文章编号:1001-9014(2009)02-0115-04

核偏最小二乘特征提取在混合气体FTIR光谱定量分析中的应用

郝惠敏[1,2], 乔聪明[2], 汤晓君[1], 刘君华[1]

(1. 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;
2. 太原钢铁公司自动化公司,山西 太原 030003)

摘要:为进一步提高FTIR光谱法实现特征吸收光谱严重重叠的甲烷、乙烷、丙烷、异丁烷、正丁烷、异戊烷以及正戊烷七组分混合气体定量分析的精度和速度,提出一种核偏最小二乘(Kernel Partial Least Square, KPLS)特征提取耦合支持向量回归机(Support Vector Regression Machine, SVR)的红外光谱定量分析新方法.首先采用KPLS方法对上述七组分混合气体的FTIR光谱进行特征提取,然后将特征提取得到的特征组分作为SVR的输入建立混合气体的定量分析模型.对标准混合气体进行定量分析的结果显示:KPLS-SVR模型的预测精度高于未进行特征提取SVR模型预测的精度,同时预测时间也减少了一半.研究表明,KPLS法可以很好地提取隐含在混合气体FTIR光谱数据与其组分浓度之间的非线性特征并有效地消除光谱数据噪声,大幅度降低数据维数,与SVR耦合可以提高红外光谱分析的精度和速度,是一种有效的红外光谱定量分析方法.

关 键 词:核偏最小二乘;支持向量回归机;特征提取;多变量校正模型;红外傅里叶变换(FTIR)
中图分类号:TE642;TH744.4 **文献标识码**:A

APPLICATION OF KERNEL PARTIAL LEAST SQUARE FEATURE EXTRACTION TO QUANTITATIVE ANALYSIS OF FTIR SPECTROSCOPY OF MULTI-COMPONENT GAS MIXTURE

HAO Hui-Min [1,2], Qiao Cong-Ming[2], TANG Xiao-Jun [1], LIU Jun-Hua [1]

(1. State Key Laboratory of Electrical Insulation for Power Equipment, Xi'an Jiaotong University, Xi'an 710049, China;
2. Taiyuan Iron and Steel Co., Automation Company, Taiyuan 030003, China)

Abstract: A new method for FTIR spectral quantitative analysis was presented. The new method couples kernel partial least squares(KPLS) feature extraction with support vector regression machine(SVR) to improve the quantitative analysis accuracy and speed of seven-component alkane gas mixtures composed of methane, ethane, propane, iso-butane, n-butane, iso-pentane, and n-pentane, whose feature absorption spectra are cross each other and overlapped seriously. Firstly, the KPLS was employed to extract feature components from the FTIR spectra of above-mentioned seven-component gas mixtures. And then, the extracted feature components were fed into SVR to create the quantitative analysis model of seven component gases. The quantitative analysis results of calibration gas mixtures show that the prediction accuracy by KPLS-SVR model is higher than that by SVR model without feature extraction processing. Meanwhile, the predicting time by KPLS-SVR model is only half of that by SVR model. The study indicates that KPLS approach can effectively extract the latent nonlinear features implied in the spectra and component concentration, eliminate the noise of FTIR spectral data, and reduce the dimension of the spectral data. Coupling with SVR, KPLS feature extraction can improve the accuracy of FTIR spectral analysis, shorten the predicting time. KPLS-SVR is a very effective method for infrared spectral quantitative analysis.

Key words: kernel partial least squares; support vector regression machine; feature extraction; multivariable calibration model; FTIR

收稿日期:2008-02-21,**修回日期**:2008-06-18 **Received date**: 2008-02-21, **revised date**: 2008-06-18
基金项目:国家自然科学基金(60276037)资助项目
作者简介:郝惠敏(1971-),女,山西太原人,西安交通大学电气工程学院博士生,研究方向:光电传感与检测,多传感信息融合. E-mail: helenwangmin@gmail.com.

