第26卷，第12期
2006年12月

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

Vol. 26, No. 12, pp2232-2235
December, 2006

光谱分析中的支持向量机方法及其性能优化

林继鹏[1,2]，刘君华[1]

1. 西安交通大学电气工程学院，陕西 西安 710049
2. 长安大学信息工程学院，陕西 西安 710054

摘 要 针对红外光谱气体分析中建立数据模型需要标定大量样本的问题，提出一种基于正则理论的支持向量机的小样本机器学习方法，该方法能在获得模型参数全局最优点的同时保证训练误差为零，因而能较好地消除光谱间的交叉敏感现象，利用其良好的非线性映射能力对多组分红外光谱仪的试验结果表明，该方法可使光谱仪的交叉灵敏度下降约81倍。针对支持向量机(SVM)没有足够的理论支持的结构参数选取比较困难的问题，提出一种基于遗传算法和交叉检验相结合的遗传支持向量机(GA_SVM)算法，利用遗传算法的随机搜索特性求取SVM的最优结构参数，在20世代即可求取光谱仪的最小均方根误差(MSE)0.018，并且在算法的前数世代，系统的MSE即已开始成倍下降。这些结果表明GA_SVM光谱仪具有更高的效率和泛华能力。

主题词 支持向量机；遗传算法；泛化能力；结构参数

中图分类号：TP18 **文献标识码**：A **文章编号**：1000-0593(2006)12-2232-04

引 言

气体检测的光谱分析是科研、生产中的一种重要分析手段，已在天然气、石油化工、尾气及烟气等诸多领域有着广泛的应用。气体光谱分析通常有3种途径：非分散型光谱分析，文献[1-3]中所采用的方法均属这种类型。另外一种是分散型光谱分析，文献[4，5]中所采用的方法均属这种类型。针对此二类分析方法，常用的建模方式有最小二乘法、K矩阵法、P矩阵法以及20世纪70年代发展起来的多元线性回归、偏最小二乘法等方法[6]。第三种分析方法是傅立叶光谱分析[7]。

本文研究的主要内容是红外分散型多组分气体分析仪进行多组分气体定量分析和降低多组分气体检测中的交叉灵敏度问题。在获取试验数据的基础上对以上方法的验证结果均不太理想；主要是因为一：这些方法是建立在均匀采样的无穷多样本理论基础上的，而实际上所观测的样本是有限的和离散的、非均匀性的；二是问题的高维特性使得样本只在输入空间形成稀疏分布，致使所建立的模型是病态的。

统计学习理论是目前针对小样本估计和预测的最佳理论，其中支持向量机作为统计学习理论的实现方法之一，对有限非均匀样本下的分类和回归问题进行了系统理论的研究，在很大程度上解决了模型选择与过学习问题、非线性问题、维数灾问题和局部极小点等问题。但支持向量机有两个不足之处：(1)没有给出核函数的选用标准；(2)没有给出在给定样本下如何选择其最优结构参数的方法和理论。由于这两点的不足，给实际工作带来诸多不便。本文根据支持向量机本身的特性，利用遗传算法和交叉检验相结合的办法，提出一种遗传-支持向量机策略，有效提高了气体分析的效率。

1 支持向量机

为了更好的理解SVM在多气体分析中的建模，可考虑噪声环境下实值函数的估计问题，估计的目标是一个没有任何先验知识的函数

$$y = g(x) + \varepsilon \tag{1}$$

其中ε是估计误差，x是一个d维输入向量，y是对应的输出电压。估计是基于有限个(n个)样本$Z^n \sim (x_i, y_i)$，$i = 1, 2, \cdots, n$来实现。其中Z^n是服从独立同分布概率$p(x_i, y_i) = p(x)p(y \mid x)$。因而对(1)的估计可表示为

$$g(x) = \int yp(y \mid x)\mathrm{d}y \tag{2}$$

学习的方法是从广义的预测函数集合$f(x, w)$选择最优函数$f(x, w_0)$使预测的期望风险最小[6]，其中$w \in \Omega$为

收稿日期：2005-07-13，**修订日期**：2005-10-20

基金项目：国家自然科学基金项目(60276037)资助

作者简介：林继鹏，1977年生，西安交通大学电气工程学院博士研究生

函数的广义参数，预测的好坏使用损失函数 $L[y, f(x, w)]$ 来表征。对于函数回归估计而言，一般采用方差损失函数。

统计学习理论认为某种学习策略所支持的函数集合 $f(x, w)$ 不一定包含(2)式所表示的回归函数。因此学习的目的是使用训练样本寻找预测函数 $f(x, w_0)$ 能最小化期望风险

