稳健极限学习机及其在近红外光谱分析中的应用

白俊健[1] 孙 群[2] 井诗博[1] 杨丽明[1]

[1]中国农业大学理学院, 北京 100083

[2]中国农业大学农学与生物技术学院, 北京 100193

摘要 极限学习机(ELM)作为一种单隐层前馈神经网络已成为大数据分析的重要工具。与传统神经网络相比,ELM具有结构简单、学习速度快和推广性较好等优势。但是,ELM的输出权值是基于最小二乘法估计的,容易夸大离群点和噪声的影响,导致其预测性能的不稳定。提出一种新的稳健的极限学习机——基于最小一乘回归的极限学习机(LAD-ELM),而且问题被转化为线性规划,能够简单、快速求解其全局最优解。进一步将LAD-ELM应用于近红外光谱数据建模,构建了基于LAD-ELM和近红外光谱数据的乌拉尔甘草种子硬实性分析系统。与传统的方法相比,在不同光谱范围的数值实验显示了提出方法的可行性和有效性,为利用近红外光谱和ELM技术进行种子硬实性研究提供了理论依据和实用方法。

关键词 光谱学; 近红外光谱; 极限学习机; 最小一乘回归; 稳健性

中图分类号 TP18; O235 **文献标识码** A

doi: 10.3788/LOP52.103002

Robust Extreme Learning Machine and Its Application in Analysis of Near Infrared Spectroscopy Data

Bai Junjian[1] Sun Qun[2] Jing Shibo[1] Yang Liming[1]

[1]*College of Science, China Agricultural University, Beijing* 100083, *China*

[2]*College of Agriculture and Biotechnology, China Agricultural University, Beijing* 100193, *China*

Abstract Extreme learning machine (ELM), as a kind of single hidden layer feedforword neural networks, is an important tool in big data analysis. Compared with traditional neural network methods, it has simple structure, high learning speed and good generalization performance. However, the output weight of ELM is estimated by the least squares estimation (LSE) method, and thus ELM network lacks of robustness since LSE is relatively sensitive to outlier. A new robust ELM based on least absolute deviations (LAD) regression, called LAD-ELM, is presented. Moreover, the proposed LAD-ELM is posed as a linear program with global optimal solution. Furthermore, the proposed LAD-ELM is directly used for near-infrared (NIR) spectral analysis, and an analysis system for hardness of licorice seeds is built based on LAD-ELM and NIR data. Compared with the traditional methods, the experimental results in different spectral regions show the feasibility and effectiveness of the proposed method. Moreover, the investigation provides theoretical support and practical method for studies on licorice seed hardness using ELM and NIR technology.

Key words spectroscopy; near-infrared spectroscopy; extreme learning machine ; least absolute deviation regression; robustness

OCIS codes 300.6340; 060.4256; 080.2720

1 引 言

近红外光谱(NIR)分析[1-3]以其快速、无损、简单等特点在分析化学领域及复杂植物样品的分析中显现了巨大的优势与潜力,已经得到了广泛的应用。近红外光谱集由不同波长下的吸光度组成,一个样本的近红外光谱在数学上可以视为一个向量,其吸光度点的个数即为向量维数。红外光是指波长在4000~12000 cm^{-1}范

收稿日期: 2015-03-12; **收到修改稿日期**: 2015-04-18; **网络出版日期**: 2015-09-23

基金项目: 国家自然科学基金(11471010,11271367)

作者简介: 白俊健(1992—),男,硕士研究生,主要从事数据挖掘方面的研究。E-mail: 944890706@qq.com

导师简介: 杨丽明(1963—),女,博士,副教授,主要从事机器学习和最优化等方面的研究。

E-mail: cauyanglm@163.com(通信联系人)

围内的电磁波。因此,光谱数据集是高维数据集。

近年来,模式识别技术已广泛应用于近红外光谱数据的分析及模型建立。常用的方法有多元线性回归、偏最小二乘[4-5]、人工神经网络[6]、支持向量机回归[7]等方法。极限学习机(ELM)[8-11]是Huang等在2006年提出的一种新型单隐层前馈神经网络。作为一种大数据学习的重要工具,极限学习机已经成为近年来的研究热点,并已成功应用于机器学习和高性能计算等多个领域。

