中国农业大学学报 2004,9(6):76～79
Journal of China Agricultural University

PLS回归法建立适应温度变化的近红外光谱定量分析模型

王 韬[1] 张录达[2] 劳彩莲[1] 李军会[1] 赵龙莲[1] 严衍禄[1]
(1. 中国农业大学 信息与电气工程学院,北京 100094; 2. 中国农业大学 理学院,北京 100094)

摘 要 研究了近红外光谱定量分析模型对于样品温度的适应性。以42个不同品种的大豆为实验材料,用2台光谱仪分别独立测定了样品在5种温度下的近红外光谱。对于2台光谱仪测定的光谱,均依据光谱信息选择部分光谱,采用PLS回归法对大豆样品的粗蛋白质和粗脂肪含量分别建立了近红外光谱定量分析模型,并以剩余样品对模型进行预测检验。4个模型的预测结果均表明:超过94%的检验样品的预测相对误差小于5%,说明了预测样品处于5～40 ℃时,模型都有较好的预测效果。

关键词 近红外光谱; PLS回归模型; 温度校正
中图分类号 O 174.22 **文章编号** 1007-4333(2004)06-0076-04 **文献标识码** A

Study on building temperature adapting near infrared spectra quantitative models with PLS regression method

Wang Tao[1], Zhang Luda[2], Lao Cailian[1], Li Junhui[1], Zhao Longlian[1], Yan Yanlu[1]
(1. College of Information and Electrical Engineering, China Agricultural University, Beijing 100094, China;
2. College of Science, China Agricultural University, Beijing 100094, China)

Abstract This paper was about the temperature adaptability to quantitative analysis model of near infrared spectrum. 42 soybean samples were used and measured the near infrared spectrum independently and separately with two spectrometers at 5 different temperatures. The spectra were then divided to calibration set and validation set according to the information of spectrum. Two quantitative analysis models with the calibration set were established and used to estimate the contents of raw protein and raw fat. The same operations were made for the remained spectra. The results showed that the relative residuals of the 94% tested samples were within a range of ±5%. The models had a high adaptation to temperature during predicting the contents of raw protein and fat in soybeans.

Key words near infrared spectra; PLS regression model; temperature calibration

近红外光谱含有丰富的含氢基团的信息,因而日益受到波谱分析学界的重视。由于近红外光谱信号强度很弱,谱峰严重重叠[1],所以必须利用多元统计的方法建立数学模型才能进行样品的分析。近红外光谱一般都有几百甚至上千个波长点,直接建立多元回归模型,则由于1)运算量太大,2)某些波长点之间存在线性关系,所建立的模型不能很好地满足预测的需要。为了消除上述2个因素的影响,一般采取的方法是利用主成分分析从近红外光谱中提取十几到几十个正交的主成分再建立模型。Geladi等[2]提出的偏最小二乘(PLS)回归方法,在提取主成分时考虑到与待分析组分的相关性,所建立的线性模型具有较高的预测精度。

在利用仪器测定近红外光谱时,任玉林等[4]研究发现所测光谱容易受到样品温度的影响。Wülfert等[5]关于温度对模型影响的研究表明,用单一温度下测得的光谱建立的PLS回归模型适应性较差,仅限于对相同温度下所测光谱进行分析,这就大大限制了模型的应用。然而,若将样品的温度作为模型的一个参数,必将加大建模工作的复杂度。

收稿日期: 2004-07-26
基金项目: 国家高技术研究发展计划资助项目(2002AA248051,2002AA243011),国家重大基础研究前期研究专项(2002CCA00800),国家科技攻关项目(2004BA210A03)
作者简介: 王韬,硕士研究生;张录达,教授,通讯作者,主要从事应用数学研究,Tel: 010-62732669,E-mail: zhangld@cau.edu.cn

基于此,本研究工作对几组不同温度下样品的近红外光谱进行测量,然后混合这些光谱建立模型。由于建模样品已包含了不同温度的信息,得到的模型在一定程度上减小了温度的影响。

1　材料与方法

1.1　样品与设备

样品为中国农科院品种资源所提供的 42 个大豆品种,测量光谱时不经任何处理。样品包括 2 个待分析组分:粗蛋白质与粗脂肪的组分含量。光谱测量仪器为常温 CCD2995 型光谱仪(以下简称常温仪器)和恒温 CCD602 型光谱仪(以下简称恒温仪器),均配有近红外光谱定量分析软件。

1.2　实验方案

1)光谱获取。利用恒温与常温仪器分别测定 42 个大豆样品在 5 种温度(5,10,20,30 和 40 ℃)下的对可见-近红外光的吸光度,谱区 643～1 250 nm,共 2 048 个波长点。考虑到光谱两端的噪声干扰严重,只取 800～1 050 nm 的共 844 个波长点进行建模和预测,同样条件下测 2 次。最终得到光谱样品 776 个,其中恒温光谱仪器测量的光谱样品 420 个,常温光谱仪器测得的光谱样品 356 个。

