第27卷,第10期
2007年10月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 27, No. 10, pp2001-2004
October, 2007

用遗传区间偏最小二乘法建立苹果糖度近红外光谱模型

李艳肖, 邹小波*, 董 英

江苏大学农产品加工研究所, 江苏 镇江 212013

摘 要 为了简化苹果糖度预测模型和提高模型的精度, 用遗传区间偏最小二乘法(GA-iPLS)建立苹果近红外光谱预测模型。应用结果表明, 整个光谱划分为40个子区间, GA-iPLS选择其中的第4, 6, 8, 11, 18号共5个子区间联合建立苹果糖度模型。遗传区间偏最小二乘法所建的模型, 其校正时的相关系数 r_c 和交互验证均方根误差 RMSECV 分别为0.962和0.334 6, 预测时的相关系数 r_p 和预测均方根误差 RMSEP 分别为0.932和0.384 2。与全光谱模型相比, 该方法建立的模型不论对校正集还是预测集, 模型的预测能力都提高了许多, 且模型得到了很大的简化: 其实际采用的波数点个数比全光谱模型采用的波数点个数大大减少, 主因子数也比全光谱少, 由此建立的模型更加简洁、数据运算量也更少。

关键词 近红外光谱; 遗传算法; 偏最小二乘法; 糖度; 苹果

中图分类号: TP242162; S123 **文献标识码**: A **文章编号**: 1000-0593(2007)10-2001-04

引 言

随着近红外光谱技术和化学计量方法的发展, 近红外光谱技术应用到农产品品质分析中越来越广泛[1-4], 借助先进的近红外光谱仪, 研究者可以在短时间内很方便地获得大量光谱数据。用近红外光谱来预测苹果的糖酸度已在许多论文中进行了报道[5-15], 目的是获得精度更高、更稳定的预测模型。由于近红外区的谱带复杂、重叠多, 通过苹果近红外光谱的分析可以看出, 光谱的总体走势比较平缓, 波峰和波谷没有剧烈的起伏。作者在利用近红外漫反射光谱技术检测苹果糖度的前期研究中也发现, 对原始光谱进行中心化处理后, 再采用偏最小二乘法(PLS)进行多变量校正所建立的模型, 出现用信噪比(SNR)高的波段比用信噪比低的波段建立PLS校正模型的预测能力明显增强, 即如何选择合适的光谱谱区的问题。为此, 本研究尝试采用采用近几年来发展起来的一种新的建模方法——区间偏最小二乘(iPLS)法[16]进行光谱建模, 并在此基础上进行改进和发展成一种光谱谱区选择和建模的方法——遗传区间偏最小二乘法(GA-iPLS), 希望能够得到性能更佳的模型。

1 遗传区间偏最小二乘的基本原理

本文所用的遗传区间偏最小二乘波长筛选法是对Nørgaard[16]提出的一种波长筛选法的改进和发展, 该法主要用于筛选偏最小二乘建模的波长区域。其算法如下:

(1) 特征波谱区间入选编码

首先将整个苹果近红外光谱等分为 s 个区间, 对这 s 个区间入选的问题, 可用一含有 s 个0/1字符(基因)的字符串(染色体串)来表示每种区间组合。字符串0和1分别代表对应区间未被选中和选中, 例如对8个区间的问题区间组合"00110101"表示第3, 4, 6, 8个区间被选中, 其余则未被选中。

(2) 适应度函数的设计

采用PLS交互验证中因变量的预测值和实际值的相关系数(r)为适应度函数。具体实施方法为, 对每个个体所选的区间进行数据重新组合, 再用PLS交互验证得到相关系数(r)。相关系数(r)的计算公式如下,

$$r = \frac{\sum_{i}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i}^{N}(x_i - \bar{x})^2 \sum_{i}^{N}(y_i - \bar{y})^2}} \quad (1)$$