乌拉尔甘草是我国的一种传统中药材,其种子具有硬实特性。硬实种子的活力和耐贮藏能力显著高于非硬实种子[1]。由于不同种源地、不同年份的甘草种子硬实率存在很大的差异,在对种子进行处理前,应事先了解种子的硬实率水平,以确定是否需要进行处理以及适宜的处理时间,因此,硬实率是种子检验的一项重要内容[1,3]。传统的测定种子硬实率的方法是浸泡法,时间相对较长。本文提出了基于ELM和近红外光谱的乌拉尔甘草种子硬实率分析方法。

2 极限学习机

ELM具有算法结构简单、学习速度快、良好的非线性处理能力和容噪能力及全局搜索和泛化性能等优势。分类问题和回归问题是它的主要研究对象。不同于传统的神经网络,ELM设置合适的隐层节点数,对输入权值和隐层偏差进行随机赋值,输出层权值通过最小二乘法得到。整个过程无需迭代,一次完成。ELM的训练速度比反向传播(BP)神经网络算法快2~3个数量级。且对一个任意无限光滑的激活函数,该网络能够以零误差逼近任意不同的线性与非线性函数。与支持向量机相比,ELM保持了与其相当的推广性能,但所需学习时间更少。因此,ELM作为一种新的挖掘技术,已成功地应用于大数据分析。

给定训练集 $D=\{(x_i,t_i)|x_i\in R^n,t_i\in R^m,i=1,2,\cdots,N\}$,具有$L$个隐层节点,激活函数为$g(x)$的单隐层前馈神经网络数学模型可以表示为

$$\sum_{i=1}^{L}\boldsymbol{\beta}_i g(\boldsymbol{w}_i\cdot x_j+b_i)=\boldsymbol{t}_j,\ j=1,2,\cdots,N, \tag{1}$$

式中 $\boldsymbol{w}_i=[w_{i1},w_{i2},\cdots,w_{in}]^{\mathrm{T}}$ 表示连接第i个隐层节点的输入权值,$\boldsymbol{\beta}_i=[\beta_{i1},\beta_{i2},\cdots,\beta_{im}]^{\mathrm{T}}$为隐层第$i$个节点到输出层的权值,$b_i$为第$i$个隐层神经元的偏移量。

具有L个隐层节点的单隐层前馈神经网络,能够零误差地逼近任意N个样本,即 $\sum_{i=1}^{L}\|\boldsymbol{y}_i-\boldsymbol{t}_i\|=0$,可以简写为一个线性系统 $\boldsymbol{H\beta}=\boldsymbol{T}$,其中

$$\boldsymbol{H}=\begin{bmatrix} g(\boldsymbol{w}_1\cdot x_1+b_1) & \cdots & g(\boldsymbol{w}_L\cdot x_1+b_L)\\ \vdots & \cdots & \vdots\\ g(\boldsymbol{w}_1\cdot x_N+b_1) & \cdots & g(\boldsymbol{w}_L\cdot x_N+b_L)\end{bmatrix}_{N\times L},\ \boldsymbol{\beta}=\begin{bmatrix}\boldsymbol{\beta}_1^{\mathrm{T}}\\ \vdots\\ \boldsymbol{\beta}_L^{\mathrm{T}}\end{bmatrix}_{L\times m},\ \boldsymbol{T}=\begin{bmatrix}\boldsymbol{t}_1^{\mathrm{T}}\\ \vdots\\ \boldsymbol{t}_L^{\mathrm{T}}\end{bmatrix}_{N\times m},$$

$\boldsymbol{H}$为神经网络的隐层输出矩阵,它的第i列代表第i个隐层神经元关于每个输入向量 $x_1,x_2,\cdots,x_N$ 的输出。Huang等指出输入权值$\boldsymbol{w}_i$与隐含层的阀值b_i不需要调整,并且隐层输出矩阵$\boldsymbol{H}$的权值与阀值可以在学习开始时进行随机赋值并保持固定。因此训练一个单隐层前馈神经网络就相当于寻求线性系统 $\boldsymbol{H\beta}=\boldsymbol{T}$ 的最小二乘问题的解,

$$\left\|\boldsymbol{H}\hat{\boldsymbol{\beta}}-\boldsymbol{T}\right\|_2^2=\min_{\beta}\left\|\boldsymbol{H\beta}-\boldsymbol{T}\right\|_2^2. \tag{2}$$

