光谱分析中 Elastic Net 变量选择与降维方法

赵安新[1,2],汤晓君[2],宋 娅[3],张钟华[2,4],刘君华[2]

(1. 西安科技大学,陕西 西安 710054; 2. 西安交通大学 电力设备电气绝缘国家重点实验室,陕西 西安 710049; 3. 中国航空工业西安飞机工业(集团)有限责任公司,陕西 西安 710089; 4. 中国计量科学研究院,北京 100013)

摘 要:在利用红外光谱进行多组分混合气体定量分析建模中,须根据各目标气体成分的光谱特点进行光谱维数降维和特征变量选择。以甲烷、乙烷、丙烷、异丁烷、正丁烷、异戊烷和正戊烷等 7 种气体为分析目标,采用最小绝对收缩和选择算子(LASSO)与弹性网络(Elastic Net)方法进行目标气体数据预处理。针对 LASSO 和 Elastic Net 方法参数优化选择的问题,采用均方误差和预测偏差最小两个准则进行参数的优化选取。对 4cm^{-1} 的实测光谱数据,采用 LASSO 和 Elastic Net 方法分别在 0.0019 和 0.0021 均方误差条件下使得维度从 2542 维分别降为 2 维和 3 维,LASSO 的交叉灵敏度最大和最小为 10.2718%和 1.4205%,Elastic Net 分别为 5.4945%和 0.7493%。结果表明:Elastic Net 在用于光谱定量分析的数据预处理中具有一定的优势,为准确建立定量分析模型奠定了基础。

关键词:气体红外光谱定量分析; 正则化算法; 特征波长选择; LASSO; Elastic Net

中图分类号:O433.4 **文献标志码**:A **文章编号**:1007-2276(2014)06-1977-05

Spectral wavelength selection and dimension reduction using Elastic Net in spectroscopy analysis

Zhao Anxin[1,2], Tang Xiaojun[2], Song Ya[3], Zhang Zhonghua[2,4], Liu Junhua[2]

(1. Xi'an University of Science and Technology, Xi'an 710054, China; 2. State Key Laboratory of Electrical Insulation and Power Equipment Xi'an Jiaotong University, Xi'an 710049, China; 3. AVIC Xi'an Aircraft Industry (group) Company Ltd, Xi'an 710089, China; 4. National Institute of Metrology, Beijing 100013, China)

Abstract: In the use of Fourier transform infrared spectroscopy to build the multi-component gases quantitative analysis model, it is necessary to reduce the dimensions and select characteristics wavelength according to the target gas spectral. Through the regularization algorithm analysis, least absolute shrinkage and selection operator (LASSO) and Elastic Net method were used to do these for seven kinds of mixed gases of methane, ethane, propane, iso-butane, n-butane, iso-pentane and n-pentane. The minimum mean square error (MSE) and prediction deviation were used as the criteria to select LASSO and Elastic Net parameters. Finally, the resolution of 4cm^{-1} measured spectral data was analyzed. The dimension of spectra were reduced from 2 542 d to 2d and 3d respectively by using LASSO and Elastic Net method

收稿日期:2013-10-12; **修订日期**:2013-11-15

基金项目:国家重大科学仪器设备开发专项(2012YQ240127);国家自然科学基金(51277144);电力设备电气绝缘国家重点实验室基金(EIPE11307)

作者简介:赵安新(1981-),男,博士,讲师,主要从事多传感数据融合及信息处理方面的研究。Email:zhaoanxin@126.com

under the condition of the MSE of 0.001 9 and 0.002 1. The cross sensitivity of maximum and minimum were 10.271 8% and 1.420 5% by LASSO method. The cross sensitivity of maximum and minimum were 5.494 5% and 0.749 3% by Elastic Net. Results show that the Elastic Net method was better in the characteristic variable selection and the spectral dimension reduction for gas spectral quantitative analysis, and it was foundation to establish the accurate quantitative analysis model.

