利用支撑向量机预报大气污染物浓度*

马晓光[1,2] 胡 非[1]

1. 中国科学院大气物理研究所大气边界层物理和大气化学国家重点实验室，北京 100029
2. 中国科学院研究生院，北京 100039

摘要 针对污染物浓度序列具有强非线性和显著长程相关性的特点，结合相空间重构和支撑向量机(SVM)构建了一个多步预报的递归模型. 对理想混沌序列和多种污染物浓度资料的实验结果表明，模型预报的准确率和效率均显著优于人工神经网络. 这种通用建模方法的优势在于对系统非线性机制的反馈十分清晰，充分发挥了SVM适用于小样本问题、映射能力强、全局最优等特点，对非线性时间序列预报适用的其他领域同样具有启发意义.

关键词 大气污染预报 支撑向量机 相空间重构 非线性

在大气污染主要发生的近地层，污染物浓度的时空分布受到气象场、排放源、复杂下垫面、物理-化学-生物耦合过程等多因素的控制，表现出强非线性特征[1~3]. 为减小污染物对生态环境和人体健康的危害，国内外均十分重视预报方法的研究，业务模式逐渐发展成目前的动力结合统计的数值预报系统，这种耦合的思路想发挥动力模式物理概念清晰和统计方法准确率较高的优势，但是却面临着例如精细下垫面参数化，多相多过程耦合，观测资料短缺等不可避免的难题，特别是在刻画复杂非线性系统特征方面还很不理想.

由于污染物浓度序列具有显著的长程相关性[2,3]，因而以人工神经网络(ANN)为主的非线性时间序列分析方法在大气环境领域的研究十分广泛[4~6]. 但是由于算法自身的缺点[7~9](例如难以避免建模对主观技巧的过分依赖，迭代过程易陷入局部最小，学习过拟合等等)，近年来ANN的理论并没有出现实质性突破[8,9]，阻碍了向业务预报的转化.

近5年内，基于统计学习理论的支撑向量机[8](SVM)逐渐成熟，它具有严格的数学基础，成功改进了ANN的以上不足，陆续在模式识别、函数估计和概率密度估计等领域取得应用[8,9]. 已有研究表明[8,10,11]，SVM对小样本条件下的非线性映射具有优势. 本文结合非线性相空间重构和SVM研究预报大气污染物浓度的方法.

1 支撑向量机映射

从信号系统的角度，预测的核心问题是确定输入($\boldsymbol{x}$)和输出($\boldsymbol{y}$)之间的映射关系. 以观测集 $D=\{x_i, y_i\}\in\mathbb{R}^N\times\mathbb{R}$，$i=1, 2, \cdots, N$ 为例，目的是在一族函数$\{f(\boldsymbol{x}, \boldsymbol{w})\}$，$\boldsymbol{w}\in\mathbb{R}^N$中估计一个最优预测函数$f(\boldsymbol{x}, \boldsymbol{w})$，实现整体期望风险 $R(\boldsymbol{w})$(即历史样本的拟合误差与预测误差)最小

$$R(\boldsymbol{w})=\int L(\boldsymbol{y}, f(\boldsymbol{x}, \boldsymbol{w}))\mathrm{d}F(\boldsymbol{x}), \qquad (1)$$

其中 L 是损失函数，表示由 $f(\boldsymbol{x}, \boldsymbol{w})$对 $\boldsymbol{y}$ 预测造成的损失.

在实际应用中，由于联合分布函数 $F(\boldsymbol{x})$未知，而且通常观测数据也极为有限，故只能以估算经验风险 $R_{\mathrm{emp}}(\boldsymbol{w})$(即拟合误差)来代替，即采用所谓经验风险最小化准则(ERM)

2003-07-22 收稿，2003-08-28 收修改稿

* 国家自然科学基金资助项目(批准号：40233030 和 40035010)

E-mail：hot_maxg@hotmail.com

$$R_{\mathrm{emp}}(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} L(\boldsymbol{y}, f(\boldsymbol{x}, \boldsymbol{w})) . \quad (2)$$

以经验风险代替期望风险这种折中的方法是包括ANN在内的传统预测模型所普遍采用的，这正是产生所谓过拟合问题的根源[8,9]. 例如ANN是依据ERM准则调整网络权值，增加网络的复杂性易提高训练样本的拟合精度，但是模型实际预测的误差反而会增大，理论分析[8]指出这是由模型的复杂程度与有限样本的数目之间不相适应造成的.