$$R(w) = \int [y - f(x, w)]^2 p(x, y)\mathrm{d}x\mathrm{d}y \tag{3}$$

但实际上无论是函数 $g(x)$，还是样本的分布 $p(x)$ 都是未知的；并且它们往往被认为是非时变的，这样才使得利用过去的数据所作的估计是有意义的，早先的模型参数估计是按照能使模型的经验风险最小来实现

$$R_{\mathrm{emp}}(w) = \frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i, w)]^2 \tag{4}$$

但实际上即使保证了预测函数的经验风险最小，也不能保证预测函数的期望风险最小，这是因为

$$R(w) \leqslant R_{\mathrm{emp}}(w) + \Phi\left(\frac{n}{h}\right) \tag{5}$$

式中 $\Phi\left(\frac{n}{h}\right)$ 表示学习的置信范围，h 是预测函数的 VC 维。根据结构风险最小化的思想，机器学习策略是保持经验风险值固定而最小化置信范围。实现思路是：通过某种事先选择的非线性映射将输入向量 x 映射到一高维特征空间 H，在这个空间中构造最优分类超平面。所以在权 w 空间中的优化问题可以描述为

$$\min_{w, b, \varepsilon} J(w, \varepsilon) = \frac{1}{2}w^T w + \frac{1}{2}\gamma\sum_{k=1}^{n}\varepsilon_k^2 \tag{6}$$

st：$y_k = w^T\varphi(x_k) + b + \varepsilon_k = g(x_k) + \varepsilon_k$

γ 是调和常数，$\varphi(x)$：$R^n \to R^{n_h}$ 高维空间映射函数。J 是均方误差和规则化量之和[7, 8]。(6)式定义的拉各朗日函数为

$$L(w, b, \varepsilon, \alpha) = J(w, \varepsilon) - \sum_{k=1}^{n}\alpha_k\{w^T\varphi(x_k) + b + \varepsilon_k - y_k\} \tag{7}$$

式中，拉各朗日乘子 $\alpha_k \in R$。对(7)式的求极值可得[9]

$$w = \sum_{k=1}^{n}\alpha_k\varphi(x_k),\ \sum_{k=1}^{n}\alpha_k = 0,\ \alpha_k = \gamma\varepsilon_k$$
$$w^T\varphi(x_k) + b + \varepsilon_k - y_k = 0 \tag{8}$$

消除 w 和 ε_k，可得矩阵形式

$$\begin{pmatrix} 0 & I^T \\ I & \Omega + \frac{1}{\gamma}I \end{pmatrix}\begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} 0 \\ y \end{pmatrix} \tag{9}$$

式中，$Y = \{y_1, y_2, \cdots, y_n\}$，$I = \{1, 1, \cdots, 1\}$，$\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_n\}$，$\Omega_{ij} = \varphi^T(x_i)\varphi(x_j)$。选择 RBF 核函数

$$\Omega_{ij} = k(x_i, x_j) = \mathrm{e}^{-\|x_i - x_j\|^2/\delta} \tag{10}$$

则最优预测函数 $g(x)$ 的最终可表示为

$$g_0(x)\sum_{k=1}^{n}\alpha_k k(x, x_k) + b \tag{11}$$

式中 α_k 和 b 可由(9)式确定。很明显，α_k，b 是调和参数 γ 以及核参数 δ 的函数。

2　遗传算法优化 SVM 策略

遗传算法是 1 种群体型操作，该操作以群体中所有的个体为对象，对于群体中的个体进行 3 种基本运算：选择、交叉和变异，结合的适应度函数构成了遗传算法的主体[10]。

