第32卷，第12期 气 象 Vol. 32 No. 12
2006年12月 METEOROLOGICAL MONTHLY December, 2006

支持向量机在大气污染预报中的应用研究

常 涛

（新疆气候中心，乌鲁木齐 830002）

提 要：支持向量机是基于统计学习理论的新一代机器学习技术，其非线性回归预测性能优越于传统统计方法。利用前一天该污染物的日均浓度、前一天地面平均风速等7个预报因子建立了基于RBF核函数支持向量回归法的大气污染预报模型，并利用十重交叉验证和网格搜索法寻找模型最优参数。乌鲁木齐大气预报实例表明：支持向量机显示出小样本时预报精度较高和训练速度快的独特优势，为空气质量预报提供一种全新的模式。

关键词：大气污染预报 支持向量机（SVM） 交叉验证 网格搜索

Application of Support Vector Machine to Atmospheric Pollution Prediction

Chang Tao

(Xinjiang Meteorological Bureau, Urumqi 830002)

Abstract: The support vector machine (SVM), a new generation machinery learning technology based on statistical theory, has been reported to have better prediction performance of non-liner regression than traditional statistical methods. An SVM regression (SVMR) model for atmospheric pollution prediction is developed according to seven forecast factors, including the daily average pollutant concentration of previous day, daily average wind speed of previous day, etc. Meanwhile, 10-fold cross-validation and grid-search methods are applied to find the best parameters of SVMR. The experimental results of Urumqi data show that SVM has the unique advantage of high prediction accuracy and training rate on small-size data sets. It suggests a new model for prediction of atmospheric pollution.

收稿日期：2006年8月8日； 修定稿日期：2006年10月15日

Key Words: atmospheric pollution prediction support vector machine (SVM) cross-validation grid-search

引 言

近20年来，大气污染预报模式的研究得到了很大的发展，从过去的统计预报模式，已发展到今天的中尺度气象预报模式、大气污染扩散模式和光化学模式相结合的空气污染预报模式和非静稳多箱格大气污染浓度预报和潜势预报系统CAPPS模式。大气预报模式主要可以归为潜势预报、统计预报及数值模式预报三类。统计预测方法多是线形模型，难以模拟复杂多变的大气污染变化。神经网络较统计方法能更好地模拟大气污染因素的非线形关系，在大气污染预报应用中取得较好结果[1]。然而，神经网络具有推广能力差、过拟合、易于陷入局部最优、寻找结构参数复杂等缺点。支持向量机(SVM)，是Vapnik开发的基于统计学习理论的新一代机器学习技术[2]，在解决小样本、非线性问题中表现出独特优势。其遵循结构风险最小化原则，预测性能和推广能力优于神经网络，因而成为应用领域研究的热点。陈永义和冯汉中[3]率先将SVM引入了气象领域。目前，SVM在气象上的应用主要是短期预报、实时短期预报业务[4-6]等方面。本文通过实例论证，探讨支持向量回归方法应用于环境空气质量预报的可行性，并利用交叉验证和网格搜索的方法确定支持向量机的超参数，从而确保模型的预测精度。

1 支持向量机回归方法的基本原理

给定 s 组样本数据 $\{x_k, y_k\}, k = 1,2,\cdots,s$，其中 $x_k \in R_m, y_k \in R$，利用一个非线性映射 Φ，将数据 x 映射到高维特征空间 G，在这个空间进行线性逼近。由统计学习理论可知，该函数具有以下形式：

$$f(x) = (\boldsymbol{\omega} \cdot \boldsymbol{\Phi}(x)) + b, \Phi: Rm \to F, \omega \in G \quad (1)$$

式中：(·) 为内积运算；b：偏置项。$\boldsymbol{\omega}$ 和 b 通过最小化下列泛函进行估计。

$$\mathrm{Rreg}[f] = \mathrm{Remp}[f] + \lambda \| \boldsymbol{\omega} \|^2 = \sum_{i=1}^{s} C(e_i) + \lambda \| \boldsymbol{\omega} \|^2 \quad (2)$$

