文章编号：0253-2239(2009)10-2800-04

遗传算法在鱼粉中肉骨粉近红外光谱检测中的应用

湛小梅[1,2] 韩鲁佳[1,2] 刘 贤[1,2] 杨增玲[1,2]

([1] 中国农业大学工学院，北京 100083；[2] 动物营养学国家重点实验室，北京 100083)

摘要 为了研究近红外光谱模型的优化方法，提高模型的精度，利用遗传算法对 64 个掺加了肉骨粉的鱼粉样品近红外光谱进行变量筛选，采用偏最小二乘法回归建模，并用 21 个样品进行外部验证。遗传算法共选取 310 个波长变量，相对于全谱的 1556 个变量减少了 80%，与全谱范围的偏最小二乘法相比，交互验证相关系数(R_{CV})从 0.80 提高到 0.90，交互验证均方根误差从 5.22%降低到 3.62%，预测相关系数(R_V)从 0.91 提高到 0.96，预测均方根误差从 3.85%降低到 2.95%，模型的稳健性和预测精度都显著提高。试验结果表明遗传算法可以改善近红外光谱法预测鱼粉中肉骨粉含量的效果。

关键词 测量；近红外光谱学；遗传算法；鱼粉；肉骨粉

中图分类号 O657.33 **文献标识码** A **doi**：10.3788/AOS20092910.2800

Genetic Algorithm Used for Predicting Meat and Bone Meal Content in Fishmeal by Near Infrared Spectroscopy

Zhan Xiaomei[1,2] Han Lujia[1,2] Liu Xian[1,2] Yang Zengling[1,2]

([1] *College of Engineering, China Agricultural University, Beijing* 100083, *China*
[2] *State Key Laboratory of Animal Nutrition, Beijing* 100083, *China*)

Abstract For the purpose of optimizing near infrared spectroscopy model, and improving the prediction result, Genetic Algorithm (GA) was used to select wavelength variables of near infrared spectroscopy for fishmeal adulterated with meat and bone meal. 310 wavelengths are selted in genetic algorithm. By contrast with all wavelengths based partial least squares(PLS), GA based PLS reduced 80% of the wavelengths, and gained much better cross validation and prediction results. Related coefficient of cross-validation R_{CV} was improved from 0.80 to 0.90, while the value of root mean square error of cross- validation (RMSECV) was reduced from 5.22% to 3.62%. The related coefficient of prediction R_V was improved from 0.91 to 0.96, while the value of root mean square error of prediction (RMSEP) was reduced from 3.85% to 2.95%. GA improved the robustness and predictability of the model. It's indicated that GA was an effective method for variable selection and could improve the prediction result of the meat and bone meal content in fishmeal by near infrared spectroscopy.

Key words measurement; near infared spectrocscpy; genetic algorithm; fishmeal; meat and bone meal

1 引 言

近红外光谱(NIRS)分析技术是国内外发展较快的一种新型定性、定量分析技术。由于其具有快速、无损、无污染等特点，已广泛应用于农业、食品、医药、石油等行业[1~4]。偏最小二乘法(PLS)是 NIR 建模的常用方法，传统观点认为 PLS 具有较强的抗干扰能力，可全波长参与多元校正模型的建立[5]。但是随着对 PLS 方法的深入研究和应用，发

收稿日期：2008-11-20；**收到修改稿日期**：2008-12-30

基金项目：国家自然科学基金(30571074)，十一五国家科技支撑计划子课题(2006BAD12B03-03)和动物营养学国家重点实验室自主研发课题资助

作者简介：湛小梅(1982—)，女，硕士研究生，主要从事近红外检测方面的研究。E-mail：chengyuanhel@cau.edu.cn

导师简介：韩鲁佳(1964—)，女，教授，博士生导师，主要从事生物质资源开发与利用等方面的研究。E-mail：hanlj@cau.edu.cn

现通过特定方法筛选变量(特征波长或波长区间)有可能得到更好的定量校正模型[6]。变量筛选一方面可以简化模型,更主要的是由于不相关变量的剔除,可以得到预测能力强、稳健性好的校正模型[7]。变量筛选方法主要有相关系数法、方差分析法、逐步回归法、无信息变量的消除法(UVE)、间隔偏最小二乘法(iPLS)、遗传算法(GA) 等[8]。

