第 26 卷 第 1 期
2006 年 1 月

光 学 学 报
ACTA OPTICA SINICA

Vol. 26, No. 1
January, 2006

文章编号：0253-2239(2006)01-0147-5

基于支持向量机的非线性荧光光谱的识别*

李素梅 韩应哲 张延炘 常胜江 申金媛

（南开大学信息技术科学学院 教育部光电信息技术重点实验室，天津 300071）

摘要： 提出将支持向量机网络应用于含不同浓度杂质气体的非线性荧光光谱的识别。由于原始光谱数据的光谱通道数目很大，首先用小波变换去噪压缩，然后采用主成分分析方法对光谱信息进行连续两次的特征提取。在保持原光谱数据主要信息基本不变的情况下，将数据维数由 3979 压缩到 514（小波变换）并提取 9 个主成分。这样，不仅减少了网络的输入维数，而且加快了网络的训练速度。实验结果表明，无论对训练样本还是未学习过的测试样本，其正确识别率均可达到 100%。网络的训练和测试速度较快，可以更有效地应用于大气杂质气体的实时监测。

关键词： 光谱学；非线性荧光光谱；支持向量机；小波变换；主成分分析

中图分类号：O433.4；TP183　　文献标识码：A

Recognition of Nonlinear Fluorescence Spectrum of Support Vector Machine Networks

Li Sumei　Han Yingzhe　Zhang Yanxin　Chang Shengjiang　Shen Jinyuan

(*Key Lab of Opto-electronics Information Technical Science, EMC, College of Information Technical Science, Nankai University, Tianjin* 300071)

Abstract: That the support vector machine network is applied to recognize the nonlinear fluorescence spectrum of impurities of different concentrations in air is proposed. Because the number of spectrum channel of the original spectrum data is large, it is cleaned up and compressed through wavelet trausform firstly, and then the principal component analysis (PCA) is used to extract the character information twice in series. It not only ensures the character of original nonlinear fluorescence spectrum, but also compresses the data number the nonlinear fluorescence spectrum from 3979 to 514, and extracts 9 principal components, which reduces the number of the input vector and improves the training speed of the network. The simulation results show that the correct recognition rates for both training spectrum samples and unlearned test spectrum samples reach 100%. So, the training and testing speed is fast enough to monitor the atmospherical impurity in air in real time.

Key words: spectroscopy; nonlinear fluorescence spectrum; support vector machine; wavelet transform; principal component analysis

1 引　　言

目前，监测大气污染的方法很多[1~4]，这些方法都要求先取样后分析，因而不具有实时性。近年来，随着超短脉冲激光技术的快速发展，脉冲激光在大气中传输会出现非线性光学自聚焦和超辐射现象[5,6]。由于气体分子与强光光场的非线性作用，大部分气体分子会被击碎分裂和发生多光子-隧道电离，从而发射具有分子光谱特性的非线性荧光光谱。不同物质具有不同特征的非线性荧光光谱，这些光谱的产生涉及的非线性效应较多，多数气体分子的非线性光谱参量尚属未知。另外，大气中污染气体的含量较低并且种类较多，它们的非线性荧光光谱会出现交叠现象，使得通过直接寻找某种杂质气体的特征光谱来分析气体成分变得非常困难。

* 国家自然科学基金（60277022，60477009）、天津市自然科学基金重点项目（023800811）、天津市科技攻关培育项目（043100811）、博士点基金资助项目（20030055022）和南开大学科技创新基金资助项目。