2.1　GA-SVM 算法

2.1.1 编码

对于一个优化问题，首先需要用编码的方式描述问题的解。设二进制编码的长度为 d，从集合 $T = \{1, 2, 3, \cdots, d\}$ 中随机选取 m 个数组成一条染色体 S，即 $S = i_1, i_2, \cdots, i_m$，其中 $1 \leqslant i_j \leqslant d$，$j = 1, 2, \cdots, m$，表示第 i_j 个特征被选取。S 中可以有相同的特征，因而实际选择的特征数 $k \leqslant m$。重复以上步骤，可生成 N 个个体，形成第一代种群。此时编码形成的解空间 Γ 容量为 $\sum_{i=1}^{m}\sum_{k=1}^{i}C_d^k$，如果 $m \ll d$，则 $\sum_{i=1}^{m}\sum_{k=1}^{i}C_d^k \ll 2^d$。因而为了提高遗传算法的效率，一般选择较小的 m。

2.1.2　适应度函数

已知 SVM 参数空间，则优化的目标是最小化均方根误差

$$\min \quad \frac{1}{N_v}\sqrt{\sum_{i=1}^{S}(y_r^i - y_e^i)^2}$$

y_r^i 是第 i 个检验样本的预测输出，y_e^i 是第 i 个检验样本的期望输出，N_v 是检验样本的数量。因而，适应度函数可采用

$$f(S) = \frac{\sqrt{\sum_{i=1}^{S}(y_r^i - y_e^i)^2}}{N_v} \tag{12}$$

2.1.3 选择算子

选择算子采用比例选择法，即个体被选择的概率与其适应度函数成正比

$$p_i = \frac{f(S_i)}{\sum_{j=1}^{N}f(S_j)} \tag{13}$$

选择的方法有轮盘赌选择法、随机遍历抽样法、局部选择法以及锦标赛选择法。本文选择最简单的轮盘赌选择法。该方法需要进行多轮选择来确定交配个体。每一轮产生一个[0, 1]均匀随机数，将该随机数作为选择指针来确定被选个体。

2.1.4　交叉和变异算子

交叉采用等概率随机确定的基因位置作为交叉点，以概率 p_c 交换 2 个个体的前后两部分，形成新的个体。变异按基因位进行，每一个基因按概率 p_m 随机进行。

2.1.5　终止判据的确定

终止判据有 2 种，一种是在给定的世代终止，另一种是在最优的世代终止。本文按第 1 种方式进行。

2.2　多组分气体分析

在多组分气体测试分析中，混合气体或多或少的会对某一被测量产生干扰；消除干扰的方法如果完全从硬件上去考虑，技术难度比较大，同时也会提高系统的成本。我们所使用的试验系统的自行研制的智能分光型多组分气体分析仪，通过组合滤光片选择不同被测成分的中心光谱，虽然对交叉

灵敏度有一定的抑制作用，但还不能满足实际的需要[10, 11]。在仪器的量程 CH_4(0～7 600 μg·mL^{-1})和 C_2H_4(0～7 800 μg·mL^{-1})内共进行了 72 次试验，分别取表 1 所示的试验数据作为 SVM 的训练样本和表 2 所示的试验数据作为 SVM 的检验样本。

Table 1 Input and output voltage of CH_4 sensor, demarcated training samples

$\Phi(CH_4)\times10^6$	不同 $\Phi(C_2H_4)$$CH_4$ 传感器输出电压/V						干扰引起波动
	50	1 000	3 000	4 500	6 000	7 800	
50	16.537	16.536	16.526	16.485	16.465	16.453	0.084
1 000	16.501	16.494	16.479	16.449	16.427	16.418	0.083
3 700	16.430	16.414	16.384	16.367	16.355	16.343	0.087
4 500	16.414	16.409	16.409	16.397	16.385	16.356	0.056
5 500	16.365	16.357	16.345	16.345	16.336	16.313	0.051
7 000	16.284	16.279	16.259	16.241	16.222	16.201	0.083

Table 2 Input and output voltage of CH_4 sensor, demarcated validating samples

$\Phi(CH_4)\times10^6$	不同 $\Phi(C_2H_4)$下 CH_4 传感器输出电压/V						干扰引起波动
	0	1 520	3 040	4 560	6 080	7 600	
0	16.546	16.538	16.472	16.458	16.431	16.385	0.161
1 520	16.416	16.409	16.381	16.364	16.355	16.317	0.099
3 040	16.342	16.332	16.320	16.310	16.302	16.275	0.067
4 560	16.327	16.312	16.287	16.285	16.244	16.239	0.088
6 080	16.304	16.280	16.266	16.254	16.224	16.226	0.078
7 600	16.268	16.264	16.250	16.230	16.196	16.188	0.080