式中, N 为样品个数; $\bar{x}$ 为交互验证预测值的均值; $\bar{y}$ 为实际测量值的均值。

(3) 初始群体

本研究的初始种群由计算机随机地产生的 m 个个体组成, 而每个个体由 s 个字符组成。

(4) 遗传操作设计

收稿日期: 2006-08-28, **修订日期**: 2006-11-29

基金项目: 教育部博士点基金项目(20040222009), 国家自然科学基金项目(30370813)和江苏省创新人才启动基金项目资助

作者简介: 李艳肖, 女, 1978年生, 江苏大学食品专业研究生 *通讯联系人 e-mail: zou_xiaobo@ujs.edu.cn

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

选择算子采用最常用的选择方法——适应度比例方法，也称转轮法，每个个体的选择概率与其适应度成比例。

交叉算子采用单点交叉方法(如图 1 所示)，参与交叉的个体概率为一个小于 1 的小数(如 0.8)。

0110| 1 交叉 01100
1100| 0 11001

Fig. 1 Crossover

变异算子采用基本变异算子，即在某个个体(字符串)中随机挑选一个或多个基因(字符)进行变异，参与变异的个体概率也为一个小于 1 的小数(如 0.1)，它通常比较小。

(5)运算终止条件

本文以遗传迭代次数达到设定的交互验证均方根误差(RMSECV)为收敛终止条件。交互验证均方根误差 RMSECV 值可按下式计算：

$$\mathrm{RMSECV} = \sqrt{\frac{\sum_{n}(\hat{y}_i - y_i)^2}{n}} \qquad (2)$$

式中，y_i 和 $\hat{y}_i$ 分别为交互验证集中第 i 个样本的糖度实测值和预测值，n 为交互验证集样本数。

(6)区间选取

本文采用的方法为，在遗传迭代后，具有最小 RMSECV 的区间组合中的所有入选区间为特征波谱区间。

2 试验方法及数据

试验选用山东水晶红富士 85 个，购回后从中随机地选取 63 个作为校正集，余下的 22 个作为预测集，将它们分别编号后置于 4 ℃冰柜中贮藏。光谱检测试验在环境温度可控的试验室(本试验环境温度控制为 24 ℃)内进行。试验前，将冰柜中取出的苹果置于试验室中 3 h，以使苹果整体温度达到与环境温度的一致。试验时，由近红外光谱仪(Nexus670 FTIR，美国 Nicolet 公司生产，配有近红外光纤附件和 ZnGas 检测器)在每个苹果的最大横径上进行光谱扫描，扫描波数范围为 4 279～9 843 cm^{-1}，扫描次数为 32 次，波数间隔为 1.924 cm^{-1}(共 2 886 个波数点)，分辨率为 4 cm^{-1}，动镜速度为 0.949 4 cm·s^{-1}，光圈为 50，以 $BaSO_4$ 作为参比材料。扫描时光纤探头与苹果表面之间间隔保持 1～3 mm 的距离。

前期研究表明，同一苹果不同部位的糖度相差可以超过 2 °brix，测量苹果上某一点或少数几点的光谱来预测整个苹果的糖度是有误差的。本研究让苹果转动起来，将每次采集的苹果光谱分散到苹果表面的多个点上。为此，本研究自行设计了一套光纤固定支架和水果载物台，其结构如图 2 所示。该装置为一对锥棍，该对锥棍通过同步轮和同步带与步进电机相连，步进电机控制其转动速度，并且该锥棍为两头大中间小，使水果转动时不会两头窜动。检测光纤探头由光纤支架固定，光纤支架位于两锥棍之间中心的正下方，使检测光纤探头正好通过两锥棍之间的间隙靠近检测的水果最大果径，检测水果置于两锥棍上。水果转动装置和光纤支架固定在底座上。检测近红外光谱时，控制水果转动速度，使得近红外仪一次检测中完成所需的扫描次数时水果正好转动一周。图 3 为所采集的苹果近红外光谱经过去均值后的结果，去均值的目的是去除每次测量光谱整体能量的影响。