式中：$e_i = f(x_i) - y_i$，s：样本容量，$C(e_i)$：损失函数，λ：规则化常数。$\| \boldsymbol{\omega} \|^2$ 反映函数 f 在高维空间平坦的复杂性。常选取线性 ε 不敏感损失函数。取经验风险为：

$$\mathrm{Remp}[f] = \frac{1}{s} \sum_{i=1}^{s} | y_i - f(x_i) |_\varepsilon \quad (3)$$

式(2)等价于求解如下的优化问题。

$$\min J = \frac{1}{2} \| \boldsymbol{\omega} \|^2 + C \sum_{i=1}^{s} (\xi_i^* + \xi_i) \quad (4)$$

$$\text{s. t.} \quad \begin{cases} y_i - (\boldsymbol{\omega} \cdot \boldsymbol{\Phi}(x)) - b \leqslant \varepsilon + \xi_i^* \\ (\boldsymbol{\omega} \cdot \boldsymbol{\Phi}(x)) + b - y_i \leqslant \varepsilon + \xi_i \\ \xi_i, \xi^* \geqslant 0 \end{cases}$$

式中：ξ_i^*，ξ_i：松弛变量。C：正规化常数，控制模型复杂度和逼近误差的折中，C 越大数据拟合度越高。ε：控制回归逼近误差管道的大小，决定对训练样本的拟合精度，值越大则支持向量越少，但精度不高。引入核函数方法将式(5)转化为：

$$\max J = -\frac{1}{2} \sum_{i,j=1}^{s} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) k(\boldsymbol{x}_i \cdot \boldsymbol{x}_j) + \sum_{i=1}^{l} \alpha_i^* (y_i - \varepsilon) + \sum_{i=1}^{s} \alpha_i (y_i + \varepsilon)$$

$$\text{s. t} \quad \begin{cases} \sum_{i=1}^{s} \alpha_i^* = \sum_{i=1}^{s} \alpha_i \\ 0 \leqslant \alpha_i \leqslant C \\ 0 \leqslant \alpha_i^* \leqslant C \end{cases} \quad (5)$$

α_i^* 和 α_i 为两组 Lagrange 乘子，即最小化 Rreg 的解。求解上述凸二次规划得到的非线性映射可表示为：

$$f(x)=\sum_{i=1}^{s}(\alpha_i-\alpha_i^*)(\boldsymbol{\phi}(x_i)\cdot\boldsymbol{\phi}(x))+b$$
$$=\sum_{i=1}^{s}(\alpha_i-\alpha_i^*)k(x_i,x_k)+b \quad (6)$$

其中：$k(x_i,x_k)=\boldsymbol{\phi}(x_i)\cdot\boldsymbol{\phi}(x)$ 是满足 Mercer 条件的核函数，对应于特征空间的点积。SVM 在计算 $f(x)$ 时，无需计算 ω 和 $\boldsymbol{\phi}(x)$ 的数值，只需计算 Lagrange 乘子以及核函 $k(\boldsymbol{x}_i\cdot\boldsymbol{x}_k)$，从而巧妙地解决了维数灾难问题，使算法的复杂度与样本维数无关。常用核函数有线性函数、多项式函数、RBF 函数、Sigmoid 函数等。

2　大气污染预报模型的建立

建立基于支持向量机的大气污染物浓度变化的预报模型，关键问题是输入模式的确定、训练样本的选取以及模型结构参数的选取。本文拟建立 PM_{10}，NO_2，SO_2 日均浓度值的预报模型。

2.1　输入因子的选取

大气污染物浓度变化主要影响因素是污染源和污染源排放的污染气象条件等。根据资料及历史经验[7,8]，确定当日污染物浓度（PM_{10}，NO_2，SO_2）预报模型的输入向量为前一天该污染物的日均浓度、前一天地面平均风速、前一天最低温度梯度、前一天平均温度梯度、前一天平均湿度、前一天平均总云量、前一天污染源的源强 7 个因子。

2.2　基于 SVM 回归的大气污染预报模型

（1）确定支持向量机的核函数类型

选择合适的核函数，可提高预测精度，降低噪声的影响。通常认为 RBF 核函数优于其他核函数，具有性能好且稳定和调节参数较少等优点[9]。因此，本文使用 RBF 核函数的支持向量回归模型。σ 对回归超平面的形成有直接影响，目前没有统一方法来确定 σ 大小。

$$k(\boldsymbol{x}_i\cdot\boldsymbol{x})=\exp\left(\frac{-\|x_i-x\|^2}{2\sigma^2}\right) \quad (7)$$

