第28卷, 第2期
2008年2月

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

Vol. 28, No. 2, pp299-302
February, 2008

层次式 SVM 子集含烃类混合气体光谱分析方法

白 鹏[1, 2], 谢文俊[3], 刘君华[2]

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

摘 要 含烃类混合气体具有组分多、组分浓度范围大的特点。为了解决海量混合气体光谱数据样本实际上是无法实现的难题, 在大量调查的基础上, 研究探索了实际工程中可能出现的混合气体分布模式, 最后确定为15种混合气体分布子模式, 共计5 500个光谱数据样本用于训练与检验。在此基础上, 按照混合气体分布子模式识别→混合气体分析→结果输出的思路, 提出了2层15子集的含烃类混合气体分析方法。多层次多子集软件集成框架以15种混合气体分布子模式为基本框架, 由于应用了基于样本关联规则及混合气体分布模式中心集的SVM快速在线分类方法, 可向原基本框架在线实时的加入新的混合气体分布子模式。实验结果显示, 混合气体组分浓度分析的最大绝对误差为0.41%, 最大平均绝对误差为0.04%。可用于其他混合气体的红外光谱分析, 具有实际应用价值。

关键词 支持向量机; 校正模型; 子集; 红外光谱; 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)02-0299-04

引 言

含烃类混合气体由甲烷、乙烷、丙烷、异丁烷、正丁烷、异戊烷、正戊烷等气体组成, 具有组分多、组分浓度范围大的特点。例如, 如果含烃类混合气体的组分为七种, 组分气体的浓度为0~100%, 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10^7个数据点。实际工作中, 构造数量如此多且分布合理的混合气体样本是不容易实现的。为了解决海量数据样本实际上是无法实现的难题; 同时, 也为了提高混合气体组分气体浓度的分析精度, 本文提出了层次式支持向量机[1, 2] (support vector machine, SVM)子集含烃类混合气体光谱分析方法。

按照混合气体分布子模式识别→混合气体分析→结果输出的思路, 确定层次式SVM子集分析方法由模式识别层和混合气体组分浓度分析层2层组成。模式识别层用于混合气体分布子模式的识别; 混合气体组分浓度分析层为具体混合气体组分浓度分析和结果输出层, 利用建立的SVM校正模型, 分析计算出具体混合气体组分浓度分析结果。本文在大量调查的基础上, 研究探索了实际工程可能出现的混合气体分布模式, 确定为15种混合气体分布子模式, 共计5 500个光谱数据样本用于训练与检验。以15种混合气体分布子模式为基本框架, 应用基于样本关联规则及混合气体分布模式中心集的SVM快速在线分类方法, 实现了含烃类混合气体组分浓度分析。

分析方法由于应用了基于样本关联规则及混合气体分布模式中心集的SVM快速在线分类方法, 可向原基本框架在线实时的加入新的混合气体分布子模式, 具有可扩展性。并具有理想的学习速度、组分浓度分析精度以及泛化能力, 可用于其他类似混合气体组分浓度的分析。

1 基本概念

混合气体分布子模式: 混合气体组分气体浓度分布的一种模式。混合气体组分浓度分布子模式的建立, 主要考虑含烃类混合气体组分浓度和组分种类分布的情况。同时, 也要考虑实际应用的具体情况。

根据上述的考虑, 对含烃类混合气体组分浓度和组分种类分布的情况进行了大量的调查研究, 在此基础上, 通过如下的技术手段建立了混合气体分布子模式。

(1) 实地调查与查阅有关的文献资料;
(2) 实地提取样本进行分析;
(3) 考虑本论文的研究与实际应用的需要;
(4) 进行统计分析的加工。

收稿日期: 2006-05-10, 修订日期: 2006-08-20
基金项目: 国家自然科学基金项目(60772016)和陕西省科技计划资助项目(2007K05-05)资助
作者简介: 白 鹏, 1967年生, 西安交通大学电气工程学院博士后 e-mail: bai-peng410@sohu.com