求解(2)式得到ELM的输出层权值。一般情况下,训练样本个数N远大于隐层节点数L,这需要求解矩阵$\boldsymbol{H}$的伪逆,即有 $\hat{\boldsymbol{\beta}}=\boldsymbol{H}^{\dagger}\boldsymbol{T}$,其中 $\boldsymbol{H}^{\dagger}=(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{T}}$ 为隐含层输出矩阵$\boldsymbol{H}$的广义莫尔逆。

最小二乘估计(LSE)是应用最为广泛的回归方法,简单实用,能在正态假定下应用统计检验理论。然而由于最小二乘是通过残差平方和来求解回归系数,容易夸大实验数据中奇异值的影响,增大统计误差,因此,最小二乘法估计容易造成ELM网络训练结果的不稳定性。

3 基于最小一乘的ELM

基于最小一乘准则(LAD)[12-13]的估计具有稳健的特性,能有效排除奇异值的干扰,可得到较稳健的估计,能有效地克服LSE的缺陷。但LAD估计中绝对值方程不便于计算,较难求解。

本文介绍了一种新的极限学习机,即基于最小一乘的极限学习机(LAD-ELM)。首先将 LAD-ELM 模型转化为线性规划问题,快速得到其全局最优解。具体来说,用 $l_{1-\mathrm{norm}}$ 替换 $l_{2-\mathrm{norm}}$,得到基于最小一乘的LAD-ELM 回归模型:

$$\left\| \boldsymbol{H}\hat{\boldsymbol{\beta}} - \boldsymbol{T} \right\|_1 = \min_{\beta} \left\| \boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{T} \right\|_1 , \tag{3}$$

引入两个变量 d^+, d^-,并且满足 $\boldsymbol{T} - \boldsymbol{H}\boldsymbol{\beta} = d^+ - d^-, d^+ \geqslant 0, d^- \geqslant 0$,则 $\left\| \boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{T} \right\|_1 = \exp(\boldsymbol{T})(d^+ + d^-),\ d^+, d^- \geqslant 0$,因此,基于最小一乘的ELM回归问题转化为最优化问题:

$$\min_{\beta, d^+, d^-} \quad \exp(T)(d^+ + d^-) \qquad \text{s.t.} \quad \boldsymbol{H}\boldsymbol{\beta} + d^+ - d^- = \boldsymbol{T} \qquad d^+ \geqslant 0, d^- \geqslant 0 . \tag{4}$$

这是一个线性规划模型,具有全局最优解,可简单快速求得 $\boldsymbol{\beta}$ 的最优解 $\hat{\boldsymbol{\beta}}$,即为LAD-ELM的输出层权值。

根据以上分析,求解LAD-ELM的算法可归纳如下:

1) 对给定训练集 $D = \{(x_i, \boldsymbol{t}_i) | x_i \in R^n, \boldsymbol{t}_i = [t_{i1}, t_{i2}, \cdots, t_{im}] \in R^m, i = 1, \cdots, N\}$,选取适当的隐层节点数 L 及激励函数 $g(x)$;

2) 随机输入隐层节点权值向量及偏移值 $(\boldsymbol{w}_i, b_i),\ i = 1, 2, \cdots, N$;

3) 计算隐层输出矩阵 $\boldsymbol{H}$;

4) 解线性规划(4)式,得到输出权值 $\hat{\boldsymbol{\beta}}$. 则 LAD-ELM 回归函数为

$$f(x) = \sum_{i=1}^{L} \hat{\boldsymbol{\beta}}_i g(\boldsymbol{w}_i \cdot x + b_i) . \tag{5}$$

