基于SVM的沙尘暴预测模型

路志英，张启孟，赵智超
（天津大学电气与自动化工程学院，天津 300072）

摘 要：根据沙尘暴天气的特点和支持向量机（support vector machine，SVM）方法在解决小样本学习问题中的优势，提出基于SVM的沙尘暴预测模型. 首先利用主成分分析法进行数据预处理，然后选择了径向基核函数，并通过分析惩罚参数和核参数对SVM分类器性能的影响，确定了参数的搜索空间，继而利用网格搜索法对其进行优化. 在此基础上，构建并实现了基于SVM的沙尘暴预测模型. 该模型与BP神经网络模型的运行结果对比表明，基于SVM的沙尘暴预报模型稳定性好，运行速度快，预报准确率提高了71.2%.

关键词：支持向量机；核函数；主成分分析；沙尘暴预测
中图分类号：TP391 **文献标志码**：A **文章编号**：0493-2137(2006)09-01110-05

Sand-Dust Storm Forecasting Model Based on SVM

LU Zhi-ying, ZHANG Qi-meng, ZHAO Zhi-chao
(School of Electrical and Automation Engineering, Tianjin University, Tianjin 300072, China)

Abstract: For the characteristics of the sand-dust storm weather and the advantages of support vector machine (SVM) in solving the learning problem with fewer samples, the sand-dust storm forecasting model based on SVM is proposed. The data is preprocessed by principal component analysis(PCA). Then the radial basic function (RBF) kernel is chosen, and the search space of the penalty parameter and the kernel parameter is defined by analyzing the influence of the two parameters on the performance of SVM classifier. And the two parameters were optimized by grid search in the search space. Lastly, the sand-dust storm forecasting model based on SVM is constructed and implemented. Results comparison between the proposed model and the BP neural networks model show that the sand-dust storm forecost model based on SVM has a better stability, faster running speed and its forecasting precision ratio is increased by 71.2%.

Keywords: support vector machine; kernel function; principal component analysis; sand-dust storm forecast

近年来，我国沙尘暴、扬沙和浮尘天气频繁发生，严重干扰着人们的正常生活，对社会经济和环境均造成一定程度的危害，使人们愈来愈认识到沙尘暴是影响大气、生态以及生存环境的重要问题. 因此，人们加强了对沙尘暴预报的研究，但由于沙尘暴预报的复杂性，一直未能取得比较满意的结果.

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[1-2]，大多采用BP神经网络预测模型，由于神经网络是以传统统计学理论为基础，而传统统计学的前提条件是要有足够多的样本，因此BP神经网络稳定性差，容易产生过学习现象，预测效果并不理想. 20世纪90年代中期发展起来的支持向量机（support vector machine，SVM）以统计学习理论（statistical learning theory，SLT）为基础，该理论着重研究小样本条件下的统计规律和学习方法，因此，SVM能较好地解决小样本、非线性和局部极小点等实际问题[3].

笔者将SVM应用于沙尘暴预测中，通过参数选择构建了基于SVM的沙尘暴预测模型，取得了比较满意的结果.

收稿日期：2005-12-05；**修回日期**：2006-05-20.
作者简介：路志英（1964— ），女，副教授，luzy@tju.edu.cn.

1　SVM 基本理论

支持向量机由线性可分情况下的最优分类面发展而来,其基本思想可用图 1 所示的 2 类线性可分情况说明.

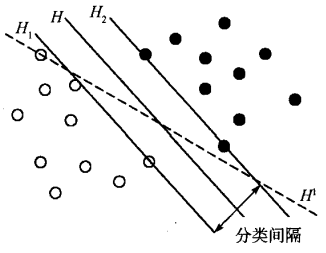

图 1　线性可分 SVM

Fig. 1　Linear separable SVM

图 1 中,实心点和空心点代表 2 类样本;H 为最优分类超平面;H_1 和 H_2 分别为过各类中离分类超平面最近的样本且平行于最优分类超平面的平面,它们之间的距离叫做分类间隔. 所谓最优分类面就是要求分类面不但能将 2 类训练样本正确分开(训练错误率为 0),而且使分类间隔最大. 距离最优分类超平面最近的样本向量称为支持向量.

