秦善知,陈　斌,陆道礼,等. 基于便携式近红外光谱仪检测梨可溶性固形物[J]. 江苏农业科学,2014,42(8):284 - 286.

基于便携式近红外光谱仪检测梨可溶性固形物

秦善知[1], 陈　斌[1], 陆道礼[1], 颜　辉[2]

(1. 江苏大学食品与生物工程学院,江苏镇江 212013; 2. 江苏科技大学生物与化学工程学院,江苏镇江 212003)

摘要:探索采用便携式近红外光谱仪,利用不同光谱预处理算法及波长优选法建立检测模型检测梨可溶性固形物含量(SSC)的可行性。比较了一阶导数(1^{st})、二阶导数(2^{nd})、多元散射校正(MSC)、标准正态变量变换(SNV)等 9 种预处理方法进行 PLS 建模的效果,确定最佳预处理方法。用相关系数法、无信息变量消除法(UVE)、向后区间偏最小二乘法(biPLS)和向后区间偏最小二乘法结合遗传算法(biPLS + GA)优选波长,用偏最小二乘法(PLS)建立梨 SSC 的定标模型,根据各个模型的校正集和预测集的相关系数(r)和交互验证均方根误差(RMSECV)、预测均方根误差(RMSEP)评价定标模型的精度和稳定性。结果表明:经过 SNV 预处理后的建模效果最好,校正集和预测集的相关系数 r 分别为 0.890 8 和 0.868 9,RMSECV 和 RMSEP 分别为 0.592 5 和 0.630 8;相较于其他 3 种波长优选法,biPLS + GA 方法不仅优选的波长数少,而且所建模型的预测效果更好,校正集和预测集的相关系数分别为 0.887 9 和 0.891 0,RMSECV 和 RMSEP 分别为 0.599 9 和 0.571 3。

关键词:便携式近红外光谱仪;梨;可溶性固形物含量;向后区间偏最小二乘法;遗传算法

中图分类号: O657.33　　**文献标志码**: A　　**文章编号**:1002 - 1302(2014)08 - 0284 - 03

中国果树栽培面积和产量均为世界第一位。2010 年中国果园面积约 1 154.39 万 hm^2,产量约 1.29 亿 t。苹果、柑桔、梨是中国最主要的三大水果,产量占到 58%[1];梨树面积、产量次于苹果、柑桔,居第 3 位,2011 年梨产量接近 1 600 万 t。梨果脆甜多汁,其糖度与成熟度、品质密切相关。成熟度高的果实糖度高、品质高、口感好,成熟度低的糖度低、品质差、口感差。糖度是评价其品质的指标之一,而糖度可以通过可溶性固形物含量(SSC)的测定获得,因此试验和研究一般以 SSC 作为衡量糖度大小的指标。

丰水梨是适合我国南方很多地区栽种的梨品种,个头较大,果形圆,外形也比较美观,果肉乳白色,细脆爽口,浓甜,汁多[2],尽管含糖量高,但肉眼很难看出来,需要使用仪器来检测。本试验使用的是 NIR256 - 2.2T2 型便携式近红外光谱仪,并使用自制光纤来解决在树检测的问题。

现有的便携式近红外光谱仪采用的是连续光源,获得的是整个谱段的近红外光谱,可以通过优选波长剔除不相关的波长[3],从而减少波长数的使用。本试验采用多种波长优选法进行建模比较,从而确定最佳波长优选方法。

1　材料与方法

1.1　仪器设备

试验用到的主要仪器和设备有:NIR256 - 2. 2T2 型近红外光谱仪(美国 Control Development 公司生产),工作波长 1 089 ~ 2 219 nm,分辨率 1 nm,检测器通道数为 256 个;阿贝折射仪(上海精科)2WAJ;石英光纤(自制),如图 1 所示,外面 1 圈 6 个为入射光纤,中间 1 个为出射光纤,芯径为 0.6 mm;白板(海洋光学)WS - 1 - SL。