Key words: gases infrared spectroscopy quantitative analysis; regularization algorithm; characteristic wavelength selection; LASSO; Elastic Net

0 引 言

傅里叶变换红外光谱分析技术由于其快速、无损、无需化学处理和能够多组分同时分析的特点广泛应用于生物医药、石油化工、电力电子等各领域物质的定性和定量分析[1-3],其中定量分析的基本原理是依据 Lamber-Beer 定律来建立定量分析模型,在利用红外光谱来建立定量分析模型中,尤其是高分辨率光谱仪器产生的高维光谱数据,首先需要对原始数据进行光谱降维和特征变量的选择[4-5]。目前,常用的光谱降维和特征变量选择方法[6-7]主要有主元分析(PCA)、最小二乘法(LS)、偏最小二乘(PLS)、多元线性回归(MIR)、逐步回归(SRA)、最小角回归(LAR)、LASSO 和 Elastic Net 等方法。主元分析和偏最小二乘算法是通过对原始变量经过正交旋转变化将原变量映射为少数主成分的多元统计方法,映射后的主成分代表原始主要信息,然后通过多元回归建立回归分析模型,其分析的本质还是线性回归,并只是对原始变量信息进行映射转换,并未改变原始变量,映射后的主成分是原始变量的线性组合。最小角回归(LAR)是 Efron 于 2004 年提出的类似于向前逐步回归(Forward Stepwise)的降维方法,与 Forward Stepwise 不同在于 Forward Stepwise 每次都是根据选择的变量子集,拟合出线性模型,根据 RSS 设计统计量对较高的模型复杂度进行惩罚,而 LAR 是每次先找出和因变量相关度最高的变量,调整变量的系数,使得变量和残差的相关系数逐渐减小,达到一定程度选择此变量作为相关性最高的变量,然后重新选择知道达到最终结果。文中针对分析的目标气体组分:甲烷、乙烷、丙烷、异丁烷、正丁烷、异戊烷和正戊烷等烷烃类气体,其在中红外波段由于其分子结构相近、分子基团相同的特点使得吸收峰严重交叠,常用的降维和特征变量选择方法不易区分和进行特征变量的选择,通过对解决不适定问题的正则化算法的分析,采用 LASSO 和 Elastic Net 对分析目标气体进行光谱降维和特征波长的选择。

1 LASSO 和 Elastic Net 基本原理

红外光谱定量分析建模中通常依据 Lamber-Beer 定律建立分析模型,在小浓度范围内[8],可以依据吸光度与浓度的正比关系建立分析模型,其分析模型如下式所示:

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e} \tag{1}$$

式中:$\boldsymbol{X}$:$\boldsymbol{n}\times\boldsymbol{p}$ 为矩阵,包含 $\boldsymbol{n}$ 个标定样本在 $\boldsymbol{p}$ 条谱线上的光谱,一般 $\boldsymbol{n}<<\boldsymbol{p}$;$\boldsymbol{\beta}$:$\boldsymbol{p}\times 1$ 为回归向量;$\boldsymbol{y}$ 为分析目标气体浓度;$\boldsymbol{e}$ 为随机误差。

公式(1)是典型的线性回归问题,通常针对线性回归问题中样本数目远大于变量即 $\boldsymbol{n}>>\boldsymbol{p}$。然而对于本文分析的问题,采用中红外光谱数据作为自变量,其数目远远大于样本数,以工业现场[9]通常使用的分辨率为 4 cm^{-1} 的光谱仪在中红外波段的谱线数量为 2 542 条,即 2 542 维数据,针对此类问题的求解方式称为病态方程,其主要问题是当数据项 $\boldsymbol{y}$ 产生较小的扰动 δ_y 时,数值解将发生较大的变化。然而正则化算法[10]正式基于这个问题,通过算法搜索一个逼近于原始问题的适定算子,得到适定算子的解即可认为该解同样逼近于原始问题解。使得常用的最小误差最小化准则转化成下式求解:

$$\min(\|\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{y}\|_{\boldsymbol{a}}^{\boldsymbol{a}}+\boldsymbol{\lambda}\|\boldsymbol{L}\boldsymbol{\beta}\|_{\boldsymbol{b}}^{\boldsymbol{b}}) \tag{2}$$