1.1　线性可分问题

设样本 $\boldsymbol{x}_i$ 为 d 维向量($i = 1,2,\cdots, k$, k 为训练样本数). 根据每个样本 $\boldsymbol{x}_i$ 属于 w_1 或者 w_2,分别令 $y_i = +1$ 或 $y_i = -1$. 组成样本集$(\boldsymbol{x}_i, y_i)$,$\boldsymbol{x}_i \in R^d$,$y_i \in \{+1, -1\}$. 设分类线方程为 $\boldsymbol{x} \cdot \boldsymbol{x} + b = 0$,则样本集满足

$$y_i(\boldsymbol{w} \cdot \boldsymbol{w}_i + b) \geqslant 0 \tag{1}$$

适当调整$\boldsymbol{w}$ 和 b,可将式(1) 改写成

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geqslant 1 \tag{2}$$

根据最优分类面的定义,可得分界面的分类间隔为 $d(\boldsymbol{w}, b) = \min\limits_{x_i | y_i = 1} \dfrac{\boldsymbol{w} \cdot \boldsymbol{x} + b}{\| \boldsymbol{w} \|} = \max\limits_{x_i | y_i = -1} \dfrac{\boldsymbol{w} \cdot \boldsymbol{x}_i + b}{\| \boldsymbol{w} \|} = \dfrac{2}{\| \boldsymbol{x} \|}$,此时,使分类间隔最大等价于使 $\| \boldsymbol{w} \|$ 最小,可将问题转化为在条件(2) 下最小化 $\dfrac{1}{2} \| \boldsymbol{w} \|^2$;由拉格朗日乘数法,又可将问题转化为比较简单的对偶问题,也就是在约束条件 $\begin{cases} \alpha_i \geqslant 0 \\ \sum\limits_i \alpha_i y_i = 0 \end{cases}$ 下最小化,即

$$\Phi(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j=1}^{k} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_i \cdot \boldsymbol{x}_j) \tag{3}$$

式中 α_i 为与每个样本对应的 Lagrange 乘子. 这是一个不等式约束下二次函数寻优的问题,存在唯一解,解中非零值的 α_i 所对应的样本就是支持向量,则最优分类函数为

$$f(\boldsymbol{x}) = \operatorname{sgn}(\boldsymbol{w} \cdot \boldsymbol{x} + b) = \operatorname{sgn}\left(\sum_{i=1}^{N_s} a_i y_i \boldsymbol{x}_i \cdot \boldsymbol{x} + b \right) \tag{4}$$

式中:N_s 为支持向量的个数;b 可由任意一个支持向量满足式(2) 中的等式求得.

1.2　线性不可分问题

如果训练样本是线性不可分的,可以在条件(3) 中加一个松弛因子 $\xi_i \geqslant 0$,变为 $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geqslant 1 - \xi_i$,显然,当分类出现错误时,$\xi_i \geqslant 1$,$\sum\limits_{i=1}^{k} \xi_i$ 是分类错误数量的一个上界. 折中考虑最少错分样本和最大分类间隔,即将目标函数变为 $\Phi(\boldsymbol{w}, b) = \dfrac{1}{2} \boldsymbol{w} \cdot \boldsymbol{w} + C \sum\limits_{i=1}^{k} \xi_i$,其中,$C$ 是一个大于零的常数,它控制对错分样本的惩罚程度,称为惩罚因子. 由拉格朗日乘数法,将问题等价为在约束条件 $\begin{cases} 0 \leqslant \alpha_i \leqslant C \\ \sum\limits_i \alpha_i y_i = 0 \end{cases}$ 下最小化式(3).