作者简介：李素梅（1975～），女，河北省人，博士，主要从事人工神经网络及其在智能信息处理中的应用研究。
E-mail：tjnklsm@163.com

收稿日期：2005-03-17；收到修改稿日期：2005-06-07

神经网络的可学习性和推广能力强的特点使得它有潜力成为复杂光谱分析的有力工具。目前，我们已在非线性荧光光谱的神经网络分析方面作了一些初步的研究工作[7,8]并取得了较好的识别结果。然而，由于传统神经网络所具有的缺陷，使得网络的训练质量和推广性没有保证。而支持向量机(support vector machine，SVM)网络采用结构风险最小化准则(structural risk minimization，SRM)，在提高学习精度(即最小化训练误差)的同时，缩小模型泛化误差的上界，即最小化模型的结构风险，从而确保模型具有最优的泛化能力[9]。因此，支持向量机网络更适用于要求实时监测的非线性荧光光谱的识别。但是，实验得到的非线性荧光光谱含有大量的冗余信息，许多无关的信息进入模型之中，不仅使识别的效率下降，而且会使模型精度和稳定性变坏。为此，本文首先采用小波变换的方法对原光谱数据进行压缩，在保证原光谱信息不受损失的前提下，去除噪声信息，大量缩减了数据的维数。然后再用主成分分析方法进行二次特征信息提取，提取出 9 个主成分，不仅进一步缩减了光谱的数据维数，而且消除了样本数据之间的相关性。最后，将两次信息提取后的数据送入支持向量机网络，不仅大大降低了网络的输入维数，而且加快了网络的学习速度。

2 算法简介

2.1 小波变换

小波变换[10,11]可同时在时域和频域上分析信号的局部特性，是一种窗口面积固定但其形状、时间窗和频率窗都可改变的时频局部化分析方法。它在低频部分具有较高的频率分辨率和较低的时间分辨率，在高频部分具有较高的时间分辨率和较低的频率分辨率，所以被誉为“数学显微镜”。正是这种特性，使小波变换具有对信号的自适应分析能力。

在对光谱信号进行小波分析时，根据小波分析后的重构信号和原始信号的均方根误差大小来比较判定小波函数的好坏，最终选定对实际问题为最优的小波函数。恢复根均方差(E_{RMSD}表示)的定义如下：

$$E_{\mathrm{RMSD}}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_{\mathrm{ri}}-s_{\mathrm{oi}})^2}, \tag{1}$$

(1)式中 s_{ri} 是重建数据；s_{oi} 是原始数据；N 是原始数据的长度。对光谱数据来说，噪声的频率一般较高，而光谱信号的频率相对较低。另外，除了选择合适的小波函数之外，分解尺度的选取也是很重要的。当选定合适的分解尺度后，就可以在不同尺度上设定阈值去除噪声，同时保留低频部分的原始光谱信息。

2.2 主成分分析

主成分分析是由 Karl Parson 在 1901 年提出的[12]，它是在基本保持原变量信息不变的前提下，通过空间变换，选择原变量中少数几个分量的线性组合来代替原变量，并揭示原变量之间关系的一种数学分析方法。线性组合的选择是以综合指标的方差大小为准则的，方差越大表示包含的信息越多。求样本数据集的主成分实际上就是求此数据集的协方差矩阵的特征值与其对应的单位正交特征向量。为了衡量各主成分的贡献，我们来考察各个主成分的累计方差贡献率，累计方差贡献率越大，丢失的数据信息就越少。

2.3 支持向量机

支持向量机是由 Vapnik 等[9]提出的一种新型的机器学习算法，它的理论基础是由 Vapnik 创建的统计学习理论(Statistic learning theory，SLT)。目前，它主要应用于模式分类和非线性回归问题中，由于其优越的学习能力，在国内外学术界受到广泛关注，已经在很多领域取得了成功应用，如文本识别[13,14]，人脸的识别[15,16]，三维物体识别[17]等。在模式分类中，支持向量机是从线性可分情况下的最优分类面发展而来的，它的核心思想是最优分类面不但能将两类样本正确分开，而且要使它们之间的分类间隔最大。实际上，我们所遇到的问题大多为非线性情况，这时可以通过非线性变换将非线性问题转化为某个高维空间的线性问题，在变换后的空间中求最优分类面[9]。下面只给出非线性情况下的支持向量机数学模型。

