基于SVM的混合气体分布模式红外光谱在线识别方法

白 鹏[1,2]，冀捐灶[3]，张发启[3]，李 彦[2]，刘君华[4]，朱长纯[1]

1. 西安交通大学电子信息工程学院，陕西 西安 710049
2. 空军工程大学理学院，陕西 西安 710051
3. 空军工程大学工程学院，陕西 西安 710038
4. 西安交通大学电气工程学院，陕西 西安 710049

摘 要 针对混合气体组分浓度分析中海量训练样本的获取、分析精度及实时在线分析等问题，将支持向量机这一新的信息处理方法和红外光谱分析法结合，提出了混合气体分布模式的概念。在此基础上，采用先进行混合气体分布模式识别，然后再进行混合气体分析的思路，在大量调查的基础上，研究探索了实际应用中可能出现的混合气体分布模式，确定60种混合气体分布模式，共计6 000个混合气体红外光谱数据样本用于模型的训练与检验。采用SMO算法实现了减量和增量的在线学习，最终建立了基于SVM的混合气体分布模式红外光谱在线识别模型。模型由模式识别和结果输出2层组成，模式识别层完成混合气体模式分布模式识别任务；结果输出层由60个SVM校正模型组成，完成具体的浓度分析任务。实验结果表明，该方法对混合气体分布模式的正确识别率不低于98.8%，可在小样本条件下对混合气体的分布模式进行在线识别，可在线实时加入新的混合气体分布模式，具有实际应用价值。

关键词 支持向量机；红外光谱；校正模型；模式识别

中图分类号：TH744.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2008)10-2278-04

引 言

混合气体红外光谱分析在天然气、石油、化工、烟气、汽车尾气、垃圾燃烧场废气监测等方面有着广泛的应用，越来越受到人们的重视[1-3]。

由于混合气体组分浓度的分布具有随意性，若要实现在较大分布范围内的混合气体分析，必须构造数量足够多，分布合理的混合气体样本供训练使用。例如，有一混合气体，由七种组分气体组成，如果每种组分气体浓度按1%的间隔标定100个点，则7组分混合气体需100^7个样本。但实际工作中，构造数量足够多且分布合理、组分浓度范围大的混合气体样本是不容易实现的，无法获得如此多海量混合气体样本。

在混合气体红外光谱分析的实验中发现，将全部的训练样本用来训练一个大的校正模型，其结果输出误差较大；如果将全部的训练样本根据混合气体分布模式分为几个小的校正模型，其结果输出误差较小。

传统的混合气体组分浓度红外光谱分析方法有曲线拟合法、K矩阵法、P矩阵法、多元线性回归法、经典最小二乘法、偏最小二乘回归[4]等。这些方法由于偏重线性问题、分析精度依赖海量的红外光谱数据样本，因而不能有效地解决上述问题。人工神经网络(artificial neural network，ANN)虽然在混合气体红外光谱分析中有应用[5,6]，但由于输入数据的维数、过学习等问题的限制，在实际应用中还是有一定的局限性。

为了解决上述的问题，本文将支持向量机(support vector machine，SVM)这一新的信息处理方法[7-9]，用于处理混合气体红外光谱数据样本。利用SVM的在线模式识别模型，先进行混合气体分布模式识别，根据模式识别模型的结果输出，确定对应的校正模型，然后再进行混合气体组分浓度分析，进而输出结果。

本文以含烃类混合气体为例，在大量调查的基础上，研究探索了实际应用中可能出现的混合气体分布模式，最后确定为60种分布模式，共计6 000个混合气体红外光谱数据样本用于训练与检验。在此基础上，通过动态调整混合气体分布模式与红外光谱数据样本分类规则知识的积累，采用SMO算法实现了减量和增量的在线学习，从而完成混合气体分布模式红外光谱在线识别，为后续的混合气体组分浓度分析奠定了基础。