式中:$\|\cdot\|_{\boldsymbol{p}}$-$\boldsymbol{p}$ 为范数;$\boldsymbol{a}$ 为回归偏差的范数,$1\leqslant\boldsymbol{a}<\infty$;$\boldsymbol{b}$ 为归向量的范数,$1\leqslant\boldsymbol{b}<\infty$;$\boldsymbol{L}$ 为正则化算子;λ 控制第 2 项相对于第 1 项的权值。

当 $\boldsymbol{a}=2$,$\boldsymbol{b}=1$ 时即为 LASSO 正则化方法[11],见公式(3)。LASSO 方法用模型回归系数的 $\boldsymbol{L}^1$ 范数作为惩罚来压缩模型回归系数,使绝对值较小的系数压

缩为 0，从而实现显著性变量选择和对应参数的估计。然而由于凸优化的性质,容易出现过拟合问题。

$$\hat{\boldsymbol{\beta}}=\arg\min_{\beta}\left(\|\mathbf{X}\boldsymbol{\beta}-\mathbf{y}\|_2^2+\lambda\|\mathbf{L}\boldsymbol{\beta}\|_1^1\right) \quad (3)$$

Zou 等作者针对此问题，提出 Elastic Net 方法，见公式(4),通过采用在 L^1 范数和 L^2 范数相结合的方式，并增加惩罚项 α，使得 Elastic Net 和 LASSO 优点相结合,以高预测精确度选择稀疏模型,同时解决群组效应问题。

$$\hat{\boldsymbol{\beta}}=\arg\min_{\beta}\left\{\left(\|\mathbf{X}\boldsymbol{\beta}-\mathbf{y}\|_2^2+\lambda\frac{(1-\alpha)}{2}\|\boldsymbol{\beta}\|_2^2+\alpha\|\boldsymbol{\beta}\|_1^2\right)\right\} \quad (4)$$

在使用 LASSO 和 Elastic Net 方法时,需要对惩罚项的系数 λ 和 α 进行合理的选定,文中对 λ 的选定依据广义交叉验证最小化的方式来确定,α 的选定依据训练样本均方误差和预测偏差均方误差最小的准则进行选定。

2　实验数据准备

实验所用仪器为傅里叶变换红外光谱仪 alpha (BRUKER Ltd)：扫描范围设置为 400~4 000cm^{-1},光谱波数分辨率为 4 cm^{-1},波数精度为 0.1 cm^{-1}。

实验所采集的目标气体为甲烷 (CH_4)、乙烷 (C_2H_6)、丙烷 (C_3H_8)、异丁烷 (iso-C_4H_{10})、正丁烷 (n-C_4H_{10})、异戊烷 (iso-C_5H_{12}) 和正戊烷 (n-C_5H_{12}) 等 7 种轻烷烃类。根据分析的需要,设定标定目标样本气分别为 0.01%、0.02%、0.05%、0.1%、0.2%、0.5%、1%。通过不同浓度单组分气体的观察，如图 1 所示,烷烃在 2 750~3 200 cm^{-1} 范围内具有较强的吸收,在低于 1%浓度范围其吸光光度和浓度成一定的线性关系。通过同浓度 0.1%的 7 组分气体的观察，如图 2 所示,由于其分子结构相近,分子基团相同,其吸收光谱严重交叠,各种目标分析气体相互干扰。

