文章编号：1006-4354 (2007) 05-0006-04

基于 LS-SVM 的新疆沙尘暴预测研究

常 涛[1]，傅玮东[1]，秦 榕[2]

(新疆气候中心，乌鲁木齐 830002，新疆气象信息中心，乌鲁木齐 830002)

摘 要：首先对新疆的沙尘暴日数和前期气候要素统计分析，然后通过相关分析提取南疆和北疆春季沙尘暴影响因子，根据沙尘暴天气的特点和支持向量机方法在解决小样本学习问题中的优势，提出并实现基于最小二乘支持向量机 (LS-SVM) 的沙尘暴预测模型。实验结果表明：所提出的沙尘暴影响因子和支持向量回归模型处理沙尘暴预测问题有一定的应用价值。

关键词：新疆；沙尘暴预测；最小二乘支持向量机；影响因子

中图分类号：P456.9 文献标识码：A

沙尘天气是一种严重的气象灾害。近年北方沙尘暴、扬沙和浮尘天气频繁发生给生态环境和人类社会造成严重损失。新疆是中国沙尘暴的多发区之一，又地处我国天气系统的上游，特别是南疆盆地的塔克拉玛干沙漠，为沙尘暴的产生提供了物质条件，是世界四大沙源之一[1]。钱正安等研究了近 50 a 来中国沙尘暴的分布及变化趋势，认为我国北方沙尘暴主要分布在河西走廊和阿拉善高原、南疆盆地南缘及内蒙古中部等三地区[2]。因此，新疆沙尘暴预报的研究及影响评估，在中国沙尘暴预报研究中占有重要位置。由于沙尘暴预报的复杂性，预报结果一直不理想。目前对沙尘暴的统计预报，主要方法有 K-最近邻法、人工神经网络法等方法[3-4]，但是 BP 神经网络稳定性差，容易产生过学习现象，需要大样本才能保证预测精度。20 世纪 90 年代中期发展的支持向量机 (SVM)，能有效地解决小样本、非线性、高维数以及局部最小等问题。它根据有限的样本信息在模型复杂性和学习能力之间寻求最佳折中，以期获得最好的泛化能力。在回归算法中表现出极好的性能，被认为是神经网络的替代方法。SVM 无论在理论上，还是在实践中，在非线性时间序列预测领域都具有优秀的表现和应用前景。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[5]等方面。本文利用新疆 24 个站点沙尘暴天气历年资料，研究南北疆沙尘暴天气特点，分析选定新疆沙尘暴春季气候影响因子，尝试使用支持向量机回归模型对新疆沙尘暴预报预警，探讨了该模型的可行性。

1 资料分析

1.1 资料选取与处理

选取 1961—2005 年北疆沙尘暴年平均日数 ＞4 d 的 9 个气象站，南疆年平均沙尘暴日数＞10 d 的 15 个气象站，利用逐日沙尘暴日数和逐日 7 种地面气象观测要素值 (表 1)，计算 1963—2005 年各站春季 (3—5 月) 沙尘暴发生日数，各站 7 种气象要素 1962—2004 年冬季 (冬季为当年 12 月—次年 2 月)、春季 (3—5 月)、夏季 (6—8 月) 平均，得到 43 a 各站春季沙尘暴日数，前期冬季、春季、夏季 7 种气象要素的平均值。北疆 9 站、南疆 15 站平均为地域平均值。

7 个气象要素分别是：气温日较差 (最高气温－最低气温) t_R，日平均风速 f_P，日平均气温 t_P，20—20 时降水量 R_{20}，日平均地面温度 t_D，地气温差 $t_Q=t_D-t_P$，日平均相对湿度 x_P。

