http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2016.0459

基于改进极限学习机的光谱定量建模方法

周美灵, 郑德智*, 娄格, 刘峥

(北京航空航天大学 仪器科学与光电工程学院, 北京 100083)

摘 要: 依据近红外光谱(NIR)产生原理,提出了粒子群优化极限学习机(PSO-ELM)算法,运用于小样本氨水浓度定量分析。通过优化极限学习机(ELM)隐藏节点参数,解决了极限学习机由于输入权值和隐含层偏差随机产生的建模结果具有随机性的问题,提高了预测模型的稳定性、精确度和泛化性能。经实验验证,优化后的 PSO-ELM 相比 ELM,模型预测集均方根误差由 0.011 66 减小至 0.003 22,预测集相关系数由 0.995 1 提高至 0.997 9。将优化后的模型预测结果与支持向量机(SVM)、BP 神经网络算法等传统方法的建模结果进行对比,优化后的 PSO-ELM 算法具有较高的精确度和良好的泛化性能,模型预测效果优于传统的定量回归分析算法。

关 键 词: 近红外光谱(NIR); 定量分析; 小样本; 极限学习机(ELM); 粒子群优化(PSO)算法

中图分类号: O433.4;TB96

文献标识码: A **文章编号**: 1001-5965(2017)06-1208-08

20 世纪 80 年代,近红外光谱技术迅速发展起来。由于该技术测试速度快、测试过程中样品无损耗、能够适应多种成分分析,而且具有结果重演性高、环保无污染等优点,逐渐引起了国内外学者的广泛关注,成为现代检测技术中的重要工具[1]。

在近红外光谱技术的发展中,化学计量学的发展与应用起到十分重要的作用。目前,常用的近红外光谱定量回归分析方法有偏最小二乘法(PLS)、支持向量机(SVM)以及 BP 神经网络(BP-ANN)等。PLS 是一种多元数据分析方法,建立在主成分分析和回归基础上,主要适合于数据量少、各变量内部高度线性相关的模型,不适应动态多变量建模[2]。SVM 是基于风险最小化的统计学习理论,可解决标准的优化问题,对小样本、非线性及高维数据具有很强的泛化能力,一经提出,便得到了广泛应用[3-4]。但 SVM 不适用于大规模数据样本,会耗费机器大量内存资源。BP-ANN 是一种按误差逆传播算法训练的多层前馈网络,其以良好的泛化能力、容错能力以及自学习、自适应能力获得了国内外学者广泛关注。但随着应用范围的扩大,BP-ANN 也暴露出不少缺点,如需要人为设置大量网络参数,网络结构模型不单一,容易陷入局部优化,而且参数需迭代调整,训练效率不高,容易出现过拟合现象[5]。

2004 年,南洋理工大学黄广斌等[6]提出一种人工神经网络模型训练新算法——极限学习机(Extreme Learning Machine,ELM),该算法输入层与隐含层之间的连接权值和阈值均随机产生,而且在模型训练过程中无需对参数进行调整,只需对隐含层神经元个数进行设置,便可避免陷入局部的最小值,从而获得唯一最优解。与传统方法相比具有简单易用、学习速度快及泛化性好的特点。此后,不少学者开始运用并不断改进 ELM 算法。Masri 等[7]运用 ELM 对土壤的近红外光谱进

收稿日期: 2016-05-31; **录用日期**: 2016-07-01; **网络出版时间**: 2016-08-25 08:33
网络出版地址: www.cnki.net/kcms/detail/11.2625.V.20160825.0833.001.html
基金项目: 国家科技支撑计划(2014BAF08B01)
* **通讯作者**: E-mail: zhengdezhi@buaa.edu.cn

引用格式: *周美灵, 郑德智, 娄格, 等. 基于改进极限学习机的光谱定量建模方法[J]. 北京航空航天大学学报, 2017, 43(6): 1208-1215. ZHOU M L, ZHENG D Z, LOU G, et al. Quantitative spectral modeling method based on improved extreme learning machine[J]. Journal of Beijing University of Aeronautics and Astronautics, 2017, 43(6): 1208-1215 (in Chinese).*

http://bhxb.buaa.edu.cn

行建模分析,并用非洲土壤属性预测数据库验证模型,获得了较好的预测性能。Guo 等[8]采用偏最小二乘支持向量机(LS-SVM)和 ELM 建立近红外光谱桃品种的聚类分析模型,预测准确率高达 100%。Huang 和 Chen[9]先后提出了核极限学习机(K-ELM)、增强增量极限学习机(EI-ELM)和凸增量的极限学习机(Cl-ELM)等改进极限学习机算法。Jin 等[10]提出了用于交叉特性学习的多任务聚类 ELM,实现了人脸特性识别。Alexandre 等[11]采用遗传算法优化 ELM 对交通噪声进行分类,改进后算法分类准确性由改进前的 74.83% 提高至 93.74%。陈媛媛等[12]采用遗传算法改进 ELM,用于 NO 与 NO_2 混合气体的定量分析,实验结果表明,改进后的模型能够快速、准确地预测混合气体各组分浓度。de Oliveira 和 Ludermir[13]采用鱼群算法改进 ELM 用于聚类分析,经过与标准数据库对比实验验证,改进后的算法具有更好的准确性和泛化性能。