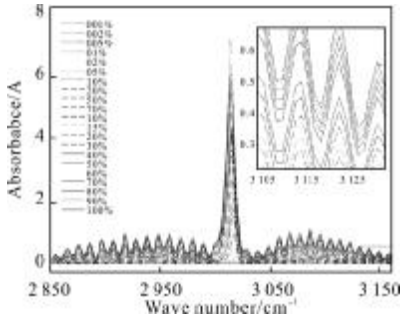

图 1 甲烷目标气体 0.01%~100%浓度的吸光度光谱

Fig.1 CH_4 absorption spectra of concentration 0.01%-100%

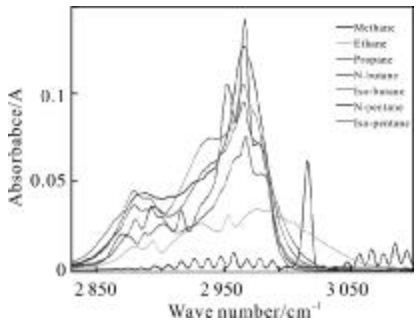

图 2 7 种目标气体 0.1%浓度的主吸收峰吸光度光谱

Fig.2 7 kinds gases first peak absorption spectra of concentration 0.1%

3　数据预处理

光谱仪器在长时间在线运行分析中，不可避免受温度、湿度和气压的影响,使检测的谱图产生畸变和漂移。文中采用课题组前期研究成果,采用分段比基线校正方法[12]对光谱数据进行预处理。该方法基本原理是找出非敏感区波段,对灵敏度较低的波段,采用线性化处理。对于第 i 种气体在第 j 个非敏感波段的灵敏度 $\mathbf{S}_{ij}$,可用下式确定：

$$\mathbf{S}_{ij}=(1-\mathrm{mean}(\mathbf{v}_{ij}))/\mathbf{c}_i \quad (5)$$

式中：$\mathbf{v}_{ij}$ 表示第 i 种气体在第 j 个非敏感波段若干连续谱线值；$\mathbf{c}_i$ 表示第 i 种气体中最高浓度。

根据估算的灵敏度，在各相邻区间内进行基线校正。

$$\begin{aligned}\mathbf{baseline}_j&=\mathrm{mean}(\mathbf{v}_j)+\mathbf{S}_j\mathbf{C}'\\ \mathbf{baseline}_{j+1}&=\mathrm{mean}(\mathbf{v}_{j+1})+\mathbf{S}_{j+1}\mathbf{C}'\\ \mathbf{rate}_j&=(\mathbf{baseline}_{j+1}-\mathbf{baseline}_j)/(\mathbf{num}_{j+1}-\mathbf{num}_j)\\ \mathbf{v}_{j,j+1}(\mathbf{num}_j:\mathbf{num}_{j+1})&=\mathbf{v}_{j,j+1}(\mathbf{num}_j:\mathbf{num}_{j+1})/(\mathbf{baseline}_j+\cdots+\\ &\mathbf{rate}_j)\times((\mathbf{num}_j:\mathbf{num}_{j+1})'-\mathbf{num}_j))\end{aligned} \quad (6)$$

式中：$\mathbf{baseline}_j$ 为第 j 个非敏感波段的基线；$\mathbf{S}_j$ 为不同组分气体在第 j 个非敏感波段的灵敏度向量；$\mathbf{C}$ 为气体浓度向量；$\mathbf{num}_j$ 表示第 j 个非敏感波段中心谱线序号。

4　计算过程

4.1　LASSO 和 Elastic Net 中参数设定

文中分段比基线校正同时实现对基线漂移和畸变进行校正，校正的效果如图 3 所示。

在通过公式(3)、(4)进行求解之前,需要确定公式中的惩罚权值 λ 和 α,文中对公式(3)采用依据广义交叉验

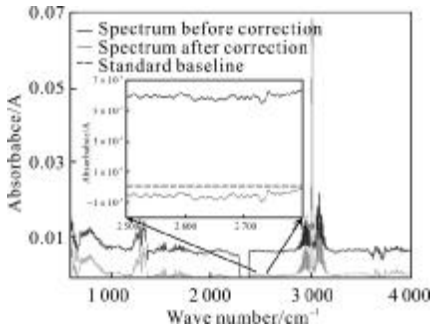

图 3 0.1%浓度甲烷气体吸光度光谱发生漂移和校正情况

Fig.3 Absorption spectra of CH_4 concentration 0.1% for baseline wander corrected and uncorrected