遗传算法是应用较广泛的一种变量筛选方法[9~14],最初是由 Holland 于 1975 年提出的,它借鉴生物界自然选择和遗传机制,利用选择、交换和突变等算子的操作,随着不断的遗传迭代,使目标函数值较优的变量被保留,较差的变量被淘汰,最终达到最优结果[15,16]。褚小立等[9]利用遗传算法对 NIR 测定石油产品有关组成的波长变量进行筛选,结果表明,通过遗传算法选取波长在简化 PLS 模型的同时也增强了所建立模型的预测能力,尤其适用于单纯 PLS 较难校正关联的体系。R. Leardi 等[10]利用傅里叶变换红外光谱法预测聚合物膜中的添加剂浓度,利用 GA 优选 PLS 回归的变量,结果表明遗传算法选择变量的预测效果与专家选择变量的预测效果相当,说明遗传算法可以可靠的选择变量,并且不需要经验。李艳肖等[13]采用 GA 筛选变量,PLS 建立近红外模型,预测苹果的糖度。光谱被划分为 40 个区间,遗传算法选取了 5 个区间,得到的 R_C 和 RMSECV 分别为 0.962 和 0.335,R_V 和 RMSEP 分别为 0.932 和 0.384,与全谱模型相比,遗传算法提高了模型的预测能力,简化了模型。

目前遗传算法已经广泛用于近红外光谱分析研究中,并且可以有效的提高近红外分析模型的精度,但还没有人用于动物饲料来源的鉴别,把遗传算法引入鱼粉中肉骨粉含量的检测,以提高近红外光谱法检测鱼粉中肉骨粉含量的精度。

2　材料与方法

2.1　样品的收集与制备

共收集 76 个纯鱼粉样品和 7 个肉骨粉样品。样品均采用旋风磨(型号 ZM100,Retsch GmbH & Co. KG,德国)粉碎,过 1 mm 筛。从纯鱼粉和肉骨粉的所有 532 个独立组合中随机选取 90 个,分别配制成肉骨粉含量为 1%～30%的鱼粉样品,含量梯度为 1%,则每个含量有 3 个不同的样品,样品配制时采用电子天平称量。每个样品充分混合均匀,密封,置于冰柜中冷藏保存。

2.2　近红外光谱的采集

光谱采集所用仪器为傅里叶变换型近红外光谱仪(Nicolet ANTARIS),光源为卤钨灯,检测器为铟镓砷(InGaAs),参比为标准陶瓷片。

光谱采集的参数为:光谱范围 10000～4000 cm^{-1}(1000～2500 nm),附件选用漫反射积分球,旋转式石英样品池。仪器分辨率选用 8 cm^{-1},增益为 2 cm^{-1},扫描 32 次。

光谱采集前先取出冷藏样品,在室温条件下放置两天,使其温度与室温平衡。为了确保装样的一致性,统一按装满样品、刮平、压紧的步骤进行。以随机顺序扫描样品,消除因扫描次序引起的偏差。每个样品扫描三次,用平均光谱作为样品的光谱,降低随机误差。

2.3　异常样品的剔除与样品集的划分

异常样品的判断采用两个指标[17]:学生残差显著异常,即化学含量值显著异常;杠杆值和学生残差都较异常,即光谱和化学含量值均异常。以杠杆值为横坐标,学生残差为纵坐标作图,以杠杆值平均值的 2 倍为阈值,大于此阈值的即为光谱值异常样品;同样以学生残差平均值的 3 倍为阈值,大于此阈值的为化学值异常样品[18]。判定光谱值和学生残差都异常的样品为异常样品,予以剔除。而学生残差是否显著异常则用 F 检验判断[17],其计算方法为

$$F_i = \frac{|\delta_i|}{\frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n}|\delta_j|} = \frac{(n-1)|\delta_i|}{\sum_{\substack{j=1\\ j\neq i}}^{n}|\delta_j|}$$

其中 δ_i 为第 i 个样品的浓度残差,n 为样品数。根据 $F_i \geqslant F_0$ 判断出学生残差显著异常的样品,F_0 取 $F_\alpha(1,n-1)$(α 为显著性水平,通常取 0.01) 或者设定为某一硬阈值(通常取 1～5),本试验中设定 $F_0=4$。

