文章编号:1005-0523(2009)04-0074-04

基于遗传算法的近红外光谱建模样品集优化研究

孙旭东,章海亮,欧阳爱国,刘燕德

(华东交通大学 机电工程学院,江西 南昌 330013)

摘要:探讨遗传算法优化方法在南丰蜜橘内部品质可溶性固形物的建模集样品选择中的应用,根据优化结果建立了南丰蜜橘可溶性固形物的快速无损检测数学模型。研究选择的波长范围为350~1 800 nm,应用遗传算法优化建模集样品,优化后建模集样品数量由89个减少到36个,累积变异系数达到99.99%,交互验证均方根误差达到最小。研究结果表明:优化后的建模样品数量大大减少,且保证了南丰蜜橘可溶性固形物的快速无损检测数学模型的稳定性。

关 键 词:近红外光谱;优化方法;遗传算法;偏最小二乘法;可溶性固形物

中图分类号:O657.33;TN247 **文献标识码**:A

近红外光谱技术因其快速、无损、样品无需前处理、准确和高效等优点,被广泛用于果蔬品质的快速无损检测中[1]。但在近红外光谱建模过程中,一般需要大量的实验样品才能建立性能稳定的数学模型。选择有代表性的建模样品是目前近红外光谱快速无损检测的关键问题之一。样品的优化与选择不但可以减少建模的工作量,而且直接影响所建立的校正模型的实用性和准确性[2]。因此,如何从大量的实验样品中,挑选代表性样品建模,是建立高稳健性和预测能力数学模型的关键。

近红外光谱建模代表性样品选择的常规方法有:含量梯度法、双向算法和Kennard-Stone法等。这些常规方法既需要大量的实验样品,又可能受样品温度和颗粒大小等随机的因素影响[3]。遗传算法是一种自适应的全局概率搜索算法,借鉴自然选择和遗传机制,利用选择、交换和突变等算子的操作,随着不断的遗传迭代,使目标函数值较优的变量被保留,较差的变量被剔除,最终达到最优的结果。遗传算法因其良好的寻优特性,已被广泛用于选择近红外光谱特征波段[4~7]。

本研究探讨应用遗传算法进行水果近红外光谱建模样品的优化选择方法的可行性,并通过119个南丰蜜橘的实验样品,运用遗传算法进行建模样品的选择与优化,并用优化后的样品建立了南丰蜜橘的可溶性固形物近红外光谱无损检测稳定数学模型。

1 材料与方法

1.1 实验材料

实验样品来自南丰果园,共选择了119个南丰蜜橘样品,其中89个南丰蜜橘作为建模样品集,供样品选择和优化使用;其余30个未参与建模的南丰蜜橘作为预测样品集,验证代表性样品选择前后所建立的数学模型的稳健性及预测能力。参照国标方法(GB/T 12295-1990),南丰蜜橘可溶性固形物含量采用折射式数字糖度计(Atago Co.Ltd.,Tokyo,Japan)测量,测量结果统计如表1所示。

1.2 光谱采集

实验采用美国ASD公司的QualitySpec & Pro光谱仪,其波长范围为350~1 800 nm,光谱采样间隔为1 nm,扫

收稿日期:2009-05-25

基金项目:国家自然科学基金(60844007);国家科技支撑计划(2008BAD96B04、2006BAD11A12-07);江西省青年基金(2008GQN0029、2007GZN0266)教育部新世纪优秀人才资助计划资助(NCET-06-0575)和江西省青年科学家培养计划(2007-130)

作者简介:孙旭东(1978-),男,吉林辽源人,硕士,助教,主要从事智能无损检测研究。

中国知网 https://www.cnki.net

描时间为100 ms/次,扫描次数为10次,光源为12 V/45 W卤钨灯,图1为南丰蜜橘的近红外光谱检测结构原理图。本次实验采用的近红外漫反射测量方式,且每个样品均在其赤道部位等间隔采集3次光谱后再取其平均值。光谱采集时尽量避免擦伤和伤疤等表面缺陷。

