研究报告(214~ 217)

应用遗传算法和 PLS 的近红外光谱预测玉米中淀粉含量的研究

沈林峰[1], 沈掌泉[2]
(1. 杭州师范大学钱江学院理学系, 浙江 杭州 310012;
2. 浙江大学 农业遥感与信息技术应用研究所, 浙江 杭州 310029)

摘 要: 以普通玉米籽粒为试验材料, 在应用遗传算法结合偏最小二乘回归法对近红外光谱数据进行特征波长选择的基础上, 应用偏最小二乘回归法建立了特征波长测定玉米籽粒中淀粉含量的校正模型. 试验结果表明, 基于 11 个特征波长所建立的校正模型, 其校正误差(RMSEC)、交叉检验误差(RMSECV)和预测误差(RMSEP)分别为 0.30%、0.35%和 0.27%, 校正数据集和独立的检验数据集的预测值与实际测定值之间的相关系数分别达到 0.9279 和 0.9390, 与全光谱数据所建立的预测模型相比, 在预测精度上均有所改善, 表明应用遗传算法和 PLS 进行光谱特征选择, 能获得更简单和更好的模型, 为玉米籽粒中淀粉含量的近红外测定和红外光谱数据的处理提供了新的方法与途径.

关键词: 近红外光谱; 偏最小二乘法; 玉米; 淀粉含量; 遗传算法; 特征选择
中图分类号: TS210.7 文献标识码: A 文章编号: 1006-3757(2008)04-0214-04

近红外光是指介于可见和中红外之间的电磁波, 一般有机物在近红外区的吸收主要是含氢基团(OH、CH、NH、SH、PH 等)的倍频和合频吸收. 由于几乎所有的有机物的一些主要结构和组成都可以在其近红外光谱中找到信号, 因此近红外光谱法被誉为分析的巨人[1].

早在 20 世纪 60 年代, 就有利用近红外光谱漫反射技术来测定农产品中的水分、蛋白质、脂肪等含量的尝试. 近年来, 随着近红外光谱技术和化学计量方法的发展, 近红外光谱技术在农产品品质分析中的应用越来越广泛[1]. 尽管近红外光谱分析技术具有测试简单、测试速度快、效率高、成本低、非破坏性等优点, 但由于它属于弱光谱信号分析技术, 因此存在一系列的技术难点: 需要对采集到的波长进行优选, 以达到提高模型预测精度和简化模型的目的[2,3].

偏最小二乘回归法(Partial Least Squares regression, PLS)是光谱多元定量校正常用的一种方法, 已被广泛应用于近红外、红外、拉曼、核磁和质谱等波谱定量模型的建立, 几乎成为光谱分析中建立线性定量校正模型的通用方法. 近年的理论与试验表明, 在 PLS 建模前先筛选波长, 消除与待分析组分无关或呈非线性关系的波长点, 将可简化模型, 并提高其预测精度和稳健性. Leardi R 提出了一种应用遗传算法结合偏最小二乘法来进行光谱特征波长筛选的方法, 他的一些研究证明这种方法是有效的[4,5].

玉米是重要的粮食作物[6], 在品质育种、资源评价、品种鉴定和分类时, 往往需要对玉米品质进行快速测定, 但常规的化学分析需要繁杂的预处理, 既费工费时又费钱.

近红外光谱分析技术可以克服传统化学分析的缺点. 尽管已有一些研究尝试应用近红外光谱技术来预测玉米的淀粉含量[6~9], 本研究尝试先应用遗传算法结合 PLS 来筛选玉米淀粉的特征波长, 然后基于这些特征波长应用偏最小二乘回归来建立校正模型, 来测定玉米籽粒中淀粉的含量.

1 实验

1.1 仪器与试剂

测定仪器: NIRsystems Model 6500 近红外光谱仪, 光谱测定范围为 1 100~ 2 498 nm, 光谱分

收稿日期: 2008-09-09; 修订日期: 2008-11-05.
作者简介: 沈林峰(1986-), 男, 杭州师范大学钱江学院在读学生, 主要从事应用化学方面的研究.

辨率为 2 nm.

试剂: 氯化钙、乙酸、硫酸锌、亚铁氰化钾.

1.2　样品处理

试验材料包括 80 份普通的玉米籽粒样品, 通过旋光法(GB5006-85)测得其淀粉含量介于 62.83% ~ 66.47% 之间, 按照 3∶1 的比例将样品集随机地分为独立的校正数据集和检验数据集, 其数据特征见表 1, 从表中可以发现, 两个数据集的特征相当接近, 均具有代表性.