（2）支持向量机预测模型的参数寻优

模型中 C、σ、ε 参数的选取，直接影响模型的预测性能和推广能力。目前尚没有通用的支持向量机参数选择模式，只能凭借经验和试验对比。多数文献随机选取，影响了模型的精度。本文利用多重交叉验证（k-fold cross validation）的方法和网格法（grid-search）[9]寻找 C 和 σ。其原理是：将训练集分成 k 个子集（样本数量大致均匀），每个子集分别作为测试集，其余子集样本作为训练集，即建模 k 次，用 k 次的平均绝对误差评估模型性能，进而确定模型的最优参数对（C，σ）。网格法是对网格上的（C，σ）点穷举搜索，C 和 σ 的步长呈指数级增长（例：$C=2^{-10}$，2^{-9}，…，2^{10}；$\sigma=2^{-10}$，2^{-9}，…，2^{10}）。不像其他启发式方法，网格搜索计算可并行进行，因而是一种较为实用有效的方法。

（3）用训练样本训练具有优化参数的支持向量机预测器，获得支持向量，确定支持向量机的结构。

（4）用训练过的支持向量预测器对测试样本预测。

3　预报试验分析

3.1　试验软件

LIBSVM 是台湾大学林智仁教授编写的软件，功能较全，提供源码，方便改进，提供 SVM 默认参数，国内外应用效果较

好[9]。

3.2 预报实验

用乌鲁木齐气象资料和同期大气环境检测资料实验。以 PM_{10} 预测为例：把 2003 年的 4、5、6 月每日共 91 组数据作为训练样本，每组数据包含 7 个输入因子和 1 个 PM_{10} 实际值。把 2004 年的 4 月每日共 30 组数据作为测试样本，每组数据包含 7 个输入因子，对每日的 PM_{10} 进行预测。另两项污染物 NO_2，SO_2 的预测方法相同。

采用 10 重交叉验证和粗细网格法寻参。先在（$C=2^{-10}$，2^{-9}，…，2^{10}；$\sigma=2^{-10}$，2^{-9}，…，2^{10}）网格内大步长寻找较优参数，得性能较优点（$C=2^1$，$\sigma=2^{-1}$），然后在此点附近网格内（$C=2^{-1}$，$2^{-0.75}$，…，2^3；$\sigma=2^{-3}$，$2^{-2.75}$，…，2^1）小步长搜索，得到最优点（$C=2^{1.75}$，$\sigma=2^{-0.75}$）。依据经验确定训练误差 $e=0.001$，并使用最优的 C，σ 值和训练样本建立预报模型，对测试样本的污染物浓度预测。最后即得到日污染物平均浓度的时间序列数据，如图 1～3 所示。

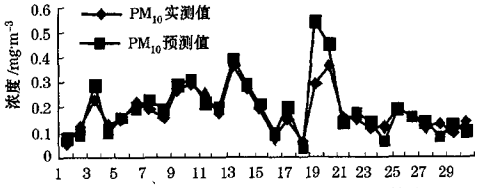

图 1 2004 年 4 月逐日 PM_{10} 实测值与预测值浓度对比

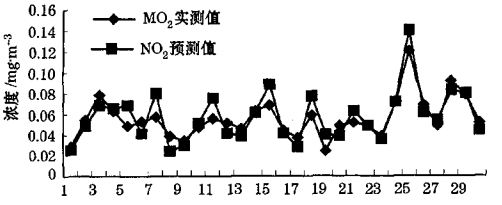

图 2 2004 年 4 月逐日 NO_2 实测值与预测值浓度对比

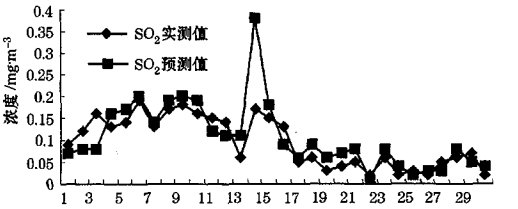

图 3 2004 年 4 月逐日 SO_2 实测值与预测值浓度对比

3.3 实验结果分析

（1）图 1～3 分别是 PM_{10}、NO_2、SO_2 的实测值和预测值的对比。由图可以看出，各污染物的实测值和预测值符合得较好，各污染物绝对预报误差较小。模型对各污染物的浓度总体变化趋势反映较敏感。