表1　可溶性固形物化学值统计表

测量指标	建模集	预测集
样品数量	89	30
可溶性固形物范围(°Brix)	9.97~18.07	9.87~17.73
可溶性固形物平均值(°Brix)	13.80	13.26
可溶性固形物标准偏差(°Brix)	1.85	1.81

光谱数据由Indico 4.0 (version 4.0, Analytical Spectral Devices, INC., USA)软件进行采集和转换,数据分析和处理软件将采用Matlab 7.0(version 7.0, Mathworks, USA)和Unscrambler (version 8.0, CAMO AS, Trondheim, Norway)来完成数据处理。

1.3　数学模型性能评价

数学模型采用外部验证对其性能进行评价,并由模型预测相关系数、预测均方根误差或交互验证均方根误差进行数学模型的评价分析。模型的相关系数越高,预测均方根误差和交互验证均方根误差越小,且建模和预测均方根误差越接近,模型的预测能力和稳健性越强。

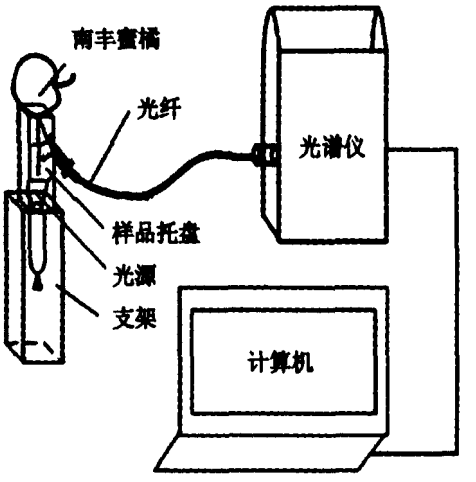

图1　南丰蜜橘近红外光谱检测示意图

2　结果与讨论

2.1　代表性样品选择与优化过程

建模集样品的选择与优化是建立稳定的数学模型的基础,近红外光谱建模过程就是将建模集样品光谱和化学值构建数学关系。因此在建模集中选择代表性样品并对样品进行优化和选择,也是建立稳健性和预测能力强的数学模型关键。本研究应用遗传算法选择代表性样品,未参与建模的样品(染色体位标为"0"),可以以某种概率(变异率),重新标记为参与建模的样品(染色体位变异为"1")。这种变异操作施行在每次建模集样品选择中(即每一代中),从而给予了未参与建模样品多次申诉的机会,这样保证了最后选出最佳的建模集样品组合。

本研究中遗传算法的运行参数选择为:种群大小为30,终止代数为100,交叉概率为0.5,变异概率为0.01。

在建模集样品优化过程中,每个样品出现的频率如图2所示,超过平均出现频率(红线)的样品作为优化后的建模集样品。当累积变异系数达到99.99%(见图3),交互验证均方根误差最小时(见图4),建模集样品数量为36个,优化后建模集样品能有效表征原建模集样品信息。

2.2　主成分分析结果

由建模集样品优化前后的第一和第二主成分得分图上可以看出:优化后的建模集样品在第一象限为10个样品,在第二象限为10个,在第三象限为7个,在第四象限为9个,优化后建模集样品数量共36个,序号分别为:4-9、18-27、43、47-52、59-60、76、78-82、85-89,而且优化后分布较为均匀(图5所示)。

2.3　校正模型建立及精度验证

在350~1 800 nm光谱范围内,本研究结合偏最小二乘法,分别建立优化前后的南丰蜜橘可溶性固形物近红外光谱检测数学模型,并分别对30个未参与建模的南丰蜜橘的可溶性固形物进行预测,建模和预测结果如图6和图7所示。优化后模型的预测性能为相关系数 $R=0.90$;预测均方差 RMSEP = 0.83。