统计学习理论则针对性地提出结构风险最小化准则[8](SRM)——经验风险和模型复杂度的最小化兼顾，在小样本情况下建立起有效的学习和推广方法. 支撑向量机正是这一思想的实践：它首先利用一个非线性映射 $\boldsymbol{\Phi}:\mathbb{R}^N \to \mathbb{R}^M (M \geqslant N)$，将输入空间映射到高维特征空间，接着在特征空间中拟合 $\{\boldsymbol{\Phi}(x_i), y_i\}$，$i = 1, 2, \cdots, N$，以向量形式表达为

$$f(\boldsymbol{x}) = (\boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{\Phi}(\boldsymbol{x})) + b , \quad (3)$$

$\boldsymbol{w}$，$\boldsymbol{\Phi}(\boldsymbol{x})$为 m 维向量，b 是阈值，$(\cdot)$表示特征空间的内积. 引入正数松弛变量 ξ 和 ξ^*，依据SRM准则，(3)式转化为带约束条件的优化问题

$$\min_{w,b,\xi,\xi^*} R(\boldsymbol{w}, b, \xi, \xi^*) = \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + \gamma\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$

$$\text{s.t.} \begin{cases} y_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Phi}(x_i) - b \leqslant \varepsilon + \xi_i^*, i = 1,2,\cdots,N \\ -y_i + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Phi}(x_i) + b \leqslant \varepsilon + \xi_i, i = 1,2,\cdots,N \end{cases} \quad (4)$$

$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$ 代表模型的复杂程度，γ 用以调整对超出拟合误差 ε 的惩罚程度. 引入Lagrange乘子 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\alpha}^*$ 构造(4)式的对偶形式(参见文献[8])，并解该凸函数的鞍点，可得预测函数

$$f(\boldsymbol{x}) = \sum_{i=1}^{N}(a_i - a_i^*)(\boldsymbol{\Phi}(x_i) \cdot \boldsymbol{\Phi}(\boldsymbol{x})) + b , \quad (5)$$

其中 a_i 不为0时对应的样本被称作支撑向量. 直接确定非线性映射 $\boldsymbol{\Phi}$ 的形式是较困难的，且计算量随特征空间维数增加呈指数倍递增. 根据Hilbert-Schmidt原理[8]，处理高维特征空间的计算问题可以避开求解空间映射 $\boldsymbol{\Phi}$ 的显式形式，即通过引入所谓核函核 $K(x_i, \boldsymbol{x}) = (\boldsymbol{\Phi}(x_i)^{\mathrm{T}} \cdot \boldsymbol{\Phi}(\boldsymbol{x}))$，将变换空间中的内积转化为原空间中某个函数的计算，从而间接求解输入空间向高维特征空间的映射 $\boldsymbol{\Phi}$，即

$$f(\boldsymbol{x}) = \sum_{i=1}^{N}(a_i - a_i^*)K(x_i, \boldsymbol{x}) + b . \quad (6)$$

任意满足泛函Mercer条件的对称函数均可作为核函数，例如多项式函数、径向基函数(RBF)等等. RBF函数因其优秀的局部逼近特性在SVM中应用最为广泛，它利用局部接收域完成函数映射：只有当输入落入输入空间的一个局部区域时，基函数才产生一个重要的非零响应，而其他情况下的响应近似为零.

构造形如(6)式映射函数的学习机器被称作支撑向量机，它将构造输入空间的非线性映射函数转化为构造高维特征空间的线性映射函数，而且通过把原问题转化为对偶问题，使得计算的复杂度不再取决于空间维数，而是取决于样本数，特别是支撑向量的个数，支撑向量机名称的由来正是强调这种在支撑向量上展开的思想.

如图1所示，标准SVM的结构形式上类似于一个神经网络：隐节点是支持向量，网络权重为 α-α^*，输出是隐节点的线性组合. 这些量均可由算法自动产生，无需像ANN构架网络过程中步步经验试算.