表 1　样本淀粉含量的分布特征

Table 1　Statistical characteristic of maize samples

	样本数	最小值 /%	最大值 /%	平均值 /%	标准差 /%
校正数据集	60	62.83	66.47	64.78	0.82
检验数据集	20	63.10	65.80	64.45	0.80

1.2　光谱数据的测定与预处理

将样品装入样品杯中进行近红外光谱的测定, 为消除样品粒度大小、均匀性不一致等因素对光谱的影响, 每个样品均重复装样测定 5 次, 然后计算其平均值. 每条光谱曲线包括 700 个波长的光谱数据, 通过计算其倒数的对数将反射率转换为吸光率, 图 1 为校正数据集中各样本的光谱曲线.

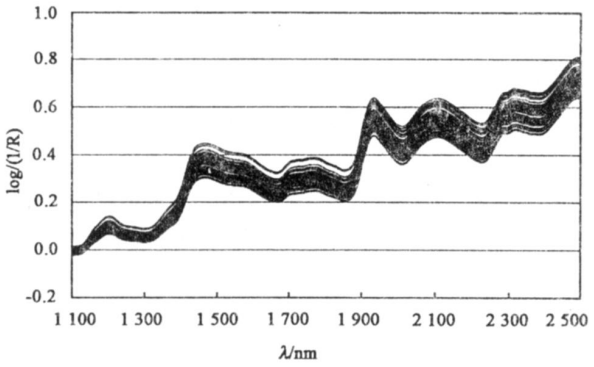


图 1　校正数据集中样本的近红外光谱曲线

Fig. 1　The absorbance near infrared spectra of maize sample in calibration dataset

考虑到近红外光谱在测定过程中的干扰等因素, 采用 Savitzky-Golay 卷积平滑与求一阶微分作为光谱数据的预处理方法. 通过试验, 选择窗口宽度为 7, 多项式次数为 3. 图 2 为经平滑后训练数据集中各样本的一阶微分光谱曲线.

1.3　基于遗传算法和 PLS 特征波长的选择

以 Leardi R 等开发的 PLS-Genetic Algorithm 工具箱为运算工具[4,5], 以 Matlab 为计算平台, 以校正集中样本的一阶微分光谱数据为基础, 进行特征波长的选择, 重复运行 5 次, 以交叉验证误差最小的作为选择的结果, 最终得到 1 552、1 554、1 556、1 558、1 748、1 750、1 752、1 754、1 756、1 758和 1 928 nm的 11 个波长组成的特征波长集(见图 2). 特征选择过程中的一些参数为: 个体数为 30, 变异率 0.01, 交叉率 0.5, 最大进化代数为 100.

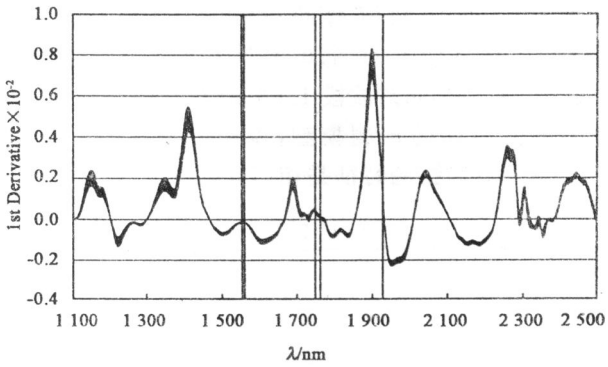


图 2　校正数据集中各样本的一阶微分光谱曲线及经 GA-PLS 选择得到的特征波长

Fig. 2　First derivative near infrared spectra and ranges of selected wavelength for maize samples in calibration dataset

1.4　校正模型建立的方法

在本研究中, 以 NΦrgaard L 等人开发的 iToolbox 工具箱[10] 中的 iPLS 作为进行 PLS 分析和建模的工具, 通过校正数据集建立基于特征波长的淀粉含量的预测模型, 在建模过程中, 应用交叉验证的方法来确定模型需包括的成分数和防止过配. 应用校正标准误 RMSEC(root mean square error of calibration)、交叉验证标准误 RMSECV(root mean square error of cross validation)和预测标准误 RMSEP(root mean square error of prediction)来衡量模型的结果.