假设在非线性情况下，样本点为 (x_i,y_i)，$i=1,2,\cdots,n$，$x\in\{+1,-1\}$，在高维空间中，其分类面方程为 $w\cdot\phi(x)+b=0$，$\phi:x\to z$ 是从输入空间到特征空间的一个映射。此时，要优化的目标函数为

$$\begin{cases}\min Q(w)=\dfrac{1}{2}w^{\mathrm{T}}w,\\ y_i\{[(w\cdot\phi(x)]+b\}-1\geqslant 0,\\ \qquad i=1,2,\cdots,n\end{cases} \tag{2}$$

利用拉格朗日优化方法可以把上述最优分类面问题转化为其对偶问题：

$$\begin{cases}\max Q(\alpha)=\displaystyle\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j),\\ 0\leqslant\alpha_i\leqslant C,\ i=1,2,\cdots,n\end{cases} \tag{3}$$

其中 $K(x_i,x_j)=\langle\phi(x_i),\phi(x_j)\rangle$ 为满足 Mercer 条件的核函数，C 是一个常数，它控制对错分样本的惩

罚程度。此时相应的分类函数为

$$f(x) = \mathrm{sgn}\Big[\sum_{i=1}^{n}\alpha_i^* y_i K(x_i, x) + b^*\Big].$$

通常，核函数的选择有很多种，常用的核函数有 d 次多项式核函数、高斯型核函数和 3 层神经网络核函数。用于识别的支持向量机网络结构如图 1 所示。

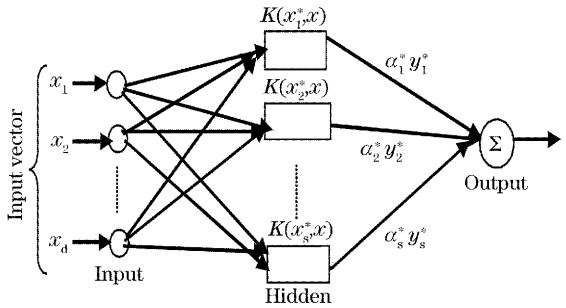

图 1　支持向量机结构示意图

Fig. 1　The illustration of model for support vector machine

3　实验结果及分析

本文分析的光谱数据是由加拿大拉维尔大学超短、超强激光研究中心所提供的。通过飞秒激光脉冲在含不同杂质成分的空气样本中的非线性作用，采用多通道分析的光谱仪得到样本的非线性荧光光谱。实验中输入脉冲的激光能量为 10 mJ，脉宽为 50fs，重复频率为10Hz。掺杂的气体分别为乙烯(Ethylene)，1－丁烯(1－Butene)和 N－丁烷(N－Butane)三种，共测得 27 个样本。掺杂气体的体积分数分别为，乙烯(25%，12.5%，6.25%，3.13%，1.56%，0.78%，0.39%，0.20%，0.10%，0.049%)；1－丁烯(25%，12.5%，6.25%，0.13%，1.56%，0.78%，0.39%，0.20%，0.10%，0.05%)；N－丁烷(12.5%，1.56%，0.78%，0.39%，0.20%，0.098%，0.049%)。可以看出，为了实验的方便，杂质气体的浓度大体上是按等比级数变化的。

3.1　小波去噪

用 Matlab 软件提供的 54 个小波函数对乙烯，1－丁烯，N－丁烷光谱作恢复根均方差平均值。对比分析表明，对于 3 种气体的荧光光谱，bior3.9 小波的恢复根均方差值总和最小。采用 bior3.9 小波函数对非线性荧光光谱进行尺度分析，然后选取低频系数跟原信号进行了对比。作为例子，图 2 是对 N－丁烷(0.098%)所进行的第一到第七层小波分解结果，其它光谱的分解结果也大致相同。从结果的对比可以看出 4 层以下的低频系数图形较好地保持了原信号的特征。比较合适的是 3 层小波分解，在完整地保留了原特征光谱的同时，光谱数据由 3979 个点压缩到 514 个点，数据量大大缩减。为了更有效地组织小波系数，避免相关系数之间的重叠，对变换后的 514 个数据又进行了主成分分析。