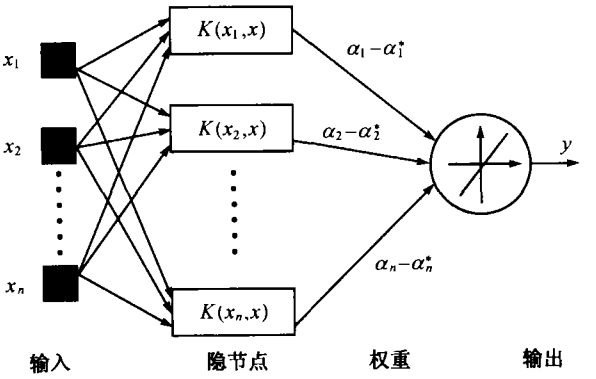

图1 SVM映射函数的结构

传统SVM求解过程是典型的二次规划问题(QP)，这种方法收敛速度慢、存储需求大. 序列最小化算法(SMO)从根本上改进了以上不足，它将QP问题分解为一系列仅包含2个参数控制的子规划问题. 由于每一个子规划单元均可精确给出，故

SMO算法无需额外矩阵存储，也无需数值迭代，可将运算效率提升3～4个数量级．故本文采用基于RBF核函数和改进型SMO算法(详细推导请参见文献[12])的支撑像向量机实现函数映射．

2 数值实验

2.1 观测资料及预处理

污染物浓度序列取自澳门特别行政区化验所和大潭山两个环境监测站：前者位于澳门半岛西北，距海岸较近，属高密度住宅和商业区，测点高度2.5 m;后者位于氹仔岛，属于山顶环境，测点高度100 m．污染物监测项目包括O_3，SO_2，NO_x和PM10，气象资料有温度、相对湿度、风速、风向、云量和露点温度．采样频率1次/h，进行了日平均处理．观测时间自1999年4月1日～2001年10月23日．考虑到日均浓度的幅值已将较多高频细节平滑，所以未进行滤波去噪．只对浓度序列进行了一阶差分处理去除周期和趋势．

2.2 动力空间重构

预报流程如图2所示．首先依据Packard[13]和Takens[14]相空间重构理论，通过自相关函数法确定迟滞时间τ，利用改进G-P算法[15]计算嵌入维数m．对于等间隔时间序列$\{x(t)\}$，$t=1,2,\cdots,N$，重构状态空间中的一点状态向量可以表示为：

$$\boldsymbol{X}(t)=\{x(t),x(t+\tau),\cdots,x(t+(m-1)\tau)\}.$$

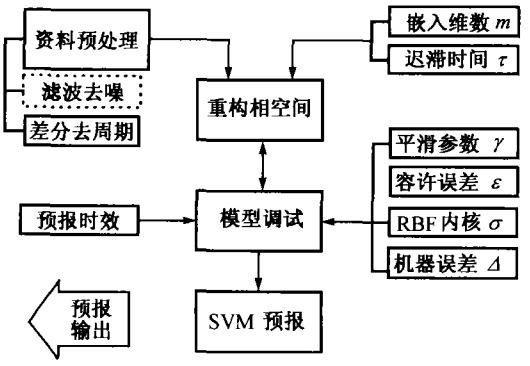

图2 非线性预报模型流程图

由SVM确定最优映射，则一步非线性时间序列预报模型可以表示为$x(t+1)=f(\boldsymbol{X}(t))$．

2.3 模型输入与输出

在模型的训练阶段，将t时刻状态空间$\boldsymbol{X}(t)$和对应的气象因子组成数据窗口作为输入项，将$t+1$时刻的浓度值作为输出项，通过滑动数据窗口完成预测函数的训练．

在预报阶段，直接把向前一步预报得到的浓度值作为已知量输入模型，采用递归的方式实现连续多步预报．

2.4 预报时效

最大Lyapunov指数能够度量系统状态的可预报性[16]．以Wolf[17]改进算法对污染物浓度序列进行计算，可预报时间尺度约为3～5 d．本文采用连续3日预报．