1.2 春季沙尘暴长期变化趋势分析

1963—2005 年北疆春季沙尘暴平均日数 2.3 d (图 1)。1963—1974 年平均为 2.2 d；1975 年最多

收稿日期：2007-05-30

作者简介：常 涛 (1966-)，男，江苏丹阳市人，汉，工程师，本科，从事气候预测与分析工作。

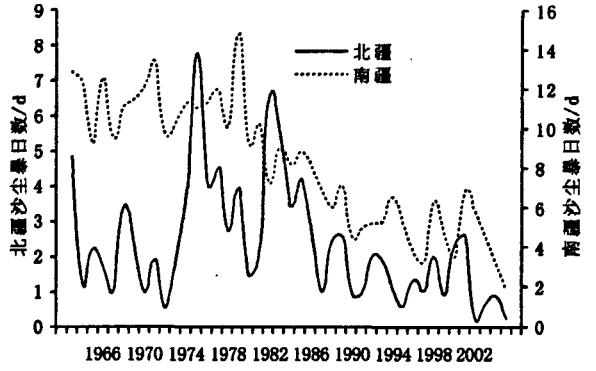

图 1　北疆和南疆春季沙尘暴日数年际变化

为 7.8 d，1982 年次多为 6.6 d；1975—1987 年具有准 6 a 的振荡周期；1988—2005 年围绕 1.4 d 上下波动，变化范围 0.2～2.6 d。南疆春季沙尘暴年平均日数 8.2 d，1963—2005 年，南疆春季沙尘明暴经历了 3 个变化阶段：1963—1980 年为明显偏多期，平均日数 11.4 d；1981—1990 年呈线性递减过程；1991—2005 年围绕 4.8 d 波动，变化范围 1.8～7.0 d，其中存在准 3～5 a 的周期。

1963—2005 年北疆春季沙尘暴日数(S)标准化数值的 5 a 滑动平均曲线表明，北疆春季沙尘暴日数的变化经历了偏少(1965—1972 年)、偏多(1973—1987 年)和偏少(1988—2003 年)3 个阶段的变化(图 2 a)。前期气候要素中，只有夏季相对湿度的变化类似于春季沙尘暴的变化，经历了干湿 3 个阶段的变化，其位向与北疆春季沙尘暴的位向相反。南疆春季沙尘暴日数标准化数值的 5 a 滑动平均曲线表明，南疆春季沙尘暴经历了偏多(1965—1985 年)和偏少(1986—2003 年)2 个阶段的变化(图 2 b)。前期气候要素中，平均风速和相对湿度的变化经历了类似的 2 个阶段的变化。南疆冬季和春季平均风速，1965—1984 年偏大，1985—2003 年偏小；夏季平均风速 1965—1985 年偏大，1987—2003 年偏小；夏季相对湿度 1965—1988 年偏干，1989—2003 年偏湿。因而前期气候要素中夏季相对湿度对来年北疆和南疆春季沙尘暴有指示意义，南疆冬季、春季和夏季平均风速对来年南疆春季沙尘暴有指示意义。

