DOI:10.13652/j.issn.1003-5788.2020.02.016

基于可见/近红外光谱和变量选择的脐橙可溶性固形物含量在线检测

Online detection of soluble solid content in navel orange based on visible / near infrared spectroscopy and variable selection

江水泉　　孙　通

JIANG Shui-quan　SUN Tong

(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:为联合可见/近红外光谱技术和变量选择方法在线检测脐橙主要内部品质指标可溶性固形物(SSC),分别选定脐橙校正集和预测集样本 141 个和 47 个,脐橙运输速度为 0.3 m/s,利用 USB4000 微型光谱仪在线采集脐橙样本的可见/近红外光谱,先分别采用无信息变量消除(UVE)和遗传算法(GA)对 650～950 nm 波段范围的波长变量进行预筛选,再分别利用竞争自适应重加权采样(CARS)及连续投影算法(SPA)对波长变量进一步筛选,并应用偏最小二乘(PLS)方法分别建立脐橙 SSC 的在线预测模型,并与原始光谱等建立的预测模型进行比较。结果表明,对于脐橙 SSC,预筛选方法 GA 优于 UVE 方法,变量选择方法 CARS 优于 SPA 方法;GA-CARS 及 GA-SPA 联合变量选择方法优于对应的单一变量选择方法 CARS 及 SPA。在上述变量选择方法中,GA-CARS 方法获得的结果最优,其所建立的脐橙 SSC 的 PLS 模型的校正集和预测集相关系数分别为 0.933 和 0.824,校正集和预测集均方根误差分别为 0.429%和 0.670%,性能优于原始光谱建立的 PLS 模型,且建模波长变量数由 1 385 个下降为 78 个,仅占原波长变量数的 5.63%。由此表明,GA-CARS 联合变量选择方法可以有效筛选脐橙 SSC 的波长变量,提高预测模型的稳定性和预测精度。

关键词:可见/近红外;变量选择;竞争自适应重加权采样;遗传算法;可溶性固形物;脐橙

Abstract: Soluble solids content (SSC) is the main internal quality index of navel orange, in order to detect the SSC of navel orange by the combination of visible/near infrared spectroscopy and variable selection method, 141 samples of calibration set and 47 samples of prediction set were used. The transportation speed of navel orange was 0.3 m/s. The visible/near infrared spectra of navel orange samples were collected online by a USB4000 micro spectrometer. Firstly, uninformative variable elimination (UVE) and genetic algorithm (GA) were used to prescreen the wavelength variables in the wavelength range of 650～950 nm, then competitive adaptive weighted sampling (CARS) and successive projections algorithm (SPA) were used to further screen the wavelength variables. Furthermore, partial least squares (PLS) method was used to establish the online prediction models of SSC of navel orange, and these prediction models were compared with the prediction model established using original spectra. The results indicate that, for SSC of navel orange, GA method is better than UVE method in pre screening, while CARS method is better than SPA method in variable selection. GA-CARS and GA-SPA combined variable selection method is better than the corresponding single variable selection methods CARS and SPA. GA-CARS method obtains the best results for SSC of navel orange among the above variable selection methods, with the correlation coefficients of PLS model of SSC of navel orange in calibration and prediction set of 0.933 and 0.824 respectively, and the root mean square errors of calibration and prediction set are 0.429% and 0.670%, respectively. The performance of GA-CARS-PLS model is better than that of PLS model established by original spectra, and the number of modeling wavelength variables reduces from 1 385 to 78, only accounting for 5.63% of the number of original wavelength variables. In conclusion, the combined variable selection method of GA-CARS can effectively screen the wavelength variables of SSC of navel orange, and improve the stability and prediction accuracy of the prediction model.

Keywords: visible/near infrared; variable selection; competitive

基金项目:江苏省重点研发专项资金(编号:BE2017302)

作者简介: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:2019-12-23

adaptive weighted sampling; genetic algorithm; soluble solids content; navel orange

