基于特征提取的极限学习机算法在可调谐二极管激光吸收光谱学中的应用

吕晓翠**，李国林，李晗，季文海*
中国石油大学(华东)信息与控制工程学院，山东 青岛 266580

摘要 采用波长为 1570 nm 的激光器分析了天然气背景下的硫化氢气体，通过自动化配气站产生了体积分数为 0～10^{-4} 的硫化氢混合气体，获取了 92 组稳定状态的光谱数据，采用极限学习机(ELM)的回归模型反演了硫化氢浓度。把非线性迭代偏最小二乘法引入到光谱预处理中，利用光谱特征参量与浓度参量建立了回归模型，采用五折交叉校验的方法对模型进行了评估。测试结果显示，光谱数据采用特征提取后的预测精度比直接用 ELM 进行回归的提升了 25%，且模型运算时间由 0.12 s 缩短到了 10 ms 以下。特征提取预处理缩短了 ELM 模型的训练时间，提高了分析仪的分析精度和实时性。

关键词 光谱学；可调谐二极管激光吸收光谱；特征提取；非线性迭代偏最小二乘；极限学习机；交叉校验
中图分类号 O433.4 **文献标识码** A **doi**：10.3788/CJL201845.0911013

Application of Feature-Extraction-Based Extreme Learning Machine Algorithm in Tunable Diode Laser Absorption Spectroscopy

Lü Xiaocui**, Li Guolin, Li Han, Ji Wenhai*
College of Information and Control Engineering, China University of Petroleum, Qingdao, Shandong 266580, *China*

Abstract The laser with a wavelength of 1570 nm is used to analyze the hydrogen sulfide gas in the background of natural gas. The gas mixture containing the hydrogen sulfide with a volume fraction of 0-10^{-4} is produced in an automatic gas mixing station, and 92 groups of the spectral data with a stable state are obtained. The regression model of extreme learning machine (ELM) is adopted for the inversion calculation of the concentration of hydrogen sulfide. The nonlinear iteration partial least square (NIPALS) algorithm is introduced into the spectral pretreatment. The ELM regression model is established by using the spectral feature vector and the concentration vector, and is evaluated by the five-fold cross validation method. The test results show that, the regression predicating accuracy of the spectral data obtained by feature extraction is improved by 25% than that by direct ELM, and the model operation time is reduced from 0.12 s to less than 10 ms. The spectral pretreatment by the feature extraction can reduce the training time of ELM model and can also improve the analysis accuracy and the real-time capability of the analyzers.

Key words spectroscopy; tunable diode laser absorption spectroscopy; feature extraction; nonlinear iteration partial least square; extreme learning machine; cross validation
OCIS codes 300.1030; 120.6200; 300.6260

1 引 言

可调谐二极管激光吸收光谱(TDLAS)技术因其灵敏度高、探测限低、响应度快、可动态测量等特点，被应用于石油化工、天然气处理、清洁能源等领域的气体过程分析[1-2]。利用 TDLAS 技术对痕量气体进行高灵敏检测时，目标气体吸收峰附近背景气体吸收峰的叠加会严重影响检测系统的检测精

收稿日期：2018-03-19；**修回日期**：2018-05-04；**录用日期**：2018-06-04
基金项目：山东省自然科学基金(ZR2017LF023)、青岛市科技惠民专项(17-3-3-89-nsh)、吉林大学集成光电子学国家重点实验室开放课题(IOSKL2017KF0)、中国石油大学(华东)研究生创新工程资助项目(YCX2018065)
* **E-mail**：jiwenhai@upc.edu.cn；** **E-mail**：xiaocuilv52@163.com

度。选用合适波长的激光器可以尽量避免多组分交叉干扰[3]，但是由于高灵敏检测对信号高增益的要求，背景组分的残余光谱也被同比放大，仍然存在多组分光谱交叉干扰的问题，这严重限制了探测底限、灵敏度、稳健性等分析仪性能。

利用化学计量学[4]的回归模型预测分析浓度可以避免多组分交叉干扰的问题。比如，使用经典最小二乘(CLS)回归[5]，将实时采集的光谱与已知浓度的参考谱拟合，可获取过程气体中待测分析物的浓度。但实时分析过程中存在的谱图形变、基线漂移等问题影响了模型的准确性。CLS方法适用于固定的气体组分的标定，但化工过程是动态变化的，实际组分与标定时所用组分偏差大，准确度因此降低。偏最小二乘(PLS)法[5-6]是另一种应用广泛的算法，它将主成分分析和线性回归拟合相结合，算法的性能比CLS的有显著提升。近年来，随着人工智能技术的快速发展，机器学习开始被应用于近红外光谱分析中。极限学习机(ELM)[7-9]具有模型结构简单、可调参量少等优点，在风力发电、电价预测、交通路牌识别等领域[10-12]获得了良好的效果。Huang等[10]将ELM核函数模型应用于交通信号灯分类；Bian等[11]实现了递进的ELM在油气近红外光谱中的应用；Javed等[12]用反双曲线正弦函数与Morlet小波函数叠加的激励函数代替了ELM原有的单一激励函数，改善了分类模型的收敛速度和稳健性。

