支持向量机分类方法在天空云量预报中的应用

熊秋芬[1] 顾永刚[2] 王 丽[2]

(1. 中国气象局培训中心,北京 100081; 2. 武汉中心气象台)

提 要: 以2001年5月1日至2004年12月31日逐日武汉市地面、高空观测资料及欧洲中心24小时预报场等资料为基础,构建了不同的训练样本集,基于支持向量机方法进行了大量多因子的随机交叉验证,从而筛选出了包含最佳预报因子的训练样本集和相应的核参数 g,建立了武汉市天空云量的预报模型。交叉验证结果表明预报模型是稳定性的、且具有较好的预报能力和推广应用能力。预报试验和实时预报的结果都显示出SVM方法对天空云量有一定的预报能力。

关键词: SVM方法 天空云量 预报 筛选因子 优化参数

Application of SVM Method to Cloud Amount Forecast

Xiong Qiufen[1] Gu Yonggang[2] Wang Li[2]

(1. Training Center, China Meteorological Administration, Beijing 100081;
2. Wuhan Central Meteorological Observatory)

Abstract: Based on the data from the surface and upper air observations in Wuhan and numerical synoptic prediction data from EC during May 1st 2001 to December 31st 2004, the different samples are made. Based on SVM method, cross-validations are performed with randomly samples to find the samples with best factors and optimization parameter g, then the models of the cloud amount forecast are built. The stability, the forecast and generalization ability of the models are also revealed by cross-validations. The results of test and real-time forecast show the forecast ability of the cloud amount forecast models by SVM.

Key Words: support vector machine(SVM) method cloud amount forecast select factors optimized parameters

资助项目:"中国气象局数值模式创新基地"开放课题、"精细化气象要素预报业务系统"和"湖北省基于数值预报产品的精细化预报系统建设"课题共同资助。

收稿日期: 2006年2月17日; 修定稿日期: 2007年4月5日

引　言

机器学习研究是从观测数据出发寻找规律,利用这些规律对未来数据或无法观测的数据进行预测。支持向量机(Support Vector Machine,简记 SVM)就是基于统计学习理论而发展起来的研究,实际应用中有限样本情况的机器学习方法[1-4]。其突出优点是基于结构风险最小化归纳原则,而不是传统统计方法的经验风险最小化原则,表现出很多优于已有方法的性能,迅速引起各领域的注意和研究兴趣,取得了大量的应用研究成果[5]。2004 年该方法首次被应用于气象要素的预报[6-7],目前已在降水、温度预报和短期气候预测中得到初步成功的应用[6-9]。

由于 SVM 是基于历史数据训练学习的一种建模方法,使用不同的训练样本、不同核参数,得到的预报模型是不同的。如何筛选预报因子、优化参数使得预报模型最优,一直是 SVM 方法应用的难点,也是该方法使用者关注的问题。

在已有的关于 SVM 方法的应用研究中[7-9],只讨论了核参数的优化而没有对预报因子进行筛选,然而训练样本中因子过多或过少都会影响模型的效果,传统的选择预报因子的方法是通过相关场分析来提取因子信息的[10-13]。此外,天空云量的预报是日常天气预报中必不可少的要素之一,目前该预报的制作主要依赖预报员的经验或 MOS 预报[10]。因此,本文就以天空云量为预报对象,基于 SVM 方法,进行预报因子的筛选和核参数的优化,然后建立天空云量的预报模型,力求为天空云量的客观预报提供新的途径。

1　SVM 分类方法基本原理简介[6]

机器学习问题可概括地表述为:给定训练样本$(x_1,y_1),(x_2,y_2),\cdots\cdots,(x_l,y_l)$,其中$x_i\in R^N$,为 N 维向量,$y_i\in\{-1,1\}$或 $y_i\in\{1,2,\cdots\cdots,k\}$,给出预报数据集:$x_{l+1}$, x_{l+2}, $\cdots\cdots,x_m$,通过训练学习建立分类模式 $M(x)$,使其不但对训练样本能够正确分类,而且具有较强的推广能力。即可以由模式对于输入的预报数据 x_i 得到正确的对应输出值 y_i。

对于训练样本集的线性二类划分问题,就是寻求函数:

$$y=f(x)=\mathrm{Sgn}((w\cdot x)+b) \quad (1)$$

使对于 $i=1,2,\cdots\cdots,l$ 满足条件:

$$y_i=f(x_i)=\mathrm{Sgn}((w\cdot x_i)+b) \quad (2)$$

其中 $w,x,x_i\in R^N,b\in R$,w,b 为待确定的参数,Sgn 为符号函数。显然$(w\cdot x)+b=0$ 为划分超平面,w 为其法方向向量。