引言

采用FTIR光谱法实现对甲烷、乙烷、丙烷、异丁烷、正丁烷、异戊烷以及正戊烷七种烷烃混合气体的浓度测量，可以克服传统色谱法分析的诸多缺点和不足. 但是，由于各组分气体吸收谱线交叉重叠异常严重，组分间存在着严重的交叉敏感，组分浓度与吸光度之间呈现非线性关系，以及系统误差和外界环境变化引起的噪声影响等因素，使得对它们的分析十分困难. 白鹏[1,2]等人将支持向量回归机[3,4](Support Vector Regression Machine，SVR)应用于红外光谱的定量分析中，实现了对上述混合气体组分浓度的定量分析，各组分预测的平均绝对误差(Mean Absolute Error，MAE)为0.132%. 该成果在七组分烷烃混合气体定量分析中取得了突破性进展，由于实际应用中，模型的预测精度仍不够理想，并且随着原始光谱数据维数的增大，SVR模型会出现计算速度减慢，参数优化困难等问题. 针对这些情况，一个可行的解决方案就是对光谱数据矩阵进行高效的特征提取[5,6]，从而充分利用光谱数据中的有用信息，并结合适当的非线性建模方法，实现以上七种气体的定量分析.

本文将核偏最小二乘(Kernel Partial Least Square，KPLS)算法[7-13]用于红外光谱的特征提取，并结合SVR建立了上述七组分混合气体的定量分析模型. KPLS-SVR模型的分析能力由模型对检验集样本的预测结果进行评定，并与未进行特征提取的SVR方法进行了比较. 研究表明，KPLS对原始光谱数据具有独特的非线性提取能力，可以消除原始光谱数据的噪声，有效降低光谱数据维数，与SVR结合可以建立准确的七组分烷烃混合气体的定量分析模型.

2 实验部分

2.1 实验仪器

采用由16位数字流量控制器(Alicat Scientific公司，美国)(量程为0～0.5SCCM到0～1500SLPM，精度为±1% Full Scale)组成的高精度配气系统，根据被测现场的气体分布模式，制备标准混合气体样本. 其中，各种组分为浓度高于99.95%的标准纯气体，稀释气体为浓度高于99.99%的氮气.

红外光谱仪为德国Bruker Optics公司生产的TENSOR27型FTIR红外光谱仪，配用气室长11cm. 扫描范围为4000 cm^{-1}～400 cm^{-1}，扫描间隔为2 cm^{-1}. 将每一个混合物样本重复扫描20次的平均值用于进一步分析.

表1 混合气体样本各组分气体的浓度范围及浓度变化间隔

Table 1 Concentration range and changing intervals of seven single-component gases in preparation gaseous mixture samples

组分气体	浓度范围(%)	最小浓度间隔 /μL·L^{-1}	最大浓度间隔(%)
甲烷	0.01～1.00	25	0.99
乙烷	0.00～0.45	25	0.45
丙烷	0.00～0.35	25	0.35
异丁烷	0.00～0.25	25	0.25
正丁烷	0.00～0.20	25	0.20
异戊烷	0.00～0.15	25	0.15
正戊烷	0.00～0.10	25	0.10

表2 九种混合气体各组分的浓度配比

Table 2 The concentration ratio of nine gas mixture samples

Samples	Concentration /μL·L^{-1}						
	CH_4	C_2H_6	C_3H_8	iso-C_4H_{10}	n-C_4H_{10}	iso-C_5H_{12}	n-C_5H_{12}
1	100	100	75	75	75	50	25
2	125	100	75	75	50	50	25
3	175	125	10	50	25	25	25
4	450	200	200	50	50	50	0
5	650	150	100	50	50	0	0
6	2750	1000	250	250	250	250	250
7	6500	1000	500	500	500	500	500
8	8500	500	500	500	0	0	0
9	10000	0	0	0	0	0	0

2.2 气体样本及光谱数据预处理

众所周知，被分析的组分浓度越低，则分析的难度越大. 文中所用混合气体样本的浓度范围为0.01～1%，气体混合物样本共1842个，每一个样本光谱的维数为1866，混合气体样本的组分浓度范围和浓度间隔情况如表1所列.

作为例子，图1给出了9个混合气体样本的光谱数据，其对应各组分气体的浓度如表2所列.

