第28卷,第6期
2008年6月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 28, No. 6, p1286-1289
June, 2008

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

郝惠敏[1,2], 汤晓君[1], 白 鹏[1,3], 刘君华[1], 朱长纯[1]

1. 西安交通大学电气工程学院, 陕西 西安 710049
2. 太原钢铁公司自动化公司, 山西 太原 030003
3. 空军工程大学理学院, 陕西 西安 710038

摘 要 提出了一种核主成分分析(KPCA)特征提取结合支持向量回归机(SVR)的红外光谱混合气体组分定量分析新方法。首先将特征吸收谱线严重重叠的混合气体光谱通过非线性变换映射到高维特征空间,然后在特征空间中再利用主成分分析法提取主成分,提取出的主成分作为SVR的输入建立校正模型,实现了甲烷、乙烷、丙烷、异丁烷、正丁烷、异戊烷以及正戊烷七种组组分特征吸收光谱严重重叠的混合气体的定量分析。用KPCA-SVR所建模型对未知浓度混合气体的七种组分预测的RMSE($\varphi\times10^{-6}$)较仅用SVR模型预测的RMSE($\varphi\times10^{-6}$)降低了一个数量级。结果表明,核主成分分析法具有很强的非线性特征提取能力,可以充分利用全光谱数据并有效地消除光谱数据噪声,降低数据维数,与支持向量回归机结合可以提高红外光谱分析的精度,缩短模型计算时间,是一种有效的红外光谱分析新方法。

关键词 核主成分分析; 支持向量回归机; 校正模型; FTIR; 定量分析

中图分类号: TE642; TH744.4 **文献标识码**: A **文章编号**: 1000-0593(2008)06-1286-04

引 言

含烃类多组分混合气体的定量分析在诸如煤矿开采、石油录井、化工分析等许多领域都有着十分重要的意义。FTIR光谱由于其分析灵敏度高、分析速度快等特点,已经在包括气体监测的许多领域得到应用[1-5]。但是,甲烷、乙烷、丙烷、异丁烷、正丁烷、异戊烷以及正戊烷等气体,其红外光谱的主、次特征吸收峰十分接近,因此它们的红外吸收谱线重叠非常严重。当它们混合时,其混合气体的红外光谱存在着严重的交叉敏感,所以对以上烃类混合气体红外光谱的定量分析存在着极大的困难。

支持向量回归机(support vector regress machine, SVR)[6,7]是基于支持向量机(support vector machine, SVM)[8-10]理论建立的一种回归估计技术,具有泛化能力强,预测准确度高等优点。白鹏[11]和林继鹏[12]等将其应用于红外光谱的定量分析中,实现了对存在严重交叉敏感的五种气体的混合气体组分浓度的定量分析,预测结果的最大均方根误差(root mean squared error, RMSE)($\varphi\times10^{-6}$)达到了1 896.00。

为了进一步提高定量分析精度,同时为了降低原始光谱数据的维数,克服原始光谱数据的维数较大时,浮现出来的SVR计算速度减慢,参数优化困难等问题,本文进一步研究了核主成分分析(kernal principal component analysis, KPCA)[13-15]方法在光谱数据特征提取中的应用。KPCA方法是Schölkopf[16]等对线性主成分分析(linear principal component analysis, LPCA)方法进行核化得到的一种新的多变量统计方法。由于核技巧的应用, KPCA较PCA具有很多优势[17-19],尤其是在提取非线性特征方面。本文将KPCA与SVR相结合,把经过KPCA特征提取得到的主成分分量作为SVR的输入建立混合气体的定量分析模型。实验表明,KPCA可以对原始光谱数据进行非线性特征提取,消除原始光谱数据的噪声,有效降低光谱数据维数,结合SVR可以建立准确的定量分析模型。通过对KPCA-SVR和SVR建模及预测结果的对比表明, KPCA-SVM方法具有更高的预测准确性,更快的建模速度和预测速度,是一种更为有效的红外光谱分析方法。

