中国农业大学学报 2006,11(2):74-77
Journal of China Agricultural University

相关成分分析法在大米直链淀粉波长选择中的应用

张巧杰[1,2] 王一鸣[2] 吴静珠[2]
(1. 北京机械工业学院 计算机及自动化系,北京 100085; 2. 中国农业大学 信息与电气工程学院,北京 100083)

摘 要 为挑选信息含量大、与样品组成或性质相关性较强的光谱区域参与建模,以提高校正模型的精度,采用相关成分分析法对大米直链淀粉的近红外光谱进行分析。结果表明:采用相关成分分析法进行波长选择后,建模波长点数减少为波长选择前的22%,模型预测值与标准值的相关系数 R 由0.921 2提高到0.973 0,交叉验证标准差(SECV)由3.404 3减小为1.977 4,预测标准差(SEP)由4.810 0减小为1.900 0,模型的预测能力得到显著提高。

关键词 近红外光谱;相关分析;波长选择;直链淀粉
中图分类号 TH744 **文章编号** 1007-4333(2006)02-0074-04 **文献标识码** A

Application of correlative component analysis in the study of selecting wavelength in apparent amylase content with near infrared spectroscopy

Zhang Qiaojie[1,2], Wang Yiming[2], Wu Jingzhu[2]
(1. Computer and Automation department, Institute of Beijing Mechanical Engineering, Beijing 100085, China;
2. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China)

Abstract Wavelength selecting can be used to select a research space with all combinations of strong correlativity wavelength and large magnitude of the concentration information as final wavelength regions to build a PLS calibration model of NIR. Correlative component analysis algorithm can be employed to identify the magnitude of the information of samples concentration by the variance of the correlative component matrix between spectral matrix and concentration matrix. The apparent amylase content test results showed that the numbers of wavelengths for building the models can be reduced to 22% of the original method. Correlation coefficient can be increased from 0.921 2 to 0.973 0, standard deviation of cross validation in calibration can be reduced from 3.404 3 to 1.977 4, and the root mean squared error in prediction was reduced from 4.810 0 to 1.900 0. The prediction precision was greatly improved by correlative component analysis algorithm.

Key words near infrared spectroscopy; correlative component analysis; wavelength selecting; apparent amylase content

近年来,近红外光谱分析技术以其不损耗样品、方便、快速、经济、无污染等特点在直链淀粉检测方面得到广泛应用。刘建学[1-2]等对不同粒度、不同类型的大米样品进行近红外光谱分析,建立了大米直链淀粉含量预测模型;谢新华[3]用近红外光谱透射技术及3种不同的回归统计分析方法建立了精米直链淀粉含量定量分析预测模型;舒庆尧[4]用近红外反射光谱技术分析比较了3种不同回归统计方法建立的精米粉直链淀粉含量校正模型的效果。这些研究均侧重于用光谱仪软件建立校正模型,尚未见有关直链淀粉特征波长选择的研究报道。

波长选择可以从得到的光谱中提取最有效的光谱信息,提高校正模型的精度和预测能力,简化运算。样品每一组分都有其特征谱区,而在特征谱区外吸收很弱;因此,挑选待测组分的特征谱区参与建模是建立稳定可靠的校正模型的重要手段。

收稿日期:2005-04-25
基金项目:国家高技术研究发展计划资助项目(2003AA209012)
作者简介:张巧杰,讲师,博士研究生,主要从事近红外光谱品质检测技术研究;王一鸣,教授,博士生导师,通讯作者,主要从事智能化检测与控制研究,E-mail:ym_wang@263.net

大米成分复杂,需要借助数学手段将其特征波长从近红外光谱中提取出来。相关矩阵分析法是从近红外光谱与大米组分含量之间相关信息量的角度提出的一种简单有效的波长选择方法。

1　基本原理

理想情况下,希望校正集样品光谱信息的最大变化是由被测样品的组成或性质的变化引起的。模型的建立是根据光谱的这种变化,而不是根据光谱的绝对强度,因此光谱变化最明显的区间也应当对应光谱信息最丰富的区域。各个波长点近红外光谱数据的方差越大,离散程度越大,则相应的样品差异也越大;因此将方差作为考察该数据点信息差异的指标,可从整体上较好地反映各波段的信息量。

