doi: 10.3788/LOP47.123001

基于遗传算法的脐橙可溶性固形物的可见/近红外光谱无损检测

薛　龙[1,2]　黎　静[1]　刘木华[1]　王　晓[1]　罗春生[1]

([1]江西农业大学工学院，江西 南昌 330045
[2]华东交通大学机电工程学院，江西 南昌 330013)

摘要　应用可见/近红外光谱结合遗传偏最小二乘法(GA-PLS)，建立了柑桔类水果可溶性固形物(SSC)的快速无损检测模型。应用光纤光谱仪采集脐橙的可见/近红外光谱，其光谱范围为350～1800 nm。把脐橙的可见/近红外光谱划分成15个光谱区间，通过GA-PLS方法，选出5个光谱区间(包含波段446个，对应波长范围为554～643 nm，1000～1088 nm，1089～1177 nm，1445～1533 nm和1623～1711 nm)建立了预测脐橙可溶性固形物的模型。验证组的最佳预测结果为相关系数和均方根误差分别为0.9132和1.2579。实验结果表明，应用GA-PLS方法选出的可见/近红外特征光谱区域，不仅提高了脐橙可溶性固形物模型的预测精度，而且使模型更加简洁。

关键词　光谱学；遗传算法；偏最小二乘法；可溶性固形物；脐橙

中图分类号　O436　　**OCIS**　300.6340 300.6550　　**文献标识码**　A

Nondestructive Detection of Soluble Solids Content on Navel Orange with Vis/NIR Based on Genetic Algorithm

Xue Long[1,2]　Li Jing[1]　Liu Muhua[1]　Wang Xiao[1]　Luo Chunsheng[1]

([1] *Engineering College, Jiangxi Agricultural University, Nanchang, Jiangxi* 330045, *China*
[2] *School of Mechanical and Electronical Engineering, East China Jiaotong University, Nanchang, Jiangxi* 330013, *China*)

Abstract　A rapid quantification technique is developed and validated for nondestructively quantifying the soluble solids content (SSC) of citrus fruits using Vis/NIR spectroscopy in conjunction with partial least squares regression (PLS) using genetic algorithm (GA). The spectral are recorded in the Vis/NIR region from 350 to 1800 nm using the fiber optic probe method. The navel orange Vis/NIR spectra data are divided into 15 intervals. Consequently, 5 subsets (wave regions 554～643 nm, 1000～1088 nm, 1089～1177 nm, 1445～1533 nm and 1623～1711 nm respectively) and 446 data points are selected quickly by GA-PLS. The best prediction results for the navel orange in predicted set are 0.9132 and 1.2579 for correlation coefficient and root mean square errors of prediction respectively. With the proposed method, a concise easily computed model can be built to select the characteristic region of Vis/NIR spectroscopy.

Key words　spectroscopy; genetic algorithm; partial least squares; soluble solids content; navel orange

1　引　　言

目前水果进入销售市场前一般按颜色、大小和外观形状等进行分级优选处理。随着人们生活水平的不断提高，对水果的品质要求也不再仅仅局限于外部品质，对内部品质如糖酸度等指标也极为看重。脐橙品质

收稿日期：2010-06-03；**收到修改稿日期**：2010-07-10

基金项目：国家自然科学基金(30760101)，新世纪优秀人才支持计划项目(NCET-09-0168)，江西省教育厅科学技术研究(GJJ08513)和江西省科技厅农业科技支撑计划(2009BNB05705)资助课题。

作者简介：薛　龙(1977—)，男，硕士，讲师，主要从事机械设计及光学检测等方面的研究。E-mail：ultimata@163.com

导师简介：刘木华(1969—)，男，博士，教授，主要从事农产品品质光学无损检测等方面的研究。

E-mail：suikelmh@sohu.com(通信联系人)

