

砂梨糖度近红外光谱波段遗传算法优化

潘璐¹, 王加华¹, 李鹏飞¹, 孙谦¹, 张勇², 韩东海^{1*}

1 中国农业大学食品科学与营养工程学院, 北京 100083

2 北京市大兴区林业局, 北京 102600

摘要 遗传算法不受搜索空间限制性假设的约束, 利用简单的编码技术和繁殖机制来解决复杂近红外光谱数据的优化问题。文章采用遗传算法的波段选择法(R-SGA)对砂梨近红外光谱进行了波段优化, 得到丰水、圆黄、黄金三种梨的R-SGA最佳因子数分别为10、12和16, 并分别建立了单一品种GA-PLS模型; 丰水梨和黄金梨的GA-PLS模型精度高于全谱PLS模型, 其模型的RMSEP分别为0.608/0.632和0.524/0.540; 圆黄梨GA-PLS模型精度(RMSEP=0.610)与全谱PLS模型(RMSEP=0.595)相当。经波段优化分析表明, 使用552个数据点建立多品种砂梨混合模型, 具有较高稳健性和预测性(RMSEC=0.627, RMSEP=0.641)。结果表明: 基于遗传算法进行波段优化可以提高砂梨糖度模型精度, 提高建模效率, 同时说明建立多品种砂梨糖度通用模型是可行的。

关键词 近红外光谱; 遗传算法; 波段优化; 糖度; 砂梨; 多品种模型

中图分类号: O657.3 S37 文献标识码: A DOI 10.3964/j.issn.1000-0593(2009)05-1246-05

引言

砂梨是我国四大梨系之一, 包括丰水梨(Hosui pear)、圆黄梨(Wonhwang pear)和黄金梨(Whangkeum bae pear)等。砂梨果肉细、汁多、味甜、爽口, 深受消费者的喜爱。糖度是砂梨主要品质指标之一, 迫切需求一种无损、快速、可在线的检测技术。

近红外光谱分析方法已广泛用于水果糖度的检测, 如苹果^[1,2]、柑橘^[3,4]、桃^[5,6]等。在建立PLS定量模型时, 为提高检测精度和简化预测模型, 常进行数据优化。基于遗传算法的波段选择方法(region selection by genetic algorithm, R-SGA)^[7]是应用较广泛的一种变量选取方法, 给出了提高分析结果准确性的一个重要途径^[8-12]。遗传算法作为一种随机优化算法, 初始参数选择对遗传结果有很大影响。在以往研究中, 遗传代数、种群大小、变异概率和交叉概率是根据当前进化代数下的最佳个体适应度来进行确定, 而对遗传算法最佳因子数的选择探讨较少; 主要采用遗传算法对单一品种物料建模进行波段优化^[13,14], 对相似物料多品种混合建模时的波段优化研究较少。

本文采用R-SGA法分别对丰水梨、圆黄梨和黄金梨近

红外光谱建模波段进行优化, 建立了GA-PLS模型, 探讨了建立砂梨糖度通用模型的可行性。采用GA-PLS模型预测标准偏差(RMSEP)结合最佳个体适应度确定遗传算法最佳因子数; 分别对比GA-PLS模型和全谱PLS模型的稳健性和预测性; 同时为建立砂梨糖度的通用模型, 减少分别建模的成本, 研究了建立混合模型的可行性。

1 实验材料与方法

1.1 实验材料

供试样品(2008年北京奥运会推荐果品评选暨第六届中国梨王擂台赛参赛样品)均产自北京大兴梨有机栽培园区。所有样品均置于2℃冷库储藏, 实验前放于室内(26℃, 相对湿度68%)8h^[15]。供试样品共有251个, 其中丰水75、圆黄99、黄金77个。

1.2 近红外光谱采集

利用ANTARES傅里叶变换近红外光谱仪(Thermo Nicolet USA)积分球附件(窗口大小为9mm)采集砂梨的近红外漫反射光谱, 光谱波长范围为11998.92~3999.64 nm⁻¹, 分辨率8 nm⁻¹, 扫描32次取平均, 使用InGaAs检测器, 光谱保存为log(1/R), R为反射比。每个样品采集赤道均分4点,