将目前能掌握和了解的含烃类混合气体组分浓度和组分种类分布的情况进行了分类，划分为 15 种混合气体分布子模式。

层次：完成特定工作任务所进行的任务合理分解和分配。

本文第 1 层是模式识别层[3]，用于 15 种混合气体分布子模式的识别；第 2 层是具体混合气体组分浓度分析和结果输出层，利用 SVM 校正模型，得出具体混合气体组分浓度分析结果[4, 5]。

SVM 子集：完成某混合气体分布子模式具体分析任务的 SVM 子集，由 SVM 校正模型组成，SVM 校正模型参数的选择与子模式有关。

SVM 子集与混合气体分布子模式一一对应，15 种混合气体分布子模式，需要 15 个 SVM 子集。每个 SVM 子集的 SVM 校正模型、SVM 校正模型参数等不同。

任务：完成特定的工作任务。具体到含烃类混合气体组分浓度分析，就是要完成对未知混合气体样本组分浓度的分析，输出结果。

子任务：SVM 子集完成的具体工作。

关联规则：混合气体分布子模式中气体组分浓度分布规则。例如，混合气体组分气体浓度的关联规则为：甲烷> 乙烷> 丙烷> 丁烷> 戊烷；甲烷> 丙烷> 乙烷> 丁烷> 戊烷。

2 层次式 SVM 子集混合气体分析原理

层次式 SVM 子集是在 SVM 基本结构[6, 7] 基础上提出的一种新的支持向量机实际应用结构。

2.1 层次式 SVM 子集结构

在层次式 SVM 子集结构中，处理复杂任务的层次、SVM 子集的数量和各子集所要处理的子任务是根据任务的复杂性，人为规定的。即人为地对一复杂任务进行合理的层次分解和分配，分解后的任务由 SVM 子集完成。这样就使得每一个 SVM 子集所处理的子任务的复杂性是基本均匀的，而且各子集所擅长处理的子任务之间的界限较明显。

设有含烃类混合气体组分浓度分析任务，其样本集记为 $S_{(0)}$，模式识别层的任务就是将待分析的混合气体样本集 $S_{(0)}$ 识别为 n 个混合气体组分浓度分布子模式，对应的支持向量机子任务集为 $S_{(1)}$，$S_{(2)}$，…，$S_{(n)}$，它们满足如下关系

$$S_{(0)} = S_{(1)} \cup S_{(2)} \cdots \cup S_{(n)} \quad (1)$$

按照混合气体分布子模式识别→混合气体分析→结果输出的思路，层次式 SVM 子集用于含烃类混合气体分析的原理如图 1 所示。

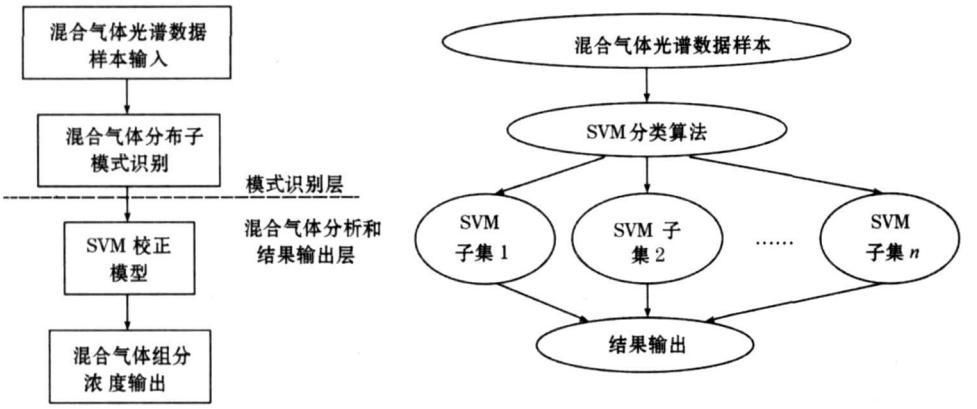

Fig 1 Principle of multi-level and SVM-subset

层次式 SVM 子集应用结构包括模式识别层和混合气体分析和结果输出层 2 层，15 个 SVM 子集，对应于 15 种混合气体分布子模式，有 15 个 SVM 校正模型。