样品集的划分采用 Kennard～Stone 方法,选取 3/4 的样品作为定标集,剩余 1/4 的样品作为验证集。

2.4　遗传算法的参数设定

遗传算法的控制参数设定为:种群大小 64,窗口宽度 10 个波长,两点交叉方式,适应度比例方法选择染色体,基本位变异算子,变异概率 0.05,适用度函数选用交互验证均方根误差(RMSECV),收敛判据选用遗传迭代次数,设定最大遗传迭代次数 100,在第 70 代时开始收敛。

2.5　变量筛选前后鱼粉中肉骨粉的 NIRS 定量分析模型的建立

模型采用 Unscrambler 9.1 (CAMO A/S,

Trond-heim, Norway)软件建立。在全谱范围内进行线性基线校正(Linear Baseline Correction),再在选定的光谱范围内用 PLS 回归建模。定标过程采用全交互验证(Full Cross Validation),设定最大的主成分因子数为 20,最优的主成分因子数根据交互验证的残差平方和最小的原则选出[19]。

3 结果与讨论

3.1 异常样品的剔除和样品集的划分

所有样品中杠杆值和学生残差都较异常样品 2 个,学生残差显著异常的样品 3 个,因此共剔除异常样品 5 个,剩余 85 个样品参与建模和验证,采用 Kennard-Stone 方法选择其中的 64 个样品(3/4)为定标集,剩余的 21 个(1/4)为外部验证集。

3.2 变量筛选前鱼粉中肉骨粉含量的 NIRS 建模结果

对光谱进行线性基线校正后,在扫描的全部光谱范围内采用 PLS 建模,并用验证集进行外部验证。交互验证的残差平方和随着偏最小二乘因子数的增加而减小,在第 17 个主成分时达到最小,为 27.27%,即前 17 个主成分的累积信息贡献率为 72.73%,因此选取 17 个主成分因子。所选因子数较多,这也许是由于鱼粉和肉骨粉非常相似,鱼粉中肉骨粉含量与近红外光谱之间的关系是非线性的,需要更多的主成分因子来解释光谱信息,Chen Quansheng 等[20]阐述了类似的道理。并且选取少于 17 个主成分因子时,外部验证结果均不如 17 个主成分因子,也说明没有过拟合。模型建立结果见表 1,得到的定标相关系数(R_C)为 0.97,定标均方根误差(RMSEC)为 1.97%,交互验证相关系数(R_{CV})为 0.80,交互验证均方根误差(RMSECV)为 5.22%,预测相关系数(R_V)为 0.91,预测均方根误差(RMSEP)为 3.85%。

3.3 遗传算法筛选变量后鱼粉中肉骨粉含量的 NIRS 建模结果

将线性基线校正后的定标集光谱数据导入 Matlab 中通过遗传算法进行变量筛选,全谱的 1556 个波长变量被分成 156 个区间,遗传算法共选取 31 个波长区间,即 310 个波长变量,变量数减少了 80%。用选择的 310 个光谱数据建立 PLS 模型,交互验证的残差平方和随着偏最小二乘因子数的增加而减小,在第 16 个主成分时达到最小,为 13.13%,即前 17 个主成分的累积信息贡献率为 86.87%,因此选取 16 个主成分因子。变量筛选后的建模结果见表 1。

表 1 遗传算法筛选变量前后的建模效果

Table 1 Calibration performance before and after variable selection using GA, calibration (n=64), validation (n=21)

Wavelength range	Wavelength number	PLS factors	R_c	RMSEC /%	R_{cv}	RMSECV /%	R_v	RMSEP /%
All	1556	17	0.97	1.97	0.80	5.22	0.91	3.85
GA	310	16	0.98	1.73	0.90	3.62	0.96	2.95