收稿日期：2007-10-18，**修订日期**：2008-01-26

基金项目：国家自然科学基金项目(60772016)资助

作者简介：白 鹏，1967年生，西安交通大学电子信息工程学院博士后 e-mail：bai-peng410@sohu.com

1　基本原理

对于满足一定分布规则的红外光谱数据样本，其分布具有聚类性[10, 11]。

对于给定的红外光谱数据样本集，不同组分浓度和种类的混合气体红外光谱数据样本在特征空间中的聚类称为混合气体分布模式。

1.1　思路

在红外光谱模式识别问题中，每个混合气体分布模式所代表的红外光谱数据样本训练集可以用相似度度量，根据不同的相似度阈值，对样本进行聚类操作，生成混合气体分布模式。由于每个混合气体分布模式代表一个对应的红外光谱数据样本训练集，全部的混合气体分布模式代表了整个训练样本集，上述过程可抽象为模式识别模型的建立问题[12, 13]。

例如，某混合气体有五种混合气体分布模式，模式识别模型的输入为包含五种混合气体分布模式的红外光谱数据样本，模式识别模型的输出为五种混合气体分布模式标识编码，根据标识编码决定使用对应的校正模型，进行具体的混合气体组分浓度分析。

1.2　训练与检验

基于 SVM 的混合气体分布模式识别模型在使用前必须进行训练与检验。训练是确定 SVM 混合气体分布模式识别模型的参数；检验是对训练过程的校验，经过训练与检验的 SVM 混合气体分布模式识别模型才可应用于实际的混合气体分布模式识别。

首先，确定模式识别模型的参数。包括 SVM 类型、核函数、惩罚因子 C 及损失函数 ε 的数值。SVM 模式识别模型的输入为包含混合气体分布模式的红外光谱数据样本集，$x_i=(x_1, x_2, \cdots, x_l)$为混合气体 l 个样本；输出 $y_i=(y_1, y_2, \cdots, y_m)$为 m 个混合气体分布模式的标识编码。例如，$y_i=(0, 1, 0, 0, 0)$表示五种混合气体分布模式中的第二种模式。当输出的数值与期望值误差满足要求时，训练结束。

然后，用另一部分样本对 SVM 模式识别模型进行检验。如果检验满足误差要求，SVM 模式识别模型的参数被最终确定，可用于实际混合气体分布模式的识别，上述过程如图 1 所示。

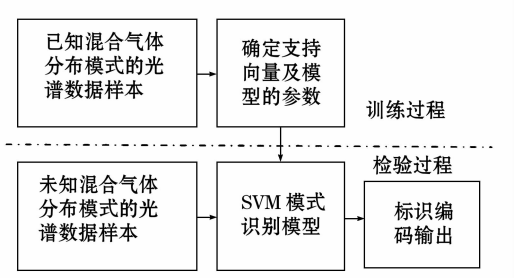

Fig. 1　The training and test flow chart of SVM calibration model

1.3　混合气体分布模式的生成

混合气体分布模式的生成，利用了模式识别中的聚类分析原理[14, 15]，先把红外光谱数据样本逐个合并成一些混合气体分布模式，然后合并混合气体分布模式，直至无法合并为止。其具体步骤如下。

设全部的红外光谱数据样本训练集为 S，大小为 l，混合气体分布模式 G 的大小设为 m，$m<l$，使用如下步骤生成混合气体分布模式 G。

(1)对红外光谱数据样本进行预处理。

(2)将每个红外光谱数据样本看成一个混合气体分布模式，l 个红外光谱数据样本有 l 个混合气体分布模式，作为初始混合气体分布模式。

(3)以初始混合气体分布模式所表示的红外光谱数据样本为基准，计算初始混合气体分布模式间的距离，将距离最近的两个混合气体分布模式合并为一个混合气体分布模式。

(4)继续选择并计算已有混合气体分布模式与其他混合气体分布模式之间的距离，并将距离最近的两个混合气体分布模式合并为一个混合气体分布模式。

(5)如果存在距离最近的混合气体分布模式，则转(4)继续并类。如果不存在距离最近的混合气体分布模式，混合气体分布模式的生成结束。

混合气体分布模式生成的过程，实质上就是基于 SVM 核函数的点积运算，在高维数据空间中根据核函数距离对红外光谱数据样本进行聚类的过程。

1.4　SVM 在线学习

实际应用中，混合气体分布模式识别的分类知识是随时间和训练样本的增多而逐渐积累的，模式识别的在线学习不是简单的增量学习。因此，模式识别分类知识的在线学习能力成为实际应用的重要特性。

