基于变量筛选的温州蜜桔品质的光谱快速检测

周 婷[1], 刘苗苗[1], 毛 飞[1], 罗 越[1], 娄淑贮[1], 张文莉[1], 孙一叶[2*]

(1. 温州大学电气与电子工程学院, 温州 325035; 2. 温州大学计划财务处, 温州 325035)

摘 要: 目的 利用可见/近红外光谱技术结合变量筛选算法建立预测模型。**方法** 采集 7 个不同批次蜜桔样本的漫透射光谱, 预处理优化后, 以无信息变量消除法(uninformative variable elimination, UVE)、竞争性自适应重加权法(competitive adaptive reweighting sampling, CARS)及其组合(UVE-CARS)共 3 种策略来进行光谱有效波段的筛选, 建立蜜桔可溶性固形物含量(soluble solid content, SSC)的偏最小二乘预测模型(partial least square, PLS)。**结果** 比较全变量模型和 3 个特征变量模型的预测性能, UVE-CARS-PLS 模型取得了最优的检测效果, 相比全变量模型, 建模变量数减少了 96.5%,其预测集相关系数 R_P 提升至 0.732, 预测集均方根误差(root-mean-square error, RMSEP)下降至 0.873 ^{0}Brix。**结论** 结合多重变量选择算法, 可以进一步压缩建模变量数, 简化模型, 提高模型预测精度, 实现区域蜜桔品质的光谱快速检测。

关键词: 蜜桔; 可见/近红外光谱; 变量选择; 可溶性固形物

Rapid spectral detection of satsuma quality in wenzhou based on variable screening

ZHOU Ting[1], Liu Miao-Miao[1], MAO Fei[1], LUO Yue[1], LOU Shu-Ning[1], ZHANG Wen-Li[1], SUN Yi-Ye[2*]

(1. *College of Electrical & Electronic Engineering*, *Wenzhou University*, *Wenzhou* 325035, *China*;
2. *Department of Planning & Finance*, *Wenzhou University*, *Wenzhou* 325035, *China*)

ABSTRACT: Objective To establish a prediction model by using visible-near infrared spectroscopy technology and variable selection algorithms. **Methods** The diffused transmission spectra of seven different batches of satsumas were collected, and then the spectra were optimized using preprocess methods. Effective spectrum bands were screened by 3 strategies, including uninformative variable elimination (UVE), competitive adaptive reweighting sampling (CARS) and its combination (UVE-CARS), and partial least squares (PLS) prediction model for the soluble solids content (SSC) of satsuma was established **Results** Comparing the prediction performance of the full variable model and the 3 characteristic variable models, the UVE-CARS-PLS model achieved the best detection effect. Compared with the full variable model, the number of modeling variables was reduced by 96.5%, and the correlation coefficient of prediction set (R_P) reached 0.732 and root mean square error (RMSEP) decreased to 0.873 ^{0}Brix. **Conclusion** Combined with the multiple variable selection algorithm, the number of modeling variables can be further compressed, the model can be simplified, the prediction accuracy of the model can be improved, and the spectral detection of regional tangerine quality can be achieved quickly.

基金项目: 大学生创新创业计划项目(JWSC2019112)、温州大学开放实验室项目(JW19SK35)

Fund: Supported by School-level Innovation & Entrepreneurship Training Program(JWSC2019112), and Open laboratory Project of Wenzhou University (JW19SK35)

***通讯作者:** 孙一叶, 硕士, 主要研究为财务信息管理与数据分析。E-mail: 16588875@qq.com

***Corresponding author:** SUN Yi-Ye, Master, Department of Planning & Finance, Wenzhou University, Wenzhou 325035, China. E-mail: 16588875@qq.com

KEY WORDS: satsuma; visible-near infrared spectroscopy; variable selection; soluble solids

1　引　言

温州蜜桔的果肉清甜微酸、饱满多汁, 富含维生素和膳食纤维等营养物质[1]。目前蜜桔内部品质的检测方法一般是: 基于经验的感官评价和破坏性的理化指标检测[2]。前者是基于果农和果商的经验判断来进行品质检测, 容易造成检测结果的不准确。后者虽能取得较高的准确性, 但却以抽检方式破坏果实取汁, 无法实施大批量的检测, 难以满足市场上蜜桔快速分级的需要。因此, 急需开发一种快速、无损的检测方法以提高蜜桔品质的检测手段。