本文基于小样本氨水近红外光谱数据,采用粒子群优化极限学习机(PSO-ELM)算法建立了氨水浓度定量分析模型,优化了 ELM 输入层与隐含层之间的连接权值和偏差,对比了 BP-ANN、SVM、ELM 以及 PSO-ELM 算法模型的精度和泛化性能。通过实验验证,改进后的 PSO-ELM 算法氨水浓度定量分析模型的稳定性、精度和泛化性能较传统方法均有显著提高。

1　近红外光谱理论

近红外光是介于可见光和中红外之间的电磁辐射波,美国材料检测协会将近红外光谱区定义为 780 ~ 2 526 nm 的区域。近红外光谱区主要反映的是有机分子中含氢基团(C—H、O—H、N—H)振动的合频和各级倍频的吸收谱线信息[14]。通过扫描样品的近红外光谱,可以得到样品分子含氢基团的特征信息。

图 1 所示为氨分子(NH_3)振动模式。NH_3 为 C_{3v} 对称性分子,具有对称伸缩振动、非对称伸缩振动、对称弯曲振动以及反对称弯曲振动 4 种振动形式,会产生 4 个基频特征谱带。气态氨一级倍频和二级倍频均为尖吸收峰,分别在 6 520 cm^{-1}(1 534 nm)和 9 650 cm^{-1}(1 046 nm)附近。NH_3 的倍频、组合频吸收峰较多,通常选择 6 609 cm^{-1}(1 513 nm)、5 084 cm^{-1}(1 967 nm)和 4 417 cm^{-1}(2 264 nm)处吸收强度较大的吸收峰建模[14]。

氨呈弱碱性,在水中主要以 NH_3 形式存在。水是近红外定量分析的主要干扰因素,在近红外谱区有较强的吸收峰,而且谱带较宽[15]。如图 2 所示为实际测得氨水近红外光谱图,氨水样本光谱在 4 484 cm^{-1}(2 203 nm)、4 892 cm^{-1}(2 044 nm)、5 962 cm^{-1}(1 706 nm)和 6 024 cm^{-1}(1 660 nm)出现较强吸收峰。由于近红外光谱峰位容易受到诱导效应、共轭效应、氢键效应、空间效应以及外部条件等多种因素的影响,氨水近红外光谱相比气态氨峰位有较大移动。

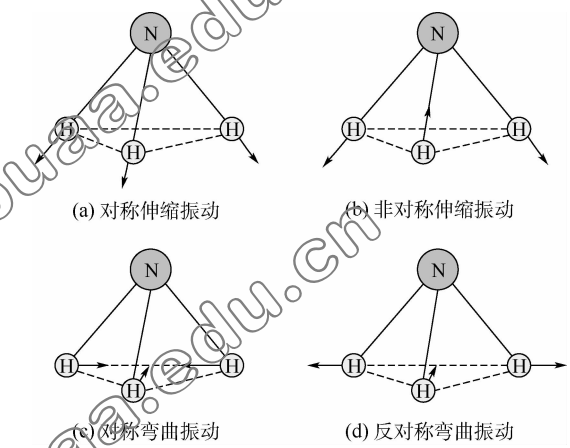


图 1　氨分子振动模式

Fig. 1　Vibration modes of ammonia molecule

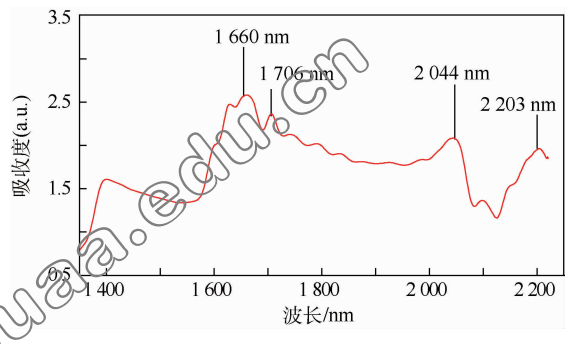


图 2　氨水近红外吸收光谱

Fig. 2　Near infrared absorption spectrum of ammonia

2　算法理论

2.1　极限学习机(ELM)

ELM 是南洋理工大学黄广斌教授提出的一种人工神经网络模型训练新算法,该算法针对单隐含层前馈神经网络(Single-hidden Layer Feed-forward Neural Network,SLFN),改进了传统训练方法训练速度慢、容易陷入局部极小值以及学习率选择敏感等缺点[6]。

图 3 所示为典型的单隐含层前馈神经网络结构图。图中:$x_1, x_2, \cdots, x_m$ 为输入;$y_1, y_2, \cdots, y_m$ 为输出;$o_1, o_2, \cdots, o_m$ 为隐含层节点;$\boldsymbol{w} = [w_{ij}]$ 为输入层与隐含层的连接权值;$\boldsymbol{\beta} = [\beta_{jk}]$ 为隐含层与输出层的连接权值。

http://bhxb.buaa.edu.cn

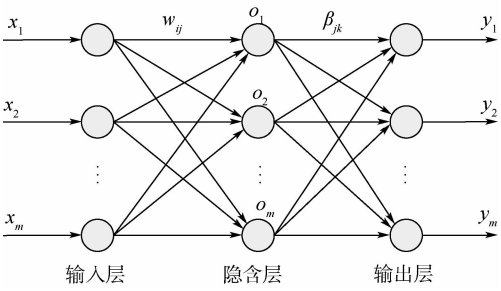


图 3 典型单隐含层前馈神经网络结构

Fig. 3 A typical single-hidden layer feedforward neural network structure

设样本数为 Q,训练集的输入矩阵为 $\boldsymbol{X}$,输出矩阵为 $\boldsymbol{Y}$:

$$\begin{cases}\boldsymbol{X}=\begin{bmatrix}x_{11} & \cdots & x_{1Q}\\ \vdots & & \vdots\\ x_{n1} & \cdots & x_{nQ}\end{bmatrix}_{n\times Q}\\ \boldsymbol{Y}=\begin{bmatrix}y_{11} & \cdots & y_{1Q}\\ \vdots & & \vdots\\ y_{m1} & \cdots & y_{mQ}\end{bmatrix}_{m\times Q}\end{cases} \tag{1}$$