优良、无籽多汁、色泽鲜艳，是世界各国竞相栽培的柑桔良种，因此对脐橙可溶性固形物（也可称为糖度）的无损检测极具实用价值。随着光谱技术和化学计量方法的发展，光谱技术在农产品品质无损检测中的应用越来越广泛，例如应用近红外光谱检测苹果可溶性固形物（SSC）[1]，茶叶品质[2,3]，脐橙的可溶性固形物[4]和油菜叶片中脯氨酸含量[5]。同时也有应用近红外光谱和高光谱技术检测农产品的农药残留[6~9]。但这些研究，很少把光谱仪采集的所有光谱数据用来建模，大都采用其中一部分光谱建立模型。波长选择的方法主要有间隔偏最小二乘法（iPLS）、逐步回归法、连续投影算法和遗传算法（GA）等。其中遗传算法被广泛应用于对象特征波长的优化选择[10~15]。并且遗传算法可与多种化学计量学多元分析方法融合，从而构造相应的适应度函数，如 PLS，多元线性回归（MLR）和支持向量机（SVM）等，以期获得更高的定量分析或者定性分析的准确度。

遗传算法以其全局最优、易实现等特点，成为目前最常用且最有效的一种波长选择方法。本文拟利用遗传算法与偏最小二乘法相融合的方法，在可见/近红外波长区域进行筛选，找出与脐橙可溶性固形物相关性最好的几个波长区域，以期提高脐橙可溶性固形物定量分析模型的准确度。

2 材料与方法

2.1 样　品

从江西南昌农大市场购买脐橙样品，从中挑选出 294 个没有腐烂和表面缺陷的脐橙。用水清洗脐橙后，在实验室条件下（温度 10 ℃，相对湿度 60%）自然风干，然后采集可见/近红外光谱数据。

2.2 设备组成

图 1 所示为可见/近红外光谱采集系统。用 QualitySpec 型可见/近红外光谱仪（ASD 公司，美国）采集水果的漫反射光谱，测量波长范围为 350～1800 nm，采样间隔为 1 nm，每次扫描 10 次。采集样品光谱前，先用标准白板（特伏龙-聚四氟乙烯）对光谱仪进行校正。从脐橙的赤道部位等距依次采集 6 幅光谱图，每幅图谱拍摄脐橙赤道部位相差约 60°，然后把 6 幅图谱平均作为脐橙最终的光谱数据。采集的脐橙光谱如图 2 所示，其中 R 为反射值。从图 2 可以看出在 350～459 nm 范围内，包含很多噪声，因此本文采用的实际光谱范围在 460～1800 nm。

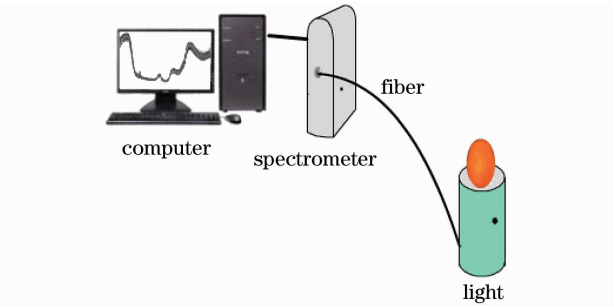

图 1 可见/近红外系统示意图

Fig. 1 Schematic diagram of Vis/NIR system

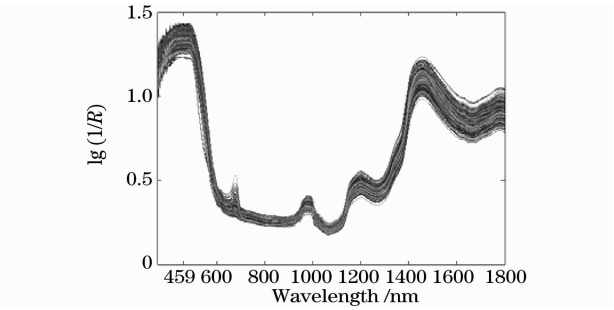

图 2 脐橙的原始光谱图

Fig. 2 Vis/NIR of navel orange

2.3 可溶性固形物测量

把采集光谱后的脐橙剥去果皮，取不同果瓣的果汁。用 PR-101α 型折射式数字糖度计（日本）测量 9 次，然后平均作为最终的测量值。共有数据 294 个，按照可溶性固形物含量的大小进行降序排列。把三个数据的前两个划分为校正组，其余一个划分为验证组，并依次类推。因此校正组共有数据 196 个，包含可溶性固形物测量值的最大值和最小值，验证组包含数据 98 个。表 1为校正组和验证组数据统计结果。