可见/近红外光谱技术作为一种简便、快捷、绿色、低成本以及无损的检测方法, 已广泛应用于食品、能源、医疗、烟草等领域[3-5], 尤其在柑橘[6], 苹果[7]和梨[8]等常见水果中广泛应用。但目前可见/近红外光谱技术检测水果品质大多是应用全波段光谱建立检测模型, 导致数据量较大, 模型计算的复杂度较高且检测精度较低。因此, 有许多研究学者应用变量筛选的方法, 选择那些包含有用信息的波段建立模型实现水果品质的快速检测。采用无信息变量消除法(uninformative variable elimination, UVE)建立了同批次的梨可溶性固形物含量(soluble solid content, SSC)的检测模型, 预测集相关系数 R_P 和均方根误差 RMSEP 分别为 0.893 和 0.158 ^{0}Brix [9]。采用 CARS 算法建立了同批次的草莓 SSC 的检测模型, R_P 和 RMSEP 分别为 0.889 和 0.359 ^{0}Brix [10]。采用连续投影算法(successive projections algorithm, SPA)建立了同批次的西瓜 SSC 的检测模型, 预测集相关系数 R_P 和均方根误差(root-mean-square error, RMSEP)分别为 0.752 和 0.883 ^{0}Brix [11]。针对以上同源批次的果蔬, 单一的变量选择算法虽能提高模型的检测精度, 但实际中不同批次果蔬仍需探究更好的方法以改善模型预测精度不够高的难题。

针对上述问题, 本研究以温州蜜桔为研究对象, 采集 7 个不同批次样本的可见/近红外光谱, 并结合 UVE、竞争性自适应重加权法(competitive adaptive reweighting sampling, CARS)及其组合(UVE-CARS)3 种变量选择方法, 建立蜜桔 SSC 的偏最小二乘预测模型(partial least square, PLS)预测模型, 比较验证多重变量筛选算法优于单一的变量选择算法, 实现压缩建模变量数, 简化模型, 提高模型预测精度, 实现温州地区蜜桔品质的快速检测, 为后续更多算法的应用和光谱设备的开发提供理论依据。

2　材料与方法

2.1　样本收集与仪器

2018 年 11 月于温州地区的不同超市, 采购新鲜蜜桔为试材, 共 7 个批次, 在实验室遴选出无缺陷、不同成熟度的 339 个样本, 于 24 ℃空调室内放置 12 h, 表面清理后编号。每批次样本以 2: 1 的比例随机划分后, 再合并于一起, 226 个样本作为校正集, 113 个样本作为预测集。

QE65 Pro 光谱仪(美国 Ocean Optics 公司); WYA-2S 数字阿贝折光仪(上海精密科学仪器有限公司)。

2.2　光谱采集

使用 QE65 Pro 光谱仪, 光谱范围为 200~1100 nm, 光谱分辨率为 7 nm, 信噪比为 1000:1, 波数点 1048 个。设置扫描次数为 4 次, 积分时间为 100 ms。参考文献[12], 以半透射形式获取蜜桔的透射光谱, 蜜桔的最大横径处对准光谱探头, 每旋转 120°采集一次, 取 3 条光谱的平均值作为该样本的最终透射光谱。

2.3　可溶性固形物的测定

取蜜桔光谱采集点处的果肉, 混合压汁, 用数字阿贝折光仪测量蜜桔 SSC 的值, 该仪器测量范围 0-高, 测量分辨率为 0.1%。参考 NY/T 2637-2014《水果和蔬菜可溶性固形物含量的测定-折射仪法》[13], 3 次测量取平均。校正集和预测集的蜜桔 SSC 的统计测量结果见表 1。

表 1　蜜桔的 SSC(^{0}Brix)的统计结果(n=3)
Table 1　Statistic results of SSC (^{0}Brix) of satsuma samples(n=3)