在采用KPLS进行特征提取之前，先对光谱数据进行中心化和标准化预处理，随后采用400 cm^{-1}

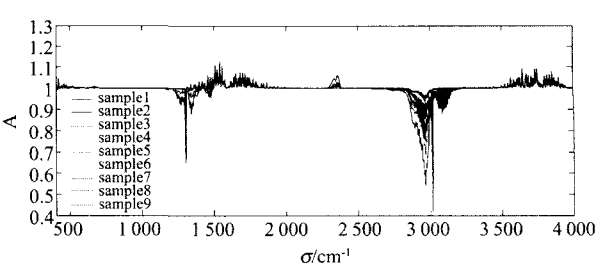

图1 9种混合气体样本的FTIR光谱

Fig. 1 FTIR spectra of nine gas mixture samples

到 4000cm^{-1} 的全部 1866 维光谱数据建立 KPLS-SVR 定量分析模型.

当未对光谱数据特征提取,采用 SVR 直接建模时,为消除光谱基线的漂移及光谱受外界环境变化产生的影响,对原始光谱数据进行归一化预处理

$$x_m = \frac{x'_m}{\frac{1}{L}\sum_{m=1}^{L} x'_m} \times A \quad , \qquad (1)$$

其中 x'_m 为原始光谱的第 m 维数据,L 为光谱数据(对应样本)个数,A (>1) 为放大倍数. 同时,为了获得更好的预测效果,经过优化选择,采用波数范围在 410cm^{-1} ~ 2070cm^{-1} 的(部分)光谱数据建立定量分析模型.

2.3 KPLS 特征提取

对原始光谱数据采用 KPLS 算法[14] 对不同组分气体分别提取其相应的特征向量. 选用 Gaussian 核作为其核函数.

2.4 模型评价标准

预测均方根误差(Root Mean Squared Error of Prediction,RMSEP)用来对定量分析模型的分析能力进行评价. 此外,超差点个数作为模型评价的另一个标准. 根据各组分气体的浓度范围,将允许的预测相对误差分成三段: ±200% (<0.01%), ±50% (0.01 ~0.1%) 和 ±20% (0.1 ~1%),预测结果超出相对误差允许范围的为超差.

3 结果与讨论

将 1842 个光谱数据分为 2 组,其中包含 1230 个样本的一组作为训练集,用来建立定量分析模型,另外一组共 612 个样本作为检验集,用来检验已建立的分析模型. 待定参数:L(Gaussian 核的宽度)、最优的 KPLS 特征组分数量以及 SVR 的优化参数(s, C,e,g,p)等均由交叉证实法(leave-ten-out cross validation)确定. 为了获得最佳的分析结果,对七种被测气体的定量分析模型参数分别优化,即每种被测气体均对应其最优的子模型.

当对原始光谱数据进行 KPLS 特征提取时,对应不同气体的最佳特征组分数(15 ~22,参见图 2)首先被确定. 实验显示:不同的特征组分数量对所建模型的预测精度影响很大. 图 2 所示为选用不同特征组分数时对应 KPLS-SVR 模型的 RMSEP. 其中,交叉检验法确定的最佳特征组分数由星号标记. 随后,当最优的 L 和 SVR 参数确定之后,KPLS-SVR 定量分析模型建立成功.

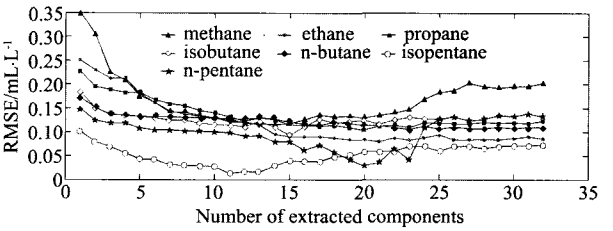

图 2　特征组分数量对 KPLS-SVR 模型预测精度的影响

Fig. 2　The effect of amount of extracted feature components on prediction accuracy of KPLS-SVR models

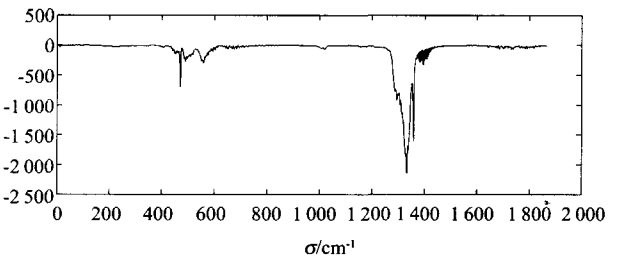

图 3　训练集样本针对甲烷 KPLS 特征提取的第一个特征组分对应的载荷向量

Fig. 3　Loading weight for the first component from KPLS procedures for methane

作为例子,图 3 所示为训练集样本针对甲烷 KPLS 特征提取的第一个特征组分对应的载荷向量.

