DOI:10.3969/j.issn.1000-1298.2010.10.027

梨可溶性固形物含量 NIR 与变量筛选无损检测*

朱伟兴[1]　江　辉[1]　陈全胜[2]　郭建光[1]

(1. 江苏大学电气信息工程学院, 镇江 212013; 2. 江苏大学食品与生物工程学院, 镇江 212013)

【摘要】 为提高利用近红外光谱技术快速检测梨可溶性固形物含量的精度和稳定性,结合区间偏最小二乘和遗传算法(iPLS-GA)来筛选校正模型中的特征光谱区和变量,通过交互验证法确定模型中的主成分因子数和筛选的变量,并以预测均方根误差(RMSEP)和相关系数(R_p)作为模型评价标准。试验结果显示:iPLS-GA 最优模型包含 5 个光谱区、50 个变量和 10 个主成分因子。最佳预测模型相关系数(R_p)和 RMSEP 分别为 0.939 8 和0.325 0,研究结果表明近红外光谱结合 iPLS-GA 算法可以准确、无损检测梨的可溶性固形物含量。

关键词: 梨　可溶性固形物含量　近红外光谱　间隔偏最小二乘　遗传算法

中图分类号: O657.33; S123　　**文献标识码**: A　　**文章编号**: 1000-1298(2010)10-0129-05

Measurement of Soluble Solids Content in Pear by FT-NIR Spectroscopy and Variable Selection

Zhu Weixing[1]　Jiang Hui[1]　Chen Quansheng[2]　Guo Jianguang[1]

(1. *College of Electronic and Information Engineering, Jiangsu University, Zhenjiang* 212013, *China*

2. *School of Food and Biological Engineering, Jiangsu University, Zhenjiang* 212013, *China*)

Abstract

In determination of soluble solids content (SSC) in pear by FT-NIR spectroscopy technique, in order to improve precision and robustness, interval partial least square coupled with genetic algorithm (iPLS-GA) was used to select the efficient spectral regions and variables in calibrating model. Selections of spectral regions and variables were implemented by the cross-validation. The performance of the final model was evaluated according to the root mean square error of prediction (RMSEP) and correlation coefficient (R_p) in prediction sets. The results of final model were achieved as follow: the optimal iPLS-GA model was obtained with 10 PLS factors, when 5 spectral regions and 50 variables were selected, respectively. R_p and RMSEP of optimal model was 0.939 8 and 0.325 0 respectively by a prediction set. This work demonstrated that NIR spectroscopy with iPLS-GA could be applied successfully to determine the SSC in pear as a precise and nondestructive method.

Key words　Pear, Soluble solids content, Near infrared spectroscopy, Interval partial least square, Genetic algorithm

引言

可溶性固形物含量(SSC)也称为糖度,是梨内部品质的重要指标之一。目前,近红外光谱(NIR)作为一种快速无损检测技术,已广泛应用于水果的 SSC 检测,但大多研究都是采用经典偏最小二乘法(PLS)建立水果的 SSC 预测模型[1~3]。近红外区域存在大量含氢基团倍频与合频的吸收,这就造成在整个近红外区域内,存在一定程度的信息冗余。梨的 SSC 并不是几个多糖或单糖成分的简单累加,而

收稿日期: 2009-12-07　修回日期: 2009-12-31

* 国家自然科学基金资助项目(30800666)和江苏省自然科学基金资助项目(BK2009216)

作者简介: 朱伟兴,教授,博士生导师,主要从事智能检测与控制、人工智能与模式识别和农业信息化技术研究,E-mail: wxzhu@ujs.edu.cn

是包含能溶于水的糖、酸、维生素和矿物质等多种成分,它是一个复杂的综合指标。在某些近红外区域,梨的光谱信息与其 SSC 之间缺乏相关性,这就造成一定的噪声信息。在 PLS 模型建立过程中,这些噪声和冗余信息的介入容易扩大估计方差,进而降低模型的精度和稳定性。