	样本量	含量范围	平均值	标准偏差
校正集	226	8.75 ~ 16.8	12.31	2.06
预测集	113	8.95 ~ 15.65	12.12	1.60

2.4　模型的建立与评价

为构建蜜桔 SSC 值与光谱信号之间的定量关系, 采用 PLS 算法[14]构建校正模型。通过对比几种预处理方法的 PLS 模型效果, 经多元散射校正(multiple scattering correction, MSC)算法[15]处理的光谱所建模型表现最佳, 可消除光谱采集过程中存在的光散射信息, 同时对光程误差进行修正, 改善基线漂移的现象。此外, 为减少建模的变量数, 采用 UVE 算法、CARS 算法及其组合 UVE-CARS 算法建立回归模型。

UVE 算法是利用 PLS 回归系数开发的用于消除无信息变量, 留下有用信息变量[16,17]。CARS 算法是通过自适应重加权采样技术, 保留回归模型中回归系数绝对值大的波长变量, 经多次筛选获得一系列波长变量子集, 消除冗余信息[18,19]。以相关系数 R、均方根误差 RMSE 来评价模型的预测性能。所有计算于 Matlab2016a 运行。

3 结果与讨论

3.1 光谱分析

图 1 为蜜桔样本的平均光谱图。因光谱两端存在部分噪声，故截取 230~950 nm 内的波段共计 971 个变量用于数据分析。由图可知，700 nm 和 825 nm 附近的波峰可能是由于 O-H 和 N-H 的伸缩振动引起的，该信号强度应该与蜜桔的 SSC 值有关。细小的光谱变化差异不足以定性样本指标的高低，需结合化学计量学方法进行数据建模、定量分析。

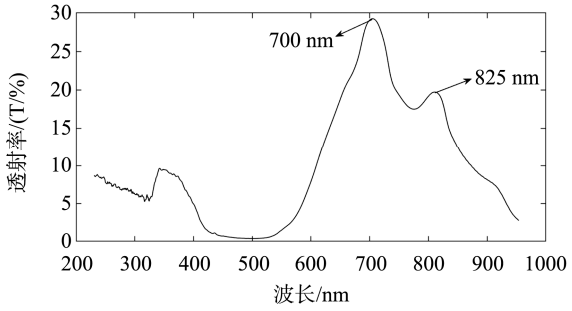

图 1 蜜桔样本光谱平均后的谱图

Fig.1 Averaged spectra of satsuma samples

3.2 全变量模型的构建与分析

采用 MSC 算法对原始光谱进行预处理，971 个变量用于建立基于 PLS 的蜜桔 SSC 的回归模型。校正集相关系数 R_{CV} 和均方根误差 RMSECV 分别为 0.691 和 1.039，预测集相关系数 R_P 和均方根误差 RMSEP 分别为 0.601 和 1.011。而由原始光谱数据建立的 PLS 模型的校正集相关系数 R_{CV} 和均方根误差 RMSECV 分别为 0.685 和 1.050，预测集相关系数 R_P 和均方根误差 RMSEP 分别为 0.600 和 1.010。比较结果可知，经光谱预处理后所建模型的预测效果较好，但这种基于全变量的 PLS 模型对蜜桔 SSC 的预测精度仍然较低，难以满足日常工业生产的需求。

3.3 变量筛选模型的构建

利用 UVE 算法人为地将随机噪声加光谱矩阵中，如图 2(a)所示，左侧为光谱波长信息，右侧为随机噪声，以噪声值为阈值，剔掉稳定性低于阈值的不提供信息的变量，留下有用信息变量。连续运行 UVE10 次，取 RMSECV 值最小的一次作为模型的运行结果，共计 263 个特征变量被筛选。

利用 CARS 算法对全变量筛选，如图 2(b)所示，在前 22 次(*处)的采样过程中，无关变量被剔除，波长变量数不断减少，RMSECV 值也不断降低，但随着采样次数增加，开始剔除有关变量，RMSECV 值上升，因此模型 RMSECV 值最小，采样次数为 22 次时的 113 个波长变量入选。

仅通过单一的 UVE 或 CARS 算法筛选的变量数仍较多，无法进一步降低模型复杂度提高模型精度。UVE 算法侧重剔除稳定性低的变量，但剩下的稳定变量不一定都有用，而 CARS 算法则侧重保留有用信息变量，因此尝试将两者的优势结合，经 UVE 筛选后的变量，CARS 继续筛选，模型运行得到共计 34 个变量。

