人工神经网络NIR定量分析方法及其软件实现*

祝诗平

【摘要】 在Visual C++环境中采用面向对象技术，开发了PCA-MBP-NIR定量分析模型软件。通过40份小麦样品的原始光谱、加噪光谱(信噪比为14 dB)与含水率所建立的PLS-NIR与PCA-MBP-NIR模型，对10份未知小麦样品的原始光谱、加噪光谱分别进行含水率的PLS-NIR与PCA-MBP-NIR预测分析。分析表明，对于含噪声的光谱，与PLS建模相比，使用PCA-MBP-NIR对未知样品预测结果具有更高的相关系数，更低的预测误差标准差。

关键词：农产品 品质检测 近红外光谱分析 主成分分析 人工神经网络

中图分类号：O433.4；S379.1 **文献标识码**：A

NIR Quantitative Analysis Method Based on Artificial Neural Network and Its Software Implementation

Zhu Shiping

(*Southwest University*)

Abstract

An artificial neural network of matrix back propagation (MBP-ANN) combined with principal component analysis (PCA) for near infrared spectroscopy (NIR) quantitative analysis method is presented, and its principles is analyzed. A PCA-MBP-NIR quantitative analysis software system is developed based on object oriented programming technology in environment of Microsoft Visual C++. The PCA-MBP-NIR model and partial least square (PLS) NIR model are built between the moisture and raw spectrum of 40 wheat samples, and the two models are also built for noise spectrum ($r^{max}=14$ dB) in the same way. The moisture of 10 unknown wheat samples are predicted by this model. Results show that, using PCA-MBP-NIR method instead of PLS-NIR for noise spectrum, the correlation coefficient of predictied values and standard values of unknown samples can be increased, and the root mean square deviation (RMSD) can be decreased.

Key words Agricultural product, Quality detection, Near infrared spectroscopy analysis, Principal component analysis, Artificial neural network

引言

近红外光谱分析(near infrared spectroscopy，简称NIR)方法在农产品内部品质检测中的应用越来越多[1]。近红外光谱分析技术的核心就是如何建立更有效的校正模型。

利用化学计量学方法，建立近红外光谱的校正模型，通常可以分为线性校正模型和非线性校正模型2类。各种多元线性校正模型主要有[2]：逐步多元线性回归、主成分回归、偏最小二乘法(partial least square，简称PLS)等方法，其中PLS应用最多。非线性校正算法主要有人工神经网络方法[3](artificial neural network，简称ANN)、非线性PLS、局部权重回归等方法。ANN因为具有抗干扰、抗噪声和强大

收稿日期：2006-04-18

* 国家自然科学基金资助项目(项目编号：30671198)、重庆市科委自然科学基金资助项目(项目编号：CSTC2005BB2211)和重庆市高等学校优秀青年骨干教师资助计划(2005)

祝诗平 西南大学工程技术学院 副教授 博士后(重庆大学)，400716 重庆市

的非线性转换能力，在近红外光谱分析中的应用较多。

本文在 Visual C++环境中实现了主成分分析(principal component analysis，简称 PCA)与基于矩阵的 BP 快速算法(matrix back propagation，简称 MBP)[4]的结合，开发了 PCA-MBP-NIR 定量分析模型软件，并应用到小麦含水率的近红外光谱建模和预测中[5~7]。

1　PCA-MBP-NIR 定量分析方法

对光谱矩阵 $\boldsymbol{A}_{n\times p}$进行主成分分解[6]，即

$$\boldsymbol{A}_{n\times p}=\boldsymbol{S}_{n\times f}\boldsymbol{F}_{f\times p}+\boldsymbol{E}_A \tag{1}$$