为了提高 NIR 检测梨 SSC 的精度,需要筛选出合理的近红外光谱区域和特征变量来建立 PLS 模型。在大多情况下,光谱区域都是通过人为筛选。如果在没有先验知识指导情况下,人为筛选主观性强,并且工作量大。鉴于此,本文尝试利用区间偏最小二乘(interval partial least square, 简称 iPLS)结合遗传算法(genetic algorithm, 简称 GA)从梨的近红外全光谱中筛选出几个有效的特征光谱区域,然后再通过 GA 从这些特征光谱区域中筛选出与梨 SSC 相关的有效变量来建立 PLS 模型,并将其与其他方法所建立的模型进行比较。

1 试验

1.1 试验材料与仪器

试验研究对象为新疆的贡梨,购于 2009 年 10 月。总样本数为 90,随机选取 60 个样本作为校正集,其余 30 个样本作为预测集。试验使用 Antaris Ⅱ傅里叶变换近红外光谱仪(美国 Thermo Scientific 公司)采集光谱,用 WYA - 2S 型数字折射仪(上海精密科学仪器有限公司)测定梨的 SSC。

1.2 试验方法

将梨分别编号后置于 4℃冰柜中贮藏。试验前,将从冰柜中取出的梨置于实验室中 3 h。试验时,保持室内温度在 25℃左右,湿度基本保持不变。每个样品在其赤道部位标记 3 处(间隔约 120°),进行光谱扫描和 SSC 测定。

Antaris Ⅱ傅里叶变换近红外光谱仪采用 InGaAs 检测器,内置参比背景,光源为 25 W 卤素灯仪器标配,光源的波长范围是 400 ~ 2 500 nm。通过漫反射式积分球附件进行光谱采集,如图 1 所示。样品承载台上套装软质黑橡胶垫圈,以防止外界光干扰。扫描波数范围为 4 000 ~ 10 000 cm^{-1}(1 557 个数据点),扫描次数 16 次,分辨率 8 cm^{-1}。采集梨 3 个标记部位的光谱,并将其平均光谱作为该样品的原始光谱。采集光谱后,将每个样品的 3 个标记部位去皮,取出果肉,用数字折射仪测定 SSC。取 3 个标记部位的 SSC 平均值作为整个梨的 SSC 值。表 1 为校正集和预测集样品的 SSC 测量结果。

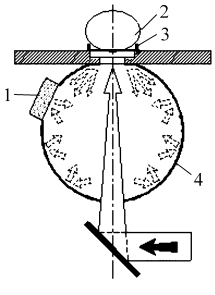

图 1 傅里叶变换近红外漫反射光谱测样示意图

Fig. 1 Schematic diagram of FT - NIR measurement diffuse-reflectance sampling

1. 检测器 2. 梨 3. 密封器 4. 漫反射积分球

表 1 梨 SSC 在校正集和预测集中的参考测量结果

Tab. 1 Reference measurement results of pear's SSC in the calibration and prediction sets

样本集	样本数	SSC/°Brix			
		平均值	最大值	最小值	标准偏差
校正集	60	11.8	14.3	9.7	0.960
预测集	30	11.7	14.0	9.8	0.958

用 iPLS - GA 建立梨的 SSC 模型时,通过交互验证法(cross-validation)来优化模型的相关参数,以 SSC 的实测值和预测值相关系数(R)、交互验证均方根误差(root mean square error of cross-validation, 简称 RMSECV)及预测均方根误差(root mean square error of prediction, 简称 RMSEP)作为评价模型的有效指标。为了显示出 iPLS - GA 模型预测性能的优越性,将其与其他方法建模所得到的结果相比较。所有数据分析在 Matlab R2007a(Mathworks, USA)软件平台上完成。