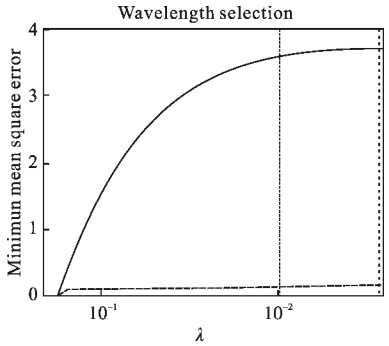

图 4 甲烷目标气体特征波长选择与 LASSO 参数之间关系

Fig.4 Wavelength selection for CH_4 vs LASSO parameter

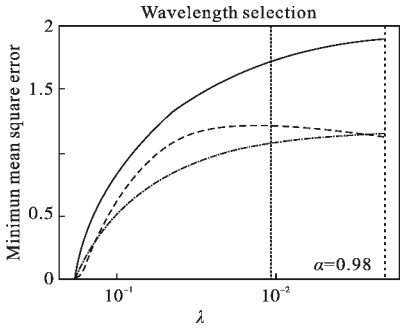

图 5 甲烷目标气体特征波长选择与 Elastic Net 参数之间关系

Fig.5 Wavelength selection for CH_4 vs Elastic Net parameter

证最小化的方式来确定惩罚权值 λ,对与公式(4)中 α 为 1 时,其实就是公式(3)也即是 LASSO 算法,因此对 α 设定 0~1 之间,采用遍历搜索,并依据训练样本均方误差和预测偏差均方误差最小的准则进行来选定。

4.2 讨论分析

按照设定的惩罚权重参数,将甲烷、乙烷、丙烷、异丁烷、正丁烷、异戊烷和正戊烷等 7 种红外光谱样本,每种气体配置 6 种浓度(小浓度范围),每组分浓度气体采集 45 组,共计 1 890 组样本。以甲烷特征气体为例,分析其参数选择过程及其最优结果,其中图 4 是甲烷气体成分选用 LASSO 方法时特征波长选择、广义交叉验证最小化与惩罚权值 λ 之间的关系;其中图 5 是甲烷气体成分选用 Elastic Net 方法最优特征波长选择、惩罚权值 α 与惩罚权值 λ 之间的关系。经过多次统计运行分析,使用 LASSO 方法在 λ 为 0.009 5,选择 2 条特征谱线时,分析结果均方误差最小,最小均方误差为 0.001 9,此时与其他气体成分的交叉灵敏度最大值为 10.271 8%,最小为 1.420 5%。使用 Elastic Net 方法在 α 为 0.980 0,λ 为 0.010 7,选择 3 条特征谱线时,分析结果均方误差最小,最小均方误差为 0.002 1,此时与其他气体成分的交叉灵敏度最大值为 5.494 5%,最小为 0.749 3%,见表 1。采用 Elastic Net 方法可以有效进行气体光谱分析中特征波长的选择和降维,同时降低气体间的交叉灵敏度,为准确建立定量分析模型建立基础。同时由于 Elastic Net 方法需要对 α 参数进行优化选择,所以运行时间会比 LASSO 方法稍长。

表 1 0.2%的甲烷检验结果及其交叉灵敏度

Tab.1 Result for 0.2% CH_4 and CSC with other gases

Gas composition	Methane	Ethane	Propane	Iso-butane	N-butane	Iso-pentane	N-pentane
LASSO Predicted concentration/%	0.200 6	-0.004 4	-0.015 3	-0.002 8	-0.020 6	-0.008 6	-0.006 8
LASSO cross sensitivity/%	-	2.203 5	7.607 3	1.420 5	10.271 8	4.307 8	3.380 4
Elastic Net Predicted concentration/%	0.200 2	-0.002 3	-0.008 2	-0.001 5	-0.011 0	-0.004 6	-0.003 6
Elastic Net cross sensitivity/%	-	1.148 9	4.095 9	0.749 3	5.494 5	2.297 7	1.798 2

5 结 论

在利用红外光谱建立定量分析模型中,为提高计算精度和减少计算量,首先需要对光谱数据进行预处理,即光谱维数降维和特征变量选择。文中根据气体红外定量分析中的需要,结合正则化算法,分别