整个算法在ANSI C++(编译器GCC 3.2)环境下实现．

3 结果与讨论

首先利用一个理想混沌系统检验SVM模型的预报效果，时间序列由四阶Runge-Kutta方法积分Mackey-Glass[18](MG)高维混沌运动方程产生，嵌入维数为6，迟滞时间为6．MG微分方程最初用来描述血细胞的调节机制，因为表现出高度的混沌特征，常被用于检验预报模型的效果．

图3是对MG系统高度混沌的运动状态做14步预报的结果(训练集500点)．从图中可以看出，预报值与实际值的趋势和幅值大小几乎重合．计算得到两者的相关系数为0.9971，平均相对误差为1.887%，对理想混沌系统预报是非常成功的．

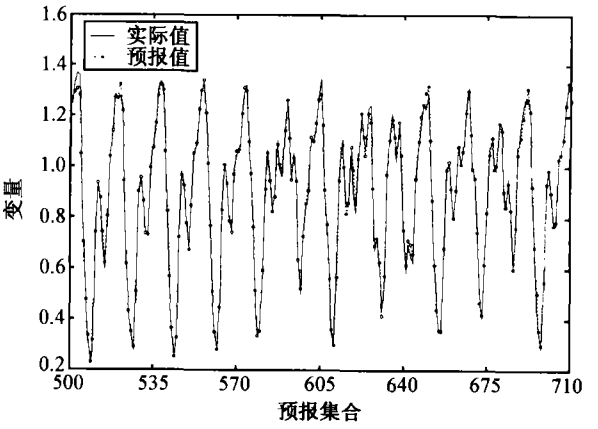

图3 SVM对MG混沌系统14步连续预报

图4是将SVM模型应用于大潭山监测站PM10浓度预报的实例. 重构嵌入维数为8，迟滞时间为2. 训练集合为1999年4月1日～2001年6月25日历史数据，以3步递归方式预报后120 d情况. 从图中可以看到，预报值与真实值的趋势吻合十分理想，能清晰地刻画出细节部分的极大值与极小值涨落. 计算得到的相关系数为0.9126，平均相对误差为18.21%，预报效果是理想的.

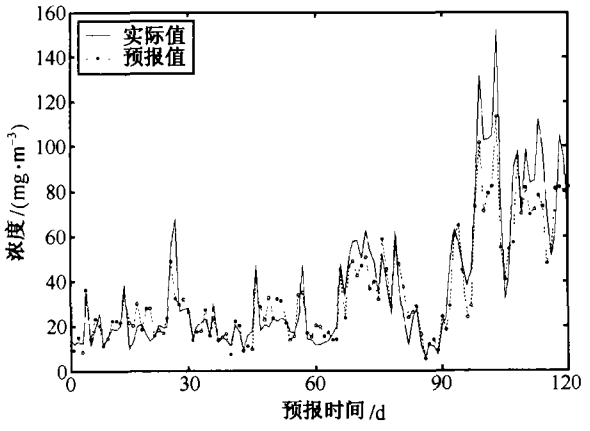

图4 SVM对PM10污染物浓度3日预报

为验证预报的稳定性，分别对8组浓度序列进行了预报，结果见表1.8组实验结果的相对误差在18.21%～23.20%之间，平均值为20.16%，可见该模型的预报能力是稳定的.

表1还给出了相同条件下传统ANN模型预报精度的结果，可以看出SVM模型的平均相对误差优于神经网络8%～9%. 实际运算过程中，SVM平均效率要快出1～2个数量级(表略). 可见SVM模型的预报性能具有明显的优势.

表1 SVM与ANN预报精度对比[a](单位:%)

MAPE		SVM	BP[b]	RBF[c]
大潭山检测站	PM10	18.21	27.55	27.67
	O_3	18.86	20.97	24.12
	SO_2	20.02	28.25	29.46
	NO_x	20.37	32.03	32.72
化验所检测站	PM10	19.71	32.66	29.90
	O_3	20.95	28.45	27.49
	SO_2	23.20	28.67	29.27
	NO_x	19.99	30.90	31.12
平均		20.16	28.67	28.97

a) 预报时间为2001年6月26日～10月23日，共120 d; b) BP网络模型结构参考文献[4]; c) RBF网络参考文献[3]