式中　n——建模样品数目

p——光谱的波长点数

f——主成分数目

$\boldsymbol{S}_{n\times f}$——主成分得分(Scores)矩阵

$\boldsymbol{F}_{f\times p}$——光谱载荷(Loading)矩阵

$\boldsymbol{E}_A$——光谱残差矩阵

设光谱数据矩阵 $\boldsymbol{A}_{n\times p}$的第 j 列的标准差为 σ_j，主成分得分矩阵 $\boldsymbol{S}_{n\times f}$的第 h 列的标准差为 σ_{S_h}，与 $\boldsymbol{S}$ 第 h 列所对应的特征值为 λ_h，则定义 f 维主成分空间的累积贡献率为

$$Q_f=\frac{\sum_{h=1}^{f}\sigma_{S_h}^2}{\sum_{j=1}^{p}\sigma_j^2}=\frac{\sum_{h=1}^{f}\lambda_h}{\sum_{j=1}^{p}\sigma_j^2} \tag{2}$$

如果 $\boldsymbol{A}_{n\times p}$是标准化的，则 $\sigma_j=1$，式(2)简化为

$$Q_f=\frac{1}{p}\sum_{h=1}^{f}\sigma_{S_h}^2=\frac{1}{p}\sum_{h=1}^{f}\lambda_h \tag{3}$$

显然 $0<Q_f\leqslant 1$，Q_f 就是描述 f 维主成分空间所携带的数据变异信息占原数据总变异信息的百分比。据此，可以预先给出一个要求的 Q_f，从而也就确定了 PCA 分解的维数 f。

PCA-MBP-NIR 定量分析分建模和预测 2 部分。PCA-MBP-NIR 定量分析原理图如图 1 所示。PCA-MBP-NIR 建模时的基本步骤[6]为：①对原始光谱矩阵和原始成分含量矩阵均进行标准化处理，得到光谱 $\boldsymbol{A}_{n\times p}$和成分含量 $\boldsymbol{C}_{n\times m}$，保存相应的均值和方差(这里 m 为成分数目)。②预先设定所要求的 f 维主成分空间的累积贡献率 Q_f，并预设主成分分解初始维数 f_0($f_0<p$)。③对 $\boldsymbol{A}_{n\times p}$进行主成分分解，分解到 f_0 维，得到主成分得分矩阵 $\boldsymbol{S}_{n\times f_0}$和载荷矩阵 $\boldsymbol{F}_{f_0\times p}$。④根据给定的累积贡献率 Q_f 确定最终的主成分数目 f，得到新的得分矩阵 $\boldsymbol{S}_{n\times f}$。⑤将新得分矩阵 $\boldsymbol{S}_{n\times f}$作输入，成分含量矩阵 $\boldsymbol{C}_{n\times m}$作为输出，构造 MBP 神经网络，显然输入层的神经元数目为 f，输出层的神经元数目为 m，确定隐含层神经元数目。⑥构造好 MBP 神经网络并对所有校正集样品进行训练后，保存 MBP 神经网络结构参数和连接权值，建立最终的 PCA-MBP-NIR 定量分析模型。