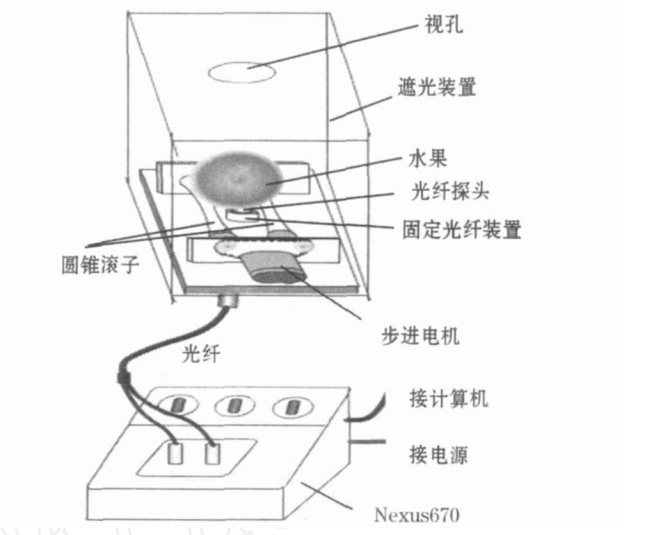

Fig. 2 The schematic of apples NIRS acquisition device

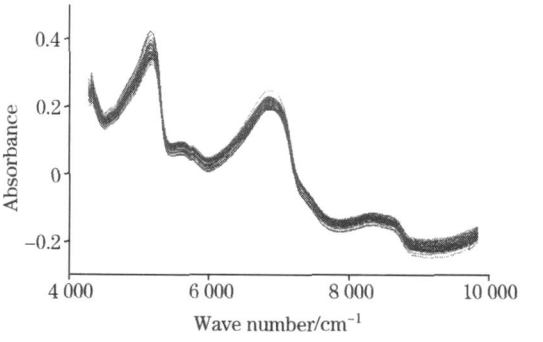

Fig. 3 The calibration "fuji" apples spectra

采集完光谱后将该苹果削皮，取可食用部分榨汁，并用手持式糖度计(WYT0-32 型，泉州韦达计量仪器厂生产)测定其糖度值，表 1 列出了被测苹果糖度实测值的变化范围、平均值、标准偏差及变异系数。

Table 1 Statistic of apple sugar content

	样本数	平均值	最大值	最小值	标准偏差	变异系数 *CV*/ %
校正集	63	12.855	16.6	9.4	1.50	11.67
预测集	22	12.795	15.8	9.0	1.473	11.51

3 试验结果与讨论

将图 3 中的光谱数据(光谱范围：4 279～9 843 cm^{-1})等分为 40 个区间(其中第 1～6 号区间每个区间波数点为 73 个，余下的区间每个区间波数点为 72)，用 Nørgaard[16]的区间偏最小二乘法进行处理，图 4 为处理后的情况。由图 4 可

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

以看出，其中建立在第 2，9，12，13 各个区间上的 PLS 模型的交互验证均方根误差 RMSECV 比全光谱模型的 RMSECV 小，说明并不是光谱数据越多越好。下面就用 GA-iPLS 法从这 40 个区间中选取特征光谱区域。设定优化参数，区间数 40，初始群体 60，最大选取变量数 40，交叉概率 0.8，变异概率 0.1，遗传迭代次数 200，交互验证均方根误差(RMSECV)0.4。图 5 为每代中最小 RMSECV 随遗传算法进化

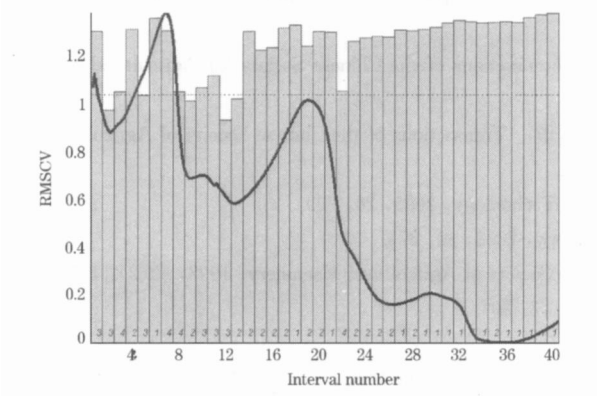

Fig 4　The RMSECV of interval model and global model

Dotted line is RMSECV (3 LVs) for global model/ Italic numbers are optimal LVs in interval model