2)挑选光谱。用离差平方和聚类法[3],从恒温仪器测量的 420 条光谱中由计算机挑选出 58 条光谱,从常温仪器测量的 356 条光谱中挑选出 56 条光谱,分别建立模型。两类光谱中剩下的光谱用于预测。对挑选出的建模光谱样品进行统计,结果列于表 1。

表 1　建模光谱样品数量分布

Table 1　The distribution of spectrum samples in calibration set

测量仪器	被测大豆样品温度/ ℃					合计
	5	10	20	30	40	
恒温仪	16	14	2	5	20	58
常温仪	4	15	5	9	23	56

1.3　建模方法

本研究利用 PLS 回归方法建立模型并预测。

1) PLS 算法原理[1～3]

PLS 法首先将应变量矩阵 $Y=(y_{ij})_{n\times m}$ 和自变量 $X=(x_{ij})_{n\times p}$ 分解成特征向量形式:

$$Y = UQ + F \tag{1}$$

$$X = TP + E \tag{2}$$

式中:U 和 T 分别为 Y 和 X 的特征因子矩阵($n\times d$ 阶,d 为抽象组分数),Q($d\times m$ 阶)和 P($d\times p$ 阶)分别为 Y 和 X 的载荷阵,F($n\times m$ 阶)和 E($n\times p$ 阶)为 Y 和 X 的残差阵。

PLS 法根据特征向量的相关性分解 Y 和 X,建立回归模型

$$U = TB + E_d \tag{3}$$

式中:E_d 为随机误差阵,B 为 d 维对角回归系数阵。

对待测样品,如果吸光度向量为 x,则含量为

$$y = x(UX)^{-}BQ \tag{4}$$

在 PLS 回归建模时,利用交叉证实方法确定 PLS 主成分的最佳维数 d。

2)模型评价指标。

(1)决定系数(R^2)

在建立 PLS 回归模型时,采用决定系数(R^2)作为一个评价指标,其计算公式[2]如下:

$$R^2 = \left(1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}\right)\times 100 \tag{5}$$

式中:分数的分子为残差平方和,即 PRESS,分母为训练集或预测集的总离差平方和。

(2)平均预测相对误差(%)(mean relative error of prediction, MREP)

平均预测相对误差是模型整体的又一个评价指标,由下面公式计算得到:

$$e_{\mathrm{mrp}} = \sqrt{\frac{\sum_{i=1}^{N}\left[\left(\hat{y}_i - y_i\right)/y_i\right]^2}{N}}\times 100 \tag{6}$$

式中:N 为样品数,$\hat{y}_i$ 为第 i 个样品的组分的回归值,y_i 为第 i 个样品的组分的观测值。

(3)预测相对残差(%)(relative residual of prediction, RRP)

这个指标是用来评价模型对单个样品的预测效果的。对于第 i 个预测样品,

$$R_{rp,i} = \frac{\hat{y}_i - y_i}{y_i}\times 100 \tag{7}$$

2　结果与分析

本研究首先以常温仪器在单温度(20 ℃)下所测的大豆样品光谱建立粗蛋白质和粗脂肪定量分析模型,并以这 2 个模型预测各温度下的光谱样品,各样品的预测相对残差见图 1,2。其中 1～84 号样品

为大豆样品在 5 ℃下的光谱样品,85～168 号样品为 10 ℃下的光谱样品,169～189 号样品为 20 ℃下的光谱样品,190～271 号样品为 30 ℃下的光谱样品,272～356 号样品为 40 ℃下的光谱样品。

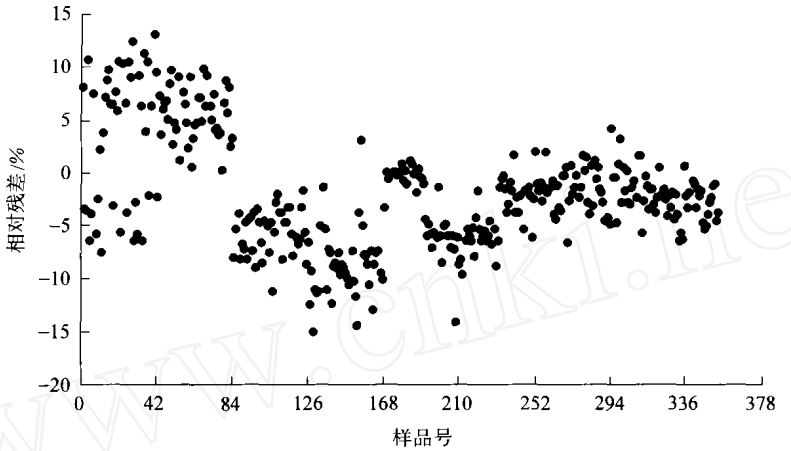

图 1 粗蛋白质预测相对残差

Fig. 1 The relative residuals of the predicted content of raw protein (shown in percentage)