4 基于LAD-ELM和NIR数据的种子识别系统

4.1 光谱采集

实验所用样品为2008年收获的乌拉尔甘草种子,产地宁夏。选取112个样品。采用德国布鲁克仪器公司生产的MPA傅里叶变换近红外光谱仪,分辨率设为4 $\mathrm{cm^{-1}}$,扫描范围为4000~12000 $\mathrm{cm^{-1}}$,扫描32次。在每粒种子胚面的不同位置采集光谱3次,以3次重复数据的平均值作为甘草种子硬实率的标准值。样品的近红外光谱如图1所示。

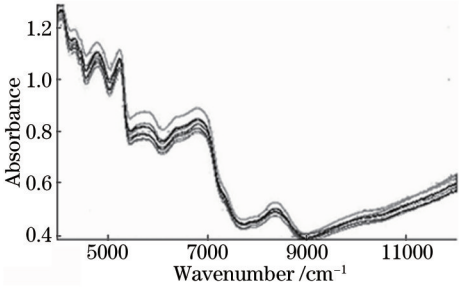

图1 乌拉尔甘草种子的近红外光谱

Fig.1 NIR spectra of licorice seeds

表1 4个谱数据集

Table 1 Four spectral data sets

Sample set	Spectral range /$\mathrm{cm^{-1}}$	Number of samples	Wavelength /nm
Set A	10,000~12,000	112	525
Set B	8000~10,000	112	525
Set C	6000~8000	112	525
Set D	4000~6000	112	525

4.2 实验设计与数据处理

实验在Matlab7.0上实现。应用软件OPUS5.5将初始光谱数字化后,在4000~12000 $\mathrm{cm^{-1}}$光谱范围内,每个样品可表示为一个2100维的列向量。在实验过程中使用Matlab Statistics Toolbox和Matlab Optimization Toolbox软件包分析实验结果,其中Matlab 函数quadprog用于求解支持向量机回归(SVR)和ELM最优化问题.为了验证提出方法的有效性,在4个不同的谱区域10000~12000 $\mathrm{cm^{-1}}$,8000~10000 $\mathrm{cm^{-1}}$, 6000~8000 $\mathrm{cm^{-1}}$,

4000~6000 cm^{-1} 进行数值实验，并分别记为 Set A, Set B, Set C 和 Set D，如表 1 所示。选取 SVR 和 ELM 作为基准算法进行比较。

在 SVR 算法中，取线性核函数，惩罚参数 C 从集合 $\{10^i | i=0,1,2,3,4\}$ 中选取，不敏感损失参数 ε 从集合 $\{10^{-i} | i=0,1,2,3\}$ 中选取，利用 10-fold 交叉实验选取回归错误最小的参数 C 及 ε。最后的优选参数为 C=1000 及 ε=0.001。

在 LAD-ELM 算法中，选择 Sigmoid 型激励函数 $g(\boldsymbol{w},b,x)=1/\{1+\exp[-(\boldsymbol{w}^{\mathrm{T}}x+b)]\}$。隐层节点个数 L 从集合 {10,20,40,60,80,100,112} 中选取。利用 10-fold 交叉实验选取回归错误最小的隐层节点个数，最后的优选值为 L=20。

在 ELM 算法中，选择的 Sigmoid 型激励函数及隐层节点个数与 LAD-ELM 相同。

4.3 算法评价标准

为了综合评估提出方法的有效性，利用下列准则评价算法[14]：校正标准偏差 $f_{\mathrm{SEC}}=\sqrt{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/(n-1)}$，预测标准偏差 $f_{\mathrm{SEP}}=\sqrt{\sum_{i=1}^{n^*}(y_i-\hat{y}_i)^2/(n^*-1)}$，回归平方和 $f_{\mathrm{SSE}}=\sum_{i=1}^{m}(y_i-\hat{y}_i)^2$，总离差平方和 $f_{\mathrm{SST}}=\sum_{i=1}^{m}(y_i-\bar{y})^2$，运行时间 t_{CPU}，其中 m 为测试样本数目，n 为校正集样本数目，y_i 和 $\hat{y}_i$ 分别为样本 x_i 的真实值和预测值，而 $\bar{y}=\frac{1}{m}\sum y_i$ 为真实值 $y_1,y_2,\cdots,y_m$ 的平均。