2　结果与分析

2.1　基于近红外光谱的玉米中淀粉含量预测模型的建立

以校正数据集中样品的淀粉含量和相应的近红外光谱的特征波长的数据为基础, 应用 iToolbox 中

的工具建立了玉米籽粒淀粉含量的预测模型. 模型所包含的成分数的确定过程见图 3, 从图中可以发现, 尽管校正集误差(RMSEC) 随着模型所包含的成分数持续降低, 但模型的交叉验证误差(RMSECV) 在所包含的成分数大于 5 后, 开始缓慢地上升, 而预测误差(RMSEP) 在成分数超过 6 以后, 随着成分数的增加, 误差快速上升, 然后缓慢上升, 因此, 根据(RMSECV) 的变化确定模型所包含的成分数为 5 时, 为最佳的预测模型. 模型的校正误差(RMSEC)、交叉检验误差(RMSECV) 和预测误差(RMSEP) 分别为 0. 30%、0. 35% 和 0. 27%. 而应用整个光谱范围所建立的预测模型的 RMSEC、RMSECV 和 RMSEP 分别为 0. 31%、0. 42% 和 0. 29%, 也就是说, 仅用不到全光谱波段 1. 6% 的波长数据, 建立了预测精度更高的预测模型, 说明基于遗传算法和 PLS 的光谱数据的特征是相当有效的.

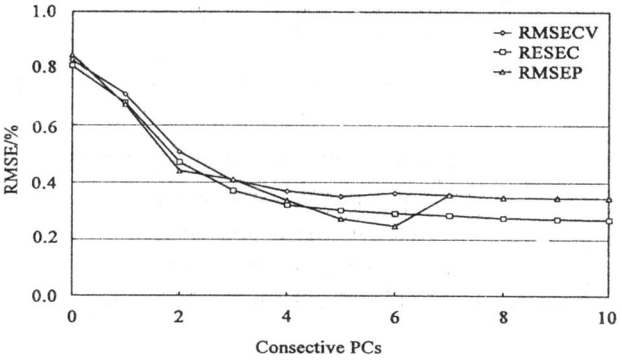
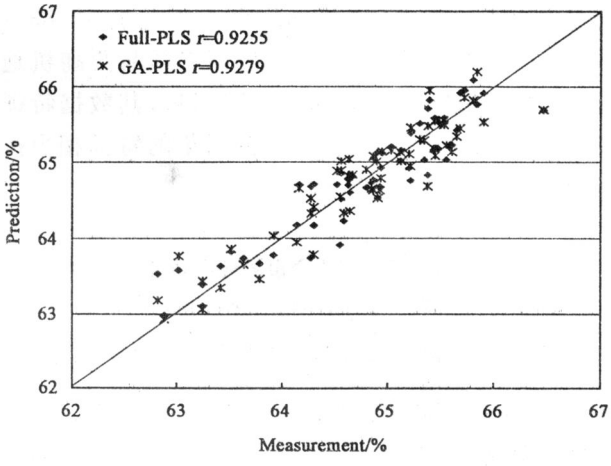


图 3 在基于特征波长的 PLS 建模过程中预测误差随模型中包含的成分数的变化情况

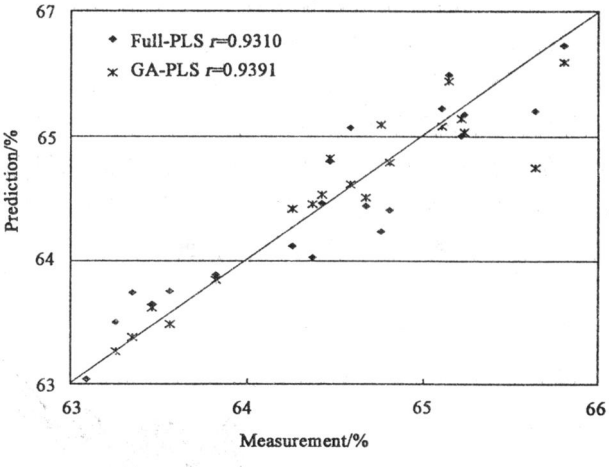
Fig. 3 Changing of RMSEC, RMSECV and RMSEP in processing of PLS by selected wavelengths with increasing of principle components

2. 2 模型的预测效果分析

图 4 分别为应用全光谱数据(图中以 Full-PLS 标记) 和根据遗传算法结合 PLS 所选择的特征光谱数据(图中以 GA-PLS 标记) 所建立的模型对校正数据集和独立的验证数据集进行预测的情况, 从中可以发现, 除个别样本外, 其预测值与实际测定值均比较接近, 其相关系数均达到较高的水平. 与基于全光谱数据的模型相比, 基于特征光谱的模型的预测精度均有一定的改善, 特别是独立的检验数据集中, 大部分样本的预测误差, 有明显的改善, 说明尽管基于全光谱和特征光谱数据的预测模型均具有较高的预测精度和可靠的预测能力, 但经过遗传算法结合 PLS 的特征波长的选择后, 不但能减少建模的复杂性, 还能提高模型的质量.