设隐含层神经元的激活函数为 $g(x)$,则网络输出 $\boldsymbol{T}$ 为

$$\boldsymbol{T}=[\boldsymbol{t}_1,\boldsymbol{t}_2\cdots,\boldsymbol{t}_Q]_{m\times Q}$$

$$\boldsymbol{t}_j=\begin{bmatrix}t_{1j}\\ t_{2j}\\ \vdots\\ t_{mj}\end{bmatrix}_{m\times 1}=\begin{bmatrix}\sum_{i=1}^{l}\beta_{i1}g(\boldsymbol{w}_i\boldsymbol{x}_j+b_i)\\ \sum_{i=1}^{l}\beta_{i2}g(\boldsymbol{w}_i\boldsymbol{x}_j+b_i)\\ \vdots\\ \sum_{i=1}^{l}\beta_{im}g(\boldsymbol{w}_i\boldsymbol{x}_j+b_i)\end{bmatrix}_{m\times 1}$$

$$j=1,2,\cdots,Q \tag{2}$$

式中:$\boldsymbol{w}_i=[w_{i1},w_{i2},\cdots,w_{in}]$;$\boldsymbol{x}_j=[x_{1j},x_{2j},\cdots,x_{nj}]^{\mathrm{T}}$。

则式(2)可表示为

$$\boldsymbol{H}\boldsymbol{\beta}=\boldsymbol{T}^{\mathrm{T}} \tag{3}$$

式中:$\boldsymbol{T}^{\mathrm{T}}$ 为矩阵 $\boldsymbol{T}$ 的转置;$\boldsymbol{H}$ 为神经网络的隐含层输出矩阵,其具体形式可表示为

$$\boldsymbol{H}(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_l,b_1,b_2,\cdots,b_l,\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_Q)=\begin{bmatrix}g(\boldsymbol{w}_1\boldsymbol{x}_1+b_1) & g(\boldsymbol{w}_2\boldsymbol{x}_1+b_2)\cdots g(\boldsymbol{w}_l\boldsymbol{x}_1+b_l)\\ g(\boldsymbol{w}_1\boldsymbol{x}_2+b_1) & g(\boldsymbol{w}_2\boldsymbol{x}_2+b_2)\cdots g(\boldsymbol{w}_l\boldsymbol{x}_2+b_l)\\ \vdots & \vdots\\ g(\boldsymbol{w}_1\boldsymbol{x}_Q+b_1) & g(\boldsymbol{w}_2\boldsymbol{x}_Q+b_2)\cdots g(\boldsymbol{w}_l\boldsymbol{x}_Q+b_l)\end{bmatrix}_{Q\times l} \tag{4}$$

在文献[16]算法的基础上,Huang 等[6]证明:在任意区间范围内,给定无限可微的激活函数 $g:\mathbf{R}\to\mathbf{R}$ 和任意小误差 $\varepsilon>0$,则总存在一个含有 $K(K\leqslant Q)$ 个隐含层神经元的单隐含层神经网络模型,在任意赋值 $\boldsymbol{w}_i\in\mathbf{R}^n$ 和 $b_i\in\mathbf{R}$ 的情况下,有 $\|\boldsymbol{H}_{N\times M}\boldsymbol{\beta}_{M\times N}-\boldsymbol{T}^{\mathrm{T}}\|<\varepsilon$。

因此,当存在无限可微的激活函数 $g(x)$ 时,单隐含层神经网络模型中的参数都无需调整,只需在训练前,通过随机的方法确定参数 w 和 b,则求下列方程组的最小二乘解便可得到隐含层与输出层间的连接权值 $\boldsymbol{\beta}$:

$$\min_{\boldsymbol{\beta}}\|\boldsymbol{H}\boldsymbol{\beta}-\boldsymbol{T}^{\mathrm{T}}\| \tag{5}$$

其解为

$$\hat{\boldsymbol{\beta}}=\boldsymbol{H}^{+}\boldsymbol{T}^{\mathrm{T}} \tag{6}$$

式中:$\boldsymbol{H}^{+}$ 为隐含层输出矩阵 $\boldsymbol{H}$ 的 Moore-Penrose 广义逆。

ELM 算法实现分为以下 3 个步骤[17]:

1)设置隐含层神经元个数,输入层与隐含层间的连接权值 $\boldsymbol{w}$ 和隐含层偏差 b 可以通过随机数产生。

2)选择无限可微的隐含层神经元激活函数,计算输出矩阵 $\boldsymbol{H}$。

3)通过求解 $\hat{\boldsymbol{\beta}}$:$\hat{\boldsymbol{\beta}}=\boldsymbol{H}^{+}\boldsymbol{T}^{\mathrm{T}}$ 计算输出层权值。