一般来说，$f_{\mathrm{SSE}}/f_{\mathrm{SST}}$ 的值越小，模型拟合效果越好。在大多数情况下，较小的 $f_{\mathrm{SSE}}/f_{\mathrm{SST}}$ 值意味着估计值与真实值之间的误差较小；f_{SEP}，f_{SEC} 以及 $f_{\mathrm{SEC}}/f_{\mathrm{SEP}}$ 值越低，则模型预测性能越好。

表 2 ELM, SVR 和 LAD-ELM 的实验结果比较

Table 2 Comparison of ELM, SVR and LAD-ELM methods

Data set	Methods	f_{SEC}	$f_{\mathrm{SSE}}/f_{\mathrm{SST}}$	f_{SEP}	$f_{\mathrm{SEC}}/f_{\mathrm{SEP}}$
Set A	ELM	0.1763	0.3912	0.2049	0.8604
	SVR	0.2423	0.3857	0.2958	0.8191
	LAD-ELM	**0.1495**	**0.3100**	**0.2029**	**0.7368**
Set B	ELM	0.0841	0.1189	0.1065	0.7897
	SVR	0.2252	0.2708	0.2733	0.8240
	LAD-ELM	**0.0705**	**0.0855**	**0.1059**	**0.6657**
Set C	ELM	0.0591	0.0857	0.0831	0.7112
	SVR	0.1587	0.1194	0.1939	0.8185
	LAD-ELM	**0.0493**	**0.0691**	**0.0821**	**0.6005**
Set D	ELM	0.0551	0.0839	0.0783	0.7037
	SVR	0.1323	0.1023	0.1931	0.6851
	LAD-ELM	**0.0452**	**0.0487**	**0.0748**	**0.6043**

表 3 ELM, SVR 和 LAD-ELM 计算时间的比较

Table 3 Comparison of ELM, SVR and LAD-ELM in terms of t_{CPU}

	Methods	Set A	Set B	Set C	Set D
Time/s	ELM	0.0140	0.0149	0.0145	0.0153
	SVR	0.0449	0.0472	0.0443	0.0457
	LAD-ELM	0.0211	0.0208	0.0220	0.0232

5 实验结果

为了评估提出的方法，在 4 个不同光谱区域上比较 LAD-ELM、传统 ELM 及 SVR，依据上述算法评价准则，10-fold 交叉实验的平均结果如表 2 所示，表 3 为 3 种方法训练时间的比较结果。

表2和表3显示LAD-ELM得到非常好的运行结果。与SVR相比，依据校正标准偏差f_{SEC}，预测标准偏差f_{SEP}和f_{SSE}/f_{SST}，LAD-ELM在4个不同NIR谱区的实验结果均明显地减小了回归误差；而且LAD-ELM的运行时间几乎是SVR的二分之一。依据f_{SEC}/f_{SEP}，在3种方法中，LAD-ELM在4个NIR谱区上取得最小f_{SSE}的同时f_{SEC}/f_{SEP}的值也最小。这些结果显示在推广能力及运行时间上LAD-ELM均优于SVR。

与传统ELM相比，LAD-ELM在4个不同NIR谱区的实验结果f_{SEC}，f_{SEP}和f_{SSE}/f_{SST}均明显地减小，因此LAD-ELM提高了预测精度。根据运行时间分析，ELM优于LAD-ELM，可能是因为ELM回归参数是通过求解一个线性方程系统得出，而LAD-ELM模型是一个线性规划问题。

6 结 论

提出了基于最小一乘回归的极限学习机，问题被转化为线性规划，能够简单、快速求解其全局最优解，构建了基于LAD-ELM和NIR数据的乌拉尔甘草种子硬实率分析系统。

通过对3种不同方法的校正标准偏差f_{SEC}，预测标准偏差f_{SEP}和f_{SSE}/f_{SST}进行比较，可以发现ELM与LAD-ELM明显优于SVR。而LAD-ELM在4个不同NIR谱区上又都略优于ELM算法。因此，LAD-ELM算法的表现最为优秀，在推广能力上均优于传统的ELM和SVR。