基于 SVM 的混合气体分布模式在线识别的学习过程就是分类知识的积累过程，在标准 SVM 中分类知识的积累单纯依赖于训练样本数量的增加。这种方式会导致训练样本数量和支持向量数量的快速增大，SVM 的训练和分类的速度也会随之快速降低[16, 17]。在采用增量学习算法中，全部或通过复杂的启发性规则将当前的非支持向量样本从训练集中删除，以换取速度的提高和减少存储压力[18, 19]。但是这必定会导致分类知识的丢失，因为当前的非支持向量同样带有分类知识，其中部分样本可能在后续训练中成为支持向量，距离当前分类面较近的样本尤是如此。

本文所述基于 SVM 的混合气体分布模式在线识别方法对新样本的学习思路如下：首先，利用新样本调整当前的混合气体分布模式，以实现分类知识的积累；然后针对混合气体分布模式的变化，在现有分布模式基础上对 SVM 模式识别模型进行在线更新。具体的在线学习方法步骤如下。

设新的样本为 S_t，混合气体分布模式 G 中在特征空间中距离最近的两个混合气体分布模式为 G_i 和 G_j，对应的核函数距离为 D_{ij}。

(1)对混合气体分布模式 G 中的所有混合气体分布模式，在特征空间中寻找与新样本 S_t 距离最近的混合气体分布模式 G_e，对应核函数距离值为 D_{te}。

(2) 如果 $D_{ij}>D_{te}$。

(a) 使用减量学习算法，将新样本 S_t 与混合气体分布模

式 G_e 合并；

(*b*) 更新混合气体分布模式 G_e 及其他分布模式的支持向量。

(3)否则 $D_{ij} \leqslant D_{te}$。

(a)使用减量学习算法，将新样本 S_t 作为新的混合气体分布模式。

从现有分布模式 SVM 中除去混合气体分布模式 G_e 对分类函数的贡献，设置对应拉格郎日乘子为 0。

(b)更新混合气体分布模式的支持向量。

(4)更新核函数矩阵，重新寻找 G 中在特征空间中距离最近的混合气体分布模式及其对应的核函数距离值 D_{ij}。

(5)若 G 中有混合气体分布模式违反 SVM 中的 KKT 条件，使用增量学习过程更新混合气体分布模式。

上述的混合气体分布模式识别在线学习方法具有便于分类知识的积累和在线学习的特点。由于支持向量直接来自于训练样本，所以分类知识的积累和在线学习过程可以通过对混合气体分布模式在特征空间中的重新聚类和支持向量的改变来实现。

分类知识的积累规则为：不能用已有混合气体分布模式正确分类的新训练样本来形成新的混合气体分布模式；能被正确分类的训练样本用于修订已有混合气体分布模式的支持向量。

2 实验结果

制作混合气体红外光谱数据样本之前，要通过如下的技术手段对混合气体分布模式情况进行分析。

(1)对实际的混合气体研究分析；(2)查阅有关的文献资料；(3)考虑实际应用的需要；(4)进行统计分析的加工。

通过上述技术手段的分析，将混合气体的分布模式划分为 60 种基本模式。对于 SVM 模式识别模型而言，红外光谱数据样本为模型的输入，60 种基本模式的标识编码为输出。

根据 60 种混合气体的分布模式，利用标准的纯气体进行配置。实验所用红外光谱仪为 Bruker 公司的 TENSOR 27 型傅里叶变换红外光谱仪，扫描范围为 4 000～400 cm^{-1}，扫描间隔 12 nm，得到 1 866 个透射光谱数据，共制作 6 000 个红外光谱数据样本。