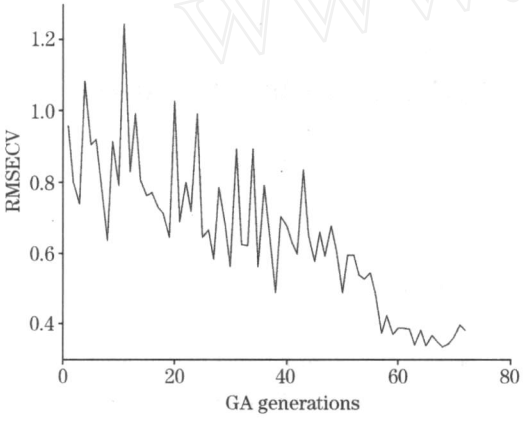

Fig 5　Minimum RMSECV values of PLS regression models with each regenerations

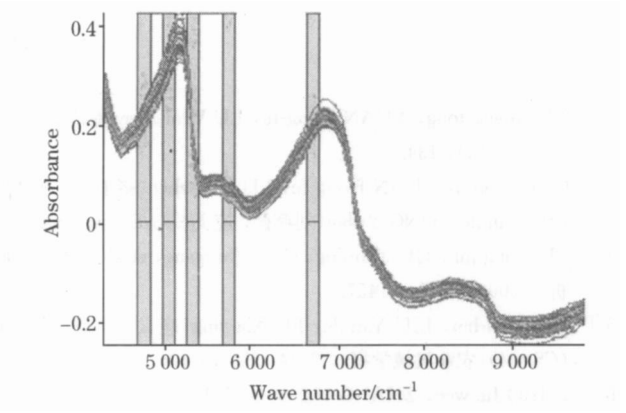

Fig 6　Optimum intervals selection was accomplished by GA-iPLS

73 代的情况。对应 PLS 模型最小 RMSECV 的光谱区间为第 4，6，8，11，18，如图 6 所示。因此 GA-iPLS 所选取的特征波谱区域为第 4，6，8，11，18 五个区间所对应的区域，此时对应的波长变量个数为 362。

为了比较遗传区间偏最小二乘法处理的效果，将所建模型分别与全光谱模型进行比较(全光谱模型的光谱范围为 4 279～9 843 cm^{-1})，结果如表 2 所示。所有模型建模过程中最佳主因子数由交互验证法(Cross-Validation) 确定，即由最小的预测残差平方和(PRESS)确定。

从表 2 可以看出，全光谱的偏最小二乘模型预测苹果的糖度的精度不高，且该模型采纳的最佳因子数为 13，这使得模型显得过于复杂。遗传区间偏最小二乘法处理所得的最佳苹果近红外光谱模型，其不论对校正集还是预测集模型的预测能力都好于全光谱模型，且该模型得到了很大的简化：其实际采用的波数点个数比全光谱模型采用的波数点个数大大减少；采纳的最佳主因子数也减少了许多，运算量也减少了许多。遗传区间偏最小二乘建立在 5 个区间上的模型，其校正时的相关系数 r_c 和交互验证均方根误差 RMSECV 分别为 0.962 和 0.334 6，预测时的相关系数 r_p 和预测均方根误差 RMSEP 分别为 0.932 和 0.384 6。

Table 2　The results after apple spectra were treated by GA-iPLS and whole spectra data model

建模方法	入选光谱/ cm^{-1}	变量个数	PLS 主因子数	交互验证均方根误差 RMSECV	校正集相关系数 r_c	预测均方根误差 RMSEP	预测集相关系数 r_p
全光谱 PLS	4 279.34～9 843.06	2 886	13	0.554 2	0.880 8	0.633 4	0.836 2
GA-iPLS (入选 5 个区间)	4 701.6～4 840.5；4 983.5～5 122.0 5 263.0～5 399.8；5 679.5～5 816.4 5 818.3～5 955.2	362	7	0.334 6	0.962	0.384 6	0.932

4　结　论

用遗传区间偏最小二乘法建立糖度的预测模型。结果发现，遗传区间偏最小二乘筛选法不仅能有效地减少建模所用的变量数，而且能有效地提高糖度模型的测量精度。通过遗传区间偏最小二乘法选取合适的光谱区间进行建模，可以减小建模运算时间，剔除噪声过大的谱区，使最终建立的农产品品质检测近红外光谱模型的预测能力和精度更高。

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

参 考 文 献

[1] XU Guang-tong, YUAN Hong-fu, LU Wan-zhen(徐广通, 袁洪福, 陆婉珍). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2000, 20(2): 134.