2.2 粒子群优化算法

粒子群优化(Particle Swarm Optimization, PSO)算法由 Kennedy 和 Eberhart[18]提出。该算法从鸟群捕食特征中得到启发。初始化一个种群,种群中的个体称为粒子。种群中的每个粒子都具有 3 个特征,即速度、位置和适应值,且随着粒子的更新和迭代不断改变。在可解空间中,每个粒子都可能是问题的一个最优解。个体粒子的位置更新取决于个体极值和群体极值,通过跟踪极值来确定在解空间中的运动方向。

假设确立了一个 D 维搜索空间,空间中有由 n 个粒子组成的种群 $\boldsymbol{X}_{\mathrm{p}}=(X_1,X_2,\cdots,X_n)$,则在 D 维搜索空间中第 i 个粒子的位置,可能是问题的一个潜在的最优解,可用一个 D 维的向量 $\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{iD})^{\mathrm{T}}$ 表示。设每个粒子位置为 $\boldsymbol{X}_i$,则根据目标函数可以算出对应的适应度值。第 i 个粒子的速度为 $\boldsymbol{V}_i=(V_{i1},V_{i2},\cdots,V_{iD})^{\mathrm{T}}$,其个体极值为 $\boldsymbol{P}_i=(P_{i1},P_{i2},\cdots,P_{iD})^{\mathrm{T}}$,种群的全局极值为 $\boldsymbol{P}_{\mathrm{g}}=(P_{\mathrm{g}1},P_{\mathrm{g}2},\cdots,P_{\mathrm{g}D})^{\mathrm{T}}$。速度和位置更新公式如下:

$$V_{id}^{k+1}=\omega V_{id}^{k}+c_1r_1(P_{id}^{k}-X_{id}^{k})+c_2r_2(P_{\mathrm{g}d}^{k}-X_{id}^{k}) \tag{7}$$

$$X_{id}^{k+1}=X_{id}^{k}+V_{id}^{k+1} \tag{8}$$

http://bhxb.buaa.edu.cn

式中:ω 为惯性权重;k 为当前的迭代次数;$d=1,2,\cdots,D$;$i=1,2,\cdots,n$;V_{id} 为粒子的速度;c_1 和 c_2 为加速度因子,为非负常数;r_1 和 r_2 为分布于[0,1]之间的随机数。为防止粒子盲目搜索,一般将其速度和位置限制在一定的区间$[-V_{\max},V_{\max}]$、$[-X_{\max},X_{\max}]$内[19]。

2.3　粒子群优化极限学习机(PSO-ELM)算法

由于 ELM 的隐藏节点参数(输入权值和偏差)随机产生,相比于很多传统的算法需要更多的隐层节点才能达到理想的精度[20]。在 ELM 中,输出权值矩阵由输入权值矩阵和偏差计算得到,可能会存在很多无效的隐含层节点,影响 ELM 的泛化性能。因此,要提高 ELM 的性能,主要目标是寻找最优的隐含层节点个数,选择恰当的隐含层节点参数。

针对上述问题,本文提出一种 PSO-ELM 算法,采用 PSO 优化 ELM 的输入层权值和隐含层偏差。流程图如图 4 所示。

PSO-ELM 具体实现步骤如下:

1)初始化种群。种群规模 P 设为 200,设隐含层节点数为 h,输入层神经元数为 m,则粒子长度 $L=h(m+1)$,种群个体即粒子,由输入权值和隐含层偏差构成。

2)计算每个粒子的适应度。采用 ELM 算法对初始化种群中的粒子进行训练,计算出每个粒子的预测集均方根误差(RMSEP),将其作为粒子群优化算法的适应度值。

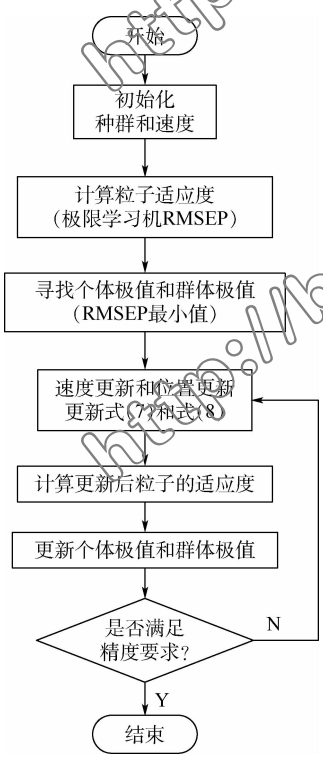


图 4　PSO-ELM 算法流程图

Fig. 4　Flowchart of PSO-ELM algorithm

3)寻找个体极值和群体极值。每次迭代后,将计算出的适应度值 PMSEP 和粒子的个体极值、群体极值进行比较,如果较好即 RMSEP 值较小,则将其作为个体极值和群体极值。

4)适当选取速度参数 c_1 和 c_1,将位置和速度限制在 ELM 隐含层节点参数取值范围$[-1,1]$内,根据式(7)更新粒子速度,式(8)更新粒子位置。

5)判断结果是否满足训练目标精度,不满足则转到步骤 2)。迭代结束时,全局最优粒子所在位置便是最优解。

3　实验设计

3.1　仪器设备

实验的检测设备是德州仪器公司生产的 DLP® NIRscan™ Evaluation Module (EVM)[21]近红外光谱仪,该仪器配置有基于 DLP 的光谱仪光学引擎。光源为频带从可见光到红外光的钨卤素灯,探测器为单元件扩展型 InGaAs 探测器,可探测长范围为 1 350 ~ 2 450 nm。