收稿日期: 2008-01-16 修订日期: 2008-04-22

基金项目: 国家科技支撑计划项目(2006BAD05A06)和国家自然科学基金项目(30571073)资助

作者简介: 潘璐, 1983年生, 中国农业大学食品科学与营养工程学院硕士研究生 e-mail: shootdoo@126.com

* 通讯联系人 e-mail: aund@au.edu.cn

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

每点作为一个样本, 共有 1004 个。所有样本按照糖度梯度排序, 依 3: 1 分为建模集和预测集, 其中糖度最大和最小样品应为建模集。

1.3 糖度值测定

采集光谱后立即测量糖度值, 采用数字式糖度折射仪 (PAL-1 型, 日本 ATAGO 株式会社) 测量对应点糖度。表 1 是糖度参考值测定结果, 从表 1 可以看出, 3 个品种砂梨糖度分布范围较宽, 预测集样品糖度信息被建模集样品糖度信息所涵盖。

Table 1 Statistic results of soluble solids content (SSC, °Brix) in pear samples				
Parameter		Hosui	Wonhwang	Wangkeumbae
Calibration set	Range	10.6~16.9	9.7~17.6	10.9~16.0
	Mean	13.5	13.7	13.4
	SD*	1.27	1.42	0.90
Validation set	Range	10.8~16.8	10.3~17.3	11.2~15.3
	Mean	13.5	13.7	13.4
	SD*	1.29	1.41	0.84

SD*: Standard deviation

1.4 遗传算法

遗传算法引用生物界物种竞争选择的进化机制, 以适应度函数为依据, 通过对群体中个体施加遗传操作, 如选择、交叉、变异, 来实现群体内个体结构重组的迭代优化^[16-17]。

本实验遗传算法计算程序基于 Visual C++ 6.0 编写。遗传算法的运行参数范围如下; 种群大小: 2~100 终止代数: 2~200 交叉概率: 0.4~0.99 变异概率: 0.0001~0.1 最佳适应值函数为: $F = R / (1 + RMSECV)$, 其中 R 为 PLS 交叉校验预测值与标准值的相关系数, RMSECV 为交叉校验均方差。经过遗传算法优化出的最佳个体染色体为 n 位 0-1 编码组合, 若基因编为 1, 建模时包括此波段; 若为 0 不包括此波段。

光谱进行多元散射校正 (MSC) 或标准归一化 (SNV) 处理后, 将整条光谱 (2 075 数据点) 分为 30 个子区间, 前 29 个子区间分别有 69 个数据点, 最后 1 个子区间有 74 个数据点, 将 30 个光谱子区间进行 0-1 二进制基因编码, 每个波段为一个基因。遗传算法的参数设定: 种群大小 30 最大繁殖代数 50 交叉概率为 0.6 变异概率为 0.01。

2 结果与分析

2.1 光谱数据预处理

由于梨表面粗糙程度不同, 梨细胞大小差异和颜色等因素影响, 导致光谱基线漂移、光散射严重, 需要对样品光谱进行预处理。采用 MSC 或 SNV 消除光谱基线漂移; 卷积平滑 (Savitzky-Golay smoothing) 用于滤除高频噪声, 处理后光谱, 基线漂移基本消除。

2.2 遗传算法优化

2.2.1 遗传算法最佳因子数选择

在遗传过程中, 因子数选择不同导致其最佳适应度函数 F 值的不同。一方面因子数越多, 拟合性能越好, R 值越大, 且 RMSECV 值越小, 相应的 F 值越高; 另一方面, 过多的因子数可能造成过度训练从而导致预测精度下降。因此仅考虑 F 值不能确定最佳因子数^[18], 本文分别计算 2~18 个因子数下遗传算法的最优解, 并采用所选波段分别建立 GA-PLS 模型考察波段选择效果。

由图 1 可见, 随 R-SGA 因子数的增加 R 值增大, RMSECV 值降低, RMSEP 值呈波动变化。丰水梨 R-SGA 因子数选为 10 时, 其 GA-PLS 模型的 RMSEP 为最小值。可见使用因子数为 10 时, 丰水梨 R-SGA 优选波段所建模型预测能力最优, 此因子数即为丰水梨 R-SGA 的最佳因子数。同样经过分析得到圆黄和黄金梨的 R-SGA 最佳因子数分别为 12 和 16。