Table 3 Fused output voltage of CH_4 sensor, validating outputs

$\Phi(CH_4)\times10^6$	不同 $\Phi(C_2H_4)$下 CH_4 传感器输出电压/V						波动/V
	0	1 520	3 040	4 560	6 080	7 600	
0	16.546	16.546	16.545	16.545	16.546	16.546	0.001
1 520	16.416	16.415	16.415	16.416	16.416	16.415	0.001
3 040	16.342	16.342	16.342	16.342	16.341	16.344	0.002
4 560	16.327	16.326	16.327	16.326	16.327	16.325	0.002
6 080	16.304	16.304	16.304	16.304	16.303	16.304	0.001
7 600	16.268	16.268	16.268	16.269	16.269	16.270	0.002

采用上面介绍的支持向量机对试验数据进行融合可以进一步改善系统的性能，将训练样本代入(10)式：C_2H_4 和 CH_4 的浓度构成输入向量 x，每组浓度下的输出电压作为对应输出量 y。整理可得式(11)，很明显式(11)是正确的，因而它必存在唯一解。求解(11)式所表示的线性方程所采用的是共轭梯度法，具体的计算步骤可参见文献[2]。根据(10)式确定(10)式和(11)式的参数 γ 和 δ，从而进一步确定(12)式的参数 α_k 和 b。

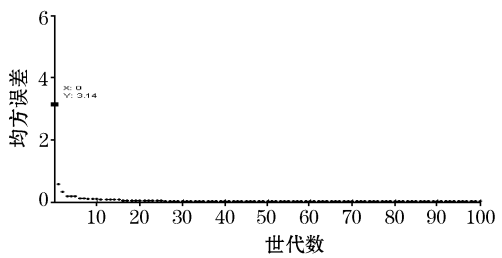

Fig. 1 Error chart of MSE *vs.* generations with $p_c=0.8$, $p_m=0.2$, $N=20$

图 1 是世代数和适应度函数 $f(S)$，即检验的均方根误差的曲线图。在第 20 代即可得到系统的最优参数 $\gamma=18.6$，$\delta=5.10$，此时对应的均方根误差为 0.018.。图 2 即为其所表示的回归曲面。

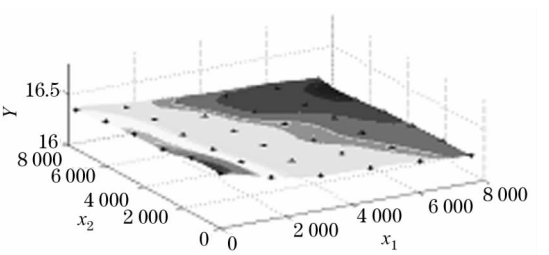

Fig. 2 Distributed diagram of regression support vector machine samples, where asterisk represents sample, $\gamma=18.6$, $\delta=5.10$

将表 2 所表示的检验样本代入(12)式：依然是 C_2H_4 和 CH_4 的浓度构成输入向量 x，可得检验结果如表 3。

融合的效果用交叉灵敏度进行检验，交叉灵敏度用引用误差定义为[12]，融合的程度采用融合前和融合后的比值表征

$$\rho = \left(\frac{|\Delta\alpha_{\max前}|}{Y_{FS}}\right)/\left(\frac{|\Delta\alpha_{\max后}|}{Y_{FS}}\right) = \frac{|\Delta\alpha_{\max前}|}{|\Delta\alpha_{\max后}|} = \frac{0.161}{0.02} = 80.5$$

可见，融合后交叉灵敏度下降了近 81 倍。需要说明的是，所采用的 CH_4 检验样本的最小值 0 $\mu g \cdot mL^{-1}$ 和最大值 7 600 $\mu g \cdot mL^{-1}$ 已经超出了训练样本的范围(50～7 000 $\mu g \cdot mL^{-1}$)，这也从另一方面表明该 SVM 具有较好的泛化能力[13, 14]。