表 1 脐橙可溶性固形物含量统计表（精度 /%）

Table 1 Statistics of soluble solids content in navel orange in samples (Brix /%)

	Number	Maximum	Minimum	Mean	Standard deviation
Calibration set	196	25.8	11.37	19.06	3.13
Prediction set	98	25.22	11.4	19.14	3.03

2.4 数据处理

由于可见/近红外光谱基线漂移和光谱的不重复等，对模型的精度影响较大，因此需要对光谱进行预处

理。通常的光谱预处理方法有多元散射校正(MSC)和一阶导数(d[1])。本文应用这两种光谱预处理方法来建立PLS模型,并根据模型的预测精度选择最佳的光谱预处理方法。图3为应用多元散射校正预处理方法后脐橙的光谱图,其光谱范围为460~1800 nm。

以校正组的相关系数(R),校正组的均方根误差(RMSEC)和验证组的均方根误差(RMSEP)来评价模型效果。建立模型的R值越大,RMSEC和RMSEP值越小,模型的效果最佳。

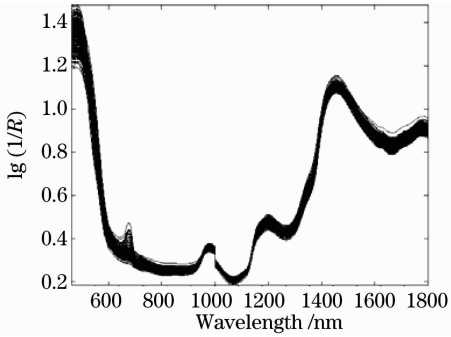

图3 多元散射校正后的光谱图

Fig. 3 Vis/NIR of navel orange using multiplicative scattering correction (MSC)

2.5 遗传算法

遗传算法是模拟生物进化机制随机优化的算法,种群中的每一个个体都表示一个可能的解,通过二进制编码的方式表示(或称之为染色体)。当编码为"1"时,表示选中;反之,当编码为"0"时,表示未选中。通过适应度函数来决定每一个个体的适应性。将其应用于波长选择,其主要步骤有染色体编码、种群初始化、适应度函数、遗传操作、算法停止条件和波长选择。其运算过程如图4所示。在建立PLS模型时首先应用校正组的数据建立模型,并根据交互验证均方根误差(RMSECV)的最小值,来确定PLS因子数。然后用建立的模型对验证组的数据进行预测,得到验证组可溶性固形物的测量值与预测值的相关系数R_p,并以此作为适应度函数的输入变量。本文的适应度函数为

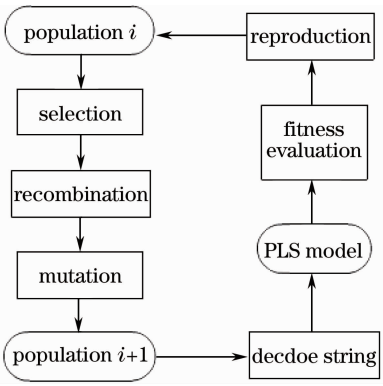

图4 遗传算法循环示意图

Fig. 4 Cycle of genetic algorithms

$$F_i = R_{pi}, \quad i = 1,2,\cdots,n \tag{1}$$

式中R_{pi}为验证组可溶性固形物的测量值与预测值的相关系数,n为种群数量。

3 结果与讨论

3.1 全光谱PLS模型

分别应用原始光谱、MSC校正后光谱和MSC结合d[1](3次多项式,4点平滑)校正后光谱,在全光谱范围内(460~1800 nm)建立了3个PLS预测模型,结果如表2所示。所有模型建模过程中最佳因子数由交互验证法确定,即最小的RMSECV对应的因子数。

表2 不同处理方法处理后的偏最小二乘校正结果

Table 2 Summary of partial least square results after being treated by different methods