试验中梨近红外光漫反射示意图如图 2 所示:光源发出的光经入射光纤传送并投射到样品后,经样品表皮及内部组织漫反射后再经出射光纤传送至近红外光谱仪,用计算机采集数据后再进行数据处理和分析。

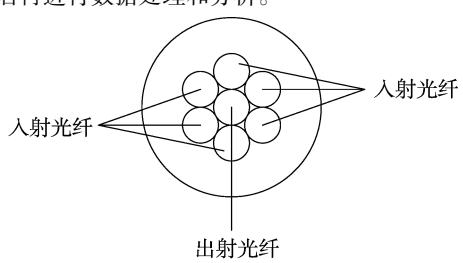

图1　光纤示意图

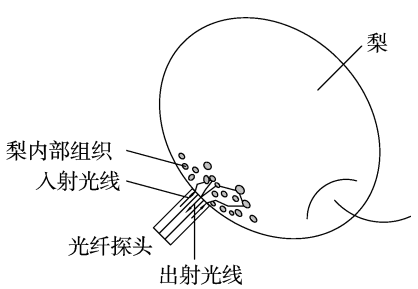

图2　梨漫反射示意图

1.2　供试材料

试验共选用 40 个不同成熟度的丰水梨,来源于江苏省镇江市句容果园。果实近圆形,略扁,皮褐色带红晕,果点大且较多。

1.3　近红外光谱测定

试验以白板为本底,在梨赤道圈上选取 4 个测试区域,采

收稿日期:2013 - 08 - 23

基金项目:江苏省镇江市科技支撑计划(农业)(编号:NY2012034)。

作者简介:秦善知(1988—),男,湖南永州人,硕士研究生,主要从事近红外无损检测技术研究。E - mail: qsz19880716@163.com。

通信作者:陈　斌,教授,博士生导师,主要从事近红外光谱分析和农产品无损检测技术的研究。E - mail:ncp@ujs.edu.cn。

用光纤探头贴近梨进行漫反射测量,经调试仪器的最佳积分时间为 0.01 s,积分次数为 10 次。其中 7 号梨第 1 个测试区域光谱异常,13 号梨前 3 个测试区域果肉呈粉末状,17 号梨第 3 个测试区域和 20 号梨第 2 个测试区域破损,剔除上述 6 个异常样本区域后,对剩余的 154 个样本区域进行主成分分析,其中主成分数选为 3,判别出 150 号样本区域(即 39 号梨第 4 个测试区域)为异常样本区域,因此总计有 7 个异常样本区域,剔除异常样本区域后,总共选取了 153 个测试区域,记为 153 个样本,此时原始光谱如图 3 所示。

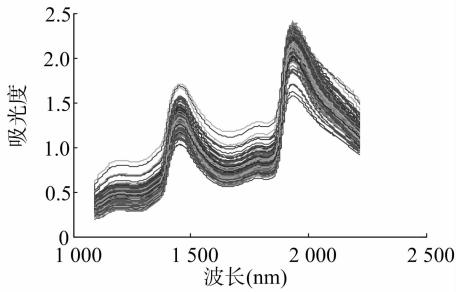

图3　梨样品原始近红外光谱

1.4　SSC 的测定与样品集划分

采用阿贝折射仪测定样本的 SSC,根据 SSC 从小到大排列,按校正集与预测集之比为 3∶1 的原则选择校正集和预测集。其中 115 个样本作为校正集,38 个样本作为预测集。