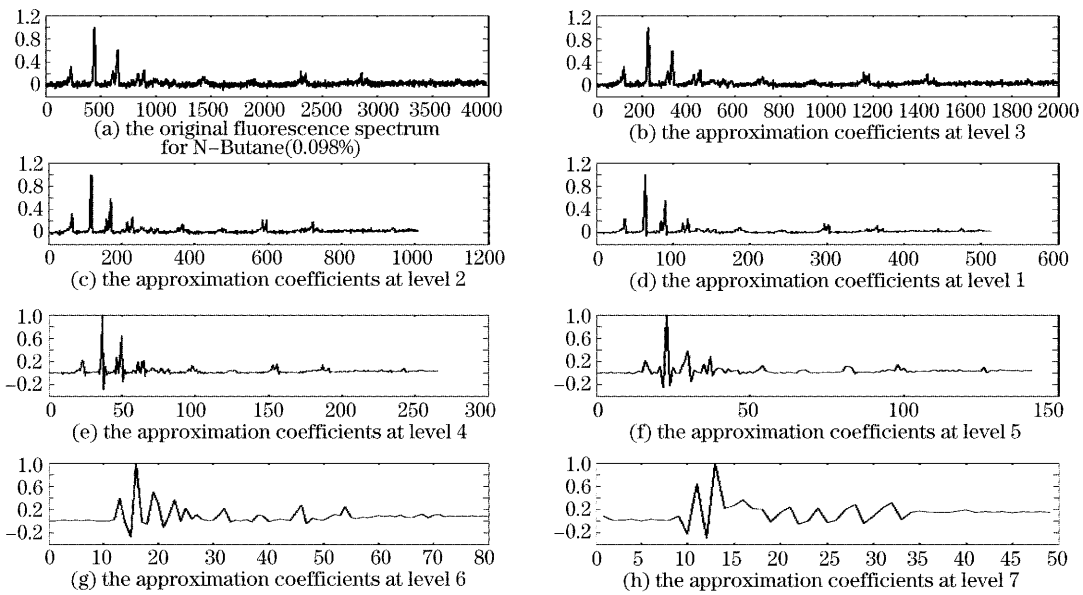

图 2　第一到第七层小波分解的结果示意图。(a)N－丁烷(0.098%)的原始光谱，采用(b)1 层，(c)2 层，(d)3 层，(e)4 层，(f)5 层，(g)6 层，(h)7 层小波分解的低频系数

Fig. 2　Illustration of the results for wavelets decomposition from the first to the seventh level. (a) The original fluorescence spectrum for N－butane (0.098%), the approximation coefficients at level 1 (b), level 2 (c), level 3 (d), level 4 (e), level 5 (f), level 6 (g) and level 7 (h)

3.2 主成分分析提取特征信息

对小波去噪后得到的514个光谱数据进行主成分分析，各个主成分的累计方差贡献率如表1所示。由表1可知，前9个主成分的累计贡献率已经达到99.403%，基本包含了光谱信息的绝大部分特征信息，因此我们认为用前9个主成分就可以表征原光谱信息。

表1 光谱数据的前9个主成分的贡献率

Table 1 The contribution rate of the first 9 principal components of the spectrum

Principal components	Contribution rate of variance /%	Cumulate contribution rate of variance /%
1	78.6441	78.6441
2	14.0703	92.7144
3	2.6648	95.3792
4	1.9887	97.3679
5	1.164	98.532
6	0.40445	98.9634
7	0.22689	99.1633
8	0.15152	99.3148
9	0.088199	99.403