1.3　非线性问题

由以上分析可以看出,通过把原问题转化为对偶问题,计算的复杂度不再取决于空间维数,而是取决于样本中支持向量数. 在对偶问题求解中,不论是寻优函数(3),还是分类函数(4),都只涉及训练样本之间的内积运算$(\boldsymbol{x}_i \cdot y_i)$,所以在高维空间实际上只需进行内积运算,如果存在一个核函数 $K(\boldsymbol{x}_i, y_i)$,满足 $K(\boldsymbol{x}_i, y_i) = (\boldsymbol{x}_i \cdot y_i)$,则内积运算就可以用原空间中的函数实现. 根据泛函理论,只要一种核函数 $K(\boldsymbol{x}_i, y_i)$ 满足 Mercer 条件,它就对应某一变换空间的内积[4].

因此,在最优分类面中采用适当的内积函数 $K(\boldsymbol{x}_i, y_i)$ 就可以实现某一非线性变换后的线性分类,而计算复杂度却没有增加,此时目标函数(3) 变为 $\Phi(\alpha) = \sum\limits_{i=1}^{k} \alpha_i - \dfrac{1}{2} \sum\limits_{i,j=1}^{k} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$,分类函数变为 $f(\dot{\boldsymbol{x}}) = \operatorname{sgn}\left(\sum\limits_{i=1}^{N_s} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b \right)$,这就是支持向量机.

1.4　核函数

核函数的选择需要满足 Mercer 条件,选择什么样的核函数,就意味着将训练样本映射到什么样的空间去进行线性划分. 目前流行的核函数有 3 种:

(1) 径向基核函数(radial basic function,RBF),即 $\exp(-\| \boldsymbol{x}_i - \boldsymbol{x}_j \|^2 / 2\sigma^2)$.

(2) 多项式核函数,即 $K(x_i, y_i) = (\gamma x_i^{\mathrm{T}} x_j + r)^d$, $r > 0$.

(3) Sigmoid 函数,即 $\tanh(Kx_i^{\mathrm{T}} \cdot x_j + \theta)$.

概括地说,支持向量机就是由支持向量确定的线性分类机,首先通过核函数定义的非线性变换将输入空间变换到一个高维空间,继而在这个空间中求最优分类面.

2 基于 SVM 的沙尘暴预测模型的建立

2.1 源数据

所采用的数据集是美国环境气象中心(NCEP)提供的,每天一个样本,每个样本的结构相同. 鉴于我国沙尘暴发生的时间和地区特点,将 NCEP 资料圈定在我国西北部从初春到初夏的范围之内(东经 70° ~ 115°,跨度45°;北纬35° ~ 55°,跨度20°). 所圈定的地域范围内生成格点场. 按照 NCEP 资料每 2.5° 记一格的数据格式,东西向跨45° 分为 18 格、南北向跨 20° 分为 8 格,东西向和南北向的交叉记为一个格点,这样,一个格点场的数据量就是 19 × 9 = 171 个,即每个格点场提供 19 × 9 的数据阵. 考虑到沙尘暴形成的基本条件,选取 4 个物理场即 500 hPa 的高度场、700 hPa 的 2 个风场和 850 hPa 的位温(θ_{se})场. 这样,一个样本的总维数为 171 × 4 = 684. 选取 1981 年至 1997 年从初春到初夏共采集到 2 027 个样本.

2.2 数据预处理方法

由于高维数据量太大,很难用来预报,必须进行数据消减. 本文应用主成分分析法(principal compment analysis, PCA)将每个样本从 684 维降为 56 维的特征向量作为支持向量机模型的训练样本集和测试样本集. 假设要压缩的数据由 N 个 d 维元组或数据向量组成. 主成分分析搜索 c 个 d 维正交向量,它们能够最佳地表示数据,其中 $c \leqslant d$. 这样原始数据就被投影到一个小得多的空间,从而产生数据压缩.

几何上,主成分分析可以看作是坐标轴的旋转,将原始坐标系的坐标轴旋转成一组新的正交坐标轴,并按照它们占原始数据方差的多少排列这些坐标轴. 为了要找到描述数据的一组数量较少的潜在变量,希望最初的几个坐标轴占原始数据中方差的大多数. 主成分分析是一种数据驱动的方法. 它没有假定数据内部存在不同的类,因此被描述成一种非监督的特征提取方法.