PCA-MBP-NIR 未知样品预测的具体步骤[7]为：①读取ANN模型文件中原始建模光谱矩阵的

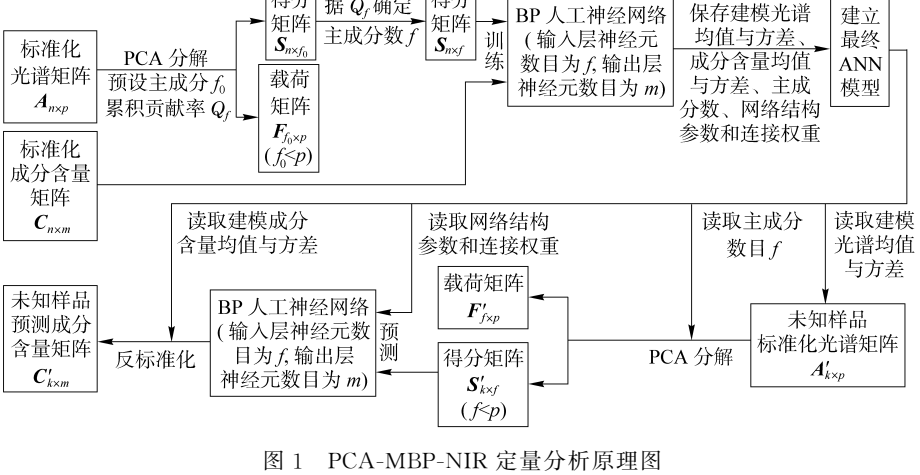

图 1　PCA-MBP-NIR 定量分析原理图

Fig. 1　Principle of PCA-MBP-NIR quantitative analysis method

均值和方差，对未知样品原始光谱矩阵进行标准化处理，得到未知样品标准化光谱矩阵 $\boldsymbol{A}'_{k\times p}$(这里 k 为未知样品数目)。②读取 ANN 模型文件中主成分数目 f，对未知样品标准化光谱矩阵 $\boldsymbol{A}'_{k\times p}$进行主成分分解，得到主成分得分矩阵 $\boldsymbol{S}'_{k\times f}$和载荷矩阵 $\boldsymbol{F}'_{f\times p}$。③读取 MBP 网络结构参数和连接权值，使用预测算法，预测未知样品标准化成分含量矩阵，进行反标准化处理(使用参与建模的原始成分含量矩阵的均值和方差)，得未知样品预测成分含量矩阵$\boldsymbol{C}'_{k\times m}$。

2　PCA-MBP-NIR 定量分析软件

系统开发环境采用 Windows XP 操作系统，开

发工具为 Visual C++ 6.0,实现了 PCA-MBP-NIR 定量分析模型软件。

在近红外光谱分析中,化学计量学算法涉及大量的向量、矩阵运算,本系统在实现时使用 C++面向对象编程技术、类封装技术,对向量、矩阵的大量运算进行封装,形成向量类和矩阵类[8],PCA-MBP-NIR 算法均调用该向量类和矩阵类的函数,大大提高了算法的开发效率,同时程序的可读性也大大增加。

3 PCA-MBP-NIR 定量分析实例

使用 Thermo 公司 GRAMS/AI 软件提供的 50 份小麦样品光谱(近红外长波段 1 000~2 617 nm,波长点数 $p=1\ 011$ 点)和含水率数据(文件 Multi.spc 及 Multi.cfl),作为算法、软件的验证数据,同时对该数据集光谱信号添加最大信噪比为14 dB的白噪声形成加噪光谱。

取所加白噪声的方差(即功率)为原始光谱矩阵中最大吸光度(max$A$)的 1%,则

$$P_n=\frac{\max A}{10^2}=\frac{0.743\ 1}{10^2}=0.007\ 431 \qquad (4)$$

50 份样品的原始光谱信号的最大功率为

$$P_a^{\max}=\max_i(P_a^i)=\max_i\left(\frac{1}{p}\sum_{j=1}^{p}|a_{ij}|^2\right)=0.189\ 2 \qquad (5)$$

式中 a_{ij}——50 份样品的原始光谱矩阵 $\boldsymbol{A}_{50\times 1\,011}$ 的第 i 行第 j 列元素

P_a^i——第 i 份样品的原始光谱信号功率

添加白噪声后,最大的信噪比(signal noise rate,简称 SNR)为

$$r^{\max}=10\lg\frac{P_a^{\max}}{P_n}=10\lg\frac{0.189\ 2}{0.007\ 431}=14\ \text{dB} \qquad (6)$$

原始光谱及加噪光谱如图 2 所示。现将 50 份样品随机选出 40 份作为校正集,余下的 10 份作为预测集。使用开发的 PCA-MBP-NIR 定量分析模型软件对原始光谱与加噪光谱分别进行建模和预测分析[6]。