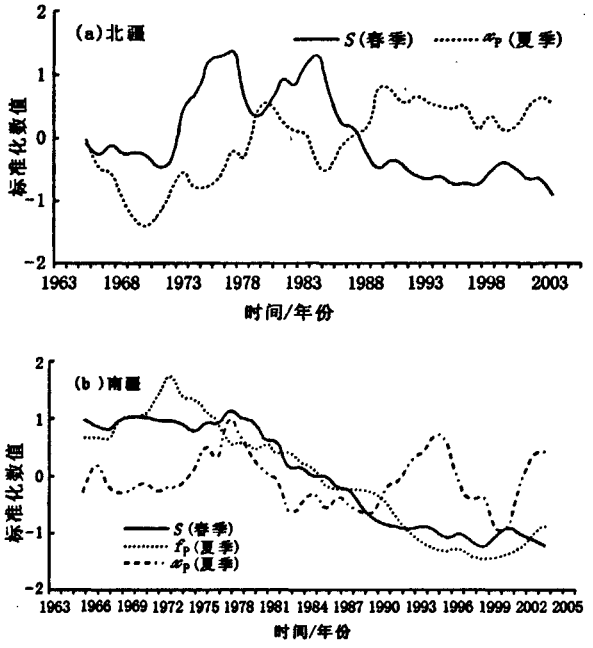

图 2　春季沙尘暴日数和气候要素标准化数值 5 a 滑动平均的年际变化

2　最小二乘支持向量机回归技术简介

给定 n 维样本向量集：(x_1, y_1)，(x_2, y_2)，…，(x_l, y_l)，$x_i \in R_n$，$y_i \in R$ 为对应的目标值，l 为样本数量。用一非线性映射 $\Phi(\cdot)$ 将数据 $\{x_i\}$ 从原 R_n 空间映射到高维特征空间进行线性回归，即 $y(x) = \omega^T \Phi(x) + b$。利用结构风险最小化原则寻找 ω，b，使 $R = \frac{1}{2} \| \omega \|^2 + c \cdot R_{\text{emp}}$ 最小化。其中 $\| \omega \|^2$ 为控制模型复杂度参数，c 为正规化参数，R_{emp} 为误差控制函数，一般用不敏感损失函数 ε。在最小二乘支持向量机中，用损失函数为误差的二次项。故优化问题为：

$$\begin{cases} \min J = \frac{1}{2}\omega^T\omega + \frac{1}{2}c\sum_{i=1}^{l}\varepsilon_i^2 \quad i = 1, 2, \cdots, l \\ s.t.: \ y_i - \omega^T\Phi(x_i) - b = \varepsilon_i \end{cases} \tag{1}$$

引入拉格朗日乘子 α_i，函数变为：

$$L(\omega, b, \varepsilon, \alpha) = \frac{1}{2}\omega^T\omega + c\sum_{i=1}^{l}\varepsilon_i^2 - \sum_{i=1}^{l}\alpha_i [\omega^T\Phi(x_i) + b + \varepsilon_i - y_i] \tag{2}$$

其中 $\alpha = \{\alpha_i | i = 1, 2, \cdots, l\}$。求解上述优化问题时，LS-SVM 的优化问题转化为求解线性方程(3)。

$$\begin{bmatrix} 0 & y_1 & \cdots & y_l \\ y_1 & y_1y_1K\ (x_1,\ x_1)\ +2/r & \cdots & y_1y_lK\ (x_1,\ x_l) \\ \vdots & \vdots & \cdots & \vdots \\ y_l & y_1y_1K\ (x_1,\ x_1) & \cdots & y_1y_lK\ (x_1,\ x_l)\ +2/r \end{bmatrix} \times \begin{bmatrix} b \\ a_i \\ \cdots \\ a_l \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ \cdots \\ 1 \end{bmatrix} \quad (3)$$

其中，$K(x,x_i)=\Phi(x_i)^T\Phi(x)$是满足 Mercer 条件的核函数。LS-SVM 具有很好的泛化能力，且训练速度要比标准的 SVM 更快。因此适用于小样本的的沙尘暴预测。

3 基于支持向量机的沙尘暴预测模型

3.1 沙尘暴预测模型影响因子分析

根据经验和第 2 节分析，气温日较差（t_R），日平均风速（f_P），日平均气温（t_P），20—20 时降水量（R_{20}），日平均地面温度（t_D），地气温差（t_Q），日平均相对湿度（x_P）是对春季沙尘暴气候影响的主要要素。对南北疆春季沙尘暴与前一年冬季、春季、夏季气候要素之间进行相关性分析（表 1）。在信度 $\alpha=0.90$ 下，相关系数的临界值 $R_0=0.254\ 3$。选取 $R>R_0$ 的气候要素作为气候影响因子。因此，北疆当年的春季沙尘暴气象影响因子 11 个：前冬的 t_R、f_P、R_{20}，前春的 t_R、f_P、x_P，前夏的 t_R、f_P、R_{20}、x_P。南疆当年的春季沙尘暴气象影响因子 16 个：前冬的 t_P、t_R、t_D、f_P，前春的 t_R、t_D、t_Q、f_P、R_{20}、x_P，前夏的 t_R、t_D、t_Q、f_P、R_{20}、x_P。首先建立基于相关系数选择因子的预测模型，并将其与基于全部因子（21 个）的预测模型的结果比较。