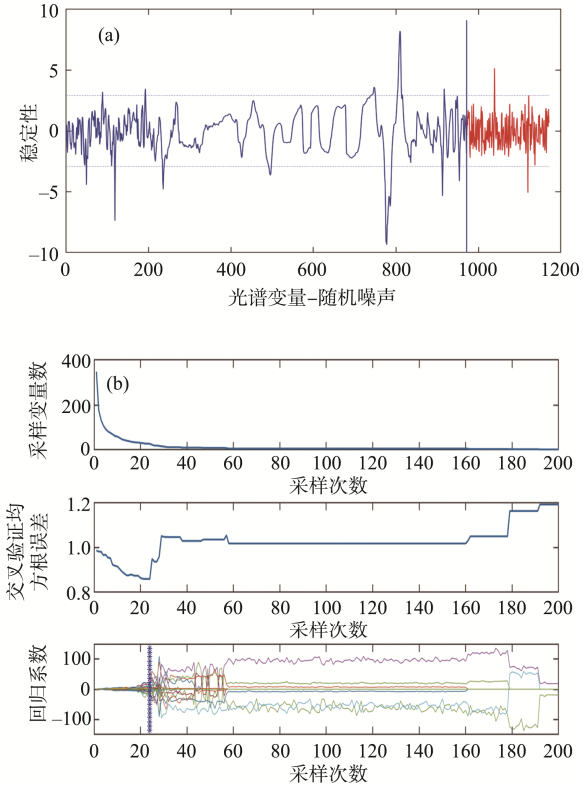

图 2 UVE 算法的光谱稳定性分布(a)及 CARS 算法的变量变化趋势(b)

Fig.2 Spectral stability distribution of UVE method (a) and the change trend of the number of CARS method (b)

3.4 不同模型结果比较

表 2 是基于 3 种变量筛选算法的模型结果比较，采用 UVE 和 CARS 算法对全波段筛选分别得到 263 个和 113 个特征变量，建模变量数分别减少了 72.9%和 88.4%。其中，UVE-PLS 模型的预测集相关系数 R_P 和均方根误差 RMSEP 分别为 0.621 和 0.998 ^{0}Brix，预测效果较优于 CARS-PLS 模型，但相比于全变量 PLS 模型，未见显著效果。而采用 UVE-CARS 算法筛选得到 34 个有效变量，建模变量数减少了 96.5%，相较于全变量 PLS 模型，其预测集相关系数 R_P 提升至 0.732，预测集均方根误差 RMSEP 下降至 0.873 ^{0}Brix。基于 3 种变量筛选算法的变量分布如图 3 所示，其中 UVE 筛选的变量零散地分布在整个光谱区域，而 CARS 筛选的变量集中于 600~850 nm 光谱波段。经过二次变量筛选后，得到的变量主要分布在 460、520、790 nm 附近，所

以这 3 个波长是这批数据的关键变量, 特别是 790 nm 附近的变量。原始光谱在 700 nm 和 825 nm 处出现吸收峰, 在近红外波段, 这些光谱变量位置主要与水果的 O-H/N-H 的倍频、三级倍频有关[20,21], 这正验证了筛选的光谱信号与糖度的分子官能团振动有关, 并且该信号强度可能与糖度也有关系。

图 4 为 UVE-CARS-PLS 模型预测集的预测值和实测值的散点图, 预测值均匀分布在等值线附近。结果表明, UVE-CARS 这种多重变量选择算法在一定程度上可简化模型提高模型精度。但单一采集 7 个不同批次的样本所建立的模型针对该地区样本仍没有较好的预测效果, 仍不具有强代表性, 模型分级筛选的精度仍存在改善空间。在后续的研究中, 应将更多批次, 不同年份的蜜桔样本数据纳入计算, 使数据量足够充分地具有强代表性。

表 2　基于 3 种变量筛选算法的模型结果比较

Table 2　Comparison of model results based on 3 different filtering variable algorithms

建模方法	变量数	主成分数	校正集		预测集	
			R_{CV}	RMSECV	R_P	RMSEP
PLS	971	8	0.691	1.039	0.601	1.011
UVE-PLS	263	11	0.716	1.004	0.621	0.998
CARS-PLS	113	8	0.673	1.064	0.619	0.998
UVE-CARS-PLS	34	11	0.704	0.949	0.732	0.873