2 结果与讨论

2.1 iPLS - GA 模型建立

2.1.1 遗传区间偏最小二乘筛选特征光谱区域

梨的原始光谱经过 SNV 预处理后再利用区间偏最小二乘法(iPLS)结合遗传算法(GA)进行特征谱区和变量的筛选。iPLS 是由 Lars Nørgaard[4~5] 于 2000 年提出的一种波长筛选法,其基本原理是将整个光谱区域划分为 n 个等宽的子区间(本研究中将全光谱区划分成 20 个等宽的子区间),然后,在每个子区间上进行偏最小二乘回归,得到 n 个局部回归模型,以交互验证均方根误差(RMSECV)作为各局部模型的精度衡量标准。遗传算法具有全局快速搜索的优点,将遗传算法和偏最小二乘法有机的结合起来,发挥各自的长处,建立更加稳定、简便和预测能力更强的模型。在模型的建立过程中,参考文献[6],通过优化,遗传算法的控制参数设定为:初始群体为 50,交叉概率 $P_c = 0.5$,变异概率 $P_m = 0.01$,遗传迭代次数为 100,以 RMSECV 值作为遗传算法的适应度函数。遗传迭代终止后,通过最低的 RMSECV 确定最终入选的区间为[2 3 9 13 17],如图 2 所示。这样,参与模型建立的光谱变量数为 390 个,剔除了大量与梨的 SSC 指标不相关的光谱

区间，大大降低了光谱变量数。

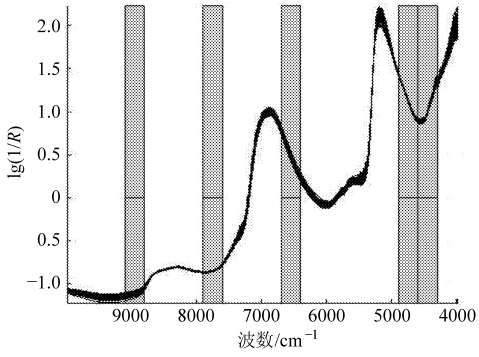

图 2　由 GA 筛选光谱所得的联合区间([2 3 9 13 17])

Fig. 2　Combination of spectral intervals selected by GA with[2 3 9 13 17]

2.1.2　遗传偏最小二乘筛选特征变量

尽管通过 iPLS 和 GA 算法能筛选出有效的特征光谱区域，可以剔除大量无关信息，大大降低参与模型建立的变量数。但是，在入选的某个子区间内，由于相邻的变量之间仍然存在高度相关性[7]，通过该方法选取的 390 个变量中仍然会存在大量的冗余信息。因此，再利用遗传偏最小二乘算法（GA－PLS）从这些优化了的联合区间[2 3 9 13 17]中选取特征变量，建立 PLS 预测模型，以降低最终复杂度。遗传算法的控制参数设定为：初始群体为 50，交叉概率 $P_c=0.5$，变异概率 $P_m=0.01$，遗传迭代次数为 100。遗传迭代终止后，根据选取频次重新排列的变量，按频数高低的顺序逐一加入到 PLS 模型中，再由所选取变量数与 RMSECV 值作图来选定最佳变量数，从而得到选定的变量，进而建立梨 SSC 指标的 PLS 模型。

图 3 为所有变量按选取频次重新排列后，交互验证均方根误差随选取变量数的逐步增加而变化的趋势图。由图可以看出最佳的变量数为 50 时，所对应的交互验证均方根误差最低为 0.272 1。因此这 50 个变量即为遗传偏最小二乘法所筛选的特征变量。图 4 显示了该运算过程中各波数点被选用的频次，所有 390 个变量被入选的频次图。从图中可以看出，被选频次超过 7 的变量为入选最终模型的特征变量。