1.5　波长优选

使用 Matlab 软件和江苏大学近红外光谱组编写的 NIR-SA 软件对数据进行处理及建模。本试验比较了使用一阶导数(1^{st})、二阶导数(2^{nd})、多元散射校正(MSC)、标准正态换(SNV)等 9 种预处理方法处理后的 PLS 建模的效果,确定最佳预处理方法。再分别使用相关系数法、无信息变量消除法(UVE)、向后区间偏最小二乘法结合遗传算法(biPLS + GA)、联合区间偏最小二乘法(siPLS + GA)优选波长,并用偏最小二乘法(PLS)建立梨 SSC 的定标模型,根据各个模型的校正集和预测集的相关系数(r)和交互验证均方根误差(RMSECV)、预测均方根误差(RMSEP)评价定标模型的精度和稳定性。

1.6　方法原理

相关系数法[4]是将校正集光谱阵中的每个波长对应的光谱参数向量与组分浓度阵中的某组分浓度向量进行相关性计算,得到每个波长的相关系数,相关系数值越大,证明该波长的光谱信息量越多。无信息变量消除法是基于偏最小二乘(PLS)的回归系数建立的波长选择算法,用于消除不提供信息的变量,减少模型变量[5],具体算法可参考[6-7]。试验中所使用的无信息变量消除法为 α - UVE。biPLS 方法是一种将数据分隔成给定数目的区间,尽管类似于 iPLS 模型,但 biPLS 模型是按顺序每次剔除 1 个区间[8],根据模型 RMSECV 值大小选出最佳子区间组合。遗传算法(GA)是模拟达尔文的遗传选择和自然淘汰生物进化过程,以目标适应度函数为判据,不断对群体进行选择、交叉、变异遗传操作,重组优化群体内的个体,实现特征变量优选的方法。优选过程是以所选特征波长变量建立 PLSR 模型的交叉验证标准差最小为目标函数进行计算[9]。

2　结果与分析

2.1　光谱最佳预处理方法的确定

9 种预处理方法中,SNV 预处理后的模型效果最好,校正集和预测集的相关系数 r 分别为 0.890 8 和 0.868 9,均方根误差 RMSECV 和 RMSEP 分别为 0.592 5 和 0.630 8。

表 1　不同预处理方法 PLS 模型的结果

模型	主因子数(个)	校正集		预测集	
		r_c	RMSECV	r_p	RMSEP
PLS	9	0.86 86	0.646 2	0.741 5	0.836 8
1^{st} Der + PLS	5	0.932 3	0.471 5	0.778 7	0.785 8
2^{nd} Der + PLS	2	0.731 5	0.889 1	0.339 2	1.184 7
SNV + PLS	8	0.890 8	0.592 5	0.868 9	0.630 8
MSC + PLS	8	0.882 3	0.613 7	0.847 8	0.665 4
1^{st}Der + SNV + PLS	6	0.960 9	0.361 2	0.765 8	0.806 7
2^{nd}Der + SNV + PLS	2	0.720 8	0.903 9	0.282 6	1.221 1
1^{st}Der + MSC + PLS	6	0.959 6	0.367 0	0.753 4	0.821 1
2^{nd}Der + MSC + PLS	2	0.720 4	0.904 4	0.282 5	1.220 3
归一化 + PLS	10	0.918 0	0.517 3	0.764 0	0.829 7

2.2　波长优选法比较

2.2.1　相关系数法波长优选　该方法是采用 Matlab 语言实现的变量选择,当阈值取为 0.25 时,预测效果较好,此时所选波长数为 648 个,模型的 r_p 和 RMSEP 分别为 0.859 1、0.634 5。

2.2.2　无信息变量法波长优选　选用 α - UVE 法进行波长优选,最佳主因子数为 8,其中 α 为 0.99。总共选取了 247 个波长数,模型的 r_p 和 RMSEP 分别为 0.852 9、0.654 3。