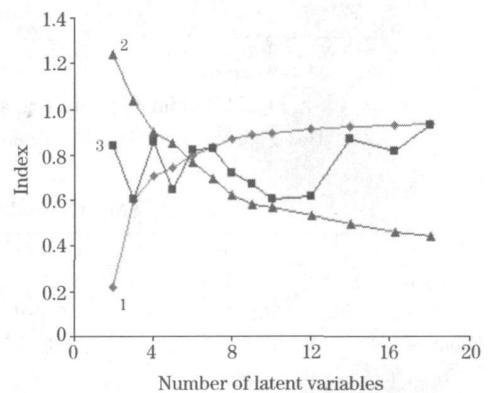


Fig 1 GA-PLS results and RMSEP value of hosui pear versus different latent variables
1: R ; 2: RMSECV; 3: RMSEP

2.2.2 最优波段选择

对于丰水、圆黄和黄金梨分别选择 R-SGA 最佳因子数, 进行遗传算法, 结果如图 2 所示。丰水梨在选择 10 个因子数时遗传结果的最佳个体染色体为: 01100101110 01000000000010110 共有 10 个子区间被选择, 合并相邻染色体中为“1”的子区间, 所保留的谱区为: 11 713.51~11 185.11, 10 915.12~10 625.85, 7 455.45~7 193.18, 6 657.06~5 862.53, 5 592.55~5 330.28 和 4 794.16~4 265.76 m^{-1} 。同理, 圆黄梨所保留的谱区为: 11 713.51~11 185.11, 10 116.74~9 056.08, 8 786.09~8 523.82, 5 326.42~5 064.15 和 4 794.16~4 265.76 m^{-1} ; 黄金梨所保留的谱区为: 11 713.51~11 185.11, 8 786.09~8 523.82, 7 455.45~6 394.79, 5 326.42~5 064.15 和 4 794.16~4 265.76 m^{-1} 。三种砂梨遗传结果所保留的数据点都为 690, 相对于全谱大大减少。

2.3 GA-PLS 模型建立

对不同品种砂梨采用不同预处理方法后 (表 2), 使用 TQ Analyst 7.2 (Thermo Nicolet USA) 软件分别建立全谱 (Full region) 模型和遗传算法波段选择 (R-SGA) 模型。建模过程中最佳因子采用交互验证法所得的预测残差平方和

(PRESS)及交互验证标准偏差 (RM SECV) 确定。

由表 2 可以看出, 丰水梨和黄金梨的 GA-PLS 模型的 RM SEP 分别为 0.608 和 0.524 均低于其 FULL-PLS 模型的 0.632 和 0.540。圆黄梨的 GA-PLS 模型的 RMSEP 为 0.610 预测精度稍差于 FULL-PLS 模型的 0.595, 但 GA-PLS 所用数据点远少于 FULL-PLS 为 690 个。遗传算法能够剔除光谱区域中信息较弱或存在多重线性关系的冗余信息, 波段优化后大大减少参与建模变量, 同时光谱信息得到了充分的利用, 从而提高模型的预测精度。

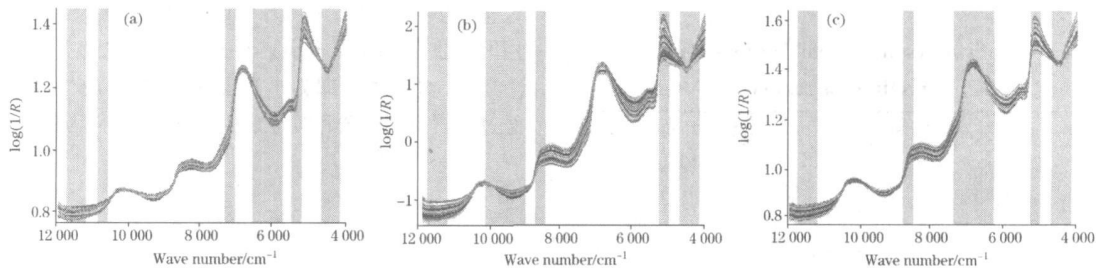


Fig. 2 Optimal spectra regions selected by R-SGA method after pretreatment

(a): Hosui pear(after MSC); (b): Wonhwang pear(after SNV); (c): Whangkeumbae pear(after MSC)