可见/近红外光谱技术是一种快速、无损、绿色的现代检测技术，其根据分析物的 C—H、C—C 及 O—H 等的合频与倍频吸收进行定性及定量分析。目前，该光谱技术已应用于玉米淀粉[1]、肉类脂肪[2]、鱼肉新鲜度[3]、茶叶种类[4]、牛奶蛋白质[5]、当归阿魏酸[6]及食用油掺假[7]等检测。对于水果可溶性固形物 SSC 检测，刘燕德等[8]利用近红外漫反射光谱技术在线检测脐橙 SSC 含量。偏最小二乘(PLS)模型的预测相关系数为0.90，预测均方根误差(RMSEP)为 0.61。韩东海等[9]建立了 3 种摆放方式的苹果 SSC 在线预测模型。对于上置式检测器而言，遮光处理和苹果摆放方式最为重要；PLS 模型的预测相关系数和 RMSEP 分别为 0.87 和 0.67。郭成等[10]采用无信息变量消除(UVE)方法优选无花果 SSC 的特征波长，并应用 PLS 方法建立无花果 SSC 的在线预测模型，其预测相关系数为 0.83～0.89，RMSEP 为 0.63～0.83°Brix。Tian 等[11]采用光谱预处理和变量选择方法对苹果 SSC 在线预测模型进行优化。随机森林方法筛选的特征波长建立的 SSC 预测模型最优，模型的预测相关系数和 RMSEP 分别为0.904 3和0.478 7。Xu 等[12]研究比较了单点和双点检测对苹果 SSC 在线检测精度的影响。此外，还有其他学者[13-16]也对水果 SSC 进行在线检测研究。综合分析上述文献可知，不少学者采用变量选择方法筛选水果 SSC 的特征变量来简化和提高预测模型性能，但基本是采用单一的变量选择方法。由于可见/近红外光谱波长变量众多，数量可达几百甚至上千，含有较多冗余及干扰变量，采用单一方法进行波长变量筛选易受冗余及干扰变量影响，从而影响 SSC 检测精度和稳定性。因此，有必要探索联合两种变量选择方法筛选 SSC 特征变量的研究。

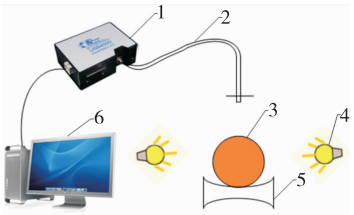
试验拟采用可见/近红外光谱技术对脐橙 SSC 含量进行在线检测。利用遗传算法(GA)和 UVE 方法对波长变量进行预筛选，在此基础上再采用竞争自适应重加权采样(CARS)及连续投影算法(SPA)进一步筛选特征波长变量，并应用 PLS 方法建立脐橙 SSC 的在线预测模型。

1　材料与方法

1.1　试验材料与检测系统

试验所用的脐橙样本购买于当地水果批发市场，脐橙质量范围为 175～327 g，数量共计 188 个。为保证校正集样本的合理性和代表性，按照脐橙样本 SSC 测量值进行排序，将最大及最小 SSC 测量值的脐橙样本直接分配到校正集，而后按 3∶1 比例并结合排序将脐橙样本分配到校正集和预测集。经分配后，校正集和预测集的脐橙样本分别为 141 个和 47 个。

试验所用的在线光谱检测系统如图 1 所示，由光谱仪、光纤、光源、输送系统及电脑等组成。光谱仪为 USB4000 微型光纤光谱仪(美国 Oceanoptics 公司)，配置 3 648 像素 CCD。光源为 2 盏 150 W 卤钨灯，功率共 300 W。光源分布在脐橙赤道两侧，光源—脐橙—光纤的角度为 90°。脐橙传输速度为 0.3 m/s。



1. 光谱仪 2. 光纤 3. 水果 4. 光源 5. 果托 6. 电脑

图 1　可见/近红外光谱在线检测系统原理图

Figure 1　Schematic diagram of on-line visible/near infrared spectrum detection system

1.2　光谱采集

样本光谱采集前，先采集暗场和参比光谱。关闭光源，所采集的光谱即为暗场光谱；以聚四氟乙烯球(直径 80 mm)为参比，在图 1 所示的在线检测系统中获得其参比光谱。对于脐橙样本，按顺序依次在线采集每个脐橙的可见/近红外光谱。暗场、参比及脐橙样本的积分时间均为 60 ms，平均采集次数设为 1，光谱平滑点数设为 6。

1.3　SSC 测定

将脐橙样本去皮，采用榨汁机破碎脐橙果肉，并通过普通快速滤纸对脐橙果汁进行过滤，然后将过滤后的果汁滴到 PR-101α 型数字折射仪(日本 Atago 公司)的测量窗口，进行脐橙 SSC 真实含量的测定。