3.2　样本材料

实验样本材料为实验用滴定标准氨水溶液。氨水溶液采用高纯氨气和蒸馏水按体积比进行配置。氨水样本浓度范围为 0 ~ 0.25,浓度梯度为 0.01。每个样本浓度取样 3 次。每次取样样本用光谱仪重复扫描 3 次,建模样本光谱取3 次扫描和 3 次取样的平均值。样本数共计 26,将样本分为预测集和训练集。为使样本具有代表性,将样本按照浓度梯度顺序等分为 5 份,随机选取每份中任意一个浓度点作为预测集样本,其余为训练集样本。预测集样本数为 5,训练集样本数为 21。

3.3　模型建立与评价

本文采用 PSO-ELM 算法建立定量分析模型。将处理后的光谱分为训练集和预测集,训练集用于建立模型,预测集用于验证模型。采用 MATLAB 2009a进行模型建立和分析。程序分为 5 个部分,产生训练集/预测集、数据归一化处理、模型训练、模型测试以及结果分析。

本文采用训练集相关系数(correlation coefficient of calibration)R_c、训练集均方根误差(Root Mean Square Error of Calibration,RMSEC)、预测集相关系数(correlation coefficient of prediction)R_p 和 RMSEP 对模型性能进行评价[22]。

http://bhxb.buaa.edu.cn

$$R_c=\frac{\sum_{i=1}^{n}(y_{ci}-\bar{y}_c)(\hat{y}_{ci}-\bar{\hat{y}}_c)}{\sqrt{\sum_{i=1}^{n}(y_{ci}-\bar{y}_c)^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_{ci}-\bar{\hat{y}}_c)^2}} \tag{9}$$

$$\mathrm{RMSEC}=\sqrt{\frac{\sum_{i=1}^{n}(y_{ci}-\hat{y}_{ci})^2}{n_c}} \tag{10}$$

$$R_p=\frac{\sum_{i=1}^{n}(y_{pi}-\bar{y}_p)(\hat{y}_{pi}-\bar{\hat{y}}_p)}{\sqrt{\sum_{i=1}^{n}(y_{pi}-\bar{y}_p)^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_{pi}-\bar{\hat{y}}_p)^2}} \tag{11}$$

$$\mathrm{RMSEP}=\sqrt{\frac{\sum_{i=1}^{n}(y_{pi}-\hat{y}_{pi})^2}{n_p}} \tag{12}$$

式中：n_c 和 n_p 分别为训练集和预测集的样本个数；y_{ci}、$\bar{y}_c$、$\hat{y}_{ci}$和$\bar{\hat{y}}_c$ 分别为训练集第 i 个样本的目标性能参考值、所有训练集样本参考值的平均值、所建模型对训练集第 i 个样本的预测值及对所有训练集样本预测值的平均值；y_{pi}、$\bar{y}_p$、$\hat{y}_{pi}$和$\bar{\hat{y}}_p$ 分别为预测集第 i 个样本的目标性能参考值、所有预测集样本参考值的平均值、所建模型对预测集第 i 个样本的预测值及对所有预测集样本预测值的平均值。

4 结果分析

4.1 光谱预处理

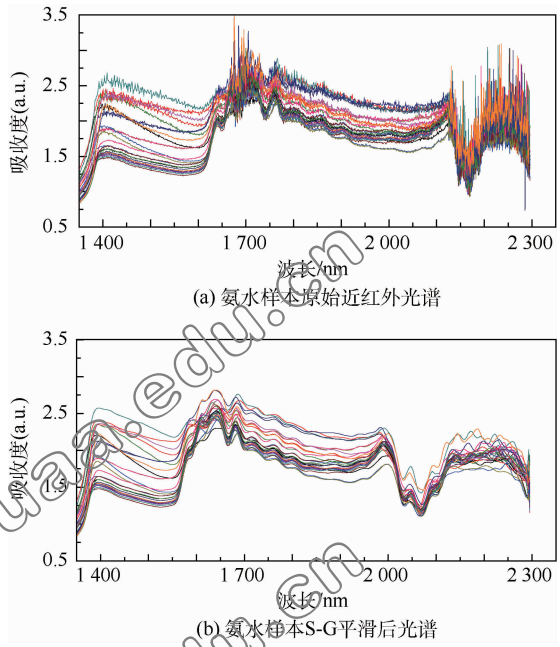
氨水样本的原始近红外光谱图如图 5(a)所示，图中各样品的近红外光谱谱峰严重重叠，且谱带较宽，无法直接进行定量分析建模，需对光谱进行预处理。

光谱仪得到的光谱信号中含有叠加的随机误差，信号平滑是消除噪声的最常用的方法。本文采用 S-G(Savitzky-Golay)滤波器对光谱进行平滑处理。S-G 滤波器是一种在时域内基于局域多项式最小二乘法拟合的滤波方法，其最大特点在于在滤除噪声的同时能够确保信号的形状、宽度不变，可尽可能保留光谱特征[23]。对原始光谱采用 25 点平滑滤波，平滑后的光谱如图 5(b)所示。氨水样本光谱吸收强度随着浓度增大逐渐增强。

4.2 建模结果分析

选择光谱吸收特征明显的 1 600 ~ 2 100 nm 谱段建立模型，用预测集对模型进行验证。为了更好地验证算法性能，本文将 PSO-ELM 建模结果与传统的 BP-ANN 和 SVM 算法建模结果进行对比。