Model	Method	Data processing	PLS factor	Spectral range /nm	Calibration set		Prediction set	
					RMSEC	R_c	RMSEP	R_p
1	PLS	None	15	460~1800	1.7722	0.9268	1.3768	0.8924
2	PLS	MSC	13	460~1800	1.2276	0.9195	1.4615	0.8789
3	PLS	MSC, first derivative	7	464~1800	1.0716	0.9393	1.3735	0.8972
4	GA-PLS	MSC, first derivative	7	554~643; 1000~1088; 1089~1177; 1445~1533; 1623~1711	1.1231	0.9330	1.2579	0.9132

从表2中可以看出,应用原始光谱数据和仅采用MSC预处理方法,在全光谱范围内建立模型预测脐橙可溶性固形物的精度小于0.9,两个模型采纳的最佳因子数分别为15和13,并且模型的输入变量为1341

个,这使得模型显得过于复杂。虽然,采用 MSC 结合 d[1] 预处理方法,在全光谱范围内建立模型预测脐橙可溶性固形物的精度稍有提高,其 R_p 为 0.8972,并且因子数也减少到 7 个,但是模型的输入变量为 1337 个(从左侧开始 4 点进行平滑,因此光谱范围 464～1800 nm),不利于水果的在线快速检测,模型预测结果如图 5所示。因此,本文在第 3 个模型的基础上,应用遗传算法来选择最佳的光谱区域,以期减少模型的输入变量,进一步降低模型的复杂程度。

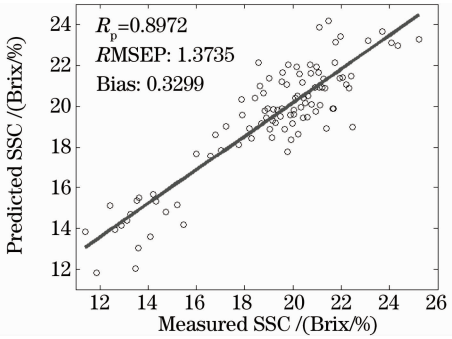

图 5 验证组中脐橙糖度测量值与预测值结果

Fig. 5 Measured SSC with predicted SSC in navel orange

3.2 遗传区间偏最小二乘法选取特征光谱区域

把光谱分成 15 个区域,前 2 个区域均包含 90 个波段,其余区域包含 89 个波段。用遗传区间偏最小二乘法从这 15 个区间中选取特征光谱区域。设定优化参量:区间数 15,初始群体 30,交叉概率 0.8,变异概率 0.1,遗传迭代次数 100。图 6 为每一代中最大的 R_p,随遗传算法进化 100 代后的结果。从图 6 可以看出,第 8,14,19,54,82 代的 R_p 值最大,其值为 0.9132。通过对这 5 代选定的特征光谱区域进行分析发现,其选定的特定光谱区域相同,分别是第 2,7,8,12 和 14 区域,对应的光谱范围是 554～643 nm,1000～1088 nm,1089～1177 nm,1445～1533 nm 和 1623～1711 nm,波段数目为 446 个。图 7 为遗传偏最小二乘法所选取的特征光谱区域。图 8 为选定的 5 个特征光谱区域所建立的 GA-PLS 模型在验证组中脐橙糖度的预测结果图。

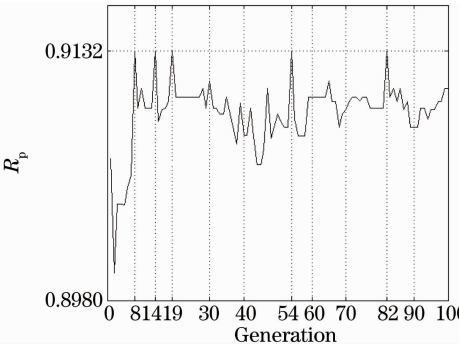

图 6 每代中最佳 R_p 随遗传代进化的情况

Fig. 6 R_p values of partial least square regression models variation with generation