采用TDLAS技术进行气体分析时，系统采集的每条光谱都含有几百甚至上千个数据点。在建立ELM模型进行浓度反演时，往往需要很多隐含层节点，存在隐含层输出矩阵维数高和共线性的问题，回归模型不稳定，而且会延长模型的分析时间。受调制吸收光谱技术的光谱分辨率和检测气体的物理状态限制，系统存在多组分光谱交叉干扰。进行痕量检测时，目标气体的光谱较弱且容易被其他气体干扰。若采用合适的特征分析的方法提取光谱特征，可以提高回归模型的精度。本文利用非线性迭代偏最小二乘(NIPALS)算法[13]对光谱数据进行特征提取，寻找少数几个由原始变量线性组合的主成分，以解释数据的结构特征。采用基于光谱特征的ELM算法模型进行回归建模和预测，比传统回归模型具有更高的精度和稳定性。

2　算法原理

2.1　光谱数据特征提取

将采集的光谱数据分为训练集光谱数据和测试集光谱数据。假设训练集光谱矩阵为 $\boldsymbol{X}_{\text{train}}(N\times m)$，硫化氢的浓度矩阵为 $\boldsymbol{Y}_{\text{train}}(N\times 1)$，其中 N 为训练集光谱的样本个数，m 为光谱数据点数。采用NIPALS法对光谱矩阵进行主成分分解，得到 r 个主成分。光谱矩阵的主成分分解可以表示为

$$\begin{cases}\boldsymbol{X}_{\text{train}}=\boldsymbol{S}\boldsymbol{P}^{\text{T}}+\boldsymbol{E}=\sum_{i=1}^{r}s_i p_i^{\text{T}}+\boldsymbol{E}\\ \boldsymbol{Y}_{\text{train}}=\boldsymbol{U}\boldsymbol{Q}^{\text{T}}+\boldsymbol{F}=\sum_{i=1}^{r}u_i q_i^{\text{T}}+\boldsymbol{F}\end{cases},\tag{1}$$

式中 $\boldsymbol{S}(N\times r)$ 和 $\boldsymbol{P}(m\times r)$ 分别为光谱矩阵 $\boldsymbol{X}_{\text{train}}$ 的得分矩阵和载荷矩阵，s_i、$p_i(i=1,2,\cdots,r)$ 为对应的矩阵元，r 为矩阵元数，T代表求转置；$\boldsymbol{U}(N\times r)$ 和 $\boldsymbol{Q}(1\times r)$ 分别为浓度矩阵 $\boldsymbol{Y}_{\text{train}}$ 的得分矩阵和载荷矩阵，u_i、$q_i(i=1,2,\cdots,r)$ 为对应的矩阵元；$\boldsymbol{E}$ 和 $\boldsymbol{F}$ 分别为光谱矩阵和浓度矩阵的残差矩阵。

定义 $\boldsymbol{W}(m\times r)$ 为光谱矩阵 $\boldsymbol{X}_{\text{train}}$ 的权重向量：

$$\boldsymbol{W}=\boldsymbol{X}_{\text{train}}^{\text{T}}\boldsymbol{U}/\boldsymbol{U}^{\text{T}}\boldsymbol{U}。\tag{2}$$

通过NIPALS得到相关矩阵向量的表达式为

$$\boldsymbol{W}=[w_1,w_2,\cdots,w_r],\boldsymbol{S}=[s_1,s_2,\cdots,s_r],$$
$$\boldsymbol{P}=[p_1,p_2,\cdots,p_r],\boldsymbol{Q}=[q_1,q_2,\cdots,q_r]。\tag{3}$$

光谱矩阵 $\boldsymbol{X}$ 的特征向量 $\boldsymbol{V}(m\times r)$ 为

$$\boldsymbol{V}=\boldsymbol{W}(\boldsymbol{P}^{\text{T}}\boldsymbol{W})^{-1}。\tag{4}$$

以 $\boldsymbol{X}_{\text{test}}(N_{\text{test}}\times m)$ 为测试集，其中 N_{test} 为测试集 $\boldsymbol{X}_{\text{test}}$ 包含的光谱样本数。通过特征向量得到投影矩阵 $\boldsymbol{X}_0(N_{\text{test}}\times r)$，其性质和作用等同于训练集得分矩阵：

$$\boldsymbol{X}_0=\boldsymbol{X}_{\text{test}}\boldsymbol{V}=\boldsymbol{X}_{\text{test}}\boldsymbol{W}(\boldsymbol{P}^{\text{T}}\boldsymbol{W})^{-1}。\tag{5}$$

2.2　基于特征提取的ELM建模与预测

将采集痕量气体的近红外光谱数据分为训练集和测试集两个部分。基于特征提取的ELM算法分析痕量气体浓度的算法流程如图1所示。提取训练集数据的最优主元，利用ELM对主元数据进行建模，将经过特征投影后的测试集光谱数据代入到模型中计算对应的浓度。