1.4　数据处理与分析

对脐橙样本光谱，先采用无信息变量消除和遗传算法分别对波长变量进行预筛选，在此基础上，再分别利用 CARS 和 SPA 方法进一步筛选波长变量；对上述方法筛选的波长变量，应用 PLS 方法分别建立脐橙 SSC 含量的在线预测模型，并比较模型性能的优劣。

无信息变量消除方法的参数设置为随机噪声矩阵的变量数为 1 385 个，与建模的光谱波长变量数一致，提取的最大主成分数为 15。UVE 算法的具体原理参见文献[17]。对于遗传算法，其种群大小及染色体长度分别为 30 和 30，变异概率及交叉概率分别为 1%和 50%，遗传代数为 100。GA 算法的具体原理参见文献[18]。在 CARS 分析中，采样次数为 50 次，提取的最大主成分数由蒙特卡罗交互验证确定。CARS 算法的具体原理参见文献[19]。在 SPA 分析中，最大及最小可选的波长变量数

分别设为 40 和 1。SPA 算法的具体原理参见文献[20]。UVE、GA、CARS 及 SPA 方法均在 Matlab 7.6.0 软件(美国 The Math Works 公司)里运行完成,PLS 方法则在 Unscrambler X 10.1 软件(挪威 CAMO 公司)里运行完成。

SSC 预测模型的性能主要由相关系数(r)、校正均方根误差(RMSEC)及 RMSEP 进行评价。预测模型的相关系数越高,RMSEC 及 RMSEP 越小,且两者较为相近,则预测模型的性能越好。

2　结果与讨论

2.1　SSC 统计分析

由表 1 可知,所有样本的 SSC 平均值、标准偏差及范围分别为 11.54%,1.19%,8.3%～14.9%。校正集和预测集样本的 SSC 平均值分别为 11.53%,11.56%,标准偏差分别为 1.20%,1.17%。校正集样本的 SSC 范围涵盖预测集样本,表明校正集样本具有一定的代表性,所建立的 SSC 预测模型能适用于预测集样本。

表 1　脐橙样本可溶性固形物的主要统计结果

Table 1　Main statistical results of soluble solids content in navel orange samples

数据集	样本数	平均值/%	标准偏差/%	范围/%
全部	188	11.54	1.19	8.3～14.9
校正集	141	11.53	1.20	8.3～14.9
预测集	47	11.56	1.17	9.2～14.4

2.2　光谱分析

由图 2 可知,所有脐橙样本的光谱形状均较为相似,表明光谱在线采集没有出现异常。脐橙样本光谱在 700 nm及 820 nm 左右有较为明显的波谷,而在 725 nm 及 830 nm 左右存在较为明显的波峰,此部分区域含有较多有用的光谱信息。样本光谱两端波动大,光谱信噪比低。为了避免引入明显的光谱噪声和丢失有用的光谱信息,后续的光谱分析在 650～950 nm 波段范围进行,该波段范围共有 1 385 个波长变量。

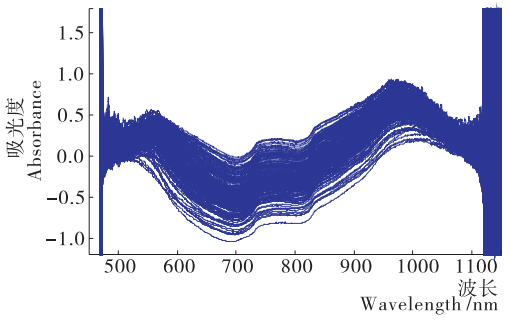


图 2　所有脐橙样本的可见/近红外光谱

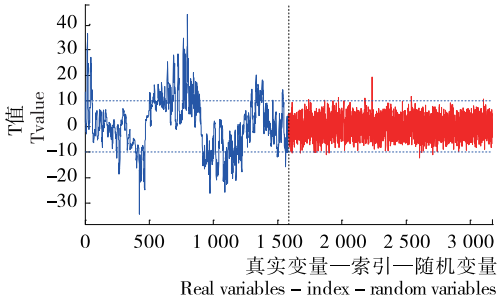
Figure 2　Visible/near infrared spectra of all navel orange samples