实际应用时，模式识别层对输入的光谱数据样本进行混合气体分布子模式识别[8, 9]，根据该光谱数据样本所属的混合气体分布子模式，分配给相应的 SVM 子集，SVM 子集按照事先获得的混合气体光谱数据样本集作为训练的数据集，经过训练后处理相应的子任务。从整体来说，应用结构对一复杂任务实现了分而治之的目的。

2.2 基于关联规则的在线分类方法

在应用 SVM 进行混合气体红外光谱分析的前期实验中发现，采用单级、单模型结构即将全部的训练样本用来训练一个大的模型，与同样训练样本根据一定的规则分为几个小模型的 SVM 子集相比，用一个大模型的结果输出误差较大，具体数据参见表 1 所示的结果。

因此，考虑如下的思路：先用一个大的模型进行分类，根据大的模型分类结果输出，确定小的模型即 SVM 子集，然后再进行混合气体分析，进而输出结果。这就涉及应用 SVM 对红外光谱数据样本进行混合气体分布子模式识别分类的问题，模式识别分类的输出为小模型所对应的 SVM 子集。

对于满足一定样本分布关联规则的光谱数据样本集，其光谱的分布具有聚类性[10, 11]。每个混合气体分布模式中心所代表的训练样本集可以根据不同的相似度度量，根据不同的相似度阈值，对训练样本集进行聚类分析，生成一定数量的混合气体分布模式中心。由于每个混合气体分布模式中心代表一个对应的光谱训练样本子集，多个混合气体分布模式中心集代表了整个训练样本集。所以，可以将混合气体分布模

式中心集作为 SVM 校正模型的训练样本，并将训练后的 SVM 校正模型应用于实际应用中。

基于关联规则的在线分类方法原理及流程如图 2 和图 3 所示。

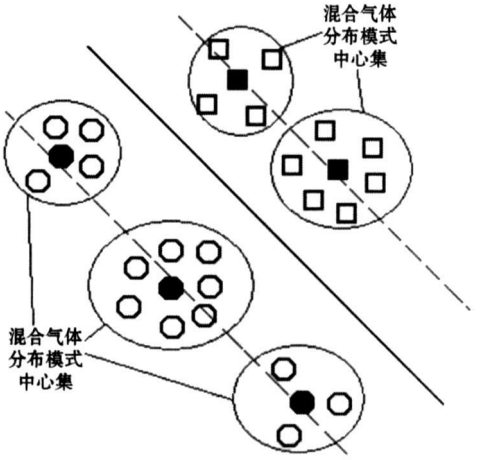

Fig 2　Principle of online categorization algorithm based on relational rule

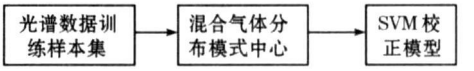

Fig 3　Flow chart of online categorization algorithm based on relational rule

3　实验结果

实验用含烃类混合气体包含甲烷，乙烷、丙烷、异丁烷及正丁烷等烃类气体。

实验用红外光谱仪为 Bruker 公司的 TENSOR27 型傅里叶变换红外光谱仪，扫描范围为 4 000~ 400 cm^{-1}，扫描间隔为 12 nm 的 1 866 个透射光谱数据，共得到 5 500 个光谱数据样本。

为分析对比，建立如图 4 所示的单级、单模型结构，用全部的 5 500 个数据样本的一半样本用来训练一个大的模型；然后用另外的一半样本对模型进行检验。

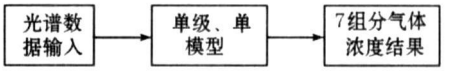

Fig 4　Structure of one-level and single model

Table 1　Experimental result of one-level and single model and multi-level SVM-subset