采用含有 L 个隐藏节点、隐含层输出权重为 $\boldsymbol{\beta}$、激发函数为 $g(\boldsymbol{S})$ 的ELM模型，光谱特征 $\boldsymbol{S}(N\times r)$ 在隐含层的输出向量表示为

$$f(\boldsymbol{S})=\sum_{j=1}^{L}\beta_j\times g(\boldsymbol{a}_i,\boldsymbol{b}_j,\boldsymbol{S})=\boldsymbol{\beta}\cdot\boldsymbol{h}(\boldsymbol{S}),\tag{6}$$

式中 β_j 为第 j 个隐含层的输出权重，$\boldsymbol{a}_i$ 为隐含节点和所有输入节点的输入权重向量，$\boldsymbol{b}_j$ 为第 j 个隐含层节点处的阈值，$\boldsymbol{h}(\boldsymbol{S})=[g(\boldsymbol{a}_1,\boldsymbol{b}_1,\boldsymbol{S}),\cdots,g(\boldsymbol{a}_r,\boldsymbol{b}_L,\boldsymbol{S})]^{\text{T}}$，$g(\boldsymbol{a}_i,\boldsymbol{b}_j,\boldsymbol{S})=g(\boldsymbol{S}\cdot\boldsymbol{a}_i+\boldsymbol{b}_j)$，$j=1,2,\cdots,L$。

选用机器学习中两种典型的激发函数

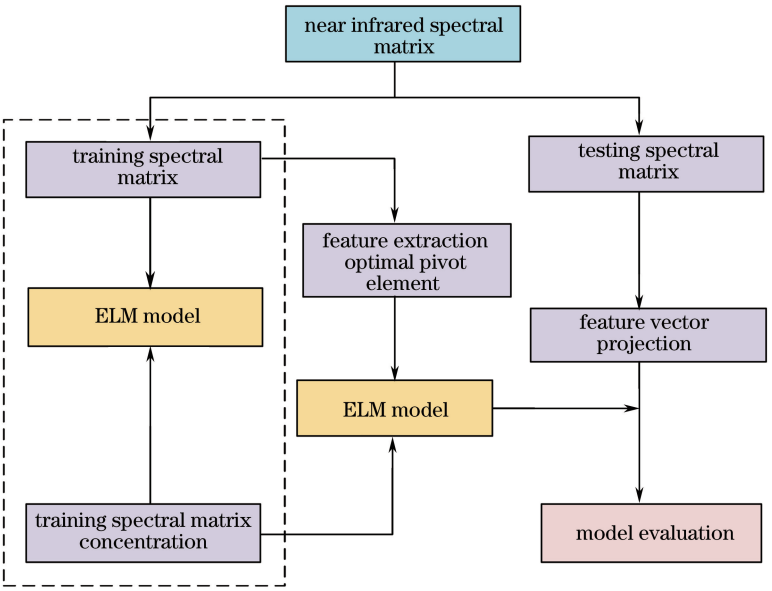

图 1　基于特征提取的 ELM 光谱分析算法流程图

Fig. 1　Flow chart of ELM algorithm for spectral analysis based on feature extraction

(Hardlim 函数和 Line 函数)验证 ELM 回归预测的性能。光谱数据与光谱浓度在 ELM 训练矩阵模型中可线性表示为

$$\boldsymbol{H}\boldsymbol{\beta}=\boldsymbol{Y}_{\text{train}}, \tag{7}$$

式中 $\boldsymbol{H}$ 为训练集光谱得分矩阵 $\boldsymbol{S}$ 通过激发函数映射的隐含层输出矩阵,其表达式为

$$\boldsymbol{H}=\begin{bmatrix}\boldsymbol{h}(s_1)\\ \vdots\\ \boldsymbol{h}(s_r)\end{bmatrix}=\begin{bmatrix}g(\boldsymbol{a}_1,\boldsymbol{b}_1,s_1) & \cdots & g(\boldsymbol{a}_L,\boldsymbol{b}_L,s_1)\\ \vdots & \ddots & \vdots\\ g(\boldsymbol{a}_1,\boldsymbol{b}_1,s_r) & \cdots & g(\boldsymbol{a}_L,\boldsymbol{b}_L,s_r)\end{bmatrix}_{r\times L}。 \tag{8}$$

光谱数据的训练集在 ELM 模型的输出权重矩阵 $\hat{\boldsymbol{\beta}}$ 可以表示为

$$\hat{\boldsymbol{\beta}}=\boldsymbol{H}^{\dagger}\boldsymbol{Y}_{\text{train}}, \tag{9}$$

式中 $\boldsymbol{H}^{\dagger}$ 为 $\boldsymbol{H}$ 的广义逆矩阵。

$\boldsymbol{X}_0$ 为测试集光谱数据经过特征提取后得到的投影矩阵,代入上述模型计算测试集对应的浓度:

$$y_{\text{p}}=\boldsymbol{H}(\boldsymbol{X}_0)\cdot\hat{\boldsymbol{\beta}}。 \tag{10}$$