将全部的红外光谱数据样本集用来训练一个大的校正模型，然后随机选择 60 种分布模式中的一种分布模式所对应的红外光谱数据样本集，用来训练一个小的校正模型，将与分布模式对应的红外光谱数据样本集作为两个模型的输入，对比两种校正模型的分析结果。用平均绝对误差值(Mean AE)来衡量实验结果的误差，实验结果如表 1 所示。

大校正模型和小校正模型的参数完全一致。从表 1 中的平均绝对误差数据可以看出，应用混合气体分布模式识别后确定的小 SVM 校正模型结果有明显提高。

对于混合气体分布模式的在线学习能力，进行已知混合气体分布模式与未知混合气体分布模式的识别实验，结果如表 2 所示。

Table 1 Experimental results of big and small SVM calibration models

气体	大校正模型	小校正模型
甲烷	0.056	0.042
乙烷	0.032	0.021
丙烷	0.031	0.018
异丁烷	0.026	0.021
正丁烷	0.017	0.013

Table 2 Experimental results of online pattern recognition

光谱数据样本个数	已知混合气体分布模式识别个数	未知混合气体分布模式识别个数
120	119	118
270	268	267
500	497	495

根据表 2 中的数据，可以得出 SVM 混合气体分布模式在线识别模型对已有混合气体分布模式红外光谱数据样本的正确识别率为：

$\min\{119/120, 268/270, 497/500\} = 99.1\%$

对于新的混合气体分布模式的正确识别率为：

$\min\{118/120, 267/270, 495/500\} = 98.8\%$。

通过实验数据的验证，可以得出采用混合气体分布模式识别→混合气体分析→结果输出思路确定的 SVM 在线模式识别模型可行，具有实际的工程意义，可应用于混合气体组分浓度分析。

3 结 论

本文的基于 SVM 的混合气体分布模式红外光谱在线识别方法解决了如下的问题。

(1)提出了混合气体分布模式识别的概念，用来解决无法获得海量混合气体样本的问题。

(2)采用先用模式识别模型进行混合气体分布模式识别，根据模式识别模型的结果输出，确定对应的校正模型，然后再进行混合气体组分浓度分析的思路，将 6 000 个混合气体红外光谱数据样本划分为 60 种混合气体分布模式，对混合气体红外光谱数据样本在线识别进行了实验验证，证明方法可行。

在实际应用中，为进一步加快在线学习速度，也可以采用遗忘策略，忽略满足当前 KKT 条件即能被当前分布模式 SVM 正确分类的训练光谱，只对违反 KKT 条件的训练光谱进行知识积累和在线学习。

参 考 文 献

[1] Blanco M, Villarroya I. Trends in Analytical Chemistry, 2002, 21(4): 240.