单纯从信息源即近红外光谱的角度,虽然考虑了波长点所含信息变化量的大小,但是这些变化可能是干扰因素所引起的,与样品被测成分的差异并无联系,属于不相关的冗余信息,对近红外光谱定量分析精度会产生影响;因此还需要考虑近红外光谱与样品被测组分含量之间的相关关系,从而挑选出与被测组分信息相关性较大的区域。为此,构造 $n \times n$阶方阵 $C = [C_1, C_2, \cdots C_n]^T$,其中 $C_i = [0, \cdots, 0, c_i, 0, \cdots 0]^T, (i = 1, 2, \cdots, n)$, c_i 为经标准化之后的样品浓度数据。进而构造相关成分矩阵 $S = XC$,其中 X 为光谱矩阵。相关成分矩阵 S 既体现了光谱矩阵的信息,也反应了浓度矩阵的信息[5];因此只需考察 S 各行向量的方差值,就可以反映出各个波长点光谱数据所体现的样品被测成分浓度差异的信息量,即可以根据方差的大小判断各波长点光谱数据的离散度以及光谱数据与浓度的相关性,从而挑选出与样品被测组分含量相关性较强的近红外光谱区间。

算法实现步骤如下:

1) 对光谱矩阵 $X_{p \times n}$、浓度矩阵 $Y_{1 \times n}$分别做中心化处理,得到 $X_p'_{\times n}$, $Y_1'_{\times n}$,即

$$X_{ij}' = X_{ij} - \overline{X}_j, \quad \overline{X}_j = \frac{1}{n}\sum_{j=1}^{p} X_{ij},$$

$$Y_i' = Y - \overline{Y}, \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

式中: p 为波长点数, n 为样品数。

2) 对 Y 进行标准化得到 Y″,即

$$Y'' = Y / R = [y_1'', y_2'' \cdots y_n'']$$

式中: R 为浓度矩阵 Y 的极值。

3) 构造 $n \times n$ 阶方阵 C

$$C = \begin{bmatrix} c_{11} & 0 & \cdots & 0 \\ 0 & c_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & c_{nn} \end{bmatrix}$$

式中: $c_{ij} = y_j'', i = j = 1, 2, \cdots, n$。

4) 构造相关成分矩阵 $S = XC = [S_1, S_2, \cdots, S_p]^T$,其中 $S_i = [s_{i1}, s_{i2}, \cdots s_{in}]$, $i = 1, 2, \cdots, p$,为相关成分矩阵 S 的行向量。

5) 计算相关成分矩阵 S 各行向量的方差,即

$$\mathrm{var}(S) = [\mathrm{var}(S_1), \mathrm{var}(S_2), \cdots, \mathrm{var}(S_p)]^T$$

设定初始方差阈值,根据校正模型的精度调整阈值,从而确定最优波段。

2　实验仪器与方法

1) 仪器。采用德国 Bruke 公司 MATRIX- I型傅里叶变换近红外光谱仪、高灵敏度 24 位数字化 PbS 检测器,光谱采集范围 4 000～12 500 cm^{-1},扫描分辨率 16 cm^{-1},波长点数 1 102,石英样品池。

2) 样品与基础数据来源。107 个大米样品由中国农业科学院作物品种资源研究所提供,选用的样品覆盖了我国主要产稻区的典型品种,直链淀粉质量分数 1.0 %～26.6 %,标准值由该所按照 GB7648—1987《水稻、玉米、谷子籽粒直链淀粉测定法》[6]测定。将样品随机分为 2 集:校正集样品 90 个,预测集样品 17 个。

3) 光谱采集。以空气为参比样品,扫描次数 64 次,采用大样品杯旋转式扫描方式,以减少样品状态对光谱的影响。3 个不同直链淀粉含量样品淀粉的

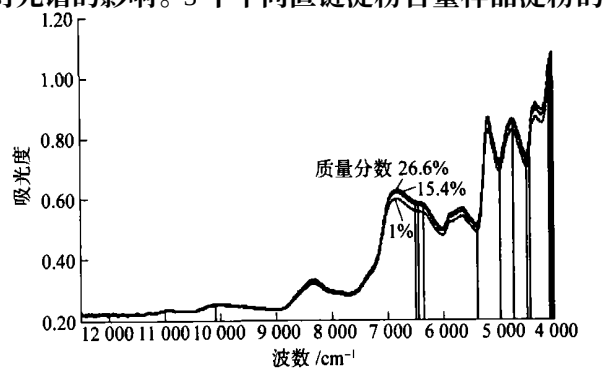

图 1　不同直链淀粉含量样品的近红外光谱及淀粉近红外谱峰

Fig. 1　Near-infrared spectra of three samples with different amylase content and starch character spectral wavelength

O —H、C —O 和 C —C 基团在近红外谱区谱峰的位置[8]见图 1,这些主要吸收峰来源见表 1。

表 1 淀粉有机基团与近红外吸收谱峰关系表

Table 1 Relationship between amylose organic radical and near-infrared spectra peak

波数/ cm^{-1}	振动方式	波数/ cm^{-1}	振动方式
10 101	3 ×O —H str.	4 762	2 ×O —H def. + 2 ×C —O str.
6 545	2 ×O —H str. (内氢键)	4 440	O —H str. + O —H def.
6 494	2 ×O —H str. (内氢键)	4 394	O —H str. + C —C str.
6 329	2 ×O —H str. (内氢键)	4 063	C —H str. + C —C str.
5 263	O —H str. + 2 ×C —O str.	4 019	C —H str. + C —C str.
5 000	2 ×O —H def. + C —O def.	4 000	C —H str. + C —C str.