3.1 模型的预测能力

表 3 中给出了 KPLS-SVR 模型与 SVR 模型对同一检验集样本的预测结果. 表中 KPLS-SVR 模型预测七种气体的 RMSEP 从 0.016 mL · L^{-1}到 0.116mL · L^{-1} 不等,明显小于 SVR 模型所得 RMSEP(从 0.048mL · L^{-1}到 0.177mL · L^{-1}). 同时,从相应的超差点数量来看,KPLS-SVR 模型的超差点数(4 ~14)也少于 SVR 模型(11 ~33)的超差点数. 这充分说明 KPLS-SVR 模型较 SVR 模型具有更高的分析精度,其中 KPLS 的特征提取功不可没. KPLS 方法在核定义的特征空间中利用协方差指导特征选择,使得提取出的特征向量可以更好地反映输入与目标之间的关系,所以结合 SVR 获得了更好的回归效果.

表 3　KPLS-SVR 和 SVR 定量分析模型对 7 种组分气体的预测结果

Table 3　The prediction results of seven component gases by KPLS-SVR and SVR models

方法	预测结果							
	评价标准	甲烷	乙烷	丙烷	异丁烷	正丁烷	异戊烷	正戊烷
KPLS-SVR	RMSEP/mL · L^{-1}	0.116	0.079	0.104	0.092	0.108	0.029	0.016
	超差个数	4	6	5	13	14	7	10
SVR	RMSEP/mL · L^{-1}	0.177	0.101	0.135	0.102	0.115	0.075	0.048
	超差个数	11	17	28	28	32	33	17

表4 模型参数及预测时间

Table 4 The parameters of modeling and the computation time of different models

方法	气体组分	模型参数		计算时间(s)	
		特征组分个数	L	建模	预测
KPLS-SVR	甲烷	15	0.61	939.35	62.28
	乙烷	21	0.6		
	丙烷	20	0.48		
	异丁烷	15	0.5		
	正丁烷	22	0.62		
	异戊烷	20	1.22		
	正戊烷	17	1.23		
SVR	组分气体	-	-	298.59	125.06

3.2 模型的计算时间

采用KPLS-SVR和SVR方法建模和预测(参数优化之后)的时间以及相应模型参数见表4. 表中显示,建立KPLS-SVR模型所用时间为939.35s,长于SVR建模时间(298.59s),但是,前者的预测时间(62.28s)少于后者(125.06s),这正是由于KPLS特征提取有效地降低了原始光谱的维数,所以才会使KPLS-SVR模型的预测时间减少. 更短的预测时间为混合气体的在线分析提供了更为有利的支持. 为了预测时间更短,预测精度更高,多花些时间用来建模是值得的.

4 结语

本文研究了KPLS方法用于FTIR光谱数据的特征提取,并将提取得到的特征组分作为SVR的输入建立了七组分烷烃混合气体的定量分析模型. 实验显示,对于交叉敏感与多重共线性问题,KPLS特征提取在高维特征空间计算得到的特征组分不但去除了噪声的影响,而且可以更为准确地反映原始光谱数据与组分浓度之间的关系. 经检验,KPLS特征提取与SVR结合所建模型较未特征提取仅用SVR所建模型表现出明显的优越性:更高的分析精度、更快的分析速度以及较KPCA特征提取更少的特征组分[15]. 研究表明,KPLS方法可以有效地提取光谱数据的有用信息,消除光谱数据的噪声,降低光谱数据的维数,是光谱特征提取的一种有效方法. 将KPLS与SVR结合,可以快速准确地实现严重交叉敏感的七组分烷烃混合气体的定量分析.