中国知网　https://www.cnki.net

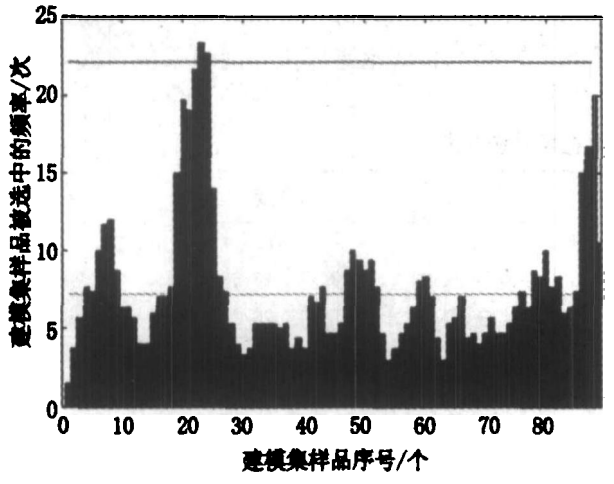

图2 遗传算法优化建模集样品的频率柱状图

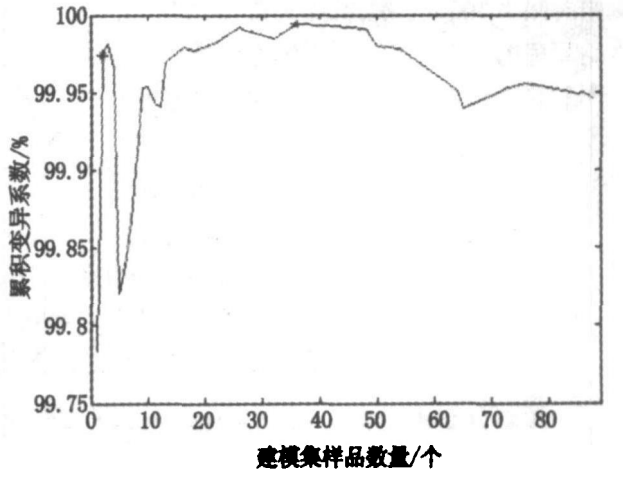

图3 累积变异系数曲线图

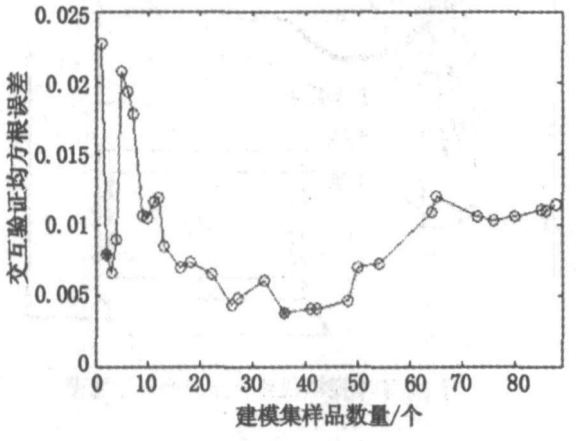

图4 交互验证均方根误差曲线图

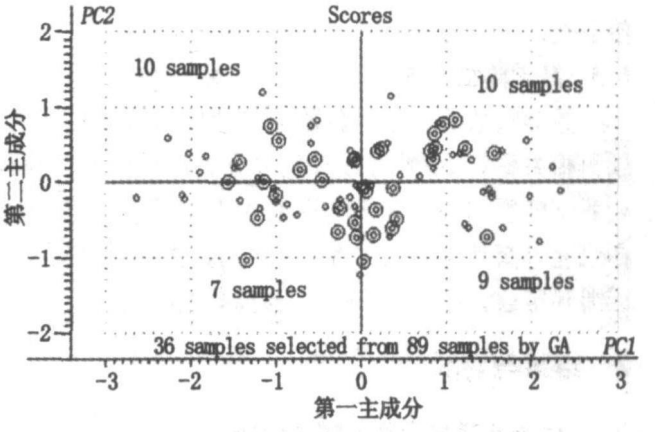

图5 建模集样品优化前后第一和第二主成分得分图

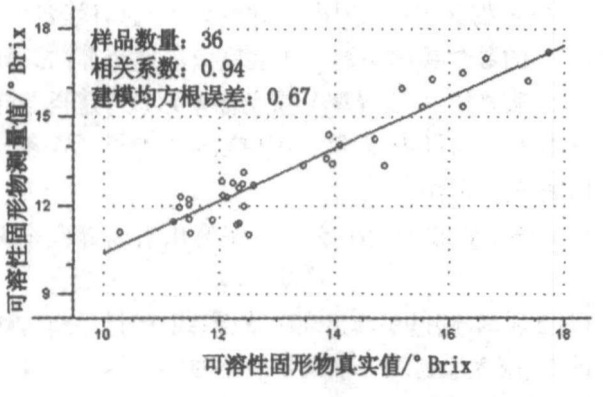

图6 建模集样品优化后的近红外光谱建模模型

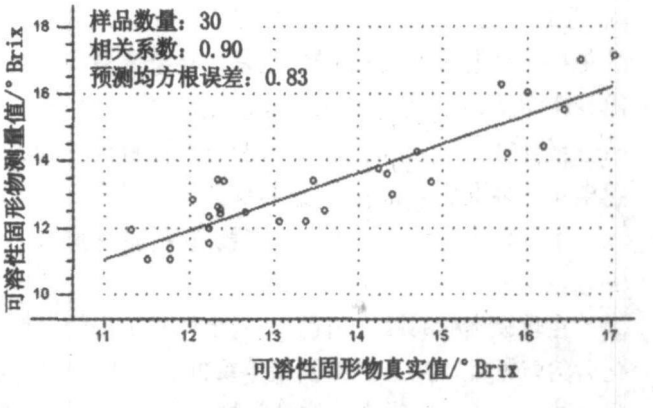

图7 30个未参与建模的南丰蜜橘可溶性固形物近红外光谱预测模型

3 结论

本研究在350～1 800 nm波长范围内，应用遗传算法优化建模集样品，研究结果为建模集样品数量可由原来的89个减少到36个，而优化后数学模型的预测能力未见显著改变，但建模样品数量的大大减少。通过本研究得出遗传算法可以有效优化近红外建模集样品，并可实现南丰蜜橘可溶性固形物含量的快速无损检测。

中国知网 https://www.cnki.net

参考文献:

[1] 李东华,徐亚民,纪淑娟,等.近红外光谱分析技术在果蔬品质无损检测方面的应用[J].农业科技与装备,2008,(1):53-54.

[2] 陆婉珍.现代近红外光谱分析技术[M].北京:中国石化出版社,2007,174-176.

[3] 吴静珠,王一鸣,张小超,等.近红外光谱分析中定标集样品挑选方法研究[J].农业机械学报,2006,37(4):80-82,101.

[4] 刘辉军,吕　进,林　敏,等.基于遗传算法的波长选择方法在绿茶近红外光谱分析模型中的应用[J].分析测试学报,2007,26(5):679-681.

[5] 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):1 316-1 321.

[6] 祝诗平,王一鸣,张小超,等.基于驰援算法的近红外光谱谱区选择方法[J].农业机械学报,2004,35(5):152-156.

[7] 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.

Study on Optimization of NIR Spectroscopy Calibration Set for Soluble Solid Content in Nanfeng Mandarin Fruit Based on GA

SUN Xu-dong, ZHANG Hai-liang, OUYANG Ai-guo, LIU Yan-de

(School of Mechanical and Electrical Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract: The objective of this research is to optimize the sample of calibration set by genetic algorithm (GA) for determining soluble solids contend (SSC) of Nanfeng mandarin fruit, then mathematical models are built for rapid determining SSC of Nanfeng mandarin fruit based on the optimization. The samples of calibration set are optimized by GA in the wavelength range of 350 ~ 1 800 nm. After optimizations, the sample number of calibration set decreases from 89 to 36, accumulative variance value reaches 99.99%, and the least root mean square error of cross validation (RMSECV) is obtained. The research results show the sample number of calibration set decreases vastly, and the stability of model for rapid determining SSC nondestructively is ensured in the Nanfeng mandarin fruit.

Key words: near infrared spectroscopy; optimization; genetic algorithm; partial least square; SSC

(责任编辑:王建华)

中国知网　https://www.cnki.net