2.2.3　biPLS 波长优选　研究采用 Matlab 语言的区间偏最小二乘包 - iToolbox - 2004,该软件在 Matlab 软件上运行,用 biPLS 优选变量的结果如表 2 所示。从表 2 可以看出,建立在 7 个子区间上的模型 RMSECV 最小,选为最佳模型,这 7 个子区间为[1 5 6 10 11 12 13],也就是区间为 1 089 ~ 1 145 nm、1 317 ~ 1 430 nm、1 602 ~ 1 827 nm 的波长数,所选区间如图 4 所示,7 个子区间的总波长数为 397 个。图 5 为 biPLS 法优选 397 个波长的 PLS 模型预测结果,其中预测相关系数 r_p 和预测均方根误差 RMSEP 分别为 0.867 7 和 0.616 1。

2.2.4　biPLS + GA 波长优选　由于单独使用 biPLS 方法优选后的 397 个波长数仍然较多,且相邻波长之间仍然存在高度相关性,因此可以利用遗传偏最小二乘算法(GA - PLS)再从这些联合区间[1 5 6 11 12 13]中优选波长。遗传算法的参数为:初始群体大小为 30,最大选取波长数为 397 个,遗传迭代次数为 100,交叉概率为 0.5,变异概率为 0.01,由选取波长数与 RMSECV 值作图选定最佳波长数。图 6 所示为所有 397 个波长按选取频率重新排列后,RMSECV 值随选取变量数的逐步增加而变化的趋势图。由图 6 可以看出,随着波长数逐渐增加,RMSECV 值开始不断减少,当波长数为 39 个时,RMSECV 首次达到最小值;而随着波长数继续增加,RMSECV 值几乎没有变化,因此选取 39 个波长作为 biPLS + GA 方法的最优波长数。

图 7 为 biPLS + GA 法优选 39 个波长的 PLS 模型预测结果,其 r_p 和 RMSEP 分别为 0.891 0 和 0.571 3。

将在39个波长下的biPLS + GA建模结果与其他3种波

表 2　biPLS 法变量优选结果

区间数(个)	子区间	RMSECV	变量数(个)
20	16	0.865 1	1 131
19	19	0.850 1	1 075
18	9	0.842 8	1 019
17	14	0.838 9	962
16	2	0.835 4	906
15	3	0.833 3	849
14	8	0.829 8	792
13	15	0.827 5	735
12	18	0.818 0	679
11	17	0.791 5	623
10	20	0.755 9	567
9	7	0.719 1	511
8	4	0.690 1	454
7	5	0.685 4	397
6	13	0.688 5	340
5	12	0.696 1	284
4	1	0.751 6	228
3	11	0.760 7	171
2	6	0.800 2	114
1	10	0.993 2	57