收稿日期: 2007-08-22, **修订日期**: 2007-11-28

基金项目: 国家自然科学基金项目(60276037)资助

作者简介: 郝惠敏, 女, 1971年生, 西安交通大学电气工程学院博士生 e-mail: helenwangmin@gmail.com

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1　实验部分

建立样本光谱数据集，对样本光谱数据进行 KPCA 特征提取，得到的主成分作为 SVR 的输入建立校正模型，用已建校正模型对预测集样本进行预测。

1.1　含烃类混合气体红外光谱数据的获得

采用由 Alicat Scientific 流量计(量程为 0～0.5 SCCM 到 0～1 500 SLPM，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，制备甲烷、乙烷、丙烷、异丁烷、正丁烷、异戊烷以及正戊烷七种气体的混合气体样气，七种气体均为浓度已知的标准气体，混合气体浓度范围为 0 %～25 %。气体经过干燥处理之后注入光谱仪的气室。采用 Bruker TENSOR27 型傅里叶变换红外光谱仪扫描样气，扫描范围为 4 000～400 cm^{-1}，扫描间隔为 12 nm，共获得 922 个样本数据，每个样本对应的光谱包含 1 866 个吸收点数据。

1.2　光谱数据的 KPCA 特征提取结合 SVR 建立回归校正模型

将样本数据分成三部分，其中 510 个样本作为校正集样本，210 个作为检验集样本，其余 202 个作为预测集样本。采用 KPCA 方法分别提取校正集样本、检验集样本和预测集样本的特征向量(也称为主成分)。

采用 Gaussian 核函数($K(x_k, x_j) = e^{-\left(\frac{\|x_j - x_j\|^2}{L}\right)}$)对波长范围在 4 000～400 cm^{-1} 的全光谱数据进行运算，得到高维特征空间内的映射数据，对其进行中心化处理，求解非零特征值，并标准化。将映射数据向特征向量上进行投影，得到样本的特征向量。

将校正集特征提取得到的主成分 $t_i (t_i \subseteq R^n, i = 1, 2, \cdots, m)$ 作为 SVR 的输入，建立校正模型，SVR 算法见参考文献[20]。采用 RMSE 评价组分浓度分析结果。程序在 Matlab7.0 环境下编写。

2　结果与讨论

2.1　KPCA-SVR 方法建立校正模型

2.1.1　特征提取得到的主成分

图 1 所示为校正集样本光谱数据经过 KPCA 特征提取得到的第一个特征向量。

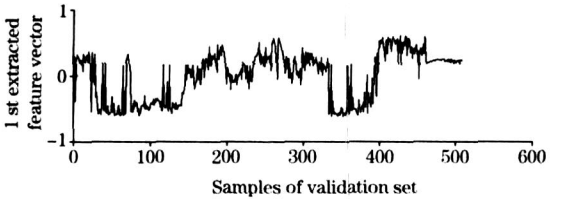

Fig. 1　First extracted feature vector of validation set

2.1.2　主成分个数对校正模型分析精度的影响

图 2 为各主成分对应的特征值。图中显示，当主成分个数超过 148 时，其对应的特征值已接近零，因此大于 148 的主成分对原始数据贡献已经很小。为了确定最佳主成分个数，要结合检验集的 RMSE 来综合考虑。

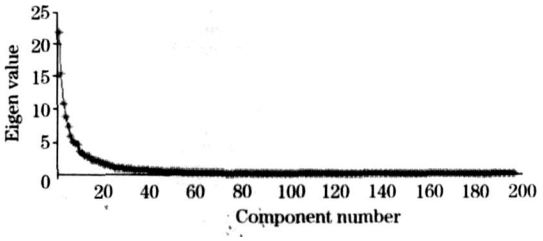

Fig. 2　Plot of eigen values of the principal components

图 3 所示为经过 KPCA 特征提取得到的主成分个数与对应检验集的 RMSE 之间的关系，图中 SVR 的相关参数 s, ε, t, g 和 C 已经过优化。由图可见，随着 KPCA 特征提取主成分个数的增加，检验集的 RMSE 在不断减小，当主元个数为 150 时，七种气体的 RMSE 均达到最小。图 3 显示主元个数超过 150 时，RMSE 随主元个数的增加有小幅度的增大，于是选取 150 个主元作为 SVR 的输入，建立校正模型。对预测集样本采用已建校正模型预测，结果见表 1。

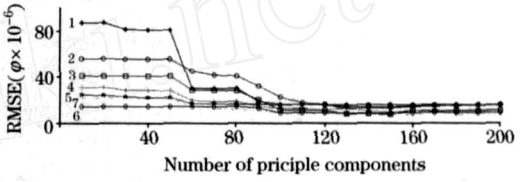

Fig. 3　Effects of the number of principal components on the RMSE of test set by KPCA feature extraction