注:"2 ×"3 ×"分别表示振动频率为基频频率的 2 或 3 倍;"str. "def. "分别表示伸缩振动和变形振动。

4)数据处理。采用偏最小二乘方法建模。偏最小二乘法和相关成分分析法算法程序均采用 MATLAB 语言编写。

3 结果与分析

使用相关成分分析法对 90 个校正集样品原光谱波长区间进行优化挑选,这里以相关成分矩阵 S 行向量方差的最小值为初始值,以其行向量方差极值的 n 等分为步长,以校正模型最佳主成分交叉校验(Cross Validation)预测值与标准值的相关系数[7]平方 R^2 最大,且交叉验证标准差(Standard Error of Cross Validation) s_c 最小(否则预测能力和精度不高)为标准评价校正模型的精度。s_c 最小可表示为 $1 + s_c$ 的倒数最大,即校正模型精度的评价标准可以表示为: $R^2/(1 + s_c)$。$R^2/(1 + s_c)$ 越大,模型精度越高。不同相关成分矩阵方差阈值与模型精度的评价标准 $R^2/(1 + s_c)$ 的关系见图 2。可以看出,当阈值为 4.754 4 ×10^{-5}时,所建校正模型的效果最好,此时经过相关成分矩阵方差选择出的波长点数为 243 个,对应的谱区见图 3。

相关成分分析法选出的 2 个频段为:6 981～6 426和 5 307～4 000 cm^{-1}。经过相关成分分析法优化的谱区,除 O —H 基团的二级伸缩振动倍频谱峰10 101 cm^{-1}由于二级伸缩振动比较弱,且受水分在10 310 cm^{-1}附近的谱峰影响比较大等未被提取出,C —H 基团的一级伸缩振动倍频谱峰 6 329 cm^{-1} 由于其振动较弱,受其他有机成分的干扰较大,未被提取出外,其他谱峰都包含在优化后的波长范围内,且大米其他组分如蛋白质、纤维素、水分、糖类等干扰组分的谱区大部分被滤除。这充分说明相关成分分析波长选择方法挑选的波长组合与被测组分浓度的相关性很强,针对性好,能有效滤除干扰谱区。

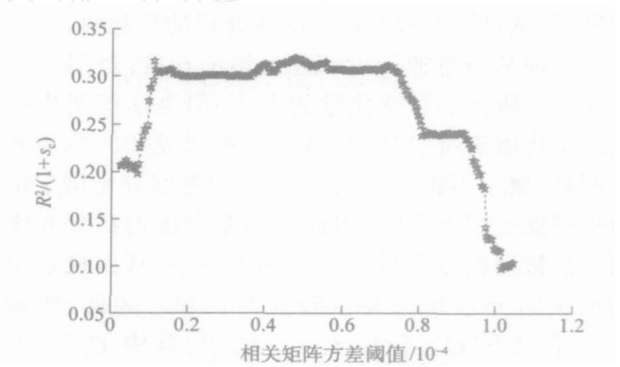

图 2 相关成分矩阵方差阈值与模型评价标准 $R^2/(1 + s_c)$ 的关系

Fig. 2 Correlative components analysis variance threshold with $R^2/(1 + s_c)$ of calibration

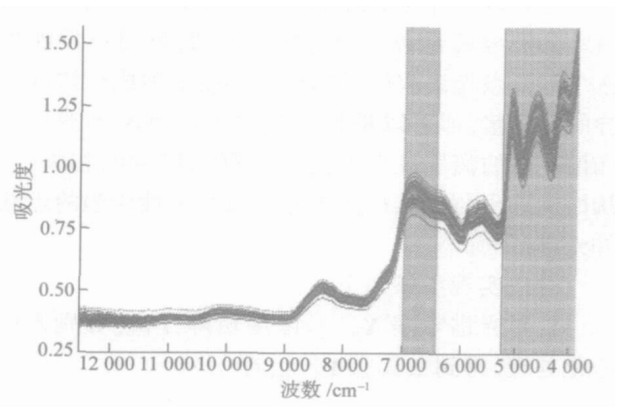

图 3 90 个大米样品经相关成分分析法波长选择后的光谱图

Fig. 3 Spectra of 90 samples after correlative components analysis selecting wavelength