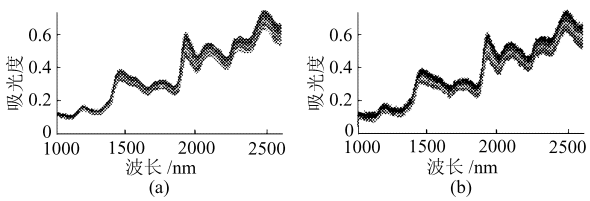

图 2 50 份小麦样品近红外光谱图

Fig. 2 Spectrum of 50 wheat samples

(a) 原始光谱 (b) 加噪光谱

将校正集 40 份小麦样品的原始光谱矩阵(经过标准化)$\boldsymbol{A}_{40\times 1\,011}$进行 PCA 分解(预设 $f_0=12$,主成分累积贡献率 $Q_f=0.99$),从图 3 可以看出,只需取 $f=4$,即通过主成分分解,可以用一个 4 维的主成分空间表示 1 011 维的原始光谱波长区间,而保留了 99%的数据信息。将主成分得分矩阵 $\boldsymbol{S}_{40\times 4}$作为输入,标准化成分含量矩阵 $\boldsymbol{C}_{40\times 1}$(含水率)作为输出,构造 MBP,显然输入层的神经元数目为 4,输出层的神经元数目为 1,经过试验 MBP 神经网络的隐含层取 1 层,该层有 4 个神经元时效果较好,MBP 网络结构简记为 4-4-1。训练神经网络的其他参数如图 4所示。

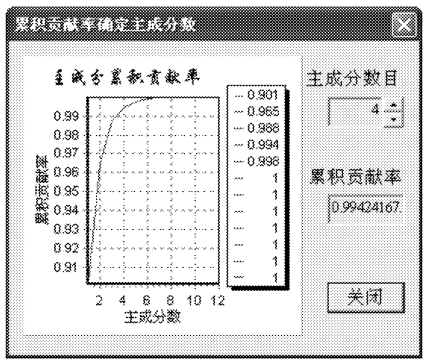

图 3 基于主成分累积贡献率确定 PCA 主成分数目的界面

Fig. 3 Determination the factor numbers of PCA based on cumulated contribution rate

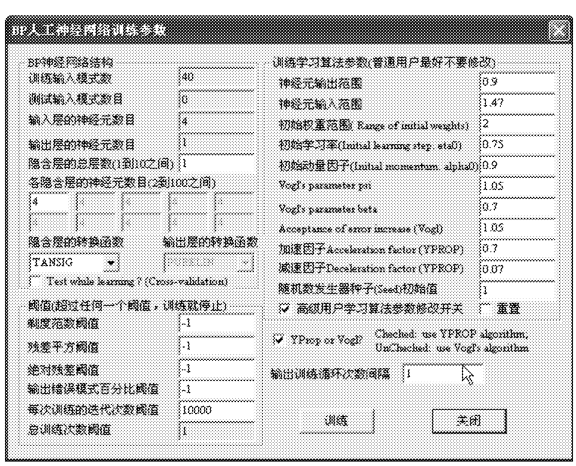

图 4 MBP 神经网络建模训练参数界面

Fig. 4 Training parameters of MBP neural network

PCA-MBP 训练过程如图 5 所示(只显示了前 200 次循环过程),可以看出,经过 80 步循环迭代(学习)后,随训练迭代次数的增加,均方根误差(δ_{RMSD})的减小已经很有限。如果继续增加训练迭代次数,将会出现过拟合现象。

使用校正集 40 份样品原始光谱建立的 PCA-MBP-NIR 定量分析模型,对余下的未参与建模的预测集 10 份样品分别进行预测,预测值与标准值的相关系数 $R^{\text{ANN}}=0.994\ 3$、预测误差标准差(均方差)

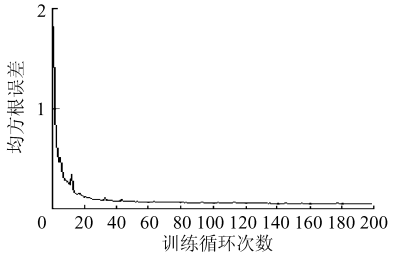

图 5　MBP 训练过程

Fig. 5　Training process of MBP

$\delta_{RMSD}^{ANN}=0.169\ 2$，图 6 显示了预测集 10 份样品使用 PCA-MBP 预测的值与标准值相关性。而使用 PLS 对 40 个样品原始光谱进行建模，对余下的未参与建模的预测集 10 样品分别进行预测，$R^{PLS}=0.996\ 3$，$\delta_{RMSD}^{PLS}=0.147\ 1$，如表 1 所示。对于原始光谱，PCA-ANN 与 PLS 的预测效果相差不大。