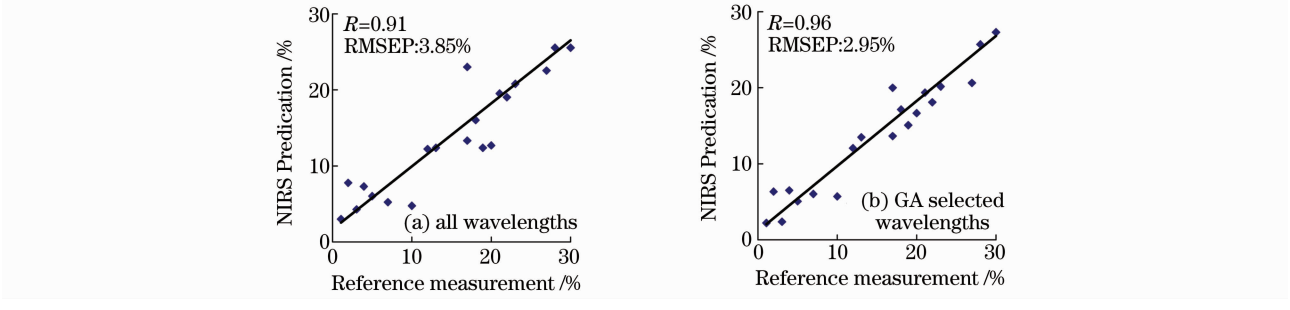

图 1 变量选择前(a)后(b)验证集样品真值与 NIRS 预测值相关关系图

Fig. 1 Correlations between real values and NIRS prediction values of samples in validation set before (a) and after (b) variable selection

由表 1 可知,遗传算法筛选变量后 NIRS 模型的定标相关系数(R_C)为 0.98,定标均方根误差(RMSEC)为 1.73%,交互验证相关系数(R_{CV})为 0.90,交互验证均方根误差(RMSECV)为 3.62%,预测相关系数(R_V)为 0.96,预测均方根误差(RMSEP)为 2.95%,由此可以看出遗传算法筛选变量后交互验证相关系数(R_{CV})和预测相关系数(R_V)明显提高,交互验证均方根误差(RMSECV)和预测均方根误差(RMSEP)明显减小。图 1 为变量筛选前后验证集样品真值与近红外预测值的相关关

系散点图，结合表 1 可以看出，遗传算法筛选变量后模型的预测效果显著提高。试验结果表明遗传算法不仅可以减少变量数，简化模型，而且提高了模型的稳健性和预测精度。

4　结　　论

利用遗传算法对掺加肉骨粉的鱼粉近红外光谱变量进行筛选，用选择的变量进行 PLS 回归建模，预测鱼粉中肉骨粉的含量。遗传算法有效的筛选了掺加肉骨粉的鱼粉近红外光谱变量，使其减少 80%；用遗传算法选择的变量建立的模型与全谱范围建立的模型相比，遗传算法筛选变量后模型的交互验证相关系数(R_{CV})和预测相关系数(R_V)明显提高，交互验证均方根误差(RMSECV)和预测均方根误差(RMSEP)明显减小。试验结果表明遗传算法结合偏最小二乘法用于鱼粉中肉骨粉近红外光谱检测中的变量筛选是有效的，不仅可以减少变量数，简化模型，而且提高了模型的稳健性和预测精度。

参　考　文　献

1 Wu Di, Huang Lingxia, He Yong *et al*.. Visible－near infrared reflection spectroscopy for crop-weed discrimination [J]. *Acta Optica Sinica*, 2008, **28**(8): 1618～1622
吴　迪，黄凌霞，何　勇 等. 作物和杂草叶片的可见-近红外反射光谱特性[J]. 光学学报，2008，**28**(8)：1618～1622

2 Liu Fei, He Yong, Wang Li. Methods for the prediction of sugar content of rice wine using visible-near infrared spectroscopy [J]. *Acta Optica Sinica*, 2007, **27**(11): 2054～2058
刘　飞，何　勇，王　莉. 黄酒糖度预测的可见近红外光谱方法研究[J]. 光学学报，2007，**27**(11)：2054～2058

3 Cen Haiyan, He Yong. Theory and application of near infrared reflectance spectroscopy in determination of food quality [J]. *Trends Food Sci. Tech.*, 2007, **18**(2): 72～83

4 Roggo Y, Pascal C, Lene M *et al*.. A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies [J]. *J. Pharm. Biomed. Anal.*, 2007, **44**(3): 683～700

5 Lu Wanzhen, Yuan Hongfu, Xu Guangtong *et al*.. Modern Near Infrared Spectroscopy Analytical Technology [M]. Beijing: Chinese Petrochemical Industry Press, 2007
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，2007

6 Chen Xiaojing, Wu Di, Yu Jiajia *et al*.. A new choice method of characteristic wavelength of visible/near infrared spectroscopy [J]. *Acta Optica Sinica*, 2008, **28**(11): 2153～2158
陈孝敬，吴　迪，虞佳佳 等. 一种用于可见-近红外光谱特征波长选择的新方法[J]. 光学学报，2008，**28**(11)：2153～2158