本文采用 Hotelling[5] 提出的方法对原数据的每个样本(684 维)进行新变量的正交变换,使新变量具有方差的极值. 得到新变量的方差,然后画出方差谱线,如图 2 所示,利用方差谱线的拐点,获取保留足够多原信息的前 56 维的新变量作为压缩后的新特征.

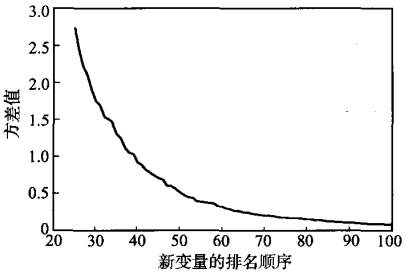

图 2 沙尘暴主成分的主差谱线

Fig. 2 Variance spectrum of PCA in sand-dust storm data

2.3 核函数的选择

本文选择 RBF 核函数来构造基于 SVM 的沙尘暴预测模型. 主要原因[6] 是:

(1) RBF 核函数只含 1 个参数 σ^2,易于参数优化,多项式核函数和 S 核函数有 2 个参数,会使参数优化更复杂,而得到的 SVM 的性能却并不比 RBF 核函数得到的 SVM 好;

(2) RBF 核函数更容易实现,因为 $0 < \exp(-\| x_i - x_j \|^2/2\sigma^2) < 1$,而多项式核函数,当次数很大时它的内积值可能趋于无穷($\gamma x_i^{\mathrm{T}} x_j + r > 1$)或者趋于 0($\gamma x_i^{\mathrm{T}} x_j + r < 1$).

2.4 参数选择

2.4.1 支持向量机参数对其性能的影响

Vapnike 等[7] 的研究表明,核参数和误差惩罚因子 C 是影响 SVM 性能的主要原因. 核参数 σ^2 主要影响样本数据在高维特征空间中分布的复杂程度,而误差惩罚因子 C 的作用是在确定的特征空间中调节学习机的置信范围和经验风险的比例. 因此要想获得推广能力良好的 SVM 分类器,首先要选择合适的 σ^2 将数据映射到合适的特征空间,然后针对该确定的特征空间寻找合适的 C 以使学习机的置信范围和经验风险具有最佳比例.

Keerthi 的研究[8] 表明对于某一确定的足够大的 C,当 $\sigma^2 \to 0$ 时会发生严重的"过学习"现象,此时径向基核 SVM 能把训练样本正确分开,但对测试样本不具有任何泛化能力;当 $\sigma^2 \to \infty$ 时会发生严重的"欠学习"现象,此时径向基核 SVM 把所有训练样本都划分到样本数较大的一类. 从核函数 $\exp(-\| x_i - x_j \|^2/2\sigma^2)$ 可以看出,σ^2 的大小完全是针对 $\| x_i - x_j \|^2$ 而言的. 因此,在实际应用中,只要 σ^2 的取值比

训练样本之间的最小距离小得多,就能达到 $\sigma^2 \to 0$ 的效果;当 σ^2 比训练样本之间的最大间隔大得多时就可以达到 $\sigma^2 \to \infty$ 的效果[9]. 基于这一考虑,实验中将 σ^2 的搜索空间确定为 $[\min(\|x_i - x_j\|^2 \times 10^{-3}), \max(\|x_i - x_j\|^2 \times 10^3)]$. 通过对训练样本的 $\|x_i - x_j\|^2$ 最大最小值的计算,确定 σ^2 的搜索空间为 $[2^7, 2^{19}]$.

在构造分类超平面方程时惩罚因子 C 的作用是对拉格朗日因子 α_i 的取值加以限制. 当惩罚因子 C 较小时,无论推广能力错误率的估计值还是实际测试的错误率都比较高;当 C 增加时,这些值急剧降低,即性能迅速提高;继续增加 C,性能的变化并不明显,当 C 增加到一定值后,性能不再随 C 的变化而变化,即在较大的范围内,推广能力对 C 的变化不敏感[10]. 因此,基于实验经验将 C 的搜索空间确定为 $[1, 2^{12}]$.