2.3　模型性能评价

NIPALS 算法对光谱矩阵特征值的选取会影响回归模型的预测精度,选取的主元过少不能全面地表示光谱信息;主元过多会导致过拟合现象。在计算光谱特征时,选择五折交叉验证的方法确定最佳主元的个数,以衡量基于特征提取的 ELM 模型的稳健性。

对所有模型求得的测试集的浓度进行性能评估,采用模型训练时间、均方根误差(y_{rmse})和决定系数(R^2)共同评估[14-15]。

均方根误差与决定系数的表达式分别为

$$y_{\text{rmse}}=\sqrt{\frac{1}{N_{\text{test}}}\sum_{i=1}^{N_{\text{test}}}(y_i-y_{\text{p}i})^2}, \tag{11}$$

$$R^2=1-\frac{\sum_{i=1}^{N_{\text{test}}}(y_{\text{p}i}-y_i)^2}{\sum_{i=1}^{N_{\text{test}}}(y_i-\bar{y})^2}, \tag{12}$$

式中 y_i 为第 i 个测试集的实际值(即硫化氢浓度的实验设定值),$y_{\text{p}i}$ 为第 i 个样本的模型预测值,$\bar{y}$ 为设定值的平均值。

3　实　　验

为了验证基于特征提取的 ELM 模型的预测效果,以分析天然气中痕量硫化氢的应用为例,结合天然气脱硫处理的国家标准,设计了天然气中痕量硫化氢(体积分数 0～10^{-4})的分析仪,并开展了测试实验。实验装置的系统框图如图 2 所示。

根据高分辨率分子透射吸收(HITRAN)数据库中硫化氢以及天然气中甲烷、乙烷、CO_2 等组分在近红外的吸收峰分布情况,结合硫化氢浓度的测量需求以及天然气组分的典型分布,TDLAS 分析平台以波长为 1570 nm 的分布反馈式(DFB)激光器[16]作为光源;利用美国Thorlabs公司的法布里-

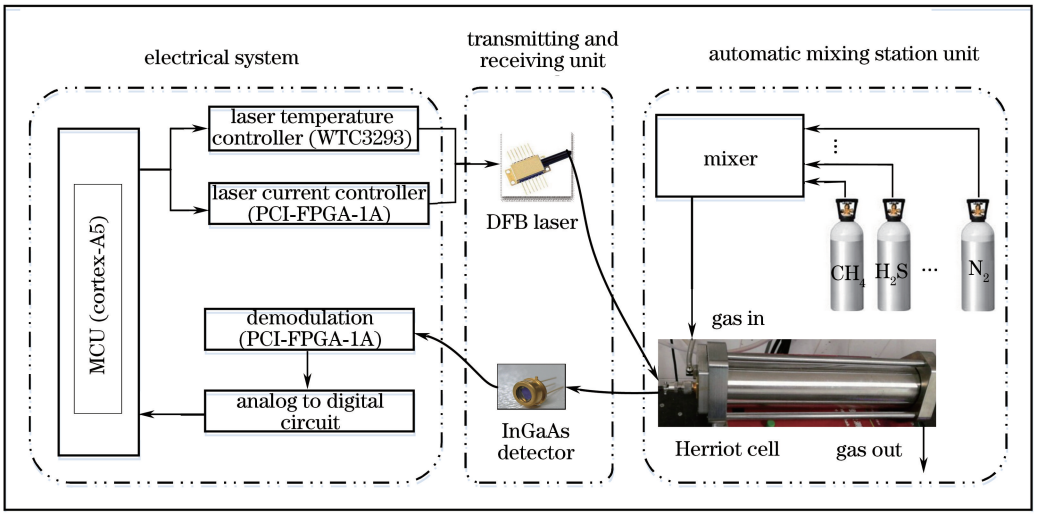

图 2　硫化氢 TDLAS 分析的实验装置

Fig. 2　Experimental device for TDLAS analysis of hydrogen sulfide

珀罗(F-P)扫描干涉仪 SA210-12B 测量驱动电流与谱峰位置的关系，得到电流调谐系数为 0.0154 nm/mA；利用北京光敏科技有限公司生产的 LSIPD-L1-L1 型 InGaAs 光电二极管探测器对透射光进行探测；以深圳米尔科技有限公司生产的 MYD-SAMA5D3X 型 Cortex-A5 作为微处理器(MCU)；利用美国 Port City Instrument 公司生产的 PCI-FPGA-1A 型外设部件互连-现场可编程门阵列(PCI-FPGA)模块实现对 TDLAS 技术可调谐半导体激光器的调制扫描，电流扫描范围为 60～90 mA，对应波长扫描范围为0.46 nm。激光经过聚焦准直进入气室，在 Herriot 气室中经多次反射到达探测器，光程为 20 m。探测到的信号经 PCI-FPGA 模块放大解调后生成表征气体吸收的二次谐波信号和反映光功率强弱的直流(DC)信号，再经模数转换发送到 MCU 进行浓度反演并建立校验模型。