REFERENCES

[1] BAI Peng, LIU Jun-Hua. Algorithm of mixed gas concentration analysis based on support vector machine and multidimensional spectrum[J]. *Control and Instruments in Chemical Industry*(白鹏,刘君华. 基于多维光谱的多组分混合气体浓度支持向量机法. **化工自动化及仪表**),2005,**32**(5):47—49.

[2] BAI Peng, XUE Wen-Jun, LIU Jun-Hua. New method of mixed gas infrared spectrum analysis based on SVM[J]. *Spectroscopy and Spectral Analysis*(白鹏,薛文俊,刘君华. 混合气体红外光谱支持向量机分析的新方法. **光谱学与光谱分析**),2007,**27**(7):1323—1327.

[3] Vapnik V N. *Statistical Learning Theory*[M]. Springer, Heidelberg,1998.

[4] DENG Nai-Yang, TIAN Ying-Jie. *New Method of Data Mining-Support Vector Machine*[M]. Beijing: Science Press(邓乃扬,田英杰. **数据挖掘中的新方法-支持向量机**. 北京:科学出版社),2004,245—254.

[5] CHENG Jie, LIU Qin-Huo, LI Xiao-Wen, *et al*. Algorithm study on soil mid-infrared emissivity extraction[J]. *J. Infrared Millim. Waves*(程洁,柳钦火,李小文,等. 土壤中红外发射率提取算法研究. **红外与毫米波学报**),2006,**34**(2):263—266.

[6] SHAO Yong-Ni, CAO Fang, HE Yong. Discrimination years of rough rice by using visible/near infrared spectroscopy based on independent component analysis and BP neural network[J]. *J. Infrared Millim. Waves*(邵咏妮,曹芳,何勇. 基于独立组分分析和BP神经网络的可见/近红外光谱稻谷年份的鉴别. **红外与毫米波学报**),2007,**26**(6):433—436.

[7] Rosipal R, Trejo L J. Kernel partial least squares regression in reproducing kernel hilbert space[J]. *Journal of Machine Learning Research*,2001,**2**(12): 97—123.

[8] Rosipal R, Trejo L J, Wheeler K. Locally-Based Kernal PLS Smoothing to Non-Parametric Regression Curve Fitting. Download from: http://www.ofai.at/~roman.rosipal/Papers/wp03.pdf.

[9] Rosipal R. Kernel partial least squares for nonlinear regression and discrimination[J]. *Neural Networks World*,2003,**13**(3):291—300.

[10] Momma M, Bennett K P. *Sparse Kernel Partial Least Squares Regression*[M]. Lecture Notes Computer Science,2003,**2777**:216—230.

[11] Tenenhaus A, Giron A, Viennet E, *et al*. Kernel logistic PLS: A tool for supervised nonlinear dimensionality reduction and binary classification[J]. *Computational Statistics & Data Analysis*,2007,**51**(9):4083—4100.

[12] YANG Hui-Hua, WANG Xing-Yu, WANG Yong. KPLS approach for network intrusion feature extraction and detection[J]. *Control and Decision*(杨辉华,王行愚,王勇,等. 基于KPLS的网络入侵特征抽取及检测方法. **控制与决策**),2005,**20**(3): 251—255.

[13] Rosipal R, Trejo L J, Matthews B. Kernel PLS-SVC for Linear and Nonlinear Classification[C]. In: Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003, Download from: http://www.hpl.hp.com/conferences/icml2003/papers/110.pdf.

[14] Taylor J S, Cristianini N. *Kernel Methods for Pattern Analysis*[M]. Cambridge University Press, Cambridge, England, 2004,187—192.

[15] HAO Hui-Min, TANG Xiao-Jun, BAI Peng, *et al*. Quantitative analysis of multi-component gas mixture based on KPCA and SVR[J]. *Spectroscopy and Spectral Analysis*(郝惠敏,汤晓君,白鹏,等. 基于核主成分分析和支持向量回归机的红外光谱多组分混合气体定量分析. **光谱学与光谱分析**),2008,**28**(6):1286—1289.