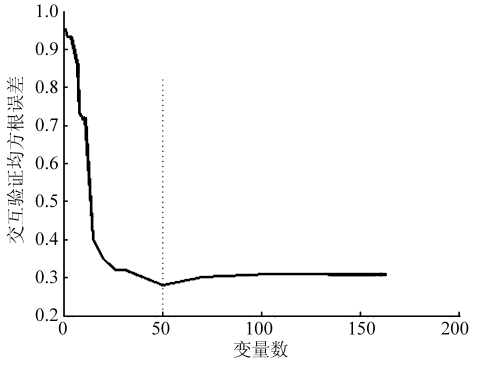

图 3　不同变量数与交互验证均方根误差的对应关系曲线

Fig. 3　Relation curves between RMSECV and variable numbers

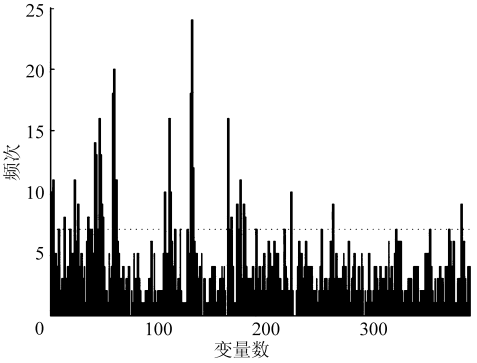

图 4　由 GA 迭代 100 次后被选用变量的频次图

Fig. 4　Cumulative frequency of variable selection after 100 times runs by GA

最终 iPLS－GA 建立模型的主成分因子数为 10 个，该模型校正集和预测集样本预测值与参考测量值之间的散点图如图 5 所示。在校正集中，其交互验证均方根误差（RMSECV）和相关系数（R_c）分别为 0.286 2、0.956 5；在预测集中，预测均方根误差（RMSEP）和相关系数（R_p）分别为 0.325 0、0.939 8。与全光谱模型相比，该模型采用的波数点由 1 557 减少到了 50，在保证精度的前提下大大简化了模型。

2.2　结果比较与分析

在预测模型建立过程中，将 iPLS－GA 模型分别与经典偏最小二乘（PLS）模型、区间偏最小二乘（iPLS）模型、向后区间偏最小二乘（biPLS）、联合区间偏最小二乘（siPLS）和 GA－PLS 模型相比较，比较的结果如表 2 所示。从表中可以看出，与 PLS、iPLS、biPLS、siPLS 和 GA－PLS 模型相比，iPLS－GA 模型的预测精度最高、稳定性最佳。

经典 PLS 模型是在全光谱区域内建立的模型，全光谱区域内含有大量与梨 SSC 无关的光谱变量，这些冗余信息的介入势必在一定程度上降低了模型的预测性能。通常情况下，在 PLS 模型建立过程中，有必要对光谱的范围进行选择，但是在没有先验知识的指导下，光谱范围的选择又具有一定的随机性[8～9]。iPLS 建模时仅选用了一个子区间来建立模型，并没有联合精度较高的几个子区间，可能会损失一部分有用的信息。反映梨 SSC 指标的近红外光谱信息主要来源于含氢基团伸缩振动的倍频或合频吸收，单独一个子区间是不能完全地提供这些倍频或合频吸收信息的[10]。而 biPLS 和 siPLS 模型是联合了几个精度较高的光谱子区间来建立的 PLS 模

型。很明显,这两个模型预测结果均优于 iPLS 模型,并且剔除了大量无用和冗余的信息。

尽管 biPLS 和 siPLS 模型已经得到了较好的预测结果,但与 iPLS－GA 模型结果相比还稍差一点。这是因为在入选的同一个子区间里相邻变量之间可能存在高度相关性,这些冗余变量难免削弱了预测模型的性能。本研究中,iPLS－GA 模型建立分两个过程:①利用 iPLS 算法将全光谱区域划分成等宽的 20 个子区间,然后通过 GA 从这 20 个子区间中筛选出精度较高的几个子区间的联合(筛选出的联合区间为[2 3 9 13 17])。②利用 GA 从联合区间中优选出有效特征变量(筛选出的有效特征变量为 50 个)。因为 iPLS－GA 所建立的模型避免了所入选的同一个子区间里相邻变量之间可能存在相关性的影响,所以该模型预测性能优于 biPLS 和 siPLS 模型。