[2] CHU Xiao-li, YUAN Hong-fu, LU Wan-zhen(褚小立, 袁洪福, 陆婉珍). Process in Chemistry(化学进展), 2004, 16(4): 528.

[3] LIU Yan-de, YING Yi-bin(刘燕德, 应义斌). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26(8): 1454.

[4] LU Yong-jun, QU Yan-ling, CAO Zhi-qiang, et al(芦永军, 曲艳玲, 曹志强, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26(8): 1457.

[5] YING Yi-bin, LIU Yan-de, FU Xia-ping(应义斌, 刘燕德, 傅霞萍). Transactions of the Chinese Society for Agricultural Machinery (CSAM) (农业机械学报), 2004, 35(6): 124.

[6] ZHAO Jie-wen, ZHANG Hai-dong, LIU Mu-hua(赵杰文, 张海东, 刘木华). Transactions of the Chinese Scoeity of Agricultural Engineering(CSAE) (农业工程学报), 2005, 21(3): 163.

[7] Ann Peirs, Jeroen Tirry, Bert Verlinden, et al. Postharvest Biology and Technology, 2003, 28: 269.

[8] Kleynen O, Leemans V, Destain M F. Postharvest Biology and Technology, 2003, 30: 221.

[9] I Wayan Budiastra, Yoshio Ikeda, Takahisa Nishizu. Journal of Japanese Society of Agricultural Machinery, 1998, 60(2): 117.

[10] Steinmetz V, Roger J M, Molto E, et al. J. Agric. Engin. Res., 1999, 73: 207.

[11] Peiris Ann, Peiris K H S, Dull G G, et al. Hort. Sci., 1999, 34: 114.

[12] Peirs Ann, Lammertyn J, Ooms K, et al. Postharvest Biology and Technology, 2000, 21: 189.

[13] Lammertyn J, Peirs Ann, De Baerdemaeker Josse, et al. Postharvest Biology and Technology, 2000, 18: 121.

[14] Lu Renfu, Guyer Daniel E, Beaudry Randolph M. Journal of Texture Studies, 2000, 31: 615.

[15] Park B, Abbott J A, Lee K J, et al. Transactions of the American Society of Agricultural Engineers, 2003, 46(6): 1721.

[16] NΦrgaard L, Saudland A, Wagner J, et al. Applied Spectroscopy, 2000, 54: 413.

Near Infrared Determination of Sugar Content in Apples Based on GA-iPLS

LI Yan-xiao, ZOU Xiao-bo*, DONG Ying

Agricultural Product Processing Research Institutes of Jiangsu University, Zhenjiang 212013, China

Abstract To improve and simplify the prediction model of sugar content, genetic algorithm interval partial least square (GA-iPLS) methods, the evolution of iPLS described by Lars NΦrgaard, were proposed and used to establish the calibration models of sugar content against apple spectra. The apple spectra data were divided into 40 intervals, among which 5 subsets, i.e. No. 4, 6, 8, 11 and 18, containing 362 data points were selected by GA-iPLS. The optimum GA-iPLS calibration model was obtained with the correlation coefficient (r_c) of 0.962, the root mean square error of cross-validation (RMSECV) of 0.334 6 and the root mean square error of prediction (RMSEP) of 0.384 6. Compared with the whole spectra data model, the data points and the factors in the GA-iPLS were decreased significantly. Consequently, the running time of the PLS model build by GA-iPLS was shorter than that of the whole spectra data model. Furthermore, the GA-iPLS model could not only improve precision, but also simplify the model.

Keywords NIR spectra; Genetic algorithm; Partial least square; Sugar content; Apple

(Received Aug. 28, 2006; accepted Nov. 29, 2006)

* Corresponding author

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net