表1 和图 6 为 BP - ANN / SVM 和 ELM / PSO -



(a) 氨水样本原始近红外光谱

(b) 氨水样本S-G平滑后光谱

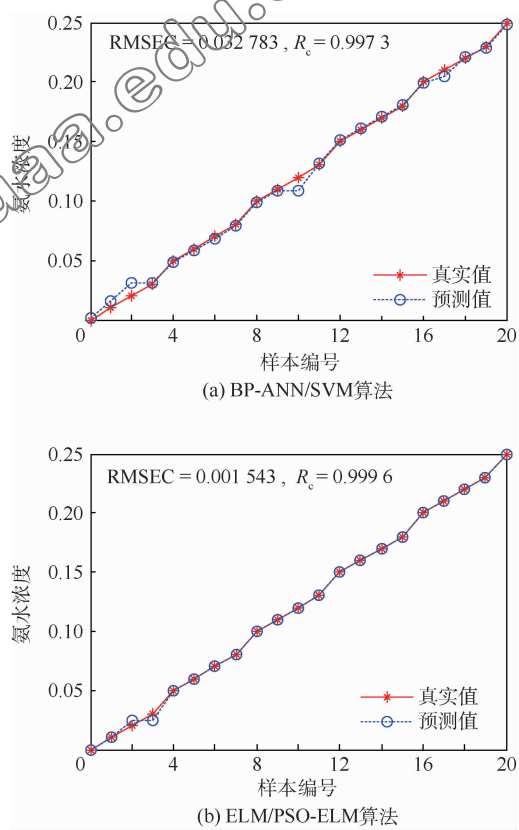
图 5 氨水近红外光谱预处理前后对比

Fig. 5 Comparison of ammonia near infrared spectrum before and after preprocessing

表 1 训练集拟合效果对比

Table 1 Comparison of imitative effects of training sets

算法类型	RMSEC	R_c
BP-ANN/SVM	0.032 783	0.997 3
ELM/PSO-ELM	0.001 543	0.999 6



(a) BP-ANN/SVM算法

(b) ELM/PSO-ELM算法

图 6 4 种算法训练集拟合结果

Fig. 6 Imitative effects of training sets of four algorithms

http://bhxb.buaa.edu.cn

ELM 训练集拟合结果对比。表 1 中 ELM/PSO-ELM 训练集均方根误差 RMSEC 为 0.001 543 远远小于 BP-ANN/SVM 训练集均方根误差 0.032 783。拟合相关系数相差 0.002 3,但均接近 1。图 6 为训练集拟合曲线,横坐标为样本编号,总共 21 个训练集样本,纵坐标为氨水体积浓度。对比图 6(a)和图 6(b)可知,ELM/PSO-ELM 对样本的利用率更高,训练集拟合效果优于 BP-ANN/SVM。

表 2 和图 7 为 BP-ANN、SVM、ELM 和 PSO-ELM 4 种算法预测结果对比。表 2 中真实值是预测集氨水样本的体积浓度,未参与训练集拟合建模。预测值为建模拟合的结果,真实值与预测值之差即为误差值。PSO-ELM 算法建模的拟合最大误差的绝对值为 0.004 61,RMSEP 值为 0.003 22,均小于其他 3 种算法,R_p 值为 0.997 9,大于其他 3 种算法,拟合效果最好。图 7(a)~图 7(d)分别为 BP-ANN、SVM、ELM 以及 PSO-ELM 预测集建模拟合结果。比较可知,图 7(d)中预测值和真实值误差最小、精度最高、拟合线性度最好。4 种算法拟合效果顺序为:PSO-ELM > ELM > BP-ANN > SVM。

表 2　4 种算法建模结果对比

Table 2　Comparison of modeling results among four algorithms

算法类型	氨水浓度		浓度误差	RMSEP	R_p
	真实值	预测值			
BP-ANN	0.04	0.064 23	0.024 23	0.014 28	0.993 8
	0.09	0.092 75	0.002 75		
	0.14	0.145 53	0.005 33		
	0.19	0.179 77	−0.010 23		
	0.24	0.220 36	−0.019 64		
SVM	0.04	0.040 84	0.000 84	0.014 55	0.980 7
	0.09	0.092 72	0.002 72		
	0.14	0.130 62	−0.009 38		
	0.19	0.189 76	−0.000 24		
	0.24	0.207 74	−0.032 26		
ELM	0.04	0.013 72	−0.026 28	0.011 66	0.995 1
	0.09	0.103 68	0.013 68		
	0.14	0.136 14	−0.003 86		
	0.19	0.189 76	−0.000 24		
	0.24	0.231 55	−0.008 45		
PSO-ELM	0.04	0.042 64	0.002 64	0.003 22	0.997 9
	0.09	0.085 39	−0.004 61		
	0.14	0.138 42	−0.001 58		
	0.19	0.194 34	0.004 34		
	0.24	0.238 40	−0.001 60		