光谱采集是在自动化配气站进行的，按照设计配成 92 组不同硫化氢浓度(体积分数，全文同)和不同组分的天然气背景，其中硫化氢的浓度为 0～10^{-4}，甲烷 70%～90%，二氧化碳 0%～3%，乙烷 0%～20%，以氮气作为平衡气体，范围在 0～30%。使用甲烷、乙烷、二氧化碳、氮气等高纯气瓶和有计量认证的以氮气为背景的含有 5.04×10^{-4} 硫化氢的标准混合气瓶，其计量可以溯源标定。配气系统使用北京七星华创公司生产的精度优于 1% 的 CS200 型数字式质量流量计，光谱采集是在气体充分混合的稳定状态下进行的。

为了提高模型的准确性，在气体混合均匀稳定时，采集多次光谱并取平均以抵消随机噪声，共获取 92 组数据，即 92 个样本。ELM 强调样本在测量范围内的随机性，以硫化氢为例，其浓度分配如图 3 所示，在取样过程中，其浓度在 0～10^{-4} 范围内是随机且宏观均匀分布的，符合模型使用的基础假设。

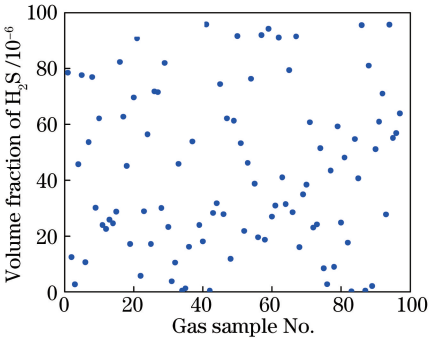

图 3　实验设计的硫化氢浓度分布

Fig. 3　Concentration distribution of hydrogen sulfide in experimental design

为了评估背景干扰的相对强度，在分析仪中采集了参考谱线，分别为 5.04×10^{-4} 的 H_2S 气体、100%浓度的 C_2H_6、100%浓度的 CH_4 和 3%浓度的 CO_2。如图 4(a)所示，满量程(5.04×10^{-4})的 H_2S 光谱吸收峰比 3%浓度的 CO_2 光谱吸收峰弱一个数量级。尽管 1570 nm 是硫化氢在近红外吸收最强的区域，但硫化氢的吸收峰相比其他气体组分的吸收结构仍然很弱，光谱干扰较大，传统的方法较难分析。

TDLAS 分析仪采用基于波长调制技术(WMS)提取二次谐波的数字解调方案，以降低测量系统中的低频噪声干扰，提高检测的灵敏度。由于采集的光谱受电源电流噪声和外部噪声的干扰较大，首先对气体检测光谱进行一系列处理(多次平

均、Savitzky-Golay 滤波[17-18]、光强归一化)，得到较为平滑的二次谐波吸收信号光谱。0～10^{-4}的 H_2S 气体在天然气背景下的 92 组光谱如图 4(b)所示，其中右侧较强峰为 CO_2 吸收峰，峰高分布缺少处对应 0～1%的浓度范围，主要是由于 CO_2 通道的质量流量计量量程偏大，无法产生 0～1%范围的 CO_2。

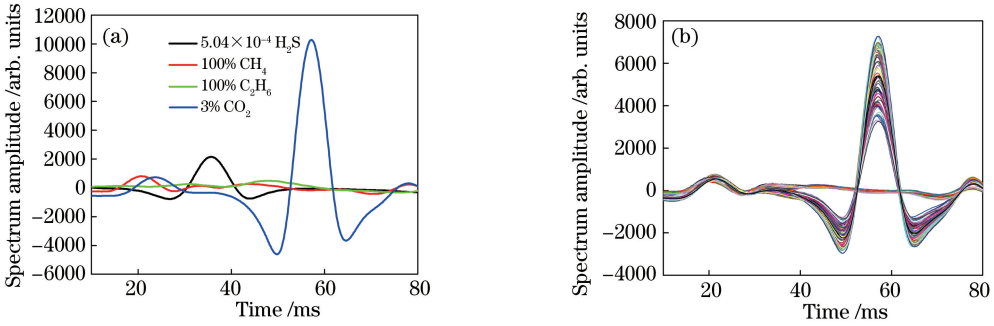

图 4　分析仪采集的光谱。(a)参考谱；(b)实测谱

Fig. 4　Spectra collected by analyzer. (a) Reference spectrum; (b) measurement spectrum

采集的 H_2S 和 CO_2 的参考谱峰值对应的点分别为 385 和 573，对应于 HITRAN 数据库的吸收峰波长为 1569.9179 nm 和 1570.0045 nm，1000 个采样点对应的光谱扫描范围为 0.46 nm，与利用激光器电流调谐系数得到的光谱扫描范围一致。半导体激光器的激光线宽一般为数兆赫兹。在电流调制的 TDLAS 分析平台上，线性区域的二次谐波信号的强度随着调制幅度的增大而增大，直至分子吸收光谱线宽的 2.2 倍。实验采用的调制电流为 6 mA，根据激光器的调谐系数，其有效线宽(即光谱分辨率)为 10.9 GHz。利用 HITRAN 数据库的分子吸收峰常数 γ_{self}(自加宽系数)和 γ_{air}(空气加宽系数)，计算在使用条件下三种气体的压力加宽和多普勒加宽，结果见表 1。三种分子的压力加宽基本上比多普勒加宽大一个数量级，故其吸收线型可以很好地近似为压力加宽主导的洛伦兹线型。其综合加宽的线宽为 4.5～5 GHz，激光器线宽约为吸收线宽的 2 倍。故在信号强度和光谱分辨率两个因素的取舍中，TDLAS 技术选择牺牲光谱分辨率以获得高信号强度，因此不可避免地带来了不同成分或同种成分的不同吸收峰的光谱叠合。

