江西农业大学学报 2010, 32(3): 0633 - 0636 http: //xuebao. jxau. edu. cn
Acta Agriculturae Universitatis Jiangxiensis E - mail: ndxb7775@sina. com

遗传算法结合区间偏最小二乘法在草莓酸度近红外光谱检测的研究

艾施荣[1],刘木华[2]

(1. 江西农业大学 软件学院,江西 南昌 330045; 2. 江西农业大学 工学院,江西 南昌 330045)

摘要:为探索近红外漫反射光谱技术快速无损检测草莓酸度的新方法,共采集了 100颗草莓漫反射近红外光谱数据(波长范围 1 000～1 800 nm)。通过采用标准正交变换(SNV)对原始光谱进行预处理后,将全光谱分为10个子区间,通过样本交互验证法优化每个子区间的最佳主成分数并计算区间对应的交互验证均方根误差(*RMSECV*),得到第 4个子区间(共 80个特征波长)对应的预测均方根误差最小。采用遗传算法对第 4子区间内的波数点进一步优选出 1 483, 1 482, 1 485, 1 460 nm 4个波数点,用这 4个波长的光谱信息建立的草莓近红外酸度模型预测集相关系数为 0. 937 5,预测集均方根误差为 0. 072。结果表明:间隔偏最小二乘法结合遗传算法能筛选出最优波长并能减少建模所用变量,提高检测精度,保证模型的稳健性。

关键词:酸度; 遗传算法; 近红外漫反射光谱; 间隔偏最小二乘法
中图分类号: S123 **文献标志码**: A **文章编号**: 1000 - 2286(2010)03 - 0633 - 04

Nondestructive Measurement of Acidity in Strawberry Using Genetic Algorithm and NIR Spectroscopy

AI Shi-rong[1], LIU Mu-hua[2]

(1. College of Software Technology, JAU, Nanchang 330045, China 2. College of Engineering, JAU, Nanchang 330045, China)

Abstract: In order to find a new method to measure the acidity in strawberry using near infrared spectroscopy, 100 strawberries was selected to collect near infrared spectroscopy. The noise of the raw strawberry was moved by SNV preprocessing method. The strawberry spectra were divided into 10 intervals, and the fourth subset containing 80 data points was selected by interval partial least square (iPLS). To improve and simplify the prediction model of acidity content, genetic algorithms was proposed to select data points. And 1 483 nm, 1 482 nm, 1 485 nm, 1 460 nm wavelengths were obtained finally. Combined with that, the prediction model was built with the prediction coefficient (*Rp*) of 0. 937 5, the root mean square error of prediction (*RMSEP*) of 0. 072. Consequently, near infrared spectroscopy could be used to measure the acidity content of strawberry.

Key words: acidity; genetic algorithm; near infrared spectroscopy; interval partial least square

近红外光谱技术的基本原理[1]是近红外光谱中包含分子中单个化学键基频震动的倍频和合频信息,主要是含氢基团 X - H(H为 C、N、O)的倍频和合频震动的叠加。其具有应用范围广、检测速度快、不破坏检测对象、重现性好等优点[2]。随着近红外光谱技术和化学计量学的发展,近红外光谱技术被

收稿日期: 2010 - 01 - 22 **修回日期**: 2010 - 03 - 24
基金项目:江西省科技厅支撑计划项目(2009BNA08500)
作者简介:艾施荣(1977 -),男,讲师,主要从事计算机科学研究, E - mail: aisrong@163. com。

广泛运用于食品和农产品品质检测。如利用近红外光谱技术对苹果[3]、南果梨[4]、猕猴桃[5]等水果内部品质指标的评价。

在近红外光谱技术研究中,国内外学者主要集中在对近红外光谱波长选择[6-7]、建模方法[8-9]及消噪方法[10]等研究。常用的波长选择方法和建模方法为间隔偏最小二乘法(iPLS)[11-12],该方法将全光谱分为一定数量的光谱子区间,通过计算每个区间的特征值如交互验证均方根误差(*RMSECV*)、预测集均方根误差(*RMSEP*)、预测集相关系数(*R*)的大小,来评价各个子区间波数点与待测指标的相关性,并选择相关性最大的子区间建立偏最小二乘(PLS)回归模型。间隔偏最小二乘法虽然能够满足建模的基本要求,但是其波长选择的方式比较简单,波长区间数凭个人经验划分。为了获取更加简洁、预测能力更好的近红外光谱模型,本文采用遗传算法对间隔偏最小二乘法选择的特征子区间进一步进行波长选择,从而优选出比子区间数量更少、代表性更好的波长,用于建立草莓酸度近红外光谱模型。