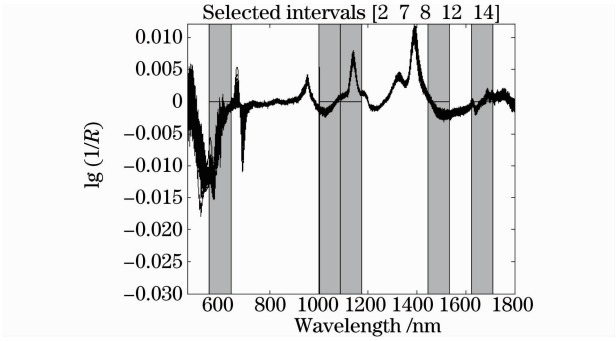

图 7 遗传偏最小二乘法所选取的特征光谱区域,光谱预处理方法为 MSC 结合 d[1]

Fig. 7 Optimal spectra regions selected by GA method after MSC combined with d[1]

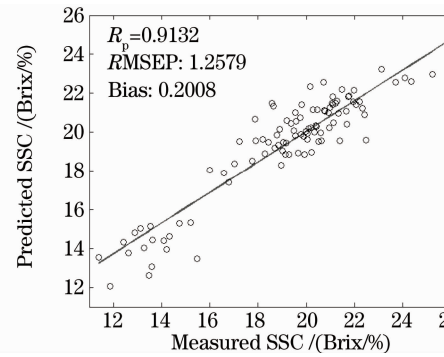

图 8 选定的 5 个特征光谱区域所建立的 GA-PLS 模型在测试组中脐橙糖度的预测结果

Fig. 8 Measured SSC with predicted SSC in navel orange using GA-PLS model with 5 characteristic spectral regions

从表 2 中的模型 4 可以看出,遗传区间偏最小二乘法处理所得的最佳脐橙可见/近红外光谱模型是建立在 5 个光谱区间(共 446 个波段)内,其校正组和预测组模型的预测能力都优于全光谱模型,且该模型得到了很大的简化:其实际使用的波段个数比全光谱模型采用的波段数目 1341 个大大减少;且采纳的最佳主因子数也减少到 7,因此其运算量也减少了许多,模型更简洁、稳健。

4 结 论

用遗传区间偏最小二乘法对脐橙可见/近红外光谱进行特征光谱区域的选取。应用 MSC 结合 d[1] 光谱

预处理方法,建立的脐橙糖度混合特征光谱区间模型(R=0.9132,RMSEC 为 1.1231,RMSEP 为 0.1279)与全谱模型(R=0.8972,RMSEC 为 1.0716,RMSEP 为 1.3735)具有相近的稳健性和适用性,但特征波段仅使用了 446 个数据点,远少于全谱所用波段数据点数。其选定的 5 个特征光谱区间分别是 554～643 nm,1000～1088 nm,1089～1177 nm,1445～1533 nm 和 1623～1711 nm。结果表明,基于遗传算法的波段选择法可用于可见/近红外脐橙可溶性固形物的分析中,并且该方法可以在一定程度上提高模型的精度和建模效率,降低模型的复杂程度。

参 考 文 献

1 Zou Xiaobo, Zhao Jiewen. Methods of characteristic wavelength region and wavelength selection based on genetic algorithm [J]. *Acta Optica Sinica*, 2007, **27**(7): 1316～1321
皱小波, 赵杰文. 用遗传算法快速提取近红外光谱特征区域和特征波长[J]. 光学学报, 2007, **27**(7): 1316～1321

2 Chen Quansheng, Zhao Jiewen, Cai Jianrong *et al*.. Estimation of tea quality level using hyperspectral imaging technology [J]. *Acta Optica Sinica*, 2008, **28**(4): 669～674
陈全胜, 赵杰文, 蔡健荣 等. 利用高光谱图像技术评判茶叶的质量等级[J]. 光学学报, 2008, **28**(4): 669～674

3 Guo Zhiming, Zhao Jiewen, Chen Quansheng *et al*.. Application of selecting wavelength regions to determination of free amino acid content in tea by FT-NIR spectroscopy[J]. *Optics Precision Engineering*, 2009, **17**(8): 1839～1844
郭志明, 赵杰文, 陈全胜 等. 特征谱区筛选在近红外光谱检测茶叶游离氨基酸含量中的应用[J]. 光学精密工程, 2009, **17**(8): 1839～1844

