

# GA-PLS 结合 PG-ANN 算法提高奶粉蛋白质模型精度

孙 谦, 王加华, 韩东海\*

中国农业大学食品科学与营养工程学院, 北京 100083

**摘 要** 提出一种偏最小二乘法(PLS)和人工神经网络(ANN)结合用于近红外光谱(NIRS)的分析方法,以提高奶粉蛋白质模型的预测精度。首先采用基于遗传算法的波长选择法(RS-GA)优化光谱数据,建立 GA-PLS 模型预测奶粉蛋白线性部分;然后在 RS-GA 法选择的波段上进行主成分分析(PCA),以主成分的得分矩阵作为 ANN 模型输入层,以 GA-PLS 预测值与真实值之差作为输出层,建立 PG-ANN 模型预测其非线性部分。最终预测结果为两个模型预测值之和,以模型的预测标准偏差(RMSEP)作为评价指标,以便考察新方法的有效性。同时建立线性的全谱模型(Fr-PLS),其 Fr-PLS、GA-PLS 和 GA-PLS+PG-ANN 模型的 RMSEP 分别为 0.511, 0.440 和 0.235。结果表明:考虑奶粉蛋白含量近红外模型的非线性部分,可以显著提高模型的预测精度,该方法也可为其它复杂体系模型精度的提高提供借鉴。

**关键词** 近红外光谱; GA-PLS; PG-ANN; 模型精度; 奶粉; 蛋白质

中图分类号: O657.3 文献标识码: A DOI: 10.3964/j.issn.1000-0593(2009)07-1818-04

## 引 言

近红外光谱(near infrared spectroscopy, NIRS)是各成分功能基团的综合反映,光谱信息相互干扰、重叠及掩盖,对于某些信息区间(波长)光谱信号并非简单的线性或非线性的。偏最小二乘法(partial least squares, PLS)是一种基于线性关系的建模方法,已广泛用于建立乳及乳制品成分的定量模型<sup>[1-4]</sup>,然而单一的线性模型往往忽略了近红外光谱信息的非线性部分,而导致模型的预测精度不是最高。对于奶粉复杂体系,建立蛋白质的近红外模型时,就需要考虑近红外光谱的线性和非线性信息部分,同时也要选择数据优化技术简化模型,提高预测精度。

在建立复杂体系目标组分模型时,需要通过波长选择提取最有效的光谱信息,简化运算;同时采用线性和非线性加和算法,以提高模型预测精度。遗传算法(genetic algorithm, GA)是应用较为广泛的一种波长选择方法,且已成功用于近红外的波长选择中<sup>[5-11]</sup>;人工神经网络(artificial neural network, ANN)算法是一种重要的非线性拟合工具,具有抗干扰、抗噪声和强大的非线性转换能力,但其输入层节点数不能太多,通常采用主成分分析(principal component analysis, PCA)选择新变量作为 ANN 输入,以减少输入层节点

数<sup>[12-17]</sup>。

本文提出了利用 PLS 模型来预测奶粉蛋白含量的线性部分,ANN 模型来预测非线性部分,将两者结合的分析方法来提高蛋白模型的预测精度。改进的遗传算法(region selection based on genetic algorithm, RS-GA)<sup>[18]</sup>用于优化近红外光谱数据,建立 GA-PLS 模型;在 RS-GA 法选择的波段上进行 PCA,以主成分得分矩阵作为 ANN 模型输入层,以 GA-PLS 预测差值作为输出层建立 PG-ANN 模型;最终预测结果为两个模型预测结果之和,以最终模型的预测标准偏差(RMSEP)作为模型的评价指标,考察该方法的可行性。

## 1 材料与方 法

### 1.1 实验材料

实验样品来自全国各地市场,收集了不同品牌、不同种类的奶粉样品。总样品数为 140 个,其中 114 个样品作为建模集,剩余样品作为预测集进行预测。

### 1.2 样品光谱采集

利用 Antaris 傅里叶变换近红外光谱仪(Thermo Nicolet, 美国)积分球附件采集奶粉样品的漫反射光谱。为消除装样差异引起的光谱变化,采用自动旋转样品杯附件,设置 8 mm 偏心距,采集光谱范围 4 000~10 000 cm<sup>-1</sup>,分辨