TDLAS 分析仪采集的每条光谱数据共 1000 个点，由于 PCI-FPGA 在扫描过程中有激光关闭和回扫环节，光谱两端不含硫化氢和天然气吸收的有效信息，故实验选择 101～800 区间的 700 个点对光谱数据进行分析。所有程序均在 MATLAB-R2014a 软件中运行。

表 1　光谱线型参数

Table 1　Spectral line parameters

Molecule	Wavenumber /cm^{-1}	γ_{self}	γ_{air}	Pressure broadening /GHz	Doppler broadening /MHz
H_2S	6369.760	0.158	0.074	4.443	403.834
CO_2	6369.408	0.088	0.069	4.192	354.970
CH_4	6370.382	0.079	0.066	4.602	588.743

4　实验数据分析与讨论

4.1　ELM 回归模型结果分析

只采用 ELM 回归模型，对采集的 92 组光谱进行分析。ELM 模型有两个变量：激发函数类别和隐含层节点个数。测试中选取不同的激发函数和隐含层节点对 TDLAS 分析仪采集的光谱数据进行分析，得到的模型性能参数见表 2。

ELM 回归模型对数据分析的结果显示，ELM 受隐含节点和激发函数种类的影响较大，激发函数类型对 ELM 的影响最为明显。由表 2 可知，当 ELM 激发函数为 line 且隐含节点为 1000 时，ELM 计算得到的决定系数为 0.9951，测量得到的均方根误差为 1.96×10^{-6}，训练时间为 0.12 s，是 ELM 算法达到的最佳效果。作为激发函数的 hardlim 函数是一种非线性函数，而在实验中无论是 0～10^{-4}的硫化氢光谱还是天然气中其他组分的干扰光谱，它们对最终过程气体的复合光谱的贡献都是线性叠加，即弱吸收假设。故 hardlim 激发函数的模型性能不如 line 函数的，因此在以下的特征提取时不再考虑该选项。随着隐含节点的个数增多，预测精度有所提高，但是当隐含节点增多到 10000 时，开始出现过拟合现象，预测精度开始变差，且隐含节点的个数与模型训练时间成正比例，隐含节点越多，训练时间越长。

表 2　ELM 算法结果

Table 2　Result by ELM

Algorithm	Excitation function	Number of nodes	Training time /s	y_{rmse} /10^{-6}	R^2
ELM	Line	200	<0.01	2.16	0.9941
		500	0.05	2.02	0.9948
		1000	0.12	1.96	0.9951
		5000	0.33	1.96	0.9951
		10000	0.70	1.97	0.9951
	Hardlim	5000	0.47	6.82	0.9409
		15000	1.84	6.77	0.9419
		20000	1.80	6.02	0.9864
		25000	3.38	6.17	0.9517

4.2　NIPALS-ELM 回归模型结果分析

TDLAS 分析系统采集的光谱数据点多，处于一个高维空间，在使用 ELM 回归分析时使用了大量的隐含节点，延长了建模时间且不利于回归模型的稳定，可以通过特征提取的方法将高维的数据映射（或变换）到低维空间来表示光谱数据。由第 3 节推导可知，图 4 中的吸收谱线为洛伦兹线型，由图 4 的参考谱线和实测吸收谱线可知，组分之间存在多组分交叉干扰，此时弱吸收的硫化氢气体的吸收谱峰不明显。

使用 NIPALS 法对原始的光谱数据进行特征提取，提取出最能代表原始数据的主成分。主成分提取的过程是寻找几个相互正交、方差最大的新变量来重新构造数据。在进行特征提取的过程中，每提取一个主成分，就会得到一个特征向量，利用前一次提取主成分后的残差矩阵进行下一次特征提取，当满足迭代条件后，再利用残差矩阵进行下一个主成分的提取。图 5 所示为模型从 92 组光谱选用最优主成分数为 3 时的光谱特征向量。第一个主成分的光谱特征向量与图4(b)所示的吸收谱线线型相似，第三个主成分的特征向量与硫化氢的光谱特征相似，故光谱不是简单的某个成分的吸收峰，而是这几个成分光谱的线性叠加。