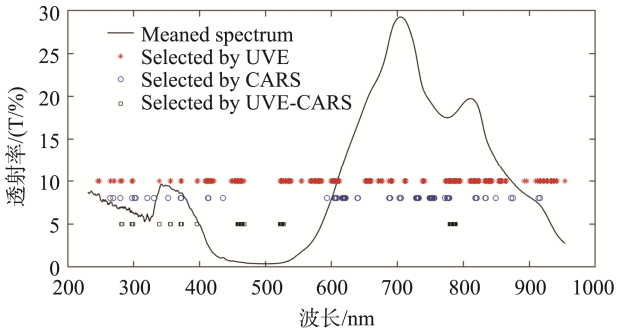

图 3　基于 3 种变量筛选算法的变量分布

Fig.3　Variable distribution based on 3 variable screening algorithms

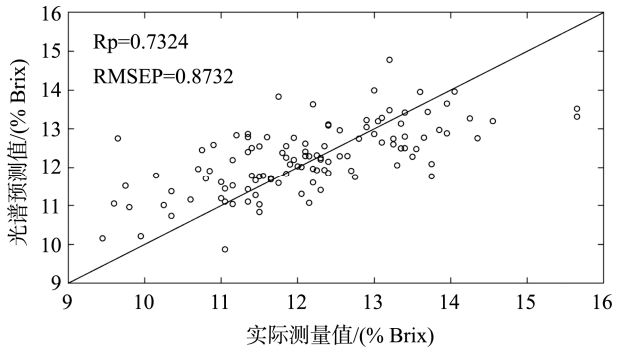

图 4　预测集的温州蜜桔 SSC 预测结果

Fig.4　Prediction results of satsuma SSC in the predicted set

4　结　论

无损地获取不同批次温州蜜桔的光谱信息, 以表征其内部品质。改进单一 UVE 和 CARS 算法, 采用多重(UVE-CARS)变量筛选算法。对比 UVE-PLS、CARS-PLS 和 UVE-CARS-PLS 模型, UVE-CARS-PLS 模型具有最好的预测性能, 但提升有限, 在今后的实践中, 应尝试不同变量筛选方法来优化多重算法组合, 减少无用变量的干扰, 提高模型精度, 实现温州蜜桔品质的光谱快速检测。

参考文献

[1] 龚晓, 张有捷, 李赤翎, 等. 几种蜜桔类水果功能性营养物质测定与比较分析[J]. 食品与机械, 2012, (2): 42-45.
Gong X, Zhang YJ, Li CL, *et al*. Determination and comparative analysis of functional nutrients of several tangerine fruits [J]. Food Mach, 2012, (2): 42-45.

[2] 袁雷明, 孙力, 林颢, 等. 基于感官品尝的柑橘糖度近红外光谱模型的简化[J]. 光谱学与光谱分析, 2013, (9): 85–89.
Yuan LM, Sun L, Lin H, *et al*. Simplification of near-infrared spectral model of citrus sugar content based on sensory tasting [J]. Spectrosc Spect Anal, 2013, (9): 85–89.

[3] Haider SZ, Lohani H, Bhandari, *et al*. Nutritional value and volatile composition of leaf and bark of cinnamomum tamala from Uttarakhand (India) [J]. J Essent Oil Bear Pl, 2018, 21(3): 732–740.

[4] Rungpichayapichet P, Mahayothee B, Khuwijitjaru P, *et al*. Non-destructive determination of *β*-carotene content in mango by near-infrared spectroscopy compared with colorimetric measurements [J]. J Food Compos Anal, 2015, 38: 32–41.

[5] Pornprasit R, Natwichai J. Prediction of mango fruit quality from NIR spectroscopy using an ensemble classification [J]. J Netw Comput Appl, 2013, 83(14): 25–30.

[6] 王平, 聂振朋, 罗君琴, 等. 无损伤检测技术在柑橘果实中的应用[J]. 浙江柑橘, 2013, 30(4): 7–10.
Wang P, Nie ZP, Luo JQ, *et al*. Application of nondestructive detecting technology in citrus fruits [J]. Zhejiang Citrus, 2013, 30(4): 7–10.