Table 2 Statistic results of models using FULL-PLS and GA-PLS

Variety	Pretreatment mean	Number of data	Latent variables	R	RM SEC	RM SEP	Notation
Hosui	MSC+ S. G.	2 075	9	0.908	0.532	0.632	FULL-PLS
	MSC+ S. G.	690	8	0.928	0.473	0.608	GA-PLS
Wonhwang	SNV	2 075	10	0.952	0.433	0.595	FULL-PLS
	SNV	690	8	0.921	0.553	0.610	GA-PLS
Whangkeumbae	MSC+ S. G.	2 075	9	0.897	0.395	0.540	FULL-PLS
	MSC+ S. G.	690	9	0.914	0.363	0.524	GA-PLS

S. G.: Savitzky-Golay smoothing

Table 3 Statistic results of mixed models using full and selected regions

Pretreatment mean	Region/ m ⁻¹	R	RM SEC	RM SEP	Notation
MSC+ S. G.	11 998.92~ 3 999.64	0.868	0.617	0.645	Full region
MSC+ S. G.	11 713.51~ 11 185.11, 8 786.09~ 8 523.82, 7 455.45~ 7 193.18, 6 657.06~ 6 394.79, 5 326.42~ 5 064.15, 4 794.16~ 4 265.76	0.863	0.627	0.641	Selected region

S. G.: Savitzky-Golay smoothing

优选波段是基于数学方法随机选择所得, 部分波段被重复选中, 说明这些区域数据信息与被检对象含量具有较高的相关性。这些波段包含了蔗糖和葡萄糖等可溶性固形物官能团的合频和倍频吸收区, 如 6 657 和 6 394 m⁻¹附近是与葡萄糖相关的 O—H 和 C—H 键一倍频吸收, 4 700, 4 400 和 4 300 m⁻¹附近是与葡萄糖相关的 C—H 键合频吸收; 4 794 m⁻¹附近是与蔗糖相关的 O—H 键合频吸收。说明遗传算法优化结果和化学结构分析具有一致性。

图 3 是使用特征波段砂梨混合模型建模结果, 其预测精度略高于全谱混合模型 (RM SEP= 0.645)。与表 2 比较而言, 混合模型的拟合性能和预测性能均劣于单一品种所建模型, 原因是砂梨品种间存在差异, 如丰水和圆黄梨分别原产于日本和韩国, 皮厚且颜色分别为浅黄褐色和暗红色, 表皮粗糙度不一样, 细胞大小、石细胞数存在差异, 果肉透明度、致

2.4 多品种混合模型建立

丰水、圆黄、黄金均属于砂梨系, 本文尝试建立了 3 品种砂梨的混合 PLS 模型, 研究建立混合通用模型的可行性。3 个品种砂梨通过遗传算法优选所得波段中有重复出现的区域, 分别是: 11 713.51~ 11 185.11, 8 786.09~ 8 523.82, 7 455.45~ 7 193.18, 6 657.06~ 6 394.79, 5 326.42~ 5 064.15 和 4 794.16~ 4 265.76 m⁻¹。选择这 6 段区域作为特征波段建立砂梨混合模型, 同时也建立全谱混合模型, 如表 3 所示。

密性不同; 而黄金梨为韩国杂交品种, 颜色为黄绿色, 皮薄而光滑, 石细胞极少。经过遗传优化后, 混合模型具有较高的稳健性和预测精度 ($R = 0.863$, $RM SEC = 0.627$, $RM SEP = 0.641$), 且具有很广的预测范围。