2.3　波长变量选择

2.3.1　UVE　采用 UVE 方法对 650～950 nm 波段范围的光谱数据进行波长变量筛选。图 3 中,波长变量的稳定值在两水平虚线之外的将作为有用信息波长变量被保留,而在两水平虚线之内的将作为冗余或噪声波长变量被去除。经 UVE 变量筛选后,共有 884 个波长变量被去除,501 个波长变量被保留。

2.3.2　GA　采用 GA 方法对 650～950 nm 波段范围的光谱数据进行波长变量筛选。图 4 中,被选中频率大于阈值的波长变量将作为有用信息波长变量被保留,而被选中频率小于阈值的波长变量将作为冗余或噪声波长变量被去除。经 GA 方法筛选后,共有 1 203 个波长变量被去除,182 个波长变量被保留。其中,在 725～755 nm 波段范围有较多的波长变量被保留。

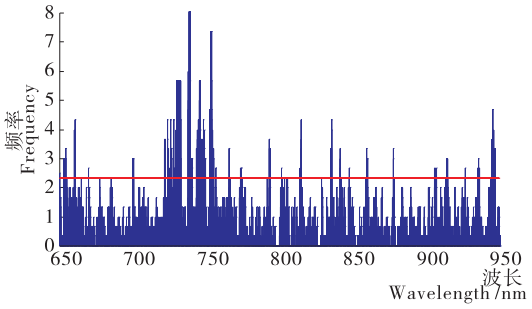
2.3.3　CARS　在 UVE 和 GA 变量预筛选的基础上,采用 CARS 方法分别对上述结果进行进一步变量筛选。对于 UVE 筛选的 501 个波长变量,经 CARS 方法筛选后,共有 187 个波长变量被保留。由图 5(a)可知,被选择的波长变量主要集中在 725～880 nm,其中 725～800 nm 波



竖虚线左侧为真实的波长变量,右侧为随机噪声变量;两水平虚线为 UVE 稳定性的阈值

图 3　脐橙 SSC 的 UVE 分析结果

Figure 3　Results of UVE analysis of SSC in navel oranges



水平横线为波长变量筛选的阈值

图 4　脐橙 SSC 的 GA 分析结果

Figure 4　Results of GA analysis of SSC in navel oranges

段范围有 61 个，801～880 nm 波段范围有80 个；而在 650～724 nm 波段范围有 21 个，881～950 nm 波段范围有25 个。对于 GA 筛选的 182 个波长变量，经 CARS 方法筛选后，共有 78 个波长变量被保留。由图 5(b)可知，被选择的波长变量主要分布在 725～800 nm，而其他波段则较少。在 650～724，725～800，801～880，881～950 nm 波段范围分别有 7，50，16，5 个。

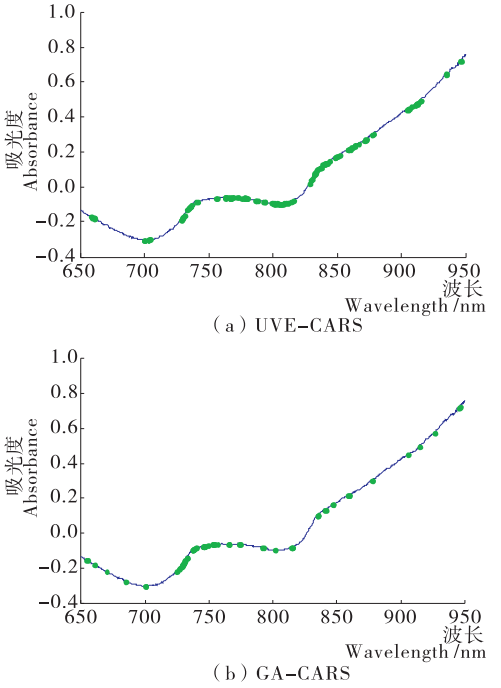


图 5　被选择波长变量的分布情况

Figure 5　Distribution of selected wavelength variables

2.3.4　SPA　在 UVE 和 GA 变量预筛选的基础上，采用 SPA 方法分别对上述结果进行进一步变量筛选。由图 6 可知，经 UVE-SPA 方法后，共有 8 个波长变量被选择，分别为 660.98，703.05，736.24，766.10，874.61，884.31，914.24，935.02 nm；经 GA-SPA 方法后，共有 16 个波长变量被选择，分别为 660.57，669.61，684.99，722.16，731.36，738.78，745.59，753.54，756.64，773.79，814.70，834.98，859.88，877.36，906.30，913.34 nm。