1: Methane; 2: Propane; 3: Ethane; 4: Isobutane; 5: *n*-butane; 6: Isopantane; 7: *n*-pentane

2.2　SVR 方法建立校正模型

为消除光谱仪基线漂移以及环境变化造成的光谱波动的影响，将原始光谱数据进行归一化处理，处理后的光谱数据直接作为 SVR 的输入进行建模。为了进一步提高模型的准确度，避免气体浓度增大造成的吸收饱和及其它因素的影响，通过优化计算后选用光谱波数范围在 2 070～410 cm^{-1} 的光谱数据进行建模。对预测集样本采用已建校正模型预测，结果见表 1。

2.3　两种建模方法预测效果的比较

2.3.1　预测精度的比较

由表 1 可见，经过 KPCA 特征提取之后，七种气体的预测精度较未特征提取，仅用 SVR 建立模型预测的精度有明显的提高，预测集的 RMSE($\varphi \times 10^{-6}$) 从 1903.00～500.06 减小到 153.01～81.72，说明 KPCA 特征提取可以提取光谱数据的非线性特征，这些非线性特征可以很好地替代原始光谱的信息，有效降低光谱数据维数。同时图 3 显示，从第一个主成分开始，随着主成分个数的增多，检验集的 RMSE 大幅度降低，直到主成分个数为 150 时，RMSE 达到最小，当主成分个数大于 150 以后，随着主元个数的增加，预测集的 RMSE 会小幅度增大，这说明 KPCA 特征提取可以消除光谱

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Table 1 Comparison of the prediction results of seven different gases using KPCA-SVR and SVR

Method	RMSE($\phi\times10^{-6}$)						
	Methane	Ethane	Propane	Isobutane	*n*-butane	Isopentane	*n*-pentane
KPCA-SVR	124.37	72.44	136.51	87.29	153.01	57.12	81.72
SVR	1 903.00	800.03	1 600.20	1 661.37	1 032.03	1 600.0	500.06

数据噪声，从而提高预测模型的预测精度。最佳数量的主成分可以最好的替代原始光谱，而主成分的个数并不是越多越好，多余的主成分反而会附带更多的噪声，影响预测精度。

2.3.2 计算速度的比较

以甲烷为例，表 2 给出了采用 KPCA-SVR 和 SVR 建模时的优化参数和计算时间。从表 2 可以看出，经过 KPCA 特征提取后再用 SVR 建模，无论是建模时间和预测时间都大幅度缩短，主要原因是通过特征提取之后，输入 SVR 数据的维数由原来的 1 866 维减少到 150 维，这样就大大提高了 SVR 运算的速度，从而提高了建模速度和预测速度。

Table 2 Comparison of the elapsed time of using KPCA-SVR and SVR

Method	Modeling parameters(CH_4)							Elapsed time/ s	
	Number of principal components	*L*	*s*	ε	*t*	*g*	*C*	Modeling	Prediction
KPCA-SVR	150	0.1	3	0.1	2	0.08	110	46.59	4.94
SVR	-	-	3	0.01	0	-	50	752.52	26.21

3 结 论

本文提出了基于 KPCA 结合 SVR 进行红外光谱多组分定量分析的新方法，通过对混合气体的红外光谱进行 KPCA 特征提取后再输入 SVR 进行学习建模，实现了特征谱线严重交叉的七种烷烃类混合气体的定量分析。研究表明，KPCA 方法具有很强的非线性提取能力和消除噪声的能力，适合于红外光谱数据的特征提取。采用 KPCA-SVR 方法的预测精度和建模速度较 SVR 方法有了大幅度的提高，在红外光谱混合气体建模中体现了独特的优越性。

参 考 文 献

[1] ZHANG Lin, ZHANG Li-ming, LI Yan, et al(张 琳, 张黎明, 李 燕, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26(4): 620.

[2] LI Hong-lei, LIU Xian-yong, ZHOU Fang-jie, et al. Proceedings of the SPIE, 2005, 5640: 692.

[3] SUN Xiu-yun, LI Yan, WANG Jun-de(孙秀云, 李 燕, 王俊德). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2003, 23(4): 739.