进一步，对于加噪光谱，使用与原始光谱相同的处理，可以得到：$R^{PLS}=0.973\ 3$，$\delta_{RMSD}^{PLS}=0.488\ 6$，$R^{ANN}=0.987\ 0$，$\delta_{RMSD}^{ANN}=0.269\ 3$，结果如表 1 所示。

表 1　PCA-MBP-NIR 与 PLS-NIR 模型的预测结果

Tab. 1　Predicted result of PCA-MBP and PLS NIR model

光谱类型	方法	主成分数或 BP 结构	未知样品预测	
			R	δ_{RMSD}
原始光谱	PLS-NIR	4	0.996 3	0.147 1
	PCA-MBP-NIR	4－4－1	0.994 3	0.169 2
加噪光谱	PLS-NIR	4	0.973 3	0.488 6
	PCA-MBP-NIR	4－4－1	0.987 0	0.269 3

从表 1 可以看出，对于具有噪声的光谱，使用 PCA-MBP-NIR 定量分析方法建立近红外光谱定量分析模型，比使用 PLS-NIR 具有更高的相关系数，更低的预测误差标准差。

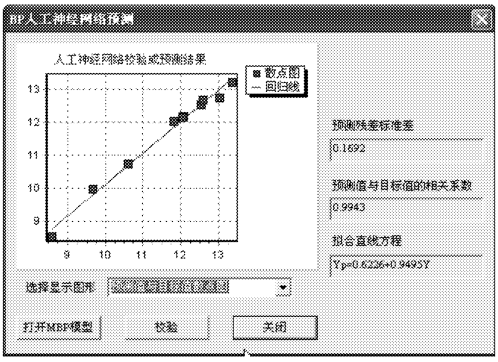

图 6　10 份未知样品 PCA-MBP 预测值与标准值相关性

Fig. 6　Relativity between the concentrations predicted by PCA-MBP-NIR and the standard of the 10 unknown samples

4　结束语

在 Visual C＋＋环境中实现了 PCA 与 MBP 的结合，开发了 PCA-MBP-NIR 定量分析模型软件。

通过对 50 份小麦样品原始光谱及加噪光谱的实例分析表明，对于含噪声的光谱，使用主成分分析进行降维，再用 MBP 神经网络进行非线性建模，比使用 PLS 进行建模，未知样品预测结果具有更高的相关系数，更低的预测误差标准差。

参　考　文　献

1　梁晓艳，吉海彦. 近红外光谱技术在农作物品质分析方面的应用[J]. 中国农学通报，2006，22(1)：366～371.

2　Wold S，Sjöström M. Chemometrics，present and future success[J]. Chemometrics and Intelligent Laboratory Systems，1998，44(1～2)：3～14.

3　李燕，王俊德，顾炳和，等. 人工神经网络及其在光谱分析中的应用[J]. 光谱学与光谱分析，1999，19(6)：844～849.

4　Anguita D，Parodi G，Zunino R. Speed improvement of the back-propagation on current-generation workstations[C]. WCNN '93，Portland，USA，1993：165～168.

5　Blanco，M Coello J，Iturriaga H，et al. NIR calibration in non-linear systems：different PLS approaches and artificial neural networks[J]. Chemometrics and Intelligent Laboratory Systems，2000，50(1)：75～82.

6　祝诗平. 近红外光谱品质检测方法研究[D]. 北京：中国农业大学，2003.

7　祝诗平，王一鸣，张小超. 基于 PCA-MBP 神经网络的 NIR 定量分析方法及其应用[C]//中国农业机械学会成立 40 周年庆典暨 2003 年学术年会论文集，北京，2003.

8　祝诗平，王一鸣，张小超. 农产品近红外光谱品质检测软件系统的设计与实现[J]. 农业工程学报，2003，19(4)：175～179.