用阈值 4.754 4 ×10^{-5}对应的波长组合建立 90 个样品的校正模型,波长选择前后校正结果见表 2。

表 2　相关成分分析法波长选择前后近红外校正模型校正结果

Table 2　Calibration results before and after selecting wavelength regions by CCA

波长选择前			波长选择后		
波长点数	相关系数 R	交叉验证标准差 s_c	波长点数	相关系数 R	交叉验证标准差 s_c
1 102	0.921 2	3.404 3	243	0.973 0	1.977 4

注: $R=\left[\sum_{i=1}^{n}(c_i^p-\bar{c}^p)(c_i^o-\bar{c}^o)\right]\Big/\left[\sqrt{\sum_{i=1}^{n}(c_i^p-\bar{c}^p)^2}\sqrt{\sum_{i=1}^{n}(c_i^o-\bar{c}^o)^2}\right]=\sqrt{R^2}$, $s_c=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(c_i^o-c_i^p)^2}$, c_i^o 和 c_i^p 分别表示样品浓度的标准值和预测值,表 3 同。

用该模型对 17 个预测集样品进行测定,结果见表 3。可以看出,波长选择前样品预测值的绝对误差最大为 - 8.72,波长选择后为 - 3.13;预测标准差 s_p 由 4.81 减小到 1.90,预测结果得到较大幅度提高。

表 3　相关成分分析法波长选择前后预测效果

Table 3　Prediction results before and after selecting wavelength regions by CCA

样品号	标准值/ %	波长选择前		波长选择后	
		预测值/ %	绝对误差	预测值/ %	绝对误差
1	13.30	12.32	0.98	10.83	2.47
2	17.20	22.59	- 5.39	18.39	- 1.19
3	16.20	22.00	- 5.80	17.75	- 1.55
4	14.90	21.64	- 6.74	17.35	- 2.45
5	13.70	18.95	- 5.25	16.49	- 2.8
6	1.00	9.72	- 8.72	2.03	- 1.03
7	16.30	22.76	- 6.46	18.99	- 2.69
8	1.70	- 1.29	2.99	- 1.50	3.2
9	1.30	4.26	- 2.96	2.22	- 0.92
10	1.40	3.24	- 1.84	4.532	- 3.13
11	15.20	15.13	0.07	16.25	- 1.05
12	15.40	22.35	6.95	14.50	0.9
13	13.10	20.33	- 7.23	14.37	- 1.27
14	18.20	17.07	1.13	18.56	- 0.36
15	17.20	17.35	- 0.15	16.77	0.43
16	19.30	22.30	- 3.00	17.27	2.03
17	15.30	17.77	- 2.47	14.79	0.51
预测标准差 s_p			4.81		1.90

注: $s_p=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(c_i^o-c_i^p)^2}$, n 为预测集样品数。

4　结　论

本研究提出的相关成分分析波长选择方法,不需要先验知识,可以选择出待测组分浓度预测效果较好的谱区,不仅简化和优化了校正模型,而且提高了所建模型的预测能力。用相关成分分析法进行波长选择后,建模的波长点数减少为原来的 22 %,模型的交叉验证标准差由 3.404 3 减小为 1.977 4,预测标准差由 4.810 0 减小为 1.900 0,模型的预测能力显著提高。

参　考　文　献

[1] 刘建学,吴守一,方如明. 大米直链淀粉含量的近红外光谱分析[J]. 农业工程学报,2000,16(3):94-96

[2] 刘建学,吴守一,方如明. 基于近红外光谱的神经网络预测大米直链淀粉含量[J]. 农业机械学报,2001,32(2):55-57

[3] 谢新华,肖昕,李晓方,等. 用近红外透射光谱技术测定精米直链淀粉含量研究[J]. 食品科学,2004,25(1):118-121

[4] 舒庆尧,吴殿星,夏英武. 用近红外光谱技术测定精米样品表观直链淀粉含量的研究[J]. 中国水稻科学,1999,13(3):189-192

[5] Chen Dezhao, Chen Yaqiu, Hu Shangxu. Correlative components analysis for pattern classification[J]. Chemometrics and Intelligent Laboratory Systems, 1996, 35: 221-229

[6] GB 7648—87 水稻、玉米、谷子籽粒直链淀粉测定法[S]

[7] 陆婉珍,袁洪福,徐广通,等. 现代近红外光谱分析技术[M]. 北京:中国石化出版社,2000

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