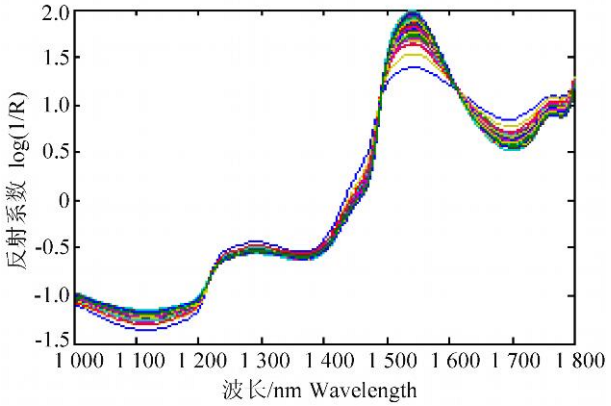

图1 原始光谱经过MSC预处理后草莓的近红外光谱

Fig 1 The NIR spectra of strawberry raw samples after MSC spectral preprocessing

1 材料和方法

1.1 试验材料

试验样品来自江西农业大学实验地种植的草莓。试验共采摘100颗草莓,为了保证模型的通用性,采摘时随机采摘不同成熟度的草莓。草莓采摘后立刻带回实验室,挑选生长正常无畸形、表面干净、无损伤的草莓随机分校正集和预测集,其中校正集80颗,预测集20颗,并逐个进行编号,上述过程保持实验室环境基本不变。

1.2 近红外光谱数据采集

实验所用的近红外光谱仪是美国ASD公司QualitySpecPro光谱仪,扫描波长范围:1 000~1 800 nm,采样间隔:1 nm;扫描次数:10次;探头视场角:45°;光源是与光谱仪配套的12 V/45 W钨卤灯。实验时,保持室内的温度和湿度基本一致,每个样本在不同时间,不同位置分别采集3次,取3次采集的平均值作为该样本的原始光谱。采用标准正交变换(SNV)对原始光谱进行预处理,图1为经过标准正交变换预处理后的光谱图。

1.3 酸度测量

草莓酸度测定参照GB PT 500911—2003《食品卫生检验方法理化部分总则》及GB PT 12456290《食品中总酸的测定方法》。使用上海精密科学仪器有限公司生产的SJ24A型实验室pH计进行测定。这种pH计具有自动温度补偿、自动校准、自动计算、显示电极的百分斜率等功能。表1为校正集和预测集样本酸度测量值。

表1 校正集和预测集样本酸度测量值

Tab.1 The measure acidity results of calibration and prediction samples

测量项 Measure items	样本数/个 Sample number	最大值 Max value	最小值 Min value	平均值 Mean value	标准偏差 Standard deviation
校正集 Calibration set	80	4.18	3.55	3.766 5	0.158 85
预测集 Prediction set	20	4.14	3.53	3.733 0	0.146 00

1.4 遗传算法(GA)优选波长基本原理[13-14]

遗传算法在近红外光谱波段选择中的主要步骤为:

(1) 编码:整个近红外光谱共包含*n*个波数点,对这*n*个波数点的入选问题,可用一含有*n*个0/1字符(基因)的字符串(染色体串)来表示每种波数点组合,字符串0和1分别代表对应波数点未被选中

和选中,例如对 10个波数点组合“0011010100”表示第 3,4,6,8个波数点被选中,其余则未被选中。

(2) 选择初始群体:假如初始群体包含 N 个个体,每一个体的染色体长度为 m,则初始群体的选择方法为随机产生 N 个 m 位的 0-1二进制数作为初始群体。

(3) 适应值函数:采用交互验证法评价模型的预测能力。评价指标为 PLS交互验证预测值与标准值的相关系数 r,以及预测标准偏差 $RMSEP$。如果 $RMSEP$ 值越小,r 值越大则校正模型的预测能力越好。为了使遗传算法对适应值较高的个体有更多的生存机会,对评价指标变换得到适应值函数为: $F = r/(1+RMSEP)$。