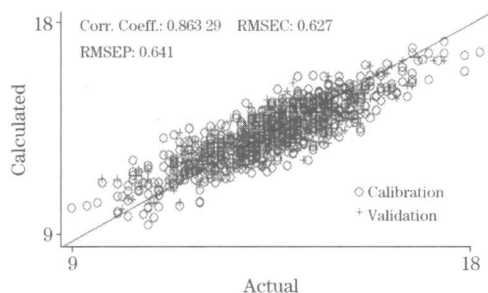


Fig. 3 Result of mixed model using selected regions

因此,建立多品种砂梨糖度的通用模型是可行的,这使得不分砂梨品种直接检测糖度及降低分类建模成本成为可能。

3 结 论

本文采用近红外光谱无损检测砂梨糖度,采用 R-SGA 进行波段优化,得到丰水、圆黄、黄金 3 个品种砂梨的 R-SGA 最佳因子数分别为 10、12 和 16,保留波段数据点均从 2 075 减少到 690。丰水梨和黄金梨的 GA-PLS 模型精度高于 FULL-PLS 模型,其 GA-PLS 模型的 RMSEP 分别为 0.608 和

0.524。圆黄梨 GA-PLS 模型精度 (RMSEP=0.610) 与 FULL-PLS 模型 (RMSEP=0.595) 相当。3 个品种砂梨特征波段混合模型 ($R=0.863$, RMSEC=0.627, RMSEP=0.641) 与其全谱混合模型 ($R=0.868$, RMSEC=0.617, RMSEP=0.645) 具有相近的稳健性和适用性,但特征波段仅使用了 552 个数据点,远少于全谱所用波长点数。结果表明,基于遗传算法的波段选择法可用于近红外砂梨糖度分析中,优选波段方法可以一定程度上提高模型精度,提高建模效率,同时建立多品种砂梨糖度通用模型是可行的。

参 考 文 献

- [1] Liu Yande, Ying Yabin, Fu Xiaping, et al. Journal of Food Engineering, 2007, 80: 986
- [2] Zou Xiaobq, Zhao Jiewen, Huang Xingyi, et al. Chemometrics and Intelligent Laboratory Systems, 2007, 87: 43
- [3] LU Huishan, FU Xiaping, XIE Lijuan, et al. (陆辉山, 傅霞萍, 谢丽娟, 等). Spectroscopy and Spectral Analysis (光谱学与光谱分析), 2007, 27(9): 1727
- [4] Cayuela J A. Postharvest Biology and Technology, 2008, 47: 75
- [5] MA Guang, FU Xiaping, ZHOU Ying, et al. (马广, 傅霞萍, 周莹, 等). Spectroscopy and Spectral Analysis (光谱学与光谱分析), 2007, 27(5): 907
- [6] Corrado Di Natale, Manuela Zude-Sasse, Antonella Magagnoli, et al. Analytica Chimica Acta, 2002, 459: 107
- [7] ZHU Shiping, WANG Yiming, ZHANG Xiaohao, et al. (祝诗平, 王一鸣, 张小超, 等). Transactions of the Chinese Society for Agricultural Machinery (农业机械学报), 2004, 35(5): 153
- [8] WANG Jiahua, SUN Xu-dong, PAN Lu, et al. (王加华, 孙旭东, 潘路, 等). Spectroscopy and Spectral Analysis (光谱学与光谱分析), 2008, 28(9): 2098
- [9] Durand A, Devos O, Rukeyser C, et al. Analytica Chimica Acta, 2007, 595: 72
- [10] ZHOU Xiaobq, ZHAO Jiewen (邹小波, 赵杰文). Acta Optica Sinica (光学学报), 2007, 27(7): 1316
- [11] LI Yanxia, ZOU Xiaobq, DONG Ying (李艳肖, 邹小波, 董英). Spectroscopy and Spectral Analysis (光谱学与光谱分析), 2007, 27(10): 2001
- [12] Christoffer Abrahamsson, Jonas Johansson, Anders Sparen, et al. Chemometrics and Intelligent Laboratory Systems, 2003, 69: 3
- [13] CHEN Bin, WANG Hao, LIN Song, et al. (陈斌, 王豪, 林松, 等). Transactions of the China Society of Agricultural Engineering (农业工程学报), 2005, 21(7): 99
- [14] CHENG Biao, CHEN Dezhao, WU Xiaohua (成飙, 陈德钊, 吴晓华). Chinese Journal of Analytical Chemistry (分析化学), 2006, 34: 123
- [15] Ann Peirs, Ni O'Sheerlinck, Bart M Nielen. Postharvest Biology and Technology, 2003, 30: 233
- [16] CHENG Biao, WU Xiaohua, CHEN Dezhao (成飙, 吴晓华, 陈德钊). Spectroscopy and Spectral Analysis (光谱学与光谱分析), 2006, 26(10): 1923
- [17] GU Xiaoyu, XU Kexin, WANG Yan (谷筱玉, 徐可欣, 汪). Spectroscopy and Spectral Analysis (光谱学与光谱分析), 2006, 26(9): 1618
- [18] Ying Yabin, Liu Yande. Journal of Food Engineering, 2008, 84: 206