2.4.2　参数优化

应用网格搜索法对参数对 (C, σ^2) 在其搜索空间进行搜索. 网格法是将 C 和 σ^2 分别取 N 个值和 M 个值,对 $N \times M$ 个 (C, σ^2) 的组合,分别训练不同的 SVM,再估计其推广识别率,从而在 $N \times M$ 个 (C, σ^2) 的组合中得到推广识别率最高的一个组合作为最优参数[11].

针对网格法需要很多时间的问题,本文采取分步搜索的方法:首先用大步长来搜索,找出结果较好的参数对,再在此参数对附近重新划出搜索空间,并缩小步长进行再搜索,如此重复几次,直至得到较满意的结果.

3　实验结果及分析

3.1　数据预处理

对源数据使用主成分分析法进行数据预处理,将得到的 56 维的特征向量作为预测模型和与 BP 网络对比实验的训练样本和测试样本,具体构成见表 1.

表 1　实验样本

Tab. 1　Experimental samples

样本采集时间	总个数	非沙尘暴日	沙尘暴日	沙尘暴日频数/%	单样本维数
1981 - 1997 (初春 ~ 初夏)	2 027	1 454	573	28.27	56

3.2　参数优化过程

本文选用样本集中的前 1 627 个样本作为训练样本集,用后 400 个样本作为测试样本集,用 CSI 值来评价预测模型的性能,待选择的参数为 (C, σ^2). 首先 C 的值分别取 $(2^0, 2^3, 2^6, 2^9, 2^{12})$, σ^2 的值分别取 $(2^7, 2^{10}, 2^{13}, 2^{16}, 2^{19})$,得出较好的 (C, σ^2) 参数对为 $(2^3, 2^{13})$,CSI 值为 0.56;然后在 $[2^3, 2^{13}]$ 邻近区域重新划定 C 的搜索区间 $[2^0, 2^6]$ 和 σ^2 的搜索区间 $[2^{11}, 2^{15}]$;并缩小网格宽度进行第 2 次网格搜索. 如此反复直至在微调参数时,CSI 趋于稳定,此时得到最好的参数对 $(4, 2^{12})$,CSI 为 0.582.

3.3　实验结果

利用支持向量机对原数据进行数据预处理后所得样本进行学习,通过参数优化构造出了基于 SVM 的沙尘暴预测模型. 整个预测模型的构建是通过 Matlab 编程实现的. 由于训练样本数的多少会影响预测模型的性能,因此本文用不同训练样本集对 SVM 模型进行训练,并和 BP 网络模型(实验样本完全相同的条件下)进行对比,实验结果如表 2 所示.

表 2　SVM 预测模型和 BP 预测模型的比较

Tab. 2　Results comparison between SVM and BP forecasting model

样本集		SVM 预测模型					BP 网络模型	
训练样本数	测试样本数	C	σ^2	支持向量数	试报 CSI	运行时间/s	试报 CSI	运行时间/s
600	1 427	16	2^{13}	362	0.416	1.09	0.260	22.98
1 000	1 027	3	2^{10}	616	0.422	1.87	0.301	49.90
1 627	400	4	2^{12}	872	0.582	2.82	0.340	48.03

表2中CSI为成功界限指数,气象上常用CSI来衡量预报模型的性能,其定义为

$$\mathrm{CSI} = \frac{c_f}{c_f + w_f} \times 100\%$$

式中:c_f 为正确报出的沙尘暴日数;w_f 为漏报与空报数之和.

3.4　实验结果分析

对比实验中所用的 BP 神经网络是经过泛化的网

络,稳定性得到很大提高,但是仍然不好.也就是说,进行重复试验,各次所得的CSI值仍有一定差异.表2中所给出BP模型的CSI值是多次实验的平均值.而SVM则有很好的稳定性,进行重复实验都能得到稳定的CSI值.