收稿日期: 2008-03-26, 修订日期: 2008-06-28

基金项目: 国家“十一五”科技支撑计划项目(2006BAD11A12-08)资助

作者简介: 孙 谦, 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: bestsq10@126.com

\* 通讯联系人 e-mail: handh@cau.edu.cn

率为 8 0 cm<sup>-1</sup>, 扫描 32 次取平均, 使用 InGaAs 检测器。样品每次装样高度为 2 cm, 每个样品采集 3 次, 3 条光谱的平均光谱作为最终的样品光谱。

1.3 样品化学值测定

奶粉蛋白质含量采用 GB/T 5009.5—2003 方法进行测定。每个样品平行测定 3 次, 所得结果平均后作为最终样品的蛋白含量。表 1 为建模集和预测集样品蛋白质含量的统计结果, 可以看出样品具有较好代表性, 分布范围较宽, 预测集样品信息很好地被建模集所涵盖。

Table 1 Statistic results of the protein content measured by the standard method

|                 | Number of samples | Mean | Max  | Min  | SD*  |
|-----------------|-------------------|------|------|------|------|
| Calibration set | 114               | 15.3 | 22.4 | 10.7 | 2.65 |
| Validation set  | 26                | 15.2 | 19.6 | 11.2 | 2.49 |

\* SD: Standard deviation

1.4 方法介绍

奶粉 NIR 光谱信息存在线性和非线性两部分, 因此, 奶粉蛋白质真实值  $C_0$  为线性部分的蛋白质含量  $C_1$  与非线性部分的蛋白含量  $C_2$  之和, 即

$$C_0 = C_1 + C_2 + \delta \tag{1}$$

线性部分的蛋白质含量  $C_1$  可由 PLS 所建立的线性模型预测得到。在建立 PLS 模型前, 通过 RS-GA 法对奶粉光谱进行谱区优化, 并在选择的波段建立遗传偏最小二乘法模型 (GA-PLS)。其遗传算法由 Visual C++ 6.0 编写。

基于主成分分析的人工神经网络模型 (PG-ANN) 来预测奶粉蛋白质含量的非线性部分。在 RS-GA 法选择的波段上进行 PCA, 以主成分的得分矩阵作为 ANN 模型输入层, 以 GA-PLS 预测差值 ( $C_0 - C_{1P}$ , 其中  $C_{1P}$  为 GA-PLS 预测值) 作为输出层建立 PG-ANN 模型。ANN 模型采用收敛速度快和学习速率高的 Levenberg-Marquardt 方法进行优化, 网络输入层与隐层、隐层与输出层节点的传递函数分别采用 Logsig 和 Purelin 函数。本文所用 ANN 算法由 Matlab 7.0 编写。

最后用所建立的 GA-PLS 和 PG-ANN 模型来对 26 个预测集样品进行预测, 结果为

$$\hat{C}_0 = C_{1P} + C_{2P} \tag{2}$$

其中:  $\hat{C}_0$ —GA-PLS+PG-ANN 预测结果,  $C_{1P}$ —GA-PLS 预测值,  $C_{2P}$ —PG-ANN 预测值。

2 结果与讨论

2.1 光谱数据预处理

多元散射校正 (multiplicative scatter correction, MSC) 和一阶导均可有效消除颗粒大小及颗粒不均等产生的基线漂移<sup>[18]</sup>。由于装样时奶粉的密实度和颗粒度的差异, 引起奶粉样品光谱产生漂移, 采用美国 Thermo Nicolet 公司 TQ 7.2 智能软件对奶粉原始光谱进行 MSC 处理以消除光谱的基线漂移。

2.2 GA 波段优化

将整个光谱区间分为 30 个子区间, 前 29 个子区间有 51 个数据点, 最后一个区间有 78 个数据点。遗传算法的参数设定为: 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。在 GA 进化过程中, 当最佳个体适应度函数  $R/(1+RMSEP)$  不再增加时, 就可以得出最优染色体组合。

经 GA 运算后的波段选择结果如图 1 所示, 其中 1 表示其代表的段被选中, 0 则未被选中。最终合并相邻的选中波段, 得到建模波段为: 3 999.64~4 192.49 cm<sup>-1</sup>, 4 983.15~5 176.01 cm<sup>-1</sup>, 5 573.26~5 766.12 cm<sup>-1</sup>, 6 360.08~6 552.93 cm<sup>-1</sup>, 7 737.00~7 929.86, 8 327.12~8 519.97 cm<sup>-1</sup>, 8 720.52~9 110.08 cm<sup>-1</sup>, 9 310.64~9 700.19 cm<sup>-1</sup>; 经 GA 波段选择后, 其数据点由原来的 1 557 个减少到了 510 个。