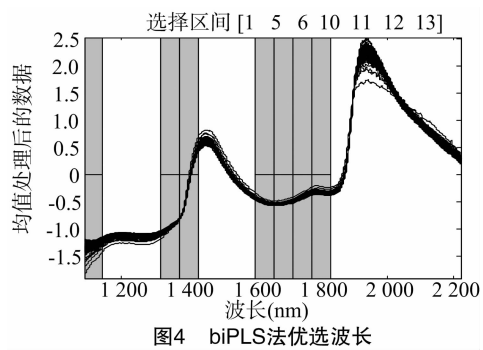

图4　biPLS法优选波长

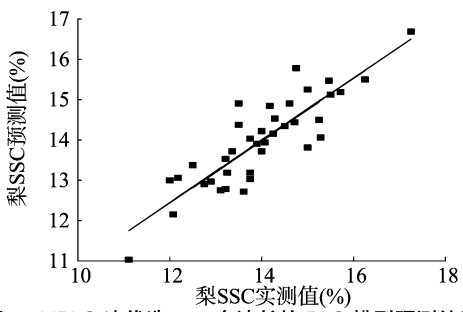

图5　biPLS 法优选 397 个波长的 PLS 模型预测结果

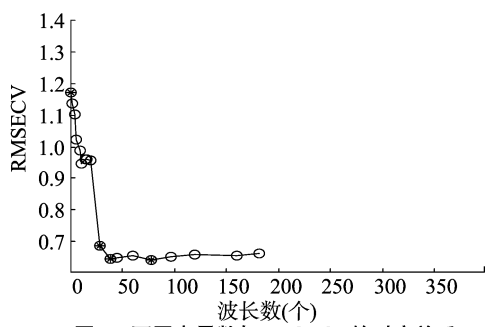

图6　不同变量数与RMSECV的对应关系

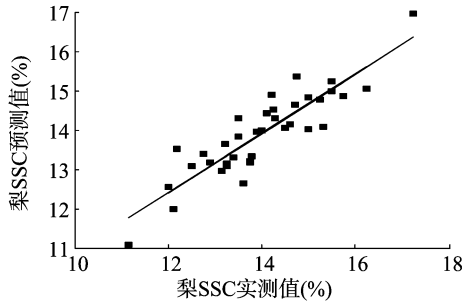

图7　biPLS+GA法优选39波长的PLS模型预测结果

长优选方法的 PLS 建模结果进行比较,结果如表 3 所示。由表 3 可以看出,biPLS + GA 所选取的波长数明显少于其他 3 种方法优选出的波长数,且 r_p 最大,RMSECP 最小,因此可以确定 biPLS + GA 优选波长的效果最好。

表 3　不同波长优选法优选波长后建模预测结果

模型	波长数(个)	校正集		预测集	
		r_c	RMSECV	r_p	RMSEP
相关系数法	648	0.885 3	0.606 4	0.859 1	0.634 5
α - UVE	247	0.865 4	0.653 5	0.852 9	0.654 3
biPLS	397	0.927 3	0.488 2	0.867 7	0.616 1
biPLS + GA	39	0.887 9	0.599 9	0.891 0	0.571 3

3　结论

研究结果表明,采用便携式近红外光谱仪建立检测梨可溶性固形物含量(SSC)的模型是可行的。SNV 预处理方法能明显提高模型的预测效果。与其他 3 种波长优选法相比,biPLS + GA 方法不仅优选的波长数最少,而且模型预测效果最好,得到的校正集和预测集的相关系数分别为 0.887 9 和 0.891 0,RMSECV 和 RMSEP 分别为 0.599 9 和 0.571 3,说明经 biPLS + GA 处理后,能明显提高模型预测梨 SSC 的效果。

参考文献:

[1]中国水果产业基本情况[J]. 世界农业热带信息,2011(12):11 - 12.
[2]代　芬,蔡博昆,洪添胜,等. 漫透射法无损检测荔枝可溶性固形物[J]. 农业工程学报,2012,28(15):287 - 292.
[3]江国兴. 丰水梨优质丰产技术要点[J]. 西南园艺,2000,28(2):15 - 16.
[4]陈　斌,王　豪,林　松,等. 基于相关系数法与遗传算法的啤酒酒精度近红外光谱分析[J]. 农业工程学报,2005,21(7):99 - 102.
[5]吴　迪,吴洪喜,蔡景波,等. 基于无信息变量消除法和连续投影算法的可见 - 近红外光谱技术白虾种分类方法研究[J]. 红外与毫米波学报,2009,28(6):423 - 427.
[6]Entner V,Massart D L,De N,et al. Elimination of uninformative variables for multivariate calibration[J]. Analytical Chemistry,1996,68(21):3851 - 3858.
[7]颜　辉. 植物油的亚油酸、亚麻酸近红外光谱融合和模型优化研究[D]. 镇江:江苏大学,2010:59 - 60.
[8]Balabin R M,Smirnov S V. Variable selection in near - infrared spectroscopy:benchmarking of feature selection methods on biodiesel data[J]. Analytica Chimica Acta,2011,692(1/2):63 - 72.
[9]马世榜,汤修映,徐　杨,等. 可见/近红外光谱结合遗传算法无损检测牛肉 pH 值[J]. 农业工程学报,2012,28(18):263 - 268.