表 1 春季沙尘暴与前期气候要素之间的相关系数

要 素		t_P	t_R	t_D	t_Q	f_P	R_{20}	x_P
北 疆	冬季	−0.177 1	0.326 3	−0.112 5	0.481 5	0.303 2	−0.363 6	−0.134 4
	春季	0.223 3	0.457 0	0.230 9	0.173 5	0.457 7	−0.215 8	−0.318 5
	夏季	0.221 0	0.362 3	0.127 7	−0.046 8	0.561 5	−0.360 2	−0.557 2
南 疆	冬季	−0.574 7	0.274 0	−0.622 7	0.142 7	0.731 2	0.035 6	0.015 9
	春季	−0.016 9	0.431 1	−0.317 0	−0.670 1	0.845 8	−0.263 3	−0.281 3
	夏季	−0.155 6	0.402 7	−0.497 6	−0.717 3	0.850 1	−0.325 3	−0.561 9

3.2 基于支持向量机的沙尘暴预测建模

使用 matlab 的 LS-SVM 工具箱函数实现模型[6]。

3.2.1 数据准备 以 1963—1984 年的数据为训练样本，1985—2006 年的数据为测试样本，南疆和北疆各有 22 个训练样本和 21 个测试样本，分别建立南疆和北疆的春季沙尘暴预测模型。基于相关系数选择因子的预测模型中，南疆训练样本数据每行包含了上年 16 个气象因子和当年春季沙尘暴实际发生日数，北疆训练样本数据每行包含了上年 11 个因子和当年春季沙尘暴实际发生日数。基于全部因子的预测模型中，南疆与北疆训练样本数据中每行包含了上年 21 个因子和当年春季沙尘暴实际发生日数。对选定的前期气候影响因子和沙尘暴日数进行标准化处理。

3.2.2 核函数确定 通常认为 $K(x,x_i)=\exp\{\|x-x_i\|^2/2\sigma^2\}$ 性能好，参数少易于优化。本文建立 RBF 核的 SVM 模型。

3.2.3 参数优化 应用网格搜索法对参数对 (c,σ) 在其搜索空间进行搜索，网格法是将 c 和 σ 分别取 N 个值和 M 个值，对 $N\cdot M$ 个 (c,σ) 的组合，分别训练不同的 SVM，从而在 $N\cdot M$ 个 (c,σ) 的组合中得到平均绝对误差最小的一个组合作为最优参数。调用 tunelssvm () 函数实现参数优化，在函数参数中指定 Leaveoneout 作为评估函数，grid-search 网格搜索法为优化函数，MAE（平均绝对误差）为成本函数。

3.3.4 用优化的参数确定 RBF 核的 SVM 模型，用训练样本训练计算支持向量等模型参数，得到 SVM 回归模型。调用 trainlssvm () 函数实现。

3.3.5 用回归模型对测试样本预测。调用 simlssvm () 函数实现。

4　实验及结论

4.1　基于全部因子的 LS-SVM 的预测结果

为观察支持向量回归模型的拟和效果，南疆和北疆的训练和测试偏差如图 3 所示。南疆、北疆的预测原始值和实际原始值的平均绝对误差分别为 0.819、0.918。南北疆预测原始值和实际原始值的偏差大多在 3 d 以内，只有个别点偏较大（南疆 2005 年，北疆 1979 年），可能是由于前期气候因子异常偏高或偏低所致。南疆 2005 年的前期（2004 年）气候因子中，冬季的气温日较差（t_R）较常年异常偏小，春季日平均相对湿度（x_P）较常年异常偏大。模型对于异常离群点的预测不太敏感。