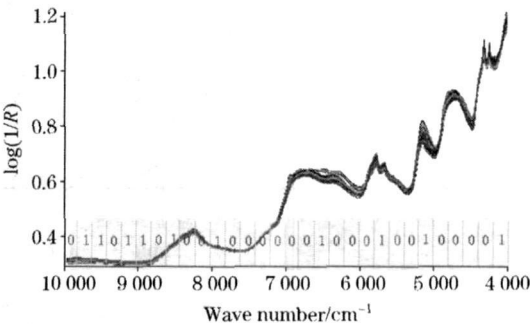


Fig 1 Result of GA region selection after MSC

2.3 PCA 压缩主成分数

由于 GA 波段选择后, 共有 510 个数据点, 不宜直接把这些光谱数据点输入神经网络建立模型, 可通过 PCA 在 RS-GA 法选择的波段上提取特征变量, 作为 ANN 模型的输入节点数。经 PCA 计算后, 其各成分的累计贡献率如图 2, 前 12 个主成分的累计贡献率已达到 99.05%, 表明这 12 个新变量能代表绝大部分原变量包含的信息。因此, 将新提取的 12 个主成分得分矩阵做为 ANN 模型的输入。

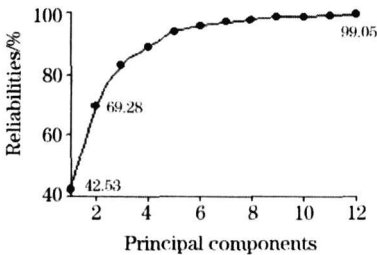


Fig 2 PCs and reliabilities

2.4 模型的建立

在 GA 选择的波段 (510 个数据点), 利用 PLS 方法建立 GA-PLS 模型, 并利用所建模型对 26 个预测集样品进行预测, 其预测结果  $C_{1P}$  如表 2 所示; 为考察 GA 算法优化波段可行性, 在全谱范围内利用 PLS 法建立了全谱 PLS 模型 (Fr-PLS), 建模参数及结果如表 3 所示。

ANN 模型的隐含层神经元经过多次训练, 最终确定最佳的节点数为 4。由于 ANN 模型只输出蛋白质的预测结

果, 故其输出层节点数为 1。因此, 奶粉蛋白质 NIR 模型的最佳网络结构为(12, 4, 1)。设定 ANN 模型的系统允许误差为 0.01, 训练迭代次数为 1 000, 最小训练速率为 0.1。采用 1.4 所述方法建立 PG-ANN 模型, 预测结果  $C_{2P}$  如表 2。

参照式(2)可得 GA-PLS+PG-ANN 模型预测结果  $\hat{C}_0$ , 如表 2。根据统计方法计算 GA-PLS+PG-ANN 模型的建模相关系数  $R$ , 建模标准偏差 RMSEC, 预测相关系数  $r$  和预测标准偏差 RMSEP(表 3)。