(a) 校正数据集



(b) 检验数据集

图 4 应用全谱和特征波长建立的模型对校正数据集和验证数据集的预测情况比较

Fig. 4 Comparison between measured versus predicted maize starch content for calibration and validation datasets by PLS with full spectra or selected wavelengths

3 结论

近红外光谱分析技术是一种快速发展的现代分析技术, 具有简单、无损、快速、高效率、低成本等优点. 但由于其属于弱信号分析技术, 因此对数据的特征选择和预测模型的建立提出了更高的要求. 本研究应用遗传算法结合 PLS 对近红外数据进行特征选择, 并以此为基础, 来建立玉米中淀粉含量的近

红外光谱预测模型, 取得了比较满意的预测结果. 与基于全光谱数据所建立的模型相比, 在较少特征波长的情况下, 模型的预测精度反而更高. 而且, 按照 Leardi 的建议, 在进行光谱特征选择时, 所包含的波长数最好不超过 200, 在本研究中, 尽管波长数达到了 700, 但仍然获得了比较满意的结果, 说明应用遗传算法结合 PLS 进行近红外光谱特征波段的选择是有效的.

参考文献:

[1] 王多加, 周向阳, 金同铭. 近红外光谱检测技术在农业和食品分析上的应用[J]. 光谱学与光谱分析, 2004, 24(4): 447-450.

[2] 刘青格, 陈斌. 基于相关分析技术的近红外光谱信息特征提取[J]. 农业机械学报, 2003, 34(3): 79-81.

[3] 谷筱玉, 徐可欣, 汪嘶. 波长选择算法在近红外光谱法中药有效成分测量中的应用[J]. 光谱学与光谱分析, 2006, 26(9): 1 618-1 620.

[4] Leardi R, Lupiáez A. Genetic algorithms applied to feature selection in PLS regression: how and when to use them[J]. Chemolab, 1998, 41: 195-207.

[5] Leardi R. Application of genetic algorithm-PLS for feature selection in spectral data sets[J]. Journal of Chemometrics, 2000, 14: 643-655.

[6] 魏良明, 严衍禄, 戴景瑞. 近红外反射光谱测定玉米完整籽粒蛋白质和淀粉含量的研究[J]. 中国农业科学, 2004, 37(5): 630-633.

[7] 蔡鑫茹, 刘广新, 焦仁海. 近红外光谱仪测定玉米子粒淀粉含量的研究[J]. 吉林农业科学, 2006, 31(6): 10-11.

[8] 朱才, 金香哲, 郭庆江. 应用固定波长近红外光谱仪测定玉米淀粉含量[J]. 分析仪器, 1993, (4): 52-54, 58.

[9] 方景春, 赵敬. 近红外分析仪在玉米质量检测中的应用[J]. 粮食加工, 2006, (3): 81-83.

[10] Nørgaard L, Saudland A, Wagner J. Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy [J]. Applied Spectroscopy, 2000, 54: 413-419.

Study on Determination of Starch Content in Maize by Near-infrared Reflectance Spectroscopy with Genetic Algorithm and PLS

SHEN Lin-feng[1], SHEN Zhang-quan[2]

(1. Department of Science, College of Qianjiang, Hangzhou Normal University, Hangzhou 310012, China; 2. Institute of Agricultural Remote Sensing and Information Application, Zhejiang University, Hangzhou 310029, China)

Abstract: Informative wavelengths were selected from the near-infrared reflectance spectroscopy (NIRS) of maize by genetic algorithm and partial least squares regression (PLS). A calibration model for determination of starch content was built by PLS based on the selected wavelengths of NIRS. The result showed that the root mean square error of calibration (RMSEC), root mean square error of cross validation (RMSECV) and root mean square error of prediction (RMSEP) derived from the calibration model based on the 11 selected wavelengths were 0.30%, 0.35% and 0.27%, respectively. And the coefficients of relationship between measurements and predictions for calibration and independent validation datasets were 0.927 9 and 0.939 0, respectively. The accuracy of prediction was better than the model based on the full NIRS data. It was proved that modeling by PLS based on the feature selection with genetic algorithm and PLS was a simpler, effective and more accurate means for determination of starch content in maize.

Key words: near-infrared reflectance spectroscopy (NIRS); partial least squares regression (PLS); maize; starch content; genetic algorithm; feature selection

Classifying number: TS210.7