[4] Evans Wayne F J, Puckrin Eldon, McMaster D. Proceedings of the SPIE, 2002, 4574; 44.

[5] Bacsik Z, McGregor J, Mink J. Food and Chemical Toxicology, 2007, 45(2): 266.

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

[7] Hoegaerts L, Suykens J A K, Vandewalle J, et al. Neurocomputing, 2005, 63: 293.

[8] Vapnik V N. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

[9] Vapnik V N. Statistical Learning Theory, New York: Wiley, 1998.

[10] Duan Kaibo, Keerthi S Sathiya, Poo Aun Neow. Neurocomputing, 2003, 51: 41.

[11] BAI Peng, LIU Jun-hua(白 鹏, 刘君华). Control and Instruments in Chemical Industry(化工自动化及仪表), 2005, 32(5): 47.

[12] LIN Ji-peng, LIU Jun-hua(林继鹏, 刘君华). Journal of Xi'an Jiaotong University(西安交通大学学报), 2005, 39(6): 586.

[13] Kim K L, Franz M O, Schölkopf B. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27: 1351.

[14] Rosipal Roman, Girolami Mark, Trejo Leonard J, et al. Neural Computing & Applications, 2001, 10(3): 231.

[15] Müller Klaus-Robert, Mika Sebastian, Ratsch Gunnar, et al. IEEE Transactions on Neural Networks, 2001, 12(2): 181.

[16] Schölkopf B, Smola A, Müller Klaus-Robert. Neural Computation, 1998, 10(5): 1299.

[17] John Shawe-Taylor, Nello Cristianini. Kernel Methods for Pattern Analysis. Cambridge: Cambridge University Press, 2004. 143.

[18] Mika Sebastian, Schölkopf Bernhard, Smola Alex, et al. Advances in Neural Information Processing System Ⅱ. Cambridge, MA: MIT Press, 1999. 536.

[19] Mika S. Advances in Neural Information Processing System Ⅱ. Cambridge, MA: MIT Press, 1999. 536.

[20] Schölkopf B, Smola A. Advances in Kernel Methods-Support Vector Learning. Cambridge, MA: MIT Press, 1999. 327.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Quantitative Analysis of Multi-Component Gas Mixture Based on KPCA and SVR

HAO Hui-min[1,2], TANG Xiao-jun[1], BAI Peng[1,3], LIU Jun-hua[1], ZHU Chang-chun[1]
1. School of Electrical Engineering, Xi'an Jiaotong University, Xi'an 710049, China
2. Taiyuan Iron and Steel Co., Automatic Company, Taiyuan 030003, China
3. Engineering Institute, Air Force Engineering University, Xi'an 710038, China

Abstract In the present paper, the authors present a new quantitative analysis method of mid-infrared spectrum. The method combines the kernel principal component analysis (KPCA) technique with support vector regress machine (SVR) to createa quantitative analysis model of multi-component gas mixtures. Firstly, the spectra of multi-component gas mixtures samples were mapped nonlinearly into a high-dimensional feature space through the use of Gaussian kernels. And then, PCA technique was employed to compute efficiently the principal components in the high-dimensional feature spaces. After determining the optimal numbers of principal components, the extracted features (principal components) were used as the inputs of SVR to create the quantitative analysis model of seven-component gas mixtures. The prediction RMSE ($\phi\times10^{-6}$) of seven-component gases of prediction set samples by use of KPCA-SVR model were respectively 124.37, 72.44, 136.51, 87.29, 153.01, 57.12, and 81.72, ten times less than that by use of SVR model. The elapsed time of modeling and prediction by using KPCA-SVR were respectively 46.59 (s) and 4.94 (s), which was consumedly less than 752.52 (s) and 26.21 (s) by using only SVR. These results show that KPCA has an excellent ability of nonlinear feature extraction. It can make the most of the information of entire spectra range and effectively reduce noise and the dimension of the spectra. The KPCA combined with SVR can improve the model's analysis precision and cut the elapsed time of modeling and analysis. From our research and experiments, we conclude that KPCA-SVR is an effective new method for infrared spectroscopic quantitative analysis.

Keywords Kernel principal component analysis; Support vector regression machine; Calibration model; Fourier transform infrared spectrum; Quantitative analysis

(Received Aug. 22, 2007; accepted Nov. 28, 2007)

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net