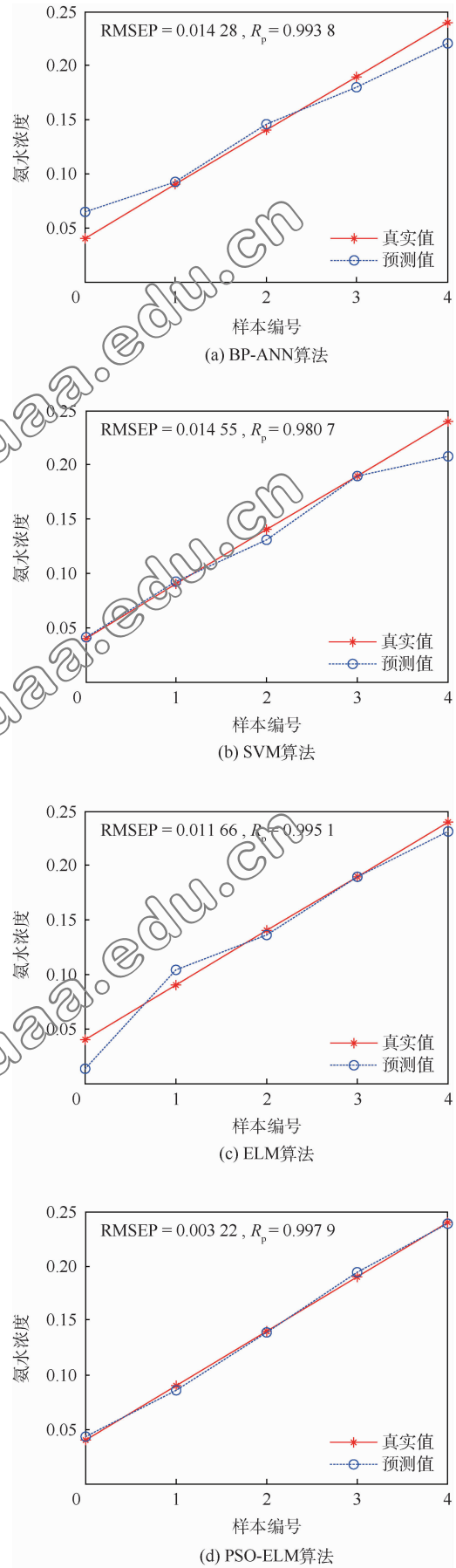


图 7　4 种算法预测集拟合结果

Fig. 7　Imitative effects of prediction sets of four algorithms

http://bhxb.buaa.edu.cn

5 结 论

本文提出了基于小样本数据的粒子群改进极限学习机算法,建立了氨水浓度的定量分析模型,可实现允许误差范围 ±0.005 的氨水浓度预测。

1) PSO-ELM 算法优化了 ELM 算法中随机产生的输入权值和隐含层偏差,减小了建模结果的随机性,提高了建模精度。

2) PSO-ELM 和 ELM 算法训练集拟合效果明显优于 BP-ANN 和 SVM 算法,提高了训练集数据点的利用率,有利于提高模型预测精度。

3) PSO-ELM 算法相比传统的 SVM、BP-ANN 算法,对于小样本数据具有更好的泛化性能和更高的回归拟合精度。

本文研究中,ELM 和 PSO-ELM 算法的隐含层节点个数和激活函数均选用经验值得出的最佳参数,未来的研究中,可结合实际样本和需求,选取出性能更好的建模参数。

参考文献(References)

[1] 王海莲,万向元,胡培松,等. 稻米脂肪含量近红外光谱分析技术研究[J]. 中国农业科学,2005,38(8):1540-1546.
WANG H L,WAN X Y,HU P S,et al. Quantitative analysis of fat content in brown rice by near infrared spectroscopy (NIRS) technique[J]. Scientia Agricultura Sinica,2005,38(8):1540-1546(in Chinese).

[2] 吴琼,原忠虎,王晓宁. 基于偏最小二乘回归分析综述[J]. 沈阳大学学报,2007,19(2):33-35.
WU Q,YUAN Z H,WANG X N. Summary of partial least squares regression[J]. Journal of Shenyang University,2007,19(2):33-35(in Chinese).

[3] CORTES C,VAPNIK V. Support-vector networks[J]. Machine Learning,1995,20(3):273-297.

[4] SUYKENS J A K,VANDEWALLE J. Least squares support vector machine classifiers[J]. Neural Processing Letters,1999,9(3):293-300.

[5] HEERMANN P D,KHAZENIE N. Classification of multispectral remote sensing data using a back-propagation neural network[J]. IEEE Transactions on Geoscience and Remote Sensing,1992,30(1):81-88.

[6] HUANG G B,ZHU Q Y,SIEW C K. Extreme learning machine:A new learning scheme of feedforward neural networks[C]// 2004 IEEE International Joint Conference on Neural Networks,2004. Piscataway,NJ:IEEE Press,2004,2:985-990.

[7] MASRI D,WOON W L,AUNG Z. Soil property prediction:An extreme learning machine approach[C]// 22nd International Conference on Neural Information Processing,ICONIP 2015. Berlin:Springer,2015:18-27.

[8] GUO W,GU J,LIU D,et al. Peach variety identification using near-infrared diffuse reflectance spectroscopy[J]. Computers and Electronics in Agriculture,2016,123(C):297-303.

[9] HUANG G B,CHEN L. Enhanced random search based incremental extreme learning machine[J]. Neurocomputing,2008,71(16):3460-3468.

[10] JIN Y,LI J,LANG C,et al. Multi-task clustering ELM for VIS-NIR cross-modal feature learning[J]. Multidimensional Systems and Signal Processing,2016,28(3):1-16.

[11] ALEXANDRE E,CUADRA L,SALCEDO-SANZ S,et al. Hybridizing extreme learning machines and genetic algorithms to select acoustic features in vehicle classification applications[J]. Neurocomputing,2015,152:58-68.