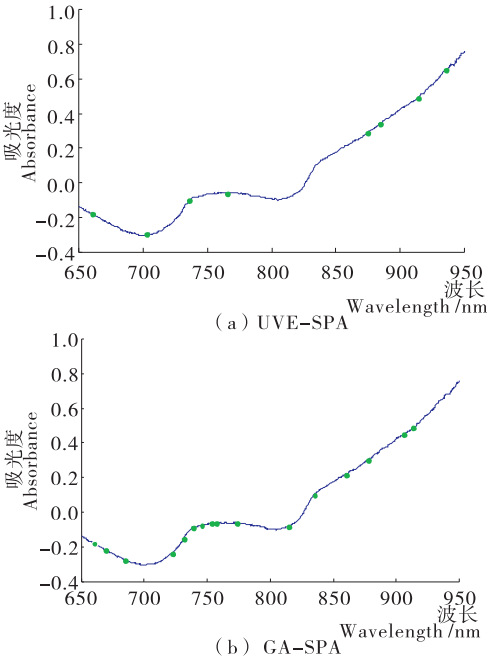


图 6　被选择波长变量的分布情况

Figure 6　Distribution of selected wavelength variables

2.4　PLS 模型建立与比较

对于 UVE-CARS、UVE-SPA、GA-CARS 及 GA-SPA 方法筛选的波长变量，应用 PLS 方法分别建立脐橙 SSC 的在线预测模型，并与直接采用 CARS 和 SPA 方法建立的预测模型及原始光谱建立的预测模型进行比较。

由表 2 可知：

(1) GA-SPA-PLS 模型的性能优于 UVE-SPA-PLS，GA-CARS-PLS 模型的性能优于 UVE-CARS-PLS，表明 GA 方法优于 UVE 方法，GA 方法更适合于光谱波长变量的预筛选。此外，UVE-CARS-PLS 模型性能优于 UVE-SPA-PLS，GA-CARS-PLS 模型性能优于 GA-SPA-PLS，CARS-PLS模型性能优于SPA-PLS，表明CARS方

表 2　不同变量选择方法下脐橙可溶性固形物的 PLS 建模结果

Table 2　The results of PLS regression of SSC of navel oranges on different variable selection methods

方法	变量数	r_c	校正均方根误差/%	r_p	预测均方根误差/%
无	1 385	0.937	0.416	0.778	0.731
UVE-CARS	187	0.835	0.656	0.657	0.876
GA-CARS	78	0.933	0.429	0.824	0.670
UVE-SPA	8	0.795	0.725	0.484	1.049
GA-SPA	16	0.891	0.541	0.734	0.793
CARS	85	0.957	0.348	0.737	0.805
SPA	17	0.830	0.655	0.529	1.004

法筛选有用波长变量更为有效。

(2) GA-CARS-PLS 模型性能优于 CARS-PLS,GA-SPA-PLS 模型性能优于 SPA-PLS,即以 GA 方法为变量预筛选的联合变量选择方法优于对应的单一变量选择方法。由此可见,对于脐橙 SSC,采用 GA 方法用于波长变量的预筛选非常必要。

此外,在所建立的预测模型中,GA-CARS-PLS 模型的性能最优,与原始光谱建立的 PLS 模型相比,其校正集相关系数略有下降,但预测集相关系数由 0.778 上升为 0.824,RMSEP 由 0.731%下降为 0.670%,且建模所用的波长变量数由 1 385 个下降为 78 个,仅占原波长变量数的 5.63%。表明 GA-CARS 联合变量选择方法能有效筛选波长变量,从而提高 SSC 预测模型的稳定性和预测精度。

3 结论

利用可见/近红外光谱技术在线检测脐橙可溶性固形物含量,采用 UVE 和 GA 方法对波长变量进行预筛选,再利用 CARS 及 SPA 方法进行波长变量筛选,并应用 PLS 方法建立 SSC 预测模型。结果表明,对于脐橙 SSC,变量预筛选方法 GA 优于 UVE,变量选择方法 CARS 优于 SPA,以 GA 为变量预筛选的联合变量选择方法优于对应的单一变量选择方法(CARS、SPA),GA-CARS 联合变量选择方法所获得的结果最优。与原始光谱相比,GA-CARS-PLS 模型的预测集相关系数由 0.778 上升为 0.824,RMSEP 由 0.731%下降为 0.670%,且建模所用的波长变量数由 1 385 个下降为 78 个,仅占原波长变量数的 5.63%。由此可见,GA-CARS 联合变量选择方法能有效筛选脐橙 SSC 的波长变量,提高 SSC 预测模型的稳定性和预测精度。

参考文献

[1] JIANG Hao, LU Jian-gang. Using an optimal CC-PLSR-RBFNN model and NIR spectroscopy for the starch content determination in corn[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2018, 196: 131-140.