非线性时间序列预报是一种典型的黑箱方法，适于解决分布概率未知、物理机制不甚明了且观测有限的复杂系统问题. 预报效果取决于模型能否反映出待求问题的物理机制. 与ANN相比较，本文构建的SVM混沌模型的预报效果更优，原因可以解释为：(1) 混沌模型以相空间重构为基础保留了动力系统本身的非线性特征，物理意义十分清晰，能够在一定程度上反映系统非线性反馈的机制[16]；性质上仍然可看作是“动力”的，较好地减少了主观性；(2) 模型通过结构风险最小准则实现了整体期望误差的控制，这等同于最小化广义误差的上边界[8]，而不是最小化训练误差；(3) 函数解通过凸函数的鞍点获得，从根本上避免了局部最小[12]；(4) 对模型复杂度的有效控制，增强了推广的稳定性.

但是预报模型也存在需进一步研究的问题. 分析误差的主要来源发现，模型对突变点的预报效果整体上不够理想. 一个主要原因是建模受到资料限制，在输入项的诸多因子中仍然缺少对当前态物理场大尺度动力学相似过程历史估计的成分. 因此在实际业务建模时，引入其他成熟预报产品以扩充输入项是十分必要的，这样可能会更好地适应系统突变等情况.

4 结论

(1) 结合相空间重构与支撑向量机映射的时间序列预报模型，物理意义清晰，预报的准确率和运行效率均显著优于人工神经网.

(2) 研制大气污染物浓度预报系统时，对于非线性时间序列预报模块，可以优先考虑SVM方法，还可以耦合传统模式组成综合预报系统.

(3) 这种通用建模方法，适合对含噪声、强非线性复杂系统进行预报，特别适合小样本学习的情况，不仅给大气污染物浓度预报提供了新的思路，也对非线性时间序列预报适用的其他领域有借鉴意义.

致谢 衷心感谢赵松年研究员和曾庆存院士的鼓励.

参 考 文 献

1 Raga G B, et al. On the nature of air pollution dynamics in Mexico

City I. Nonlinear analysis. Atmos Environ, 1996, 30: 3987
2 Lee C K, et al. Fractal analysis of temporal variation of air pollutant concertration by box counting. Environmental Modelling & Software, 2003, 18: 243
3 Anh V V, et al. Multifractal analysis of Hong Kong air quality data. Environmetrics, 2000, 11: 139
4 Gardner M W, et al. Artificial neural networks-A review of applications in atmospheric sciences. Atmos Environ, 1998, 32, 2627
5 刘 罡, 等. 大气污染物浓度的神经网络预报. 中国环境科学, 2000, 20(5): 429
6 万显烈, 等. 利用人工神经网络对大气中 O_3 浓度进行预测. 中国环境科学, 2003, 23(1): 110
7 Yao X. Evolving artificial neural netwoks. Proceedings of the IEEE, 1999, 87(9): 1423
8 Vapnik V N 著. 张学工译. 统计学习理论的本质. 北京: 清华大学出版社, 2000. 5～155
9 张学工. 关于统计学习理论与支撑向量机. 自动化学报, 2000, 26(1): 32
10 Mukherjee S, et al. Nonlinear prediction of chaotic time series using a support vector machines. In: Principe J, et al. eds. IEEE Workshop on Neural Networks for Signal Processing VII. IEEE Press, 1997. 511
11 Müler K R, et al. Using support vector machines for time series prediction. In: Scholkopf C, et al. eds. Advances in Kernel Methods. MIT Press, 1999. 242
12 Shevade S K, et al. Improvements to SMO algorithm for SVM regression. IEEE Trans on Neural Networks, 2000, 11(5): 1188
13 Packard N H, et al. Geometry from a time series. Phys Rev Lett, 1980, 45: 712
14 Takens F. Detecting strange attractors in turbulence. Phys Rev Lett, 1981, 79(8): 1475
15 Theiler J. Efficient algorithm for estimating the correlation demension from a set of discrete points. Phys Rev A, 1987, 36: 4456
16 Kantz H, et al. Nonlinear Time Series Analysis. Cambridge: Cambridge University Press, 1997. 29～108
17 Wolf A. Determing Lyapunov exponents from a time series. Physica D, 1985, 16: 285
18 Mackey M C, et al. Oscillation and chaos in physiological control systems. Science, 1977, 2: 287