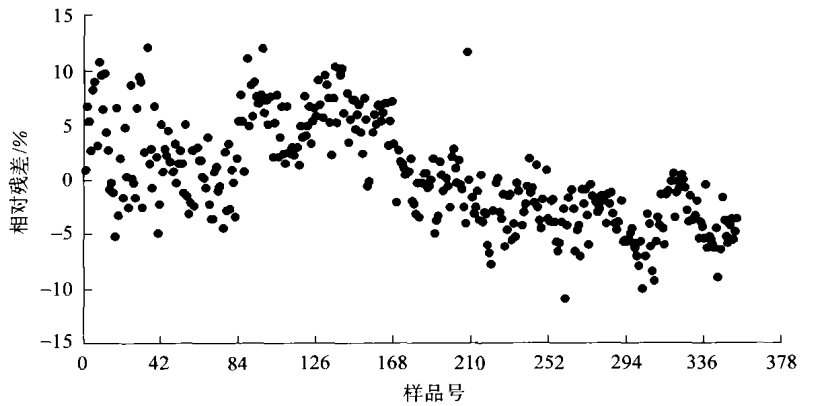

图 2 粗脂肪预测相对残差

Fig. 2 The relative residuals of the predicted content of raw fat (shown in percentage)

从图中数据可见,20 ℃样品所建模型预测其相对残差较大,特别是 5 和 10 ℃下所测光谱样品。

针对单温度模型的上述缺点,本研究建立了 4 个混合温度的定量分析模型:1) 恒温-粗蛋白质(即选用恒温仪器测量的光谱,分析组分为大豆的粗蛋白质含量)模型;2) 恒温-粗脂肪模型;3) 常温-粗蛋白质模型;4) 常温-粗脂肪模型。4 个模型预测结果的相对残差分布,即相对残差在某个区间的光谱样品个数与样品总数的百分比,见表 2。结果显示,预测相对残差在 ±5 %以内的样品所占比例均在 94 %以上。这表明 4 个模型对温度的适应性强,对于较宽温度范围(5～40 ℃)内测得的大豆光谱样品都能有较好的预测结果。

表 2 模型预测相对残差分布情况

Table 2 The distribution of relative residuals from four models %

相对残差区间	恒温粗蛋白模型	恒温粗脂肪模型	常温粗蛋白模型	常温粗脂肪模型
$(-\infty, -5) \cap (-5, +\infty)$	5.80	4.70	3.00	4.00
$[-5,5]$	94.20	95.30	97.00	96.00
$[-4,4]$	88.43	87.60	91.33	90.33
$[-3,3]$	81.49	75.97	82.33	75.33
$[-2,2]$	60.50	56.91	60.33	54.67
$[-1,1]$	30.39	30.94	31.33	29.00
合计	100.00	100.00	100.00	100.00

从表 3 可以看到,4 个模型中,有 3 个模型的校正决定系数在 90 以上,预测结果的决定系数中亦有 3 个在 90 以上,表明预测模型有较好的预测结果。同时,4 个模型的平均相对误差均小于 2.1 %,也表明了预测模型的温度适应性好。另外,利用处于不同温度状态的 2 个仪器测得的光谱建立的模型,其平均相对误差差别不大,这表明仪器的温度状态对测量、建模、预测的结果影响不大。

表 3　模型总体评价指标

Table 3　The overall evaluation indices of models

评价指标	校正或检验	恒温粗蛋白质模型	恒温粗脂肪模型	常温粗蛋白质模型	常温粗脂肪模型
决定系数 R^2	校正	87.07	94.72	92.31	95.41
	检验	89.04	91.28	90.44	91.19
平均相对误差/ %	校正	1.91	1.70	1.92	1.75
	检验	1.94	2.04	2.06	2.08

3　结论与讨论

混合 5 种不同温度下测定的光谱所建立的大豆样品粗蛋白、粗脂肪定量分析模型可以适用于较宽的温度范围,对各温度下所测光谱的预测有较好的结果。

从所得结果可以看出,仪器所处温度状态对建模、预测没有大的影响,因此在测量光谱时,可以不考虑测量仪器的温度状态。

本研究主要从总体上对模型的预测效果进行探讨,然而利用离差平方和聚类法挑选出的光谱样品,在不同温度下的分布很不平均。如果人为地使建模光谱样品在不同温度下分布较平均,将会对模型的建立与预测产生什么样的影响?这一点有待进一步研究。

建模光谱样品在总光谱样品中所占比例都比较小(均小于 1/5),信息量小是否会影响模型建立与预测效果?如果增加建模光谱样品的数量,模型的预测效果是否会有改善?亦有待研究。

感谢徐志龙,侯瑞峰,王南,郭亮,王琳,万利峰,张海艳,涂少雄,伍芳翠等实验人员提供光谱数据。

参 考 文 献

[1] 齐小明,张录达,杜晓林,等. PLS-BP 法近红外光谱定量分析研究[J]. 光谱学与光谱分析,23(5):870～872

[2] Geladi P, Kowalski B R. Partial least squares: a tutorial [J]. Anal Chem Acta, 1986,185:1～17

[3] 于秀林, 任雪松 编著. 多元统计分析[M]. 北京: 中国统计出版社,1999. 61～97

[4] 任玉林,张滨,郭晔,等. 用主成分回归进行生理盐水的非破坏分析[J]. 数理医药学杂志, 11(2):165～167

[5] Florian Wülfert, Wim Th Kok, Age K Smilde. Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models. Anal Chem, 1998,70(9):1761～1767