由表2可知,无论SVM模型还是BP神经网络模型,当训练样本增大时,预测模型的性能都有不同程度地提高.这说明当训练样本少时,不能包括所有沙尘暴类型,预测模型对一些类型的沙尘暴缺乏学习的机会,故不具备识别能力.适当地增加训练样本数目,会增加沙尘暴的类型,使预测模型预报水平提高.但由于总样本数有限,训练样本越多,则测试样本会相应减少,而测试样本过少,预测模型性能就得不到全面的检验,实验中保证了有一定数目的测试样本.

训练样本集和测试样本集完全相同时,无论在CSI值、还是运行时间以及稳定性上,SVM模型的学习能力明显比BP神经网络要强,特别是在小样本情况下,即当训练样本数很小时,SVM表现出了更突出的学习能力.由此可见本文所构造的基于SVM的预测模型在性能上比以往方法有了很大提高.

4 结 论

(1)支持向量机在气象方面的应用还处于试验探索阶段.实验表明,基于SVM的预测模型比传统方法(BP神经网络)在稳定性、准确率和运行速度等方面有了很大改进.

(2)SVM是通过对历史样本的学习来建立预测模型的.因此SVM预测模型的建立过程就是对历史样本的自学习记忆过程,由于多的训练样本可能会包含多的典型样例,因此训练样本增多会使预测模型的性能更好.

(3)实验中的支持向量数占训练样本总数的比例较大,如果能通过改进参数优化过程,构造一个支持向量数相对较少的最优或广义最优分类面,则会得到更高性能的预测模型.

参考文献:

[1] 王汉芝,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.
Wang Hanzhi, Liu Zhenquan, Wang Ping. Apply of fuzzy neural networks with fuzzy weights to the forecasting of sand-dust storm[J]. *Journal of Tianjin University of Science and Technology*,2005,20(2):64—67(in Chinese).

[2] 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.
Zhao Zhichao. A Study of Intelligent Dust Storm Forecast System Based on Data Mining[D]. Tianjin: School of Electrical and Automation Engineering, Tianjin University, 2005 (in Chinese).

[3] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报,2000,26(1):32—42.
Zhang Xuegong. Introduction to statistical learning theory and support vector machines[J]. *Acta Automatica Sinica*, 2000,26(1):32—42(in Chinese).

[4] Vapnik V. *The Nature of Statistical Learning Theory*[M]. New York: Springer Verlag, 1995.

[5] Hotelling H. Analysis of a complex of statistical variables into principal components[J]. *Journal of Educational Psychology*, 1993, 24:417—444.

[6] Keerthi S S, Lin Chih-Jen. Asymptotic behaviors of support vector machines with Gaussian kernel[J]. *Neural Computation*, 2003,15(7):1667—1689.

[7] Vapnik V. *Statistical Learning Theory*[M]. New York: John Wiley and Sons,Inc,1998.

[8] Keerthi S S. Efficient tuning of SVM hyperparameters using radius/margin bound and iterative algorithm[J]. *IEEE Transactions on Neural Networks*,2002,13(5):1225—1229.

[9] 杨 旭,纪玉波,田 雪.基于遗传算法的SVM参数选取[J]. 辽宁石油化工大学学报, 2004,24(1):54—58.
Yang Xu, Ji Yubo, Tian Xue. Parameters selection of SVM based on genetic algorithm[J]. *Journal of Liaoning University of Petroleum and Chemical Technology*, 2004,24(1): 54—58(in Chinese).

[10] Lin Kuan-Ming, Lin Chih-Jen. A study on reduced support vector machines[J]. *IEEE Transactions on Neural Networks*, 2003,14(6):1449—1459.

[11] Hsu Chih-Wei, Chang Chih-Chung, Lin Chih-Jen. A Practical Guide to Support Vector Classification[R]. UK: School of Electronics and Computer Science, University of Southampton, 2000.