3.3 支持向量机网络杂质气体识别

从以上光谱信息的提取中可以发现，原光谱信息经小波变换后由3979个数据点压缩到514个点；而被压缩后，514个光谱信息数据点的绝大部分信息都集中到了前9个主成分中。将主成分分析方法得到的9个主成分作为支持向量机网络的输入。由于支持向量机通常适用于二分类的样本，而这里要识别的气体种类有三种，属于多目标的分类问题。为此，选用两个二分支持向量机网络和聚类编码方法[18]对气体进行识别。三种气体的编码分别为：乙烯：(−1，−1)；1−丁烯：(1，−1)；N−丁烷：(1，1)。也就是说，对其中一个网络，当输出为1时，气体为乙烯，当输出为−1时，气体为1−丁烯和N−丁烷；而对另一个网络，输出1时，气体为N−丁烷，输出为−1时，气体为乙烯和1−丁烯。这样对27个光谱数据，我们选取10个样本作为训练样本，其余17个样本作为测试样本。两个二分类支持向量机网络的核函数均选取高斯型核函数。即，

$$K(x,x_i)=\exp\left[-\frac{|x-x_i|^2}{2\sigma^2}\right].$$

在训练过程中两个支持向量机二分网络的最优参量 C,σ 分别为60，1.6和80，6。两个网络支持向量的个数均为8。仿真实验结果表明，无论是对10个训练样本还是对未学习过的17个测试样本，其正确识别率均达到100%。

4 结 论

本文对小波变换压缩后的非线性荧光光谱数据采用主成分分析方法进行二次信息特征提取，并采用结构风险最小化的支持向量机网络进行识别。仿真实验结果表明，通过对非线性荧光光谱信息的两次特征提取，不仅对训练样本而且对未学习过的测试样本都达到了100%的正确识别率，而且由于有效地去除非线性荧光光谱信息的噪声，大大简化了网络的输入维数，减少了网络的训练时间，能够更有效地应用于大气杂质气体的实时监测。

参 考 文 献

1 Zhou Bing, Liu Wenqing, Qi Feng *et al.*. Study on differential optical absorption spectrometry for atmospheric pollutants monitoring[J]. *Research of Environmental Sciences*, 2001, **14**(5): 23～26 (in Chinese)
周 斌，刘文清，齐 峰 等. 差分光学吸收光谱法测量大气污染气体的研究[J]. 环境科学研究，2001，**14**(5)：23～26

2 F. Evangelisti, A. Baroncelli, P. Bonasoni *et al.*. Differential optical absorption spectrometer for measurement of tropospheric pollutants[J]. *Appl. Opt.*, 1995, **34**(15): 2737～2744

3 Zhao Zhong, Wang Rongzong, Sun Tianhui. Portable gas leakage monitor with mass spectrometer and film technology[J]. *Chinese Space Science and Technology*, 1999, **4**(8): 46～49 (in Chinese)
赵 忠，王荣宗，孙天辉. 质谱薄膜联用小型气体泄漏检测仪研制[J]. 中国空间科学技术，1999，**4**(8)：46～49

4 Liu Wenqing, Cui Zhicheng, Dong Fengzhong. Optical and spectroscopic techniques for environmental pollution monitoring [J]. *Optoelectronic Technology & Information*, 2002, **15**(5): 1～12 (in Chinese)
刘文清，崔志成，董凤忠. 环境污染监测的光学和光谱学技术[J]. 光电子技术与信息，2002，**15**(5)：1～12

5 N. Akozbek, M. Scalora, C. M. Bowden *et al.*. White-light continuum generation and filamentation during the propagation of ultra-short laser pulses in air[J]. *Opt. Commun.*, 2001, **191**(3): 353～362

6 Hu Xueyuan, Zhong Fangchuan, Deng Jian *et al.*. Ultra-short intense laser pulse propagating in atmosphere: Behavior of self-focusing[J]. *Acta Optica Sinica*, 2001, **21**(6): 641～646 (in Chinese)
胡雪原，钟方川，邓 建 等. 超短强激光脉冲在大气传播中的自聚焦行为[J]. 光学学报，2001，**21**(6)：641～646

7 Shen Jinyuan, Han Yingzhe, Chang Shengjiang *et al.*. Neural network analysis and application of nonlinear fluorescence spectra [J]. *Acta Optica Sinica*, 2004, **24**(7): 1000～1003 (in Chinese)
申金媛，韩应哲，常胜江 等. 非线性荧光光谱的神经网络分析及其应用[J]. 光学学报，2004，**24**(7)：1000～1003