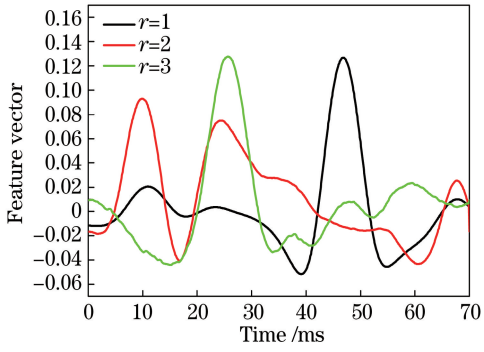

图 5　光谱矩阵的特征向量

Fig. 5　Feature vector of spectral matrix

根据各个主成分的方差贡献率选择最优主成分的个数，确定相应的特征向量，获取待测光谱矩阵在特征向量上的投影。ELM 模型利用最优主成分的得分向量与光谱浓度矩阵建立定量回归模型，然后利用待测光谱的投影矩阵预测光谱浓度，基于特征提取的 ELM 回归分析结果见表 3。

表 3　基于特征提取算法的 ELM 回归结果

Table 3　Regression result by ELM based on feature extraction

Algorithm	Excitation function	Number of hidden nodes	Training time /s	y_{rmse} /10^{-6}	R^2
NIPALS-ELM	Line	10	<0.01	1.55	0.9969
		20	<0.01	1.46	0.9973
		30	<0.01	1.46	0.9973
		200	<0.01	1.46	0.9973

利用基于特征提取的 ELM 算法对光谱数据进行分析，训练时间均低于 10 ms。特征提取后的近红外光谱数据代表了数据的全部特征且不含冗余信息，故在 ELM 算法中即使隐含节点个数无限增大，也不会出现过拟合现象。隐藏节点个数达到 20 时，模型的方差和决定系数均达到了最佳，其中模型方差为 1.46×10^{-6}，线性度拟合系数达到 0.9973，完全满足工业应用的 $0\sim10^{-4}$ 范围内 2% 的精度要求。该方法克服了直测谱数据维度高和共线性强的缺点，提高了在线分析仪的分析速率；对于激发函数为 line 的 ELM，当隐含节点大于最优的主成分个数时，模型不受隐含节点个数的影响，具有相同的预测

效果。

不同化学计量学方法的回归效果见表 4，模型的相关性如图 6 所示。

表 4　几种回归算法的比较

Table 4　Comparison among several kinds of regression algorithms

Algorithm	$y_{rmse}/10^{-6}$	R^2
CLS	18.19	0.5799
PLS	1.90	0.9954
ELM (line)	1.96	0.9951
NIPALS-ELM (line)	1.46	0.9973

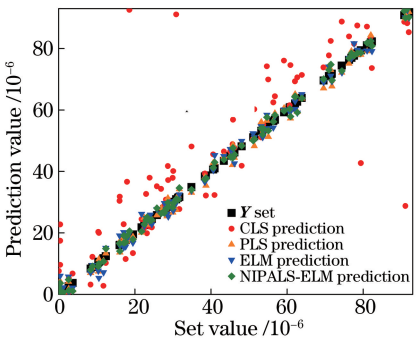

图 6　四种回归模型的预测值与设定值对比

Fig. 6　Comparison between predicted value and set value for four regression models

由于实验中设置的背景气体组分的变化范围超出了 CLS 算法的误差容许范围，CLS 算法的均方根误差达到了 1.8×10^{-5}，比基于特征提取的 ELM 回归算法的高一个数量级，因此气体参考谱回归建模的方法满足不了实时测量的需要，这也证明了 ELM 的稳健性。

PLS 算法使用光谱的特征信息和浓度进行建模，即逆校正（模型包含气体光谱与气体浓度相关的数据）的方法，有利于预测光谱浓度，因而 PLS 算法的均方根误差比直接校正的 CLS 算法的误差小。利用 PLS 算法得到的结果与 ELM 回归模型的数据相近。借助 NIPALS 法对直测谱进行特征提取，降低了原始光谱的复杂度，解决了传统 ELM 算法中光谱数据维度高、线性度强、计算量大的问题，进一步提高了分析的准确性。

由表 3、4 可知，光谱数据经过特征提取后，利用 ELM 模型定量回归分析的预测性能提高了 25.5%。由图 6 可知，基于特征提取的 ELM 回归相比于其他三种模型有更好的精度。

5　结　论

TDLAS 分析系统采集的光谱变量数据多，当使用 ELM 模型直接分析光谱数据时，需要较多的隐含层节点，延长了 TDLAS 分析仪的分析时间，限制了其在实时在线分析系统中的使用。提出了一种用于近红外光谱定量分析的基于特征提取的 ELM 回归模型，特征提取后的光谱数据去除了光谱的冗余信息，更能代表光谱的数据特征，更有利于 ELM 充分发掘数据光谱的浓度信息。

将基于特征提取的 ELM 回归模型应用于痕量硫化氢气体 TDLAS 检测中，实验结果表明，与 PLS 算法和 ELM 回归模型相比，基于特征提取的 ELM 回归模型在准确性、快速性方面具有显著的优势，可以应用于在线监测，具有重要的工程应用价值。

参　考　文　献

[1] Yao L, Liu W Q, Liu J G, *et al*. Research on open-path detection for atmospheric trace gas CO based on TDLAS [J]. Chinese Journal of Lasers, 2015, 42(2): 0215003.
姚路, 刘文清, 刘建国, 等. 基于 TDLAS 的长光程环境大气痕量 CO 监测方法研究[J]. 中国激光, 2015, 42(2): 0215003.