3种不同方法中，SVR算法的运行速度最慢，ELM的运行速度最快，而LAD-ELM算法的运行速度介于SVR和ELM算法之间。LAD-ELM算法虽慢于ELM但是二者的差距并不明显，ELM的运算速度仅略优于LAD-ELM算法。

因此，与传统的方法相比，在不同光谱范围的数值实验显示了本文提出方法的可行性和有效性，为利用NIR和ELM技术进行种子硬实性研究提供了理论依据和实用方法。

参 考 文 献

1 Han Liangliang, Mao Peisheng, Wang Xinguo, *et al.*. Study on vigor test oat seeds with near infrared reflectance spectroscopy[J]. Journal of Infrared and Millimeterwaves, 2008, 27(2): 86-90.
韩亮亮, 毛培胜, 王新国, 等. 近红外光谱技术在燕麦种子活力测定中的应用研究[J]. 红外与毫米波学报, 2008, 27(2): 86-90.

2 Cao Pengfei, Li Hongning, Luo Yanlin, *et al.*. Selection of feature bands for phaseolus vulgaris leaves based on multi-spectral imaging[J]. Laser & Optoelectronics Progress, 2014, 51(1): 011101.
曹鹏飞, 李宏宁, 罗艳琳, 等. 基于多光谱成像选取四季豆叶片的特征波段[J]. 激光与光电子学进展, 2014, 51(1): 011101.

3 Yang L M, Sun Q. Recognition of the hardness of licorice seeds using a semi-supervised learning method[J]. Chemometrics and Intelligent Laboratory Systems, 2012, 114: 109-115.

4 Nie F P, Meng G F, Pan C H, *et al.*. Discriminative least squares regression for multiclass classification and feature selection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2012, 23(11): 1738-1754.

5 Yu Xiaoya, Zhang Yujun, Yin Gaofang, *et al.*. Feature wavelength selection of phytoplankton fluorescence spectra based on partial least squares[J]. Acta Optica Sinica, 2014, 34(9): 0930002.
余晓娅, 张玉钧, 殷高方, 等. 基于偏最小二乘回归的藻类荧光光谱特征波长选取[J]. 光学学报, 2014, 34(9): 0930002.

6 Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: Theory and application[J]. Neurocomputing, 2006, 70(1-3): 489-501.

7 Vapnik V N. Statistical Learning Theory[M]. New York: Wiley-Interscience, 1998.

8 Wang Y G, Cao F L, Yuan Y B. A study on effectiveness of extreme learning machine[J]. Neurocomputing, 2011, 74(16): 2483-2490.

9 Zhang Haidong, Li Guirong, Li Ruocheng, *et al.*. Determination of tea polyphenols content in puerh tea using near-infrared spectroscopy combined with extreme learning machine and GA-PLS algorithm[J]. Laser & Optoelectronics Progress, 2013, 50(4): 043001
张海东, 李贵荣, 李若诚, 等. 近红外光谱结合极限学习机和GA-PLS算法检测普洱茶茶多酚含量[J]. 激光与光电子学进展, 2013, 50(4): 043001.

10 Horata P, Chiewchanwattana S, Sunat K. Robust extreme learning machine[J]. Neurocomputing, 2013, 10(2): 31-44.

11 Yuan Y B, Wang Y G, Cao F L. Optimization approximation solution for regression problem based on extreme learning

machine[J]. Neurocomputing, 2011, 74(16): 2475-2482.

12 Cao H R, Liu X. System identification based on the least absolute criteria[J]. Techniques of Automation and Applications, 2009, 28(7): 8-10.

13 Gu Yuemin. Least absolute deviation curve fitting[J]. Journal of Tongji University, 2011, 39(9): 1377-1382.
顾乐民. 曲线拟合的最小一乘法[J]. 同济大学学报, 2011, 39(9): 1377-1382.

14 Peng X. TSVR: An efficient twin support vector machine for regression[J]. Neural Networks, 2010, 23(3): 365-372.

栏目编辑：吴秀娟