结构类型	误差指标	甲烷	乙烷	丙烷	异丁烷	正丁烷	异戊烷	正戊烷
单级、单模型结构	最大绝对误差值 Max AE/%	0 499	0 278	0 278	0 182	0 228	0 164	0 088
	平均绝对误差值 Mean AE/%	0 056	0 032	0 041	0 026	0 017	0 008	0 006
层次式SVM 子集结构	最大绝对误差值 Max AE/%	0 408	0 251	0 267	0 154	0 251	0 155	0 074
	平均绝对误差值 Mean AE/%	0 042	0 021	0 028	0 016	0 013	0 007	0 006

单级、单模型结构结果输出的误差较大，与层次式 SVM 子集结果输出比较如表 1 所示。

从表 1 的实验数据可以看出，采用层次式 SVM 子集应用结构的结果明显好于单级、单模型结构的结果。

4　结　论

通过对实验数据的验证，采用混合气体分布子模式识别→混合气体分析→结果输出思路确定的层次式 SVM 子集含烃类混合气体光谱分析方法可行。分析方法的最大绝对误差为 0. 41%，最大平均绝对误差为 0 04%。既可应用于实际含烃类混合气体组分浓度分析，也可应用于其他类似混合气体组分浓度分析，为混合气体组分浓度分析提供了新的方法，具有实际的工程意义。

参 考 文 献

[1] Vapnik V N. Statistical Learning Theory. New York: John Wiley & Sons Inc, 1998.

[2] Vapnik V N. IEEE Trans on Neural Networks, 1999, 10(5): 988.

[3] DING Hui, LIU Jun-hua, SHEN Zhong-ru(丁 晖, 刘君华, 申忠如). Chinese Journal of Scientific Instrument(仪器仪表学报), 2001, 22(6): 592.

[4] XU Guang-tong, YUAN Hong-fu, LU Wan-zhen(徐广通, 袁洪福, 陆婉珍). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2000, 20(2): 134.

[5] ZHANG Lu-da, SU Shi-guang, WANG Lai-sheng, et al(张录达, 苏时光, 王来生, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(1): 33.

[6] BAI Peng, LIU Jun-hua(白 鹏, 刘君华). Chinese Journal of Scientific Instrument(仪器仪表学报), 2006, 27(10): 1242.

[7] 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.

[8] CHANG Chih-chung, LIN Chih-jen. http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf.

[9] WANG Ling, MU Zhi-chun, GUO Hui(王 玲, 穆志纯, 郭 辉). Journal of University of Science and Technology Beijing(北京科技大学学报), 2007, 29(8): 855.

[10] 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.

[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.

Method of Infrared Spectrum Analysis of Hydrocarbon Mixed Gas Based on Multilevel and SVM-Subset

BAI Peng[1,2], XIE Wen-jun[3], LIU Jun-hua[2]

1. Science Institute, Air Force Engineering University, Xi'an 710051, China
2. School of Electrical Engineering, Xi'an Jiaotong University, Xi'an 710049, China
3. Engineering Institute, Air Force Engineering University, Xi'an 710038, China

Abstract The hydrocarbon mixed gas was characterized by multi-component and varied density. In order to deal with the difficulties that can not be actually solved with mass mixture gas spectrum data samples, 15 kinds of subset patterns were determined on the basis of investigations and studies, which needed 5 500 spectrum data samples for training and testing. On the basis of this, a method of hydrocarbon mixed gas infrared spectrum analysis based on 2-levels and 15 SVM-subsets was proposed in the light of the idea of working pattern recognition→mixture gas analysis→the final result output. In order to solve the problem of new subset working pattern, the SVM online categorization algorithm based on spectrum data relational rule was used. The experimental results show that the component concentration maximal deviation is 0.41% and the maximal average deviation is 0.04%. The method can be used in other mixture gas infrared spectrum analyses, and has the theoretic and application value.

Keywords Support vector machine; Calibration model; Subset; Infrared spectrum; Quantitative analysis

(Received May 10, 2006; accepted Aug. 20, 2006)