3　结　论

本文将遗传算法应用于支持向量机结构参数的优化，提出一种遗传支持向量机算法，该算法在红外多组分气体定量分析建模交叉敏感抑制实验中，分别能在第 20 世代和第 68 世代就能取得均方根分别为 0.018 和 0.032 的结果。并且可以看出，在遗传算法的激励下，系统的均方根前数代随着参数的进一步优化就开始了成倍的下降，这充分说明遗传算法可以在支持向量机结构参数优化过程中有较高的效率。通过将支持向量机对多组分气体组分浓度的分析，试验结果表明，SVM 可以有效的降低交叉灵敏度近 81 倍，因而在实际应用中，该方法具有一定的理论意义和实践价值。

参　考　文　献

[1] ZHANG Yong-huai, LIN Ji-peng, LIU Jun-hua(张永怀，林继鹏，刘君华). Chinese Journal of Scientific Instrument(仪器仪表学报), 2004, 26(4): 45.

[2] ZHANG Yong-huai, LIU Jun-hua(张永怀，刘君华). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2004, 24(3): 376.

[3] Pontil M, Verri A. IEEE Trans. On Pattern Analysis and Machine Intelligence, 1998, 20: 637.

[4] ZHANG Guang-jun, LÜ Jun-fang, ZHOU Xiu-yin(张广军，吕俊芳，周秀银). Journal of Beijing University of Aeronautics and Astronautics(北京航空航天大学学报), 1996, 22(6): 655.

[5] ZHANG Guang-jun, LÜ Jun-fang, ZHOU Xiu-yin(张广军，吕俊芳，周秀银). Chinese Journal of Scientific Instrument(仪器仪表学报), 1997, 18(2): 134.

[6] Hara H, Kishi N, Iwaoka H. Optical MEMS, 2000 IEEE/LEOS International Conference on, 2000. 139.

[7] Burges C J C. Data Mining Knowl. Discovery, 1998, 22: 1.

[8] Rojo-ALVAREZ J L, de Prado-Cumplido M, Anibal R. IEEE Trans on Signal Processing, 2004, 52: 155.

[9] Childers W, Thompson E L Jr, Harris D B, et al. Atmospheric Environment, 2001, (35): 1923.

[10] LIN Ji-peng, LIU Jun-hua(林继鹏，刘君华). Journal of Jilin University(吉林大学学报), 2005, 23(4): 521.

[11] LIN Ji-peng, LIU Jun-hua(林继鹏，刘君华). Journal of Xi'an Jiaotong Universtity(西安交通大学学报), 2005, 38(6): 787.

[12] LIN Ji-peng, LIU Jun-hua, LING Zhen-bao(林继鹏，刘君华，凌振宝). Journal of Jilin University(吉林大学学报), 2004, 22(5): 453.

[13] Zhang Y H, Zhou J L, Lin J P, et al. Group Technology & Production Modernization, 2003, 20(1): 51.

[14] Amatoa F D, De Rosa M. Optics and Lasers in Engineering, 2002, 37: 533.

Support Vector Machine and Optimized Method for Spectral Analysis

LIN Ji-peng[1, 2], LIU Jun-hua[1]

1. School of Electrical Engineering, Xi'an Jiaotong University, Xi'an　710049, China

2. School of Information Engineering, Chang'an University, Xi'an　710054, China

Abstract　According to support vector machine based on the regularization theory, a small scale machine study theory was proposed to solve the problem of multi-gas analysis, which is mainly restricted by the lack of experimental samples. With its well nonlinear mapping ability, the training error was decided to be zero and global optimal parameters were obtained, hence the cross-sensitivity of spectrum is preferably eliminated. In multi-component gas analysis, the results show that the cross-sensitivity decreased to 1/81. A method based on genetic algorithm and cross-validation was proposed to solve the parameters selection of support vector machine(SVM), which still lacks theory support. The optimal structure parameters were achieved by genetic random search algorithm, the mean square error(MSE)0.018 of the spectrometer was achieved in 20th generation, and MSE decreased by multi-times in the fore generations. This hints that the genetic algorithm SVM is more efficient and has better generalizing ability.

Keywords　Support vector machine; Genetic algorithm; Prediction ability; Structure parameter

(Received Jul. 13, 2005; accepted Oct. 20, 2005)