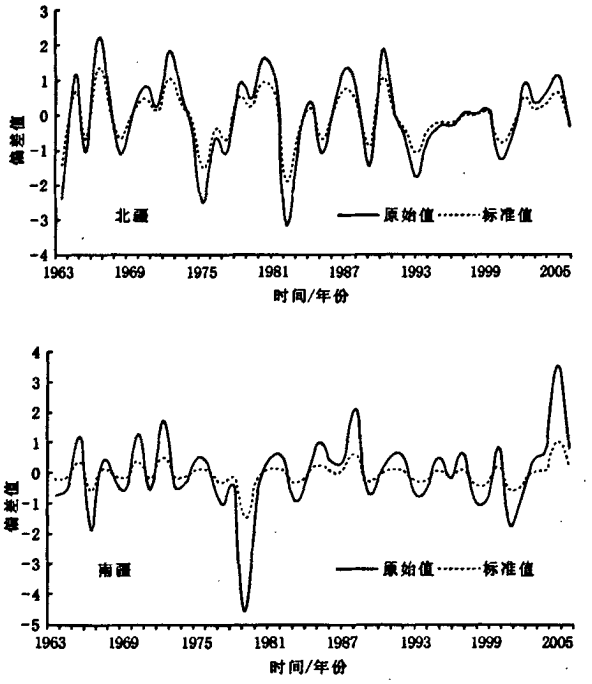

图 3　南、北疆模型偏差值

若定义预测值和实际值的标准化数值的绝对值偏差≤0.5 为正确，则北疆的正确率为 79%，南疆的正确率为 91%。南疆的预测效果明显好于北疆，这主要因为南疆的春季平均沙尘暴日数（8.19 d）明显大于北疆的春季平均沙尘暴日数（2.34 d）。

4.2　基于回归法选择预报因子的预测结果

预测原始值和实际原始值的平均绝对误差南疆为 0.861，北疆为 0.950（图略）。基于回归法选择预报因子的预测误差稍微大于基于 21 个预报因子的预测结果，这是由于回归法挑选出的因子是对预报目标有线性依赖关系，可能遗漏一些对目标有影响的预报因子。由于本文备选因子不多，并没有显著影响到 LS-SVM 预测模型的计算速度，所以基于 21 个预报因子的 LS-SVM 预测模型是可行的。

综上所述，对新疆春季沙尘暴历史资料进行分析，通过回归分析提取的春季沙尘暴影响因子有一定的解释能力。而基于 21 个预报因子的 LS-SVM 模型对新疆春季沙尘暴预测具有较好的精度。支持向量回归模型处理小样本的沙尘暴预测问题有一定的应用价值。但是沙尘暴预测是一个受多种因素影响的复杂问题，本文针对预报模型进行了粗浅探讨，还有一些细节因素需要纳入以便对模型进一步修正，尤其是预报因子的选择需要更加深入细致的探讨。

参考文献：

[1]　王炜，方宗义．沙尘暴天气及其研究进展综述[J]．应用气象，2004，15（3）：376-381.

[2]　钱正安，宋敏红，李万元．近 50 年来中国北方沙尘暴的分布及变化趋势分析[J]．中国沙漠，2002，22（2）：106-111.

[3]　工汉芝，刘振全，王萍．模糊权的神经网在沙尘暴预报中的应用[J]．天津科技大学学报，2005，20（2）：64-67.

[4]　赵智超．基于数据挖掘的沙尘暴智能系统的研究[D]．天津：天津大学电气与自动化工程学院，2005.

[5]　冯汉中，陈永义．处理非线性分类和回归问题的一种新方法—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.

[6]　Pelckmans K. LS-SVMlab Toolbox User's Guide version 1.5 [EB/OL]．[2003-02-05]．http://www.esat.kuleuven.ac.be/sista/lssvmlab/.