分析数据后可知，PM_{10} 误差的来源主要受特殊天气的影响，例如沙尘暴及扬尘天气，气象预报的偏差也在不同程度上造成误差。图 1 看出 4 月 19—20 日的扬尘天气对 PM_{10} 浓度预测造成较大影响。4 月中旬以来新疆北部地区气温持续上升，近半个月没有降水，地表疏松干燥，这也是扬尘、扬沙及沙尘暴天气产生的主要原因。

SO_2 的误差较突出。乌鲁木齐市的 SO_2 污染主要为燃煤型，随着冬季采暖期的结束和污染排放源的减少而有很大程度削减。又源于其自身水溶性较 NO_2 强的特点，因此受环境空气状况影响较大，一旦有风雨天气，其衰减辐度很大，引起相应的误差。

乌鲁木齐市的 NO_2 污染水准较之 SO_2 并不低，但是由于冬季被 SO_2 的污染所掩盖。夏季来临，SO_2 的污染水平有所降低，加之气温和日照等气象条件较冬季大为改善，对于二次污染物 NO_2 的生成有利，因此其污染程度也就体现无遗。从 NO_2 浓度误差统计中得知：NO_2 的误差主要受特殊天气（风、雨等）的影响。

（2）PM_{10}、NO_2、SO_2 的实测值和预

测值的线性相关系数为 0.795、0.778、0.702（图略）。这表明 SVMR 模型处理大气污染物的非线性问题具有优势。由于本文对不同污染物均选取相同的气象因子作为训练样本，因此在对不同污染物的预测中，预测值与实测值间的相关性存在一定差异。

4 结　论

（1）RBF 核函数的支持向量回归模型能很好捕捉大气污染物浓度与其影响因子的非线性关系，具有预报精度较高和训练速度快的优点。但是大气污染物浓度预报的准确率受到预报模式本身、气象预报和环境预报准确率的影响，对重大天气变化的预报尚存一定局限。

（2）多重交叉验证法和网格搜索法是寻找支持向量机模型参数的有效方法，也是确保模型预测精度的关键。

（3）由于资料所限，对不同污染物均选取相同的气象因子，具有一定的局限性。今后改进的方向是采取定性和定量的方法筛选大气污染物浓度的影响因子。

（4）支持向量机运用于大气污染预报的研究尚处于试验探索阶段，本文仅作了粗浅探讨。模型的推广还必须考虑样本容量、气候环境诸多因素进一步改进。

参考文献

[1] 金龙. 人工神经网络技术发展及在大气科学领域的应用 [J]. 气象科技，2004，32 (6)：12-13.

[2] Vapnik V. N.，张学工译. 统计学习理论的本质 [M]. 北京：清华大学出版社，2000.

[3] 冯汉中，陈永义. 处理非线性分类和回归问题的一种新方法）(2) ——支持向量机方法在天气预报中的应用 [J]. 应用气象学报，2004，15 (3)：356-365.

[4] 冯汉中，陈永义. 支持向量机回归方法在实时业务预报中的应用 [J]. 气象，2005，31 (2)：41-44.

[5] 车怀敏. 用支持向量机方法作德阳降水预报 [J]. 四川气象，2005 (2)：13-15.

[6] 冯汉中，杨淑群，刘波. 支持向量机 (SVM) 方法在气象预报中的个例试验 [J]. 四川气象，2005，(2)：9-12.

[7] 王俭，胡筱敏，郑龙熙. 基于模型的大气污染预报方法的研究 [J]. 环境科学研究，2002，15 (5)：62-65.

[8] 熊忠华，陈琦，郑秀梅. 基于遗传算法的人工神经网络大气环境评价 [J]. 环境科学与技术，2005，28 (4)：82-84.

[9] LIBSVM：a Library for support Vector Machines [OL]. Chih _ Chung，http：//www. csie. ntu. edu. tw.