(4) 复制:复制的策略是以“轮盘赌”的方式进行正比选择。

(5) 交叉:采用的交叉方式为普通单点交叉方式。

(6) 变异:变异方式是以一定概率产生发生变异的基因数,用随机方法选出发生变异的基因。如果所选的基因的编码为 1,则变为 0;反之编码为 0,则变为 1。本文选取基本变位算子。

重复 (4)—(6)至最大繁殖代数时停止。

2 结果与讨论

2.1 间隔偏最小二乘法选择特征子区间

为了提取同草莓酸度密切相关的近红外光谱特征波长,将全光谱 (共 800个波长)分为 10个特征子区间,根据交互验证法确定最佳因子数并分别计算 10个特征子区间对应的交互验证均方根误差 ($REMSCV$)。计算结果表明,当主成分因子数为 8时,第 4个子区间 (波长范围 1 481～1 560 nm)取得的交互验证均方根最小,该特征子区间光谱模型对应的校正集相关系数为 0.895 7,校正集均方根误差为 0.087;预测集相关系数为 0.863 8,预测集均方根误差为 0.098。间隔偏最小二乘法优选特征子区间结果如图 2所示 (图中灰色条带是优选波长区间)。

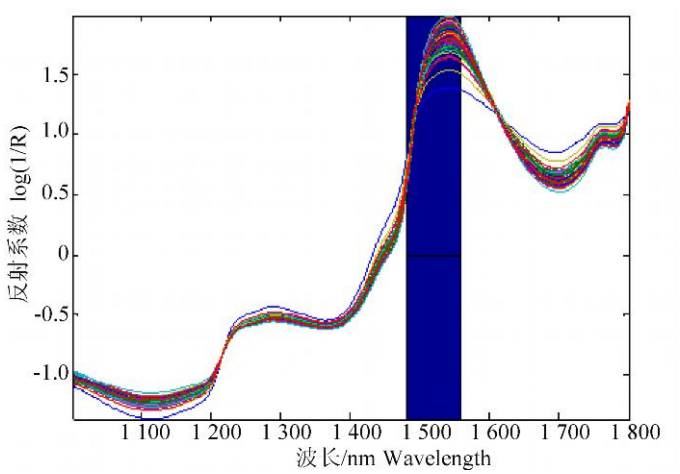

图 2　间隔偏最小二乘法优选特征子区间

Fig 2　Interval partial least square optimizing feature subsets

2.2 遗传算法优选特征波长

间隔偏最小二乘法虽然能够建立草莓酸度近红外光谱模型,但是建模所用的变量较多,模型过于复杂,不能保持模型的稳健性。为得到更加简洁、可靠、预测能力更强的近红外光谱模型,本文对间隔偏最小二乘法选择的特征子区间所包含的波长,引入遗传算法进一步对特征波长进行优选。遗传算法的基因长度为 80(总波长数为 80),遗传代数为 100次。由于遗传算法是一种随机搜索算法,为了保证优化结果的可靠性,共运行了 10次遗传算法,特征波长选择结果如表 2所示。

表 2表示 10次遗传算法优选特征波长结果,其中第 1列表示程序运行次数,第 2列表被遗传算法优选出来的特征波长,第 3列表示入选波长 (即表 2中第 2列)入选的次数。只有当入选次数大于或者等于 5次,才被认为是特征波长。从表 2可以看出,1 483,1 482,1 485,1 460 nm这 4个波数点每次都被选入,而且每次被选择的次数都很高,因此,这 4个波长被认为是草莓酸度的特征波长。根据这 4个波长建立的草莓酸度近红外光谱模型,见式 (1),其预测集相关系数为 0.937 5,预测均方根误差为 0.072,优于间隔偏最小二乘法建立的近红外光谱模型。

$$Y = -6.637X_1 + 2.897X_2 + 2.07X_3 + -3.076X_4 + 7.89 \quad (r = 0.937\ 5) \tag{1}$$

