 81.

[10] Sun Tong, Lin Hong-jian, Xu Hui-rong, et al. Effect of fruit moving speed on predicting soluble solids content of ‘Cuiguan’ pears (Pomaceae pyrifolia Nakai cv. Cuiguan) using PLS and LS-SVM regression. Postharvest Biology and Technology, 2009, 51(1): 86-90.

[11] 韩东海, 常冬, 宋曙辉, 等. 小型西瓜品质近红外无损检测的光谱信息采集[J]. 农业机械学报, 2013, 44(7): 174-178.

[12] Li Jiang-bo, Huang Wen-qian, Zhao Chun-jiang, et al. A comparative study for the quantitative determination of soluble solids content, pH and firmness of pears by Vis/NIR spectroscopy [J]. Journal of Food Engineering, 2013, 116(2): 324-332.

[13] 欧阳爱国, 谢小强, 刘燕德. 苹果可溶性固形物近红外在线光谱变量优选[J]. 农业机械学报, 2014, 45(4): 220-225.

[14] Moller S M, Travers S, Bertram H C, et al. Prediction of post-harvest dry matter, soluble solids content, firmness and acidity in apples (cv. Elshof) using NMR and NIR spectroscopy: a comparative study[J]. European Food Research and Technology, 2013, 237(6): 1 021-1 024.

[15] Jie Deng-fei, Xie Li-juan, Rao Xiu-qin, et al. Using visible and near infrared diffuse transmittance technique to predict soluble solids content of watermelon in an on-line detection system[J]. Postharvest Biology and Technology, 2014, 90: 1-6.

[16] Zhao Xin, Zhu Qi-bing, Huang Min, et al. An IGA—PLSP method for FT—NIR wavelength selection for measuring soluble solid content of citrus fruits[J]. Analytical Methods, 2013, 5(18): 4 811-4 817.

[17] 孙通, 许文丽, 胡田, 等. 基于 UVE—ICA 和支持向量机的南丰蜜桔可溶性固形物可见—近红外检测[J]. 光谱学与光谱分析, 2013, 33(12): 3 235-3 239.

[18] 张纯, 张海东, 江水泉. 用混合线性分析法建立苹果糖度近红外光谱预测模型[J]. 食品与机械, 2006, 22(6): 83-85, 126.

[19] Li Hong-dong, Liang Yi-zeng, Xu Qing-song, et al. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration[J]. Analytica Chimica Acta, 2009, 648(1): 77-84.

[20] Centner V, Massart D L, Denoord O E, et al. Elimination of uninformative variables for multivariate calibration[J]. Analytical Chemistry, 1996, 68(21): 3 851-3 858.

[21] Galvão R K H, Araújo M C U, Fragoso W D, et al. A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm. Chemometrics and Intelligent Laboratory Systems, 2008, 92(1): 83-91.