[2] PRIETO N, DUGAN M E R, JUÁREZ M, et al. Using portable near-infrared spectroscopy to predict pig subcutaneous fat composition and iodine value[J]. Canadian Journal of Animal Science, 2018, 98(2): 221-229.

[3] 陶瑞, 史智佳, 贡慧, 等. 傅里叶变换近红外光谱技术快速检测金枪鱼新鲜度[J]. 肉类研究, 2017, 31(4): 43-49.

[4] LIU Yun, PENG Qing-wei, YU Jian-cheng, et al. Identification of tea based on CARS-SWR variable optimization of visible/near-infrared spectrum [J]. Journal of the Science of Food and Agriculture, 2020, 100(1): 371-375.

[5] WANG Yan-zeng, GUO Wen-chuan, ZHU Xin-hua, et al. Effect of homogenisation on detection of milk protein content based on NIR diffuse reflectance spectroscopy[J]. International Journal of Food Science and Technology, 2019, 54(2): 387-395.

[6] 雷晓晴, 王秀丽, 李耿, 等. 近红外光谱法快速测定当归中 7 种成分的含量[J]. 中草药, 2019, 50(16): 3 947-3 954.

[7] 冼瑞仪, 黄富荣, 黎远鹏, 等. 可见和近红外透射光谱结合区间偏最小二乘法(iPLS)用于橄榄油中掺杂煎炸老油的定量分析[J]. 光谱学与光谱分析, 2016, 36(8): 2 462-2 467.

[8] 刘燕德, 翟建龙. 脐橙可溶性固形物的在线近红外光谱检测[J]. 西北农林科技大学学报: 自然科学版, 2014, 42(3): 186-190.

[9] 韩东海, 周恩洋, 戚淑叶. 苹果糖度在线检测降低杂散光影响研究[J]. 农业机械学报, 2015, 46(11): 260-265.

[10] 郭成, 梁梦醒, 江明珠, 等. 在线检测无花果中可溶性固形物的近红外漫透射技术研究[J]. 江苏科技大学学报: 自然科学版, 2018, 32(2): 285-290, 297.

[11] TIAN Xi, FAN Shu-xiang, LI Jiang-bo, et al. Comparison and optimization of models for SSC on-line determination of intact apple using efficient spectrum optimization and variable selection algorithm[J]. Infrared Physics & Technology, 2019, 102: 102979.

[12] XU Xiao, MO Jian-can, XIE Li-juan, et al. Influences of detection position and double detection regions on determining soluble solids content (SSC) for apples using on-line visible/near-infrared (Vis/NIR) spectroscopy[J]. Food Analytical Methods, 2019, 12(9): 2 078-2 085.

[13] SHEN F, ZHANG B, CAO C J, et al. On-line discrimination of storage shelf-life and prediction of post-harvest quality for strawberry fruit by visible and near infrared spectroscopy [J]. Journal of Food Process Engineering, 2018, 41(7): e12866.

[14] 刘燕德, 朱丹宁, 吴明明, 等. 玉露香梨可溶性固形物近红外漫透射光谱在线检测[J]. 食品与机械, 2016, 32(10): 115-119, 163.

[15] 孙通, 江水泉. 基于可见/近红外光谱和变量优选的南水梨糖度在线检测[J]. 食品与机械, 2016, 32(3): 69-72.

[16] 陆辉山, 陈鹏强, 闫宏伟, 等. 基于近红外光谱漫透射技术的苹果可溶性固形物含量在线检测[J]. 食品与机械, 2015, 31(3): 40-42.

[17] 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.

[18] LEARDI R. Application of genetic algorithm-PLS for feature selection in spectral data sets[J]. Journal of Chemometrics, 2000, 14(5/6): 643-655.

[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] 陈斌, 孟祥龙, 王豪. 连续投影算法在近红外光谱校正模型优化中的应用[J]. 分析测试学报, 2007, 26(1): 66-69.