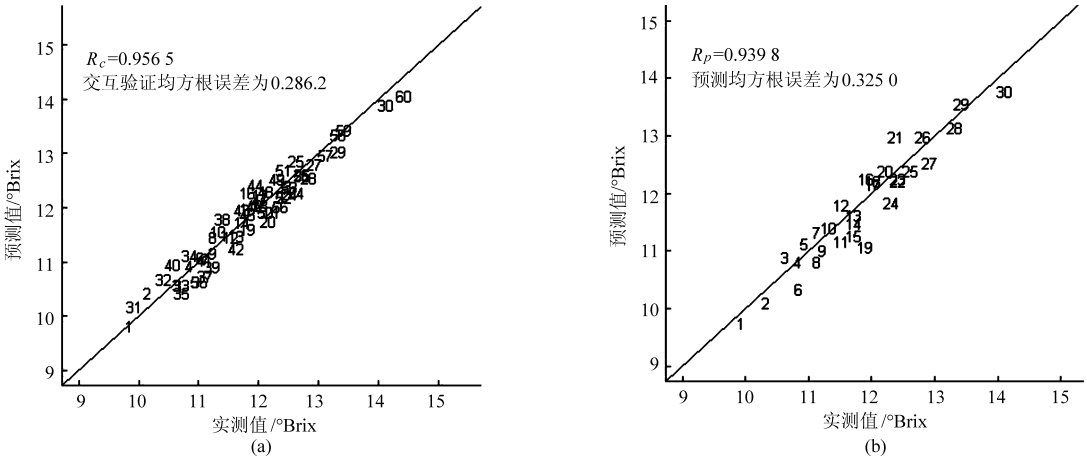

图 5 优化模型的校正集和预测集样本预测值和实测值的相关图

Fig. 5 Reference measurement *vs* NIR prediction in calibration set and prediction set

(a) 校正集 (b) 预测集

表 2 不同 PLS 模型下的结果与分析

Tab. 2 Results and comparison from different PLS models

模型	被选区间	变量数	主成分因子数	校正集		预测集	
				R_c	RMSECV	R_p	RMSEP
PLS	全光谱	1 557	15	0.886 9	0.489 1	0.853 8	0.574 5
iPLS	2	78	6	0.877 4	0.500 4	0.849 6	0.615 6
biPLS	[13 17 9 14 4 2 19 3 7]	701	13	0.925 5	0.343 9	0.874 1	0.553 3
siPLS	[3 7 16]	234	11	0.945 3	0.308 9	0.931 0	0.379 3
GA－PLS	全光谱	167	13	0.937 2	0.330 2	0.898 5	0.456 9
iPLS－GA	[2 3 9 13 17]	50	10	0.956 5	0.286 2	0.939 8	0.325 0

由于梨 SSC 是一个复杂的综合指标,遗传算法又具有快速、随机的特点。如果是在全光谱中来筛选有效特征变量,可选择的特征变量数会增加。但从研究结果来看,从全光谱中选择有效特征变量来建立的 GA－PLS 模型预测性能并不优于 iPLS－GA 模型,并且建模所需的时间更长。

3 结束语

利用近红外光谱技术快速无损检测梨的可溶性固形物含量。在模型的建立过程中,利用区间偏最小二乘结合遗传算法(iPLS－GA)对梨的近红外光谱进行特征光谱区域和特征波长的选取。结果发现,应用该方法所建立的模型来快速、无损检测梨的 SSC 是可行的;与其他模型相比,iPLS－GA 模型显示出明显的优越性,它不仅有效地减少了建模所需的变量数和时间,而且也进一步提高了模型的稳定性和预测精度。

参 考 文 献

1 Park B, Abbott J A, Lee K J, et al. Near-infrared diffuse reflectance for quantitative and qualitative measurement of soluble solids and firmness of delicious and Gala apples[J]. Transactions of the ASAE, 2003, 46(6): 1 721 ~ 1 731.