[2] WANG Hai-shui, WANG Dong-mei, XI Shi-quan(王海水，汪冬梅，席时权). Analysis and Testing Technology and Instruments(分析测试技术与仪器), 2002, 8(3): 136.
[3] LIAN Chen-zhou, LÜ Zi-an, XU Xu-chang(连晨舟，吕子安，徐旭常). Environmental Monitoring in China(中国环境监测), 2004, 20(2): 17.
[4] LU Wan-zhen, YUAN Hong-fu, XU Guang-tong, et al(陆婉珍，袁洪福，徐广通，等). Modern Near Infrared Spectroscopic Analysis Techniques(现代近红外光谱分析技术). Beijing: China Petrochemical Press(北京：中国石化出版社), 2000.
[5] 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.
[6] JI Hai-yan, YAN Yan-lu(吉海彦，严衍禄). Journal of Instrumental Analysis(分析测试学报), 1999, 18(3): 12.
[7] Vapnik V N. The Nature of Statistical Learning. New York: Springer, 1995.
[8] Vapnik V N. Statistical Learning Theory. New York: John Wiley & Sons Inc, 1998.
[9] CHANG Chih-chung, LIN Chih-jen. http://www.csie.ntu.edu.tw/～cjlin/papers/libsvm.pdf.
[10] DU Pei-jun, FANG Tao, TANG Hong, et al(杜培军，方　涛，唐　宏，等). Acta Photonica Sinica(光子学报), 2005, 34(2): 293.
[11] PANG Shi-ping, ZHENG Xiao-ling, HE Ying, et al(庞士平，郑晓玲，何　鹰，等). Advances in Marine Science(海洋科学进展), 2007, 25(1): 91.
[12] BAI Peng, XIE Wen-jun, LIU Jun-hua(白　鹏，谢文俊，刘君华). Opto-Electronic Engineering(光电工程), 2006, 33(8): 37.
[13] LIN Ji-peng, LIU Jun-hua(林继鹏，刘君华). Acta Photonica Sinica(光子学报), 2006, 35(3): 408.
[14] SUN Su-qin, TANG Jun-ming, YUAN Zi-min, et al(孙素琴，汤俊明，袁子民，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2003, 23(2): 258.
[15] NIE Lei, LUO Guo-an, CAO Jin, et al(聂　磊，罗国安，曹　进，等). Acta Pharmaceutica Sinica(药学学报), 2004, 39(2): 136.
[16] YING Wei, WANG Zheng-ou, AN Jin-long(应　伟，王正欧，安金龙). Computer Engineering(计算机工程), 2006, 32(16): 74.
[17] PENG Bin-bin, JIN Xiang-yu, XU Xiao-gang, et al(彭彬彬，金翔宇，徐晓刚，等). Computer Science(计算机科学), 2003, 30(3): 75.
[18] ZENG Wen-hua, MA Jian(曾文华，马　健). Journal of Xiamen University(Natural Science)(厦门大学学报·自然科学版), 2002, 41(6): 687.
[19] XIAO Rong, WANG Ji-cheng, SUN Zheng-xing, et al(萧　嵘，王继成，孙正兴，等). Journal of NanJing University(Natural Sciences)(南京大学学报·自然科学版), 2002, 38(2): 152.

Method of Infrared Spectrum On-Line Pattern Recognition of Mixed Gas Distribution Based on SVM

BAI Peng[1,2], JI Juan-zao[3], ZHANG Fa-qi[3], LI Yan[2], LIU Jun-hua[4], ZHU Chang-chun[1]
1. School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an　710049, China
2. Science Institute, Air Force Engineering University, Xi'an　710051, China
3. Engineering Institute, Air Force Engineering University, Xi'an　710038, China
4. School of Electrical Engineering, Xi'an Jiaotong University, Xi'an　710049, China

Abstract　In order to solve the difficulties that the spectrum training data samples of the massive mixed gas cannot be actually obtained, the analysis precision is low and it is not real time online analysis in the analysis of mixed gas component concentration, the support vector machine, a new information processing method, was used in the mixed gas infrared spectrum analysis, and the concept of mixed gas distribution pattern was proposed in the present paper. Based on the thought that the mixed gas distribution pattern recognition is carried out first, and then the analysis work of mixed gas component concentration is done, sixty kinds of mixed gas distribution pattern were determined after investigation and study, and 6 000 mixed gas spectrum data samples were used for model training and testing. Sequential minimal optimization algorithm was applied to realize the decrement and the increase of online learning, and finally the model of infrared spectrum online pattern recognition of mixed gas distribution based on SVM was established. The model structure is composed of 2 levels, pattern recognition level and result output level. The pattern recognition level completes the task of mixed gas distribution pattern recognition; while the result output level is composed of 60 SVM calibration models, and it completes the task of mixed gas concentration analysis. Experimental results show that the correct recognition rate of mixture gas distribution pattern is not lower than 98.8%, and that the method can be used for online recognition of mixed gas distribution pattern under the conditions of small samples of a mixed gas, and can add new mixed gas online, and it has the practical application value.

Keywords　Support vector machine; Infrared spectrum; Calibration model; Pattern recognition

(Received Oct. 18, 2007; accepted Jan. 26, 2008)