采用 LASSO 和 Elastic Net 方法对气体红外光谱进行降维和特征变量的选择。经过对 4 cm^{-1} 的实测光谱数据的检验,采用 LASSO 和 Elastic Net 方法分别在 0.001 9 和 0.002 1 均方误差条件下使得光谱维度从 2542 维分别降为 2 维和 3 维,LASSO 计算方法的交叉灵敏度最大和最小分别为 10.271 8%和 1.420 5%, Elastic Net 计算方法分别为 5.494 5%和 0.749 3%。LASSO 和 Elastic Net 方法比较适合应用与文中所分析的目标混合气体, 在吸收峰交叠相对比较严重的波数段特征波长选取 Elastic Net 具有一定的优势,然而由于 Elastic Net 需要进行参数 α 的优化选取,其运行时间比 LASSO 方法稍长,但随着并行计算的应用,该问题可以进一步优化。

参考文献:

[1] Materazzi S, Vecchio S. Recent applications of evolved gas analysis by infrared spectroscopy (IR-EGA)[J]. Applied Spectroscopy Reviews, 2013, 48(8): 654-689.

[2] Sepman A V, den Blanken R, Schepers R, et al. Quantitative fourier transform infrared diagnostics of the gas-phase composition using the HITRAN database and the equivalent width of the spectral features [J]. Appl Spectrosc, 2009, 63(11): 1211-1222.

[3] Xu Xiaojing, Huang Wei. Application of spectral imaging in forensic science [J]. Infrared and Laser Engineering, 2012, 41(12): 3280-3284. (in Chinese)
许小京, 黄威. 光谱成像技术在物证鉴定领域的应用[J]. 红外与激光工程, 2012, 41(12): 3280-3284.

[4] Kalivas J H. Multivariate calibration, an overview [J]. Analytical Letters, 2005, 38(14): 2259-2279.

[5] Kunz M R, Ottaway J, Kalivas J H, et al. Impact of standardization sample design on Tikhonov regularization variants for spectroscopic calibration maintenance and transfer [J]. Journal of Chemometrics, 2010, 24(3-4SI): 218-229.

[6] Zeng T, Wen Z, Wen Z, et al. Weighted fusion of multiple models for wavelength selection[J]. Appl Spectrosc, 2013, 67(7): 718-723.

[7] Zou H, Hastie T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society Series B-statistical Methodology, 2005, 67(Part 2): 301-320.

[8] Tang Xiaojun, Zhang Lei, Wang Erzhen, et al. An improved characteristic spectral selection method for multicomponent gas quantitative analysis based on tikhonov regularization[J]. Spectroscopy and Spectral Analysis, 2012, 32(10): 2730-2734. (in Chinese)
汤晓君, 张蕾, 王尔珍, 等. 一种改进型多组分气体的 Tikhonov 正则化特征光谱提取方法 [J]. 光谱学与光谱分析, 2012, 32(10): 2730-2734.

[9] Wang Gaofeng, Zhao Yiqiang, Yang Dong. Data acquisition of 1 024-pixel long linear infrared detectors[J]. Infrared and Laser Engineering, 2012, 41(8): 1990-1994. (in Chinese)
王高峰, 赵毅强, 杨栋. 1 024 元长线列红外探测器的数据采集技术[J]. 红外与激光工程, 2012, 41(8): 1990-1994.

[10] Friedman J H, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent [J]. Journal of Statistical Software, 2010, 33(1): 1-22.

[11] Dyar M D, Carmosino M L, Breves E A, et al. Comparison of partial least squares and lasso regression techniques as applied to laser-induced breakdown spectroscopy of geological samples [J]. Spectrochimica Acta Part B-atomic Spectroscopy, 2012, 70: 51-67.

[12] Tang Xiaojun, Wang Jin, Zhang Lei, et al. Spectral baseline correction by piecewise dividing in fourier transform infrared gas analysis [J]. Spectroscopy and Spectral Analysis, 2013, 33(2): 334-339. (in Chinese)
汤晓君, 王进, 张蕾, 等. 气体光谱分析应用中傅里叶变换红外光谱基线漂移分段比校正方法 [J]. 光谱学与光谱分析, 2013, 33(2): 334-339.