Table 2 Statistic results of the validation set

| N $\alpha$ | $C_0$ | $C_{1P}$ | $C_{2P}$ | $\hat{C}_0$ | N $\alpha$ | $C_0$ | $C_{1P}$ | $C_{2P}$ | $\hat{C}_0$ |
|------------|-------|----------|----------|-------------|------------|-------|----------|----------|-------------|
| 1          | 17.2  | 17.66    | -0.37    | 17.29       | 14         | 13.5  | 13.17    | 0.32     | 13.49       |
| 2          | 12.1  | 12.46    | -0.43    | 12.03       | 15         | 11.2  | 11.80    | 0.09     | 11.89       |
| 3          | 18.5  | 18.56    | -0.04    | 18.52       | 16         | 13.8  | 13.43    | 0.14     | 13.57       |
| 4          | 15.4  | 16.01    | -0.24    | 15.77       | 17         | 13.4  | 13.38    | 0.10     | 13.48       |
| 5          | 18.0  | 17.29    | 0.60     | 17.89       | 18         | 11.6  | 11.97    | -0.27    | 11.70       |
| 6          | 15.1  | 15.13    | -0.04    | 15.09       | 19         | 12.7  | 12.94    | 0.02     | 12.96       |
| 7          | 14.1  | 14.07    | 0.05     | 14.12       | 20         | 14.7  | 14.76    | 0.03     | 14.79       |
| 8          | 19.6  | 18.87    | 0.45     | 19.32       | 21         | 12.4  | 12.71    | 0.12     | 12.83       |
| 9          | 13.2  | 13.19    | -0.03    | 13.16       | 22         | 13.2  | 13.15    | 0.03     | 13.18       |
| 10         | 17.3  | 16.31    | 0.96     | 17.27       | 23         | 17.7  | 17.29    | 0.22     | 17.51       |
| 11         | 19.2  | 20.01    | -0.54    | 19.47       | 24         | 16.8  | 16.52    | 0.25     | 16.77       |
| 12         | 15.7  | 15.12    | 0.04     | 15.16       | 25         | 18.8  | 18.91    | -0.23    | 18.68       |
| 13         | 13.0  | 13.15    | 0.02     | 13.17       | 26         | 16.0  | 15.70    | 0.26     | 15.96       |

Table 3 Statistic results of the calibration results

| Method        | Pretreatment | Factors/ BP structure | Calibration |       | Validation |       |
|---------------|--------------|-----------------------|-------------|-------|------------|-------|
|               |              |                       | $R$         | RMSEC | $r$        | RMSEP |
| Fr-PLS        | MSC+1D*      | 9/-                   | 0.986       | 0.433 | 0.978      | 0.511 |
| GA-PLS        | MSC+1D       | 11/-                  | 0.990       | 0.369 | 0.980      | 0.440 |
| GA-PLS+PG-ANN | MSC+1D+PCA   | 11/12-4-1             | 0.996       | 0.185 | 0.995      | 0.235 |

\* 1D: First order Derivative

表 3 结果表明, GA-PLS 模型优于 Fr-PLS 模型, RMSEC 分别为 0.433 和 0.369, RMSEP 分别为 0.511 和 0.440。GA-PLS 模型的建模数据点由全谱的 1 557 减少到 510, 通过 RS-GA 波段选择后不仅减少了建模数据点, 且克服了光谱数据与成分含量多重共线性问题, 提高了模型的预测精度。

GA-PLS+PG-ANN 模型相对 GA-PLS 模型, 考虑了非线性部分。其模型稳健性和预测精度都有很大提高, 其  $R$  和  $r$  分别由 0.990 和 0.980 提高到 0.996 和 0.995, RMSEC 和 RMSEP 由 0.369 和 0.440 降到 0.185 和 0.235(表 3)。通过 RS-GA 波段选择后, 所选波段还存在非线性信息, 将非线性部分加入到线性部分后, 其模型的预测精度可以得到明显的提高。

3 结 论

本文引入了线性和非线性建模方法联合预测蛋白质含量

的 NIR 分析方法。采用 RS-GA 法优化光谱数据, 建立 GA-PLS 模型预测其线性部分; 在 RS-GA 基础上进行 PCA, 建立 PG-ANN 模型预测非线性部分; 以最终模型的 RMSEP 作为模型的评价指标。

采用所建模型对 26 个预测集样品进行预测, 结果显示: Fr-PLS 模型、GA-PLS 模型和 GA-PLS+PG-ANN 模型的 RMSEP 分别为: 0.511, 0.440 和 0.235, 预测精度逐步提高。表明 RS-GA 可有效减少建模数据量, 提高预测精度; 且线性和非线性建模方法联合使用, 可显著提高模型预测精度, 证明了该方法的可行性, 同时为其它复杂体系模型精度的提高提供借鉴。

参 考 文 献

[ 1 ] WU Jing-zhu, WANG Yi-ming, ZHANG Xiao-chao, et al(吴静珠, 王一鸣, 张小超, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2007, 27(9): 1735.  
[ 2 ] WEI Shao-hua, YANG Li, WU Jian-guo(魏少华, 杨莉, 吴建国). Journal of Dairy Science and Technology(乳品科学与技术), 2003,