(1)式中:Y 为草莓酸度;X_1,X_2,X_3,X_4 分别是波长为 1 483,1 482,1 485,1 460 nm 4个波长处的近红外光谱数据。

3 结　论

本文利用近红外漫反射光谱分析技术检测草莓酸度,采用间隔偏最小二乘法提取同草莓酸度密切

表 2 10次遗传算法优选特征波长结果

Tab. 2 Feature wavelength results when using ten times genetic algorithms

运行次数 Running times	入选波数点 /nm Selecting optimal wavelength	波数点入选次数 Selected times of wavelength
1	1 483、1 482、1 485、1 460、1 503、1 486	41、26、22、14、9、5
2	1 483、1 482、1 485、1 460、1 503	38、33、23、17、5
3	1 483、1 482、1 485、1 460、1 503、1 484、1 486	38、29、19、15、8、7
4	1 483、1 482、1 485、1 460、1 503、1 484	47、27、15、10、10、7
5	1 483、1 482、1 485、1 460、1 503	39、35、13、12、9
6	1 483、1 482、1 485、1 460、1 484、1 481	41、31、14、13、6、6
7	1 483、1 482、1 485、1 460、1 486、1 484	32、32、23、14、6、5
8	1 483、1 482、1 485、1 460、1 503、1 484	36、29、17、13、10、5
9	1 483、1 482、1 485、1 460、1 503、1 484	36、31、17、12、7、5、5
10	1 483、1 482、1 485、1 460	47、29、14、8

相关的光谱特征区间,在该区间内进一步应用遗传算法进行筛选,对入选波长点的光谱数据与草莓的酸度建立回归模型。

结果表明,将全光谱分为 10个特征子区间,第 4个子区间 (波长范围 1 481～1 560 nm)取得的交互验证均方根最小,在该子区间建立的草莓酸度近红外光谱模型的校正集相关系数为 0. 895 7,校正集均方根误差为 0. 087;预测集相关系数为 0. 863 8,预测集均方根误差为 0. 098。在该区间内使用遗传算法进一步优选出 1 483, 1 482, 1 485, 1 460 nm 4个特征波长,由这 4个波长建立的草莓酸度近红外光谱模型的预测集相关系数为 0. 937 5,预测均方根误差为 0. 072,优于间隔偏最小二乘法建立的模型。其原因是间隔偏最小二乘法中波长区间数凭个人经验划分,建模所用的变量较多,模型过于复杂,不能保持模型的稳健性;遗传算法在组合优化问题上具有很大的搜索优势,在间隔偏最小二乘法优选的子区间内进一步筛选出与草莓酸度最相关的特征波长,减少建模所用变量,提高检测精度,保证模型的稳健性。

参考文献:

[1]严衍禄. 近红外光谱分析基础与应用 [M]. 北京:中国轻工业出版社, 2005.

[2]毕卫红,付兴虎,王魁荣,等. 水果品质近红外检测技术的研究现状与发展 [J]. 激光与光电子学进展, 2006, 43(4): 3 - 7.

[3] Zou Xiao-bo, Zhao Jie-wen. Use of FT - NIR spectrometry in non - invasive measurements of soluble solid contents (SSC) of 'Fuji' apple based on different PLS models[J]. Chemometrics and Intelligent Laboratory Systems, 2007, 87(1): 43 - 51.

[4]周亚凤,马岩松,张平. 南果梨采收前与褐变有关的生理生化指标的变化 [J]. 沈阳农业大学学报, 2001, 32(4): 263 - 265.

[5]陈香维,岳田利,杨公明. 猕猴桃品质光谱无损检测技术研究进展 [J]. 农业工程学报, 2006, 22(8): 240 - 245.

[6]Norgaard R, 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.

[7]Xu Lu, Zhan Wen-Jun. Comparison of different methods for variable selection[J]. Analytica Chimica Acta, 2001, 446: 477 - 483.

[8]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 delicous and Gala apples[J]. Transactions of the ASAE, 2003, 46(6): 1721 - 1731.

[9]Zhao Jie-wen, Chen Quan-sheng, Huang Xing-yi, et al. Qualitative identification of tea categories by near infrared spectroscopy and support vector machine[J]. Journal of Pharmaceutical and Biomedical Analysis, 2006, 41(4): 1198 - 1204.

[10]闵顺耕,谢秀娟,周学秋. 近红外漫反射光谱的小波变换滤波 [J]. 分析化学, 1998, 26(1): 34 - 37.

[11]Zou Xiao-bo, Zhao Jie-wen, Li Yan-xiao. Using genetic algorithm interval partial least squares selection of the optimal near infrared wavelength regions for determination of the soluble solids content of 'Fuji' apple[J]. Journal of Near Infrared Spectroscopy, 2007, 15(3): 153 - 159.

[12]王惠文. 偏最小二乘回归方法及其应用 [M]. 北京:国防工业出版社, 1999.

[13]褚小力,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, 2001, 29(4): 437 - 442.

[14]邹小波,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.