4 Liu Yande, Chen Xingmiao, Ouyang Aiguo. Non-destructive measurement of soluble solid content in gannan navel oranges by visible/near-infrared spectroscopy[J]. *Acta Optica Sinica*, 2008, **28**(3): 478～481
刘燕德, 陈兴苗, 欧阳爱国. 可见/近红外光谱法无损检测赣南脐橙可溶性固形物[J]. 光学学报, 2008, **28**(3): 478～481

5 Sun Guangming, Liu Fei, Zhang Fan *et al*.. Determination of proline in herbicide-stressed oilseed rape leaves based on near infrared spectroscopy[J]. *Acta Optica Sinica*, 2010, **30**(4): 1192～1196
孙光明, 刘 飞, 张 帆 等. 基于近红外光谱技术检测除草剂胁迫下油菜叶片中脯氨酸含量的方法[J]. 光学学报, 2010, **30**(4): 1192～1196

6 Xue Long, Li Jing, Liu Muhua. Detecting pesticide residue on navel orange surface by using hyperspectral imaging[J]. *Acta Optica Sinica*, 2008, **28**(12): 2277～2280
薛 龙, 黎 静, 刘木华. 基于高光谱图像技术的水果表面农药残留检测试验研究[J]. 光学学报, 2008, **28**(12): 2277～2280

7 Li Jing, Xue Long, Liu Muhua *et al*.. Recognition of navel orange contaminated by omethoate based on Vis-NIR spectroscopy[J]. *Transactions of the CSAE*, 2010, **26**(2): 366～369
黎 静, 薛 龙, 刘木华 等. 基于可见-近红外光谱识别氧乐果污染的脐橙[J]. 农业工程学报, 2010, **26**(2): 366～369

8 Li Jing, Xue Long, Liu Muhua. A Vis/NIR spectrum recognition method of pesticide contamination on fruit surface[J]. *Optics & Optoelectronic Technology*, 2010, **8**(2): 27
黎 静, 薛 龙, 刘木华. 水果表面农药污染的可见/近红外光谱识别法[J]. 光学与光电技术, 2010, **8**(2): 27

9 S. Satanwong, S. Kawano. Rapid determination of fungicide contaminated on tomato surface using the DESIR-NIR: a system for ppm-order concentration[J]. *J. Near Infrared Spectrosc*, 2005, **13**(3):169～175

10 J. Li, L. Xue, M. Liu *et al*.. Recognition of different pesticide contamination in Navel oranges based on spectra technology [J]. *Acta Agriculturae Universitatis Jiangxiensis*, 2010, **32**(4): 0723～0728
黎 静, 薛 龙, 刘木华 等. 基于光谱技术识别不同农药污染脐橙的研究[J]. 江西农业大学学报, 2010, **32**(4): 0723～0728

11 R. Leardi, L. Nφrgaard. Sequential application of backward interval PLS and genetic algorithms for the selection of relevant spectral regions[J]. *J. Chemomelr.*, 2004, **18**(11): 486～497

12 Qiang Fei, Ming Li, Bin Wang *et al*.. Analysis of cefalexin with NIR spectrometry coupled to artificial neural networks with modified genetic algorithm for wavelength selection[J]. *Chemometr. Intell. Lab. Syst.*, 2009, **97**: 127～131

13 Jagdish C. Tewari, Vivechana Dixit, Byoung-Kwan Cho *et al*.. Determination of origin and sugars of citrus fruits using genetic algorithm, correspondence analysis and partial least square combined with fiber optic NIR spectroscopy [J]. *Spectrochim. Acta A*, 2008, **71**: 1119～1127

14 R. Leardi. Application of genetic algorithm-PLS for feature selection in spectral data sets[J]. *J. Chemometr. Intell. Lab. Syst.*, 2000, **14**(5): 643～655

15 R. E. Fan, P. H. Chen, C. J. Lin. Working set selection using second order information for training SVM[J]. *J. Mach. Learn. Res.*, 2005, **6**: 1889～1918