8 Shen Jinyuan, Su xiaoxing, Chang Shengjiang *et al.*. A new method for gas component analysis in air[J]. *J. Optoelectronics · Laser*, 2003, **14**(9): 954～957 (in Chinese)
申金媛，苏晓星，常胜江 等. 一种用于大气中杂质气体识别的新方法[J]. 光电子·激光，2003，**14**(9)：954～957

9 Simon Haykin. *Neural Networks: A Comprehensive Foundation* (second edition)[M]. Beijing: Tsinghua University Press, 2001. 318～350

10 Li Jianping. *Wavelet Analysis and Information Transmission* [M]. Beijing: China Industry Press, 2004. 1～94 (in Chinese)

李建平. 小波分析信息传输基础[M]. 北京：国防工业出版社，2004. 1～94

11 Tang Yuanyan, Wang Ling. *Wavelets Analysis and Text Character Recognition* [M]. Beijing: Science Press, 2004. 1～193 (in Chinese)
唐远炎，王　玲. 小波分析与文本文字识别[M]. 北京：科学出版社，2004. 1～193

12 Lindsay I. Smith. A Tutorial on Principal Components Analysis [R]. 2002. 1～27.
http://kybele.psych.cornell.edu/～edelman/Psych-465-Spring-2003/PCA-tutorial.pdf

13 Lu Zengxiang, Li Yanda. Interactive support vector machine learning algorithm and its application [J]. *J. Tsinghua University* (*Science and Technolgy*), 1999, **39**(7): 93～97 (in Chinese)
卢增祥，李衍达. 交互支持向量机学习算法及其应用[J]. 清华大学学报，1999，**39**(7)：93～97

14 Edgar Osuna, Robert Freund, Federico Girosi. Training support vector machines: an application to face detection[C]. *In Proc. of CVPR '97*, Purrto Rico, 1997. 1～8

15 Bernd Heisele, Purdy Ho, Tomaso Poggio. Face Recognition with Support Vector Machines: Global versus component-based approach. http://cbcl.mit.edu/projects/cbcl/publications/ps/iccv2001.pdf

16 Corinna Cortes, Vladimir Vapnik. Support vector networks, machine learning[J]. 1995, **20**(3): 273～297

17 V. Blanz, B. Scholkopf, H. Bulthoff *et al.*. Comparison of view-based object recognition algorithms using realistic 3D models[C]. *Artificial Neural Networks ICANN '96*, Springer Lecture Notes in computer Science, 1996, Berlin, **1112**: 251～256

18 Zhang Yanxin, Liu Yue, Chen Tianlun. The method of optics correlation and neural networks for multi-objects[J]. *Laser J.*, 2000, **21**(3): 44～46 (in Chinese)
张延炘，刘　玥，陈天伦. 大量目标识别的光学相关与神经网络融合方法[J]. 激光杂志，2000，**21**(3)：44～46

(上接封底)

二、征文要求

1) 论文详细摘要，内容包括题目、姓名、单位、通讯地址、邮编。文中可含一定的图、表。

2) 详细摘要用 Word2000 或 Word XP 排版，一式两份，并通过电子邮件提交电子版一份。

3) 征文截稿日期：2006 年 6 月 20 日(以寄方邮戳为准)。

4) 论文录用与发表：论文详细摘要在 2006 年 7 月中旬经专家审稿录用后，即发会议正式通知，并将优秀论文推荐在《应用光学》上发表。

5) 来稿请寄：南京理工大学电光学院 431 教研室，地址：南京市孝陵卫 200 号，邮编：210094。来稿请在信封上注明：年会征文。

6) 联系人：高志山　沈　华　Tel：025-84315427，84315433，84317569
E-mail：Edward_bayun@163.com

三、关于产品介绍与展销

欢迎企事单位、公司在论文集上刊登广告。会议期间将举办产品展销、信息发布和产品介绍、交流业务，请尽早来函联系。截止日期：2006 年 7 月 1 日(以寄方邮戳为准)。

中国光学学会光学测试专业委员会

2005 年 11 月 1 日