[2] Gao Y W, Zhang Y J, Chen D, *et al*. Measurement of oxygen concentration using tunable diode laser absorption spectroscopy [J]. Acta Optica Sinica, 2016, 36(3): 0330001.
高彦伟, 张玉钧, 陈东, 等. 基于可调谐半导体激光吸收光谱的氧气浓度测量研究[J]. 光学学报, 2016, 36(3): 0330001.

[3] Pan W D, Zhang J W, Dai J M, *et al*. Tunable diode laser absorption spectroscopy system for trace ethylene detection[J]. Spectroscopy and Spectral Analysis, 2012, 32(10): 2875-2878.
潘卫东, 张佳薇, 戴景民, 等. 可调谐半导体激光吸收光谱技术检测痕量乙烯气体的系统研制谱[J]. 光谱学与光谱分析, 2012, 32(10): 2875-2878.

[4] Bakeev K A. Process analytical technology: Spectroscopic tools and implementation strategies for the chemical and pharmaceutical industries[M]. Yao Z X, Chu X L, Su H, *et al*, Transl. 2th Edition. Beijing: China Machine Press, 2014: 322-356.
Bakeev K A. 过程分析技术：针对化学和制药工业的光谱方法和实施战略[M]. 姚志湘, 褚小立, 粟晖, 等, 译. 2 版. 北京：机械工业出版社, 2014: 322-356.

[5] Yang Y H, Li G L, Li X P, *et al*. Partial least squares algorithm application in TDLAS based trace H_2S analyses in natural gas[J]. Acta Photonica Sinica, 2017, 46(2): 0230002.
杨雅涵, 李国林, 李小鹏, 等. 基于 TDLAS 技术的

天然气中痕量硫化氢分析中的PLS算法应用[J]. 光子学报, 2017, 46(2): 0230002.

[6] Chen X L, Dong F Z, Wang J G, *et al*. Slag quantitative analysis based on PLS model by laser-induced breakdown spectroscopy[J]. Acta Photonica Sinica, 2014, 43(9): 093002.
陈兴龙, 董凤忠, 王静鸽, 等. PLS算法在激光诱导击穿光谱分析炉渣成分中的应用[J]. 光子学报, 2014, 43(9): 093002.

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

[8] Huang G B. An insight into extreme learning machines: Random neurons, random features and kernels[J]. Cognitive Computation, 2014, 6(3): 376-390.

[9] Huang G B, Zhou H M, Ding X J, *et al*. Extreme learning machine for regression and multiclass classification[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2012, 42(2): 513-529.

[10] Huang Z Y, Yu Y L, Gu J, *et al*. An efficient method for traffic sign recognition based on extreme learning machine [J]. IEEE Transactions on Cybernetics, 2017, 47(4): 920-933.

[11] Bian X H, Zhang C X, Tan X Y, *et al*. A boosting extreme learning machine for near infrared spectral quantitative analysis of diesel fuel and edible blend oil samples[J]. Analytical Methods, 2017, 9(20): 2983-2989.

[12] Javed K, Gouriveau R, Zerhouni N. SW-ELM: A summation wavelet extreme learning machine algorithm with a priori parameter initialization[J]. Neurocomputing, 2014, 123: 299-307.

[13] Wang C X, Hu J, Wen C L. A nonlinear PLS modeling method based on extreme learning machine [J]. Proceeding of the 34th Chinese Control Conference, 2015: 15440382.

[14] Malegori C, Marques N E J, Freitas S T, *et al*. Comparing the analytical performances of Micro-NIR and FT-NIR spectrometers in the evaluation of acerola fruit quality, using PLS and SVM regression algorithms[J]. Talanta, 2017, 165: 112-116.

[15] Hu B. Filtering optimization for on-line spectral data and TDLAS system integration[D]. Tianjin: Tianjin University, 2018: 8-21.
胡波. 在线光谱数据的滤波优化与TDLAS系统集成[D]. 天津: 天津大学, 2010: 8-21.

[16] Wang Q, Guo J J, Chen W, *et al*. Widely tunable distributed feedback semiconductor lasers with constant power and narrow linewidth[J]. Chinese Journal of Lasers, 2017, 44(1): 0101004.
王琪, 郭锦锦, 陈伟, 等. 功率稳定且波长可调谐的窄线宽分布式反馈半导体激光器[J]. 中国激光, 2017, 44(1): 0101004.

[17] Otto M. Chemometrics[M]. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA, 2007, 51-56.

[18] Zou D B, Chen W L, Du Z H, *et al*. Selection of digital filtering in the escaping ammonia monitoring with TDLAS [J]. Spectroscopy and Spectral Analysis, 2012, 32(9): 2322-2326.
邹得宝, 陈文亮, 杜振辉, 等. 数字滤波方法在TDLAS逃逸氨检测中的应用[J]. 光谱学与光谱分析, 2012, 32(9): 2322-2326.