[12] 陈媛媛,张记龙,赵冬娥. 基于极限学习机的混合气体 FTIR 光谱定量分析[J]. 中北大学学报(自然科学版),2011,32(5):636-641.
CHEN Y Y,ZHANG J L,ZHAO D E. Quantitative analysis of mixed gas FTIR spectrum based on extreme learning machine[J]. Journal of North University of China (Natural Science Edition),2011,32(5):636-641(in Chinese).

[13] DE OLIVEIRA J F L,LUDERMIR T B. Homogeneous ensemble selection through hierarchical clustering with a modified artificial fish swarm algorithm[C]// 2011 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI). Piscataway,NJ:IEEE Press,2011:177-180.

[14] WORKMAN J,WEYER L. 近红外光谱解析实用指南[M]. 褚小立,许育鹏,田高友,译. 北京:化学工业出版社,2009:6-17.
WORKMAN J,WEYER L. Practical guide to interpretive near-infrared spectroscopy[M]. CHU X L,XU Y P,TIAN G Y,translated. Beijing:Chemical Industry Press,2009:6-17(in Chinese).

[15] 严衍禄. 近红外光谱分析基础与应用[M]. 北京:中国轻工业出版社,2005:128-130.
YAN Y L. Near infrared spectrum analysis and application[M]. Beijing:China Light Industry Press,2005:128-130(in Chinese).

[16] POGGIO T,GIROSI F. A theory of networks for approximation and learning[M]. Cambridge:Massachusetts Institute of Technology,1989:25-39.

[17] 陈林伟. 基于极限学习机的骨髓细胞识别技术研究[D]. 杭州:中国计量学院,2014:44-49.
CHEN L W. Research on recognition of bone marrow cells based on extreme learning machine[D]. Hangzhou:China Jiliang University,2014:44-49(in Chinese).

[18] KENNEDY J,EBERHART R. Particle swarm optimization[J]. Swarm Intelligence,2007,1(1):33-57.

[19] 石山,刘德鹏,李成茂. 基于改进粒子群优化算法的飞机作动系统功率调度[J]. 北京航空航天大学学报,2016,42(10):2024-2030.
SHI S,LIU D P,LI C M. Power dispatch of actuator of aircraft based on improved particle swarm optimization algorithm[J]. Journal of Beijing University of Aeronautics and Astronautics,2016,42(10):2024-2030(in Chinese).

[20] 李永强. 基于粒子群优化的极限学习机的 XML 文档分类中的研究与应用[D]. 沈阳:东北大学,2013:27-33.
LI Y Q. Research and application of XML classification based

http://tbxb.buaa.edu.cn

on extreme learning machine with particle swarm optimization [D]. Shenyang: Northeastern University, 2013: 27-33 (in Chinese).

[21] PRUETT E. Techniques and applications of programmable spectral pattern coding in Texas Instruments DLP spectroscopy [C]// Proceedings of SPIE-The International Society for Optical Engineering. Bellingham, WA: SPIE, 2015.

[22] 刘小丽. 红外光谱在药物分析中的应用研究[D]. 西安:西北大学,2013:6-8.
LIU X L. The application of infrared spectroscopy in pharmaceutical analysis [D]. Xi'an: Northwest University, 2013: 6-8 (in Chinese).

[23] GORRY P A. General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method [J]. Analytical Chemistry, 1990, 62(6): 570-573.

作者简介:

周美灵　女,硕士研究生。主要研究方向:近红外光谱分析。

郑德智　男,博士,副教授,博士生导师。主要研究方向:传感器敏感机理及检测系统。

Quantitative spectral modeling method based on improved extreme learning machine

ZHOU Meiling, ZHENG Dezhi*, LOU Ge, LIU Zheng

(School of Instrumentation Science and Opto-electronics Engineering, Beijing University of Aeronautics and Astronautics, Beijing 100083, China)

Abstract: According to the principle of near infrared spectrum (NIR), an optimized extreme learning machine algorithm with particle swarm optimization (PSO-ELM) was proposed and used in ammonia concentration quantitative analysis of small sample. By optimizing the hidden node parameters of extreme learning machine (ELM) algorithm, the problem of randomly generated input weight and hidden layer of ELM leading to random modeling results has been solved. At the same time, the model stability, accuracy and generalization performance were improved. Through the experimental verification, by the optimized PSO-ELM, compared to ELM, the root mean square error of prediction set reduces to 0.003 22 from 0.011 66 and the correlation coefficient of prediction increases from 0.995 1 to 0.997 9. After comparing the optimized model prediction results with the modeling results of traditional support vector machine (SVM) regression and BP neural network algorithm, optimized PSO-ELM offers high accuracy and excellent generalization performance. Model prediction effect is superior to the traditional quantitative regression analysis algorithm.

Keywords: near infrared spectrum (NIR); quantitative analysis; small sample; extreme learning machine (ELM); particle swarm optimization (PSO) algorithm

http://bhxb.buaa.edu.cn

Received: 2016-05-31; **Accepted**: 2016-07-01; **Published online**: 2016-08-25 08:33

URL: www.cnki.net/kcms/detail/11.2625.V.20160825.0833.001.html

Foundation item: National Key Technology Research and Development Program of China (2014BAF08B01)

* **Corresponding author**. E-mail: zhengdezhi@buaa.edu.cn