- (2): 57.
- [ 3 ] ZHENG Lǐ mǐn, ZHANG Lǐ dǎ, GUO Huǐ yuán, et al(郑丽敏, 张录达, 郭慧媛, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2007, 27 (11): 2224.
- [ 4 ] PI Fǔ wéi, WANG Yǎn lǐ ng, LU Chǎo, et al(皮付伟, 王燕玲, 鲁超, 等). Modern Scientific Instruments(现代科学仪器), 2006, (4): 34.
- [ 5 ] LI Yǎn xiǎo, ZOU Xiǎo bō, DONG Yǐ ng(李艳肖, 邹小波, 董英). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2007, 27 (10): 2001.
- [ 6 ] Zhang W, Xiang Bingren. Analytica Chimica Acta, 2007, 585(1): 55.
- [ 7 ] GU Xiǎo yu, XU Kě xīn, WANG Yǎn(谷筱玉, 徐可欣, 汪瞰). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26 (9): 1618.
- [ 8 ] Ying Yibin, Liu Yande. Journal of Food Engineering, 2008, 84: 206.
- [ 9 ] LIU Huǐ jūn, Lǚ Jīn, Lǐ N Mīn, et al(刘辉军, 吕进, 林敏, 等). Journal of Instrumental Analysis(分析测试学报), 2006, 26(5): 679.
- [ 10 ] Durand A, Devos O, Ruckebusch C, et al. Analytica Chimica Acta, 2007, 595: 72.
- [ 11 ] LI Xiǎo lǐ, HÉ Yǒng, QIU Zhēng jūn(李晓丽, 何勇, 裘正军). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2007, 27 (2): 279.
- [ 12 ] Shao Yongni, He Yong, Feng Shuijuan. Food Research International, 2007, 40: 835.
- [ 13 ] Roman M B, Ravilya Z S, Ekaterina I L. Chemometrics and Intelligent Laboratory Systems, 2007, 88: 183.
- [ 14 ] Li Xiaoli, He Yong, Fang Hui. Journal of Food Engineering, 2007, 81: 357.
- [ 15 ] Lǚ Huǐ yī ng, RĒN Yǔ lín, LIU M í ng yā ng(吕慧英, 任玉林, 刘名扬). Chemical Journal of Chinese Universities(高等学校化学学报), 2007, 28(5): 847.
- [ 16 ] SHAO Yǒ ng nǐ, HÉ Yǒ ng(邵咏妮, 何勇). Journal of Infrared Millimeter Waves(红外与毫米波学报), 2006, 25(6): 478.
- [ 17 ] CHEN Jiān, CHEN Xiǎo, Lǐ Wēi, et al(陈建, 陈晓, 李伟, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2008, 28(8): 1806.
- [ 18 ] CHU Xiǎo lǐ, YU ĀN Hō ng fū, LU Wǎ n zhēn(褚小立, 袁洪福, 陆婉珍). Progress in Chemistry(化学进展), 2004, 16(4): 528.

## Improving the Prediction Model of Protein in Milk Powder Using GA-PLS Combined with PG-ANN Arithmetic

SUN Qian, WANG Jia hua, HAN Dong hai\*

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

**Abstract** The present paper presents a new NIR analysis method with partial least square regression (PLS) and artificial neural network (ANN) to improve the prediction precision of the protein model for milk powder. First, an efficient method named region selecting by genetic algorithms (RS-GA) was used to select the calibration region, and then the GA-PLS model was made to predict the linear part of the protein content in milk powder. And then in the region selected by RS-GA method, principal component analysis (PCA) was calculated. The principal components were taken as the input of ANN model. The remnant values by subtracting the standard values and the GA-PLS validation values were regarded as the output of ANN. The ANN model was made to predict the nonlinear part of the protein content. The final result of the model was the addition of the two model's validation values, and the root mean squared error of prediction (RMSEP) was used to estimate the mixed model. A full region PLS model (Fr-PLS) was also made, and the RMSEP of the Fr-PLS, GA-PLS and GA-PLS+PG-ANN model was 0.511, 0.440 and 0.235, respectively. The results show that the prediction precision of the protein model for milk powder was largely improved when adding the nonlinear part in the NIR model, and this method can also be used for other complex material to improve the prediction precision.

**Keywords** NIRS; GA-PLS; PG-ANN; Model accuracy; Milk powder; Protein

(Received Mar. 26, 2008; accepted Jun. 28, 2008)

\* Corresponding author