7 Nurns D A, Ciurczak E W. Handbook of Near infrared Analysis [M]. New York: Marcel Dekker, 2001

8 Chu Xiaoli, Yuan Hongfu, Lu Wanzhen. Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique [J]. *Progress In Chemistry*, 2004, **16**(4): 528～542
褚小立，袁洪福，陆婉珍. 近红外分析中光谱预处理及波长选择方法进展与应用[J]. 化学进展，2004，**16**(4)：528～542

9 Chu Xiaoli, Yuan Hongfu, Wang Yanbin *et al*.. Variable selection for partial least squares modeling by genetic algorithms [J]. *Chinese Journal of Analytieal Chemistry*, 2001, **29**(4): 437～442
褚小立，袁洪福，王艳斌 等. 遗传算法用于偏最小二乘方法建模中的变量筛选[J]. 分析化学，2001，**29**(4)：437～442

10 Leardi R, Seasholtz M B, Pell R J. Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data [J]. *Anal. Chim. Acta.*, 2002, **461**(2): 189～200

11 Ghasemi J, Niazi A, Leardi R. Genetic-algorithm-based wavelength selection in multicomponent spectrophotometric determination by PLS: application on copper and zinc mixture [J]. *Talanta*, 2003, **59**(2): 311～317

12 Zou Xiaobo, Zhao Jiewen. Methods of characteristic wavelength region and wavelength selection based on genetic algorithm [J]. *Acta Optica Sinica*, 2007, **27**(7): 1316～1321
邹小波，赵杰文. 用遗传算法快速提取近红外光谱特征区域和特征波长[J]. 光学学报，2007，**27**(7)：1316～1321

13 Li Yanxiao, Zou Xiaobo, Dong Ying. Near infrared determination of sugar content in apples based on GA-iPLS [J]. *Spectroscopy and Spectral Analysis*, 2007, **27**(1): 2001～2004
李艳肖，邹小波，董　英. 用遗传区间偏最小二乘法建立苹果糖度近红外光谱模型[J]. 光谱学与光谱分析，2007，**27**(1)：2001～2004

14 Ying Yibin, Liu Yande. Nondestructive measurement of internal quality in pear using genetic algorithms and FT-NIR spectroscopy [J]. *J. Food Eng.*, 2008, **84**(2): 206～213

15 Liu Huijun, Lü Jin, Lin Min. Application of characteristic wavelength selection method in NIR model of green tea based on genetic algorithms [J]. *J. Instrumental Anal.*, 2007, **26**(5): 679～681
刘辉军，吕　进，林　敏. 基于遗传算法的波长选择方法在绿茶近红外光谱分析模型中的应用[J]. 分析测试学报，2007，**26**(5)：679～681

16 Zeng Libo, He Zhiping. Study on the application of genetic algorithm for synchronous selection of wavelength and spectral data pretreatment method in near-infrared spectrometric analysis [J]. *Analytical Instrumentation*, 2006, **3**: 23～26
曾立波，贺志平. 遗传算法在近红外光谱分析波长及预处理方法同步选择中的应用[J]. 分析仪器，2006，3：23～26

17 Zhu Shiping, Wang Yiming, Zhang Xiaochao *et al*.. Region selection method of near infrared spectrum based on genetic algorithm [J]. *Transactions of The Chinese Society of Agricultural Machinery*, 2004, **35**(5): 152～156
祝诗平，王一鸣，张小超 等. 基于遗传算法的近红外光谱谱区选择方法[J]. 农业机械学报，2004，**35**(5)：152～156

18 Niu Zhiyou. The NIRS Analysis of Fish Meal, Concentrate Supplement and MBM Content Inside [D]. Beijing: China Agricultural University, 2005
牛智有. 鱼粉、精料补充料及其中肉骨粉含量的近红外漫反射光谱分析[D]. 北京：中国农业大学，2005

19 Li Minzan. Spectral analysis technology and application [M]. Beijing: Science Press, 2006
李民赞. 光谱分析技术及其应用[M]. 北京：科学出版社，2006

20 Chen Quansheng, Zhao Jiewen, Chaitep Sumpun *et al*.. Simultaneous analysis of main catechins contents in green tea (Camellia sinensis (L.)) by Fourier transform near infrared reflectance (FT-NIR) spectroscopy [J]. *Food Chemistry*, 2009, **113**(4): 1272～1277