[7] Giovanelli G, Sinelli N, Beghi R, *et al*. NIR spectroscopy for the optimization of postharvest apple management [J]. Postharvest Biol

Technol, 2014, 87: 13–20.

[8] Suh SR, Lee KH, Yu SH, *et al*. Comparison of performance of measuring method of VIS/NIR spectroscopic spectrum to predict soluble solids content of 'Shingo' pear [J]. Biosyst Eng, 2011, 36(2): 130–139.

[9] 王铭海, 郭文川, 商亮, 等. 基于近红外漫反射光谱的多品种桃可溶性固形物的无损检测[J]. 西北农林科技大学学报, 2014, 42(2): 142–148.
Wang MH, Guo WC, Shang L, *et al*. Nondestructive detection of soluble solids in peach varieties based on near infrared diffuse reflectance spectra [J]. J Northwest A & F Univ, 2014, 42(2):142–148.

[10] 李江波, 彭彦昆, 陈立平, 等. 近红外高光谱图像结合 CARS 算法对鸭梨 SSC 含量定量测定[J]. 光谱学与光谱分析, 2014, (5): 1264–1269.
Li JB, Peng YK, Chen LP, *et al*. Near-infrared hyperspectral image combined with CARS algorithm for quantitative determination of SSC in “Ya” pear [J]. Spectrosc Spect Anal, 2014, (5): 1264–1269.

[11] 钱曼, 黄文倩, 王庆艳, 等. 西瓜检测部位差异对近红外光谱可溶性固形物预测模型的影响[J]. 光谱学与光谱分析, 2016, (6): 1700–1705.
Qian M, Huang WQ, Wang QY, *et al*. The effect of the difference of watermelon detection position on the prediction model of soluble solids in near infrared spectroscopy [J]. Spectrosc Spect Anal, 2016, (6): 1700–1705.

[12] Yuan LM, Cai JR, Sun L, *et al*. Nondestructive measurement of soluble solids content in apples by a portable fruit analyzer [J]. Food Anal Method, 2016, 9(3): 785–794.

[13] NY/T 2637-2014 水果和蔬菜可溶性固形物含量的测定-折射仪法[S].
NY/T 2637-2014 Determination of soluble solid content of fruits and vegetables-Refractometer method [S].

[14] Nanni MR, Cezar E, Junior CADS, *et al*. Partial least squares regression (PLSR) associated with spectral response to predict soil attributes in transitional lithologies [J]. Arch Agron Soil Sci, 2017, 64(5): 682–695.

[15] Peterson HM, Bang HH, Geller D, *et al*. Abstract 699: Clinical feasibility of chemotherapy monitoring for bone sarcoma patients with diffuse optical spectroscopic imaging [J]. Cancer Res, 2017, 77(13 S): 699.

[16] Centner V, Massart DL, De Noord OE, *et al*. Elimination of uninformative variables for multivariate calibration [J]. Anal Chem, 1996, 68(21): 3851–3858.

[17] Yuan LM, Chen XJ, Lai YJ, *et al*. A novel strategy of clustering informative variables for quantitative analysis of potential toxics element in tegillarca granosa using laser-induced breakdown spectroscopy [J]. Food Anal Method, 2018, 11(5): 1405–1416.

[18] Li HD, Liang YZ, Xu QS, *et al*. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration [J]. Anal Chim Acta, 2009, 648(1): 77–84.

[19] Zhang XY, Li, QB, Zhang GJ. Calibration transfer without standards for spectral analysis based on stability competitive adaptive reweighted sampling [J]. Spectrosc Spect Anal, 2014, 34(5): 1429–1433.

[20] Nicolai BM, Beullens K, Bobelyn E, *et al*. Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review [J]. Postharvest Biol Tec, 2007, 46(2): 99–118.

[21] Zou XB, Zhao JW, Povey MJW, *et al*. Variables selection methods in near-infrared spectroscopy [J]. Anal Chim Acta, 2010, 667(1–2): 14–32.

(责任编辑：韩晓红)

作者简介

周　婷，硕士研究生，主要研究方向为光谱信息的检测与控制技术。
E-mail: 2334986022@qq.com

孙一叶，会计师，主要研究为农业财务信息管理与数据分析。
E-mail: 16588875@qq.com