Region Optimization of SSC Model for *Pyrus Pyrifolia* by Genetic Algorithm

PAN Lu¹, WANG Jiahua¹, LI Pengfei¹, SUN Qian¹, ZHANG Yong², HAN Donghai^{1*}

1. College of Food Science and Nutritional Engineering, China Agricultural University, Beijing 100083, China

2. Daxing Forestry Administration, Beijing 102600, China

Abstract Genetic algorithm is widely used in NIRS data optimization, which is not limited by searching space. The region selection by genetic algorithms (R-SGA) was applied in building calibration model of soluble solid content (SSC) of *Pyrus pyrifolia*, and the number of variables used to build calibration was further reduced from 2 075 to 690 in all of 3 models. Studies were performed to build GA-PLS

models by different R-SGA latent variables and the optimal R-SGA latent variables of Hosui W onhwang and W hangkeum bae pear were 10, 12 and 16, respectively. The R-SGA procedure was found to perform well (RMSEP = 0.608 and 0.524 for Hosui and W hangkeum bae pear respectively), leading to calibration models that significantly outperform those based on full-spectrum analyses (RMSEP = 0.632, 0.540). The prediction precision of GA-PLS models was similar to FULL-PLS for W onhwang pear, with RMSEP of 0.610/0.595. In addition, the selected regions from R-SGA methods were used to build mixed model of 3 *Pyrus pyrifolia* varieties. The results indicated that the prediction precision of GA-PLS model was close to that of the full spectrum model, with RMSEP of 0.641 and 0.645, respectively. This work proved that the R-SGA could find optimal values for several disparate variables associated with the calibration model and that the PLS procedure could be integrated into the objective function driving the optimization, and it was feasible to build a universal model of different *Pyrus pyrifolia* varieties.

Keywords FT-NIR spectroscopy; Genetic algorithms; Region optimization; Soluble solid content (SSC); *Pyrus pyrifolia*; Mixed model

(Received Jan. 16, 2008; accepted Apr. 22, 2008)

* Corresponding author

第 4 届全国实验室管理科学研讨会征文通知

由中国分析测试协会主办、《分析试验室》编辑部承办的“第 4 届全国实验室管理科学研讨会”定于 2009 年 7 月在甘肃省兰州市召开。

会议内容包括实验室管理科学、实验室认证认可、实验室比对、实验室信息管理系统 (LMS)、实验室仪器设备运行及考核、实验室技术人员培训等有关方面的学术研讨。会议将邀请有关专家做大会报告, 欢迎全国各行业的实验室管理人员及实验室科研工作人员参加会议。

会议征文请在 2009 年 5 月 31 日前通过邮局邮寄、电子邮件发送到会议筹备组 (三日内收到筹备组的回复邮件方为发送成功), 并注明联系人、详细通信地址、联系电话、传真号码及 e-mail 地址。征文经专家审稿后部分录用征文可在《分析试验室》期刊上正式发表。

欲参加会议的各界人士, 也请于 2009 年 5 月 31 日前与会议筹备组联系, 以便继续为您寄发下一轮通知。

筹备组地址: 100088 北京新外大街 2 号《分析试验室》编辑部实验室管理会议筹备组

电话: 010-82241919, 82013328 传真: 010-82013328

e-mail: ana-inf@263.net

联系人: 孙臣良 田春霞

中国分析测试协会

2009 年 1 月 4 日