2　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.
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. (in Chinese)

3　赵杰文,张海东,刘木华. 简化苹果糖度预测模型的近红外光谱预处理方法[J]. 光学学报,2006,26(1):136~140.
Zhao Jiewen, Zhang Haidong, Liu Muhua. Preprocessing methods of near-infrared spectral for simplifying prediction model of sugar content of apples[J]. Acta Optica Sinica, 2006, 26 (1): 136~140. (in Chinese)

4　Norgaard L, Saudland A, Wagner J, et al. Interval partial least squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy[J]. Applied Spectroscopy, 2000, 54(3): 413~419.

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

6　褚小立,袁洪福,王艳斌,等. 遗传算法用于偏最小二乘方法建模中的变量筛选[J]. 分析化学研究简报,2001,29(4):437~442.
Chu Xiaoli, Yuan Hongfu, Wang Yanbin, et al. Variable selection for partial least squares modeling by genetic algorithms [J]. Chinese Journal of Analytical Chemistry, 2001, 29(4): 437~442. (in Chinese)

7　邹小波,朱曾,赵杰文. 基于间隔偏最小二乘法的农产品近红外光谱谱区选择方法[J]. 现代科学仪器,2007(1):86~88.
Zou Xiaobo, Zhu Zeng, Zhao Jiewen. Selection of the efficient wavelength regions in agricultural product NIR spectroscopy based on interval partial least-squares (iPLS)[J]. Modern Scientific Instruments, 2007(1): 86~88. (in Chinese)

8　褚小立,袁洪福,陆婉珍. 近红外分析中光谱预处理及波长选择方法进展与应用[J]. 化学进展,2004,16(4):528~542.
Chu Xiaoli, Yuan Hongfu, Lu Wanzhen. Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique[J]. Progress in Chemistry, 2004, 16(4): 528~542. (in Chinese)

9　Pereira A F C, Pontes M J C, Gambarra F F, et al. NIR spectrometric determination of quality parameters in vegetable oils using iPLS and variable selection[J]. Food Research International, 2008, 41(4): 341~348.

10　Chen Q S, Zhao J W, Liu M H, et al. Determination of total polyphenols content in green tea using FT-NIR spectroscopy and different PLS algorithms[J]. Journal of Pharmaceutical and Biomedical Analysis, 2008, 46(3): 568~573.

11　孙旭东,章海亮,欧阳爱国,等. 南丰蜜桔可溶性固形物近红外特征波段选择[J]. 农业机械学报,2009,40(7):129~132.
Sun Xudong, Zhang Hailiang, Ouyang Aiguo, et al. Selection of NIR characteristic wavelength bands for soluble solids content in Nanfeng mandarin fruit[J]. Transactions of the Chinese Society for Agricultural Machinery, 2009,40(7):129~132. (in Chinese)

(上接第 **128** 页)

8　Chang B S, Randall C S. Use of subambient thermal analysis to optimize protein lyophilization [J]. Cryobiology, 1992, 29(5): 632~656.

9　Hancock B C, Shamblin S L, Zografi G. Molecular mobility of amorphous pharmaceutical solids below their glass transition temperatures[J]. Pharm. Res., 1995, 12(6): 799~806.

10　Duddu S P, Dal Monte P R. Effect of glass transition temperature on the stability of lyophilized formulations containing a chimeric therapeutic monoclonal antibody[J]. Pharm. Res., 1997, 14(5): 591~595.

11　Zuo J G, Hua T C, Liu B L, et al. Thermal analysis of tertiary butyl alcohol/sucrose/water ternary system[J]. CryoLetters, 2005, 26(5): 289~296.

12　Knopp S A, Chongprasert S, Nail S L. The relationship between the TMDSC curve of frozen sucrose solutions and collapse during freeze-drying [J]. J. Therm. Anal., 1998, 54(2): 659~672.

13　Fonseca F, Passot S, Lieben P, et al. Collapse temperature of bacterial suspensions: the effect of cell type and concentration [J]. CryoLetters, 2004, 25(6): 425~434.

14　Meister E, Gieseler H. Freeze-dry microscopy of protein/sugar mixtures: drying behavior, interpretation of collapse temperatures and a comparison to corresponding glass transition data[J]. J. Pharm. Sci., 2009, 98(9): 3 072~3 087.