对于线性可分离的问题,满足条件形如(1)的线性决策函数是不唯一的。图 1 给出二维情况下满足条件的划分直线的分布区域图。落在虚线区域内的任一直线都可作为决策函数。谁是最优的决策函数,就要对其进行判断。

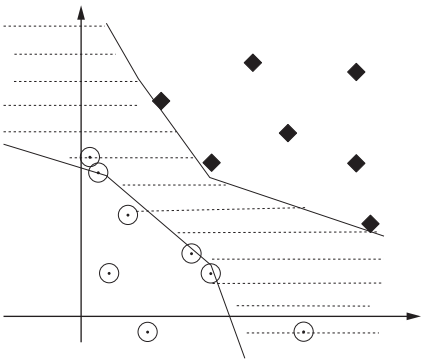

图 1　划分直线的分布区域图

Vapnik 提出一个间隔最大化原则。所谓间隔最大化原则是指寻求使间隔达到最大的划分为最优,即是对 w,b 寻优,求得最大间隔:$\underset{w,b}{\mathrm{Max}}(\mathrm{Min}(\|x-x_i\|:x\in R^N,(w\cdot x)+b=0,i=1,\cdots,l))$,对应最大间隔的划分超平

面称为最优划分超平面，简称为最优超平面，如图 2 中的 L。图 2 中两条平行虚线 l_1，l_2（称为边界）距离之半就是最大间隔。

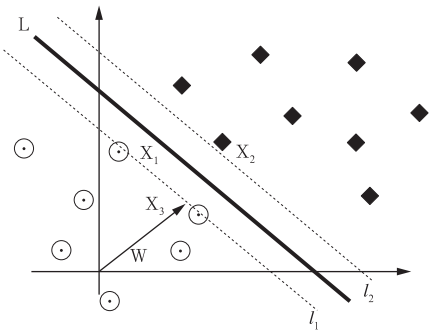

图 2 最优划分超平面示意图

最大间隔和最优超平面只由落在边界上的样本点完全确定，而不依赖于所有点，称这样的样本点为支持向量，如图 2 中的 x_1，x_2，x_3 样本点。

对于给定的训练样本集，根据相关的理论和算法，最终获得的线性支持向量机为：

$$M(x) = \mathrm{Sgn}((w^* \cdot x) + b^*) = \mathrm{Sgn}(\sum_{S.V.} \alpha_i^* y_i (x \cdot x_i) + b^*) \quad (3)$$

其中 α_i^*，b^* 为确定最优划分超平面的参数；$(x \cdot x_i)$ 为两个向量的点积，$S.V.$ 为支持向量。

对于线性不可分的情况，通过非线性映射 φ，把样本集映射入一个高维乃至于无穷维的特征空间，使在样本空间中高度非线性问题在高维空间中应用线性分类的方法得以实现。

由于在特征空间中采用的是线性分类方法，所以在特征空间中的最优划分超平面分类函数的形式为：

$$M(x) = \mathrm{Sgn}((w^* \cdot \varphi(x)) + b^*) = \mathrm{Sgn}(\sum_{i=1}^{l} \alpha_i^* y_i (\varphi(x) \cdot \varphi(x_i)) + b^*) \quad (4)$$

与式(3)相比，这里只是用 $\varphi(x)$ 和 $\varphi(x_i)$ 代替了 x 和 x_i。根据 Mercer 定理，式(4)最终转变为：

$$M(x) = \mathrm{Sgn}((w^* \cdot \varphi(x)) + b^*) = \mathrm{Sgn}(\sum_{S.V.} \alpha_i^* y_i K(x \cdot x_i) + b^*) \quad (5)$$

式(5)就是 SVM 方法确定的最终非线性分类的决策函数。与式(3)相比，这里只是用 Mercer 核函数的计算代替了点积的计算，在整个求解过程中不需要知道非线性映射的显式表达式，而是通过支持向量来表达预报因子与预报对象的关系。

2 训练样本集的构建和预报因子的筛选

2.1 训练样本集的构建

天空云量是云遮蔽天空视野的成数，即有云天空占天空总面积的比例。我国云量计量单位采用 10 成制。天空无云，云量为 0；天空完全为云所遮蔽，云量为 10。本文以武汉市单站日平均总云量（每日 02、08、14 和 20 时 4 个时次总云量的平均值 N）为预报对象，约定日平均总云量＜4 为“晴天到少云”，日平均总云量≥8 为“多云到阴天”。由于武汉市属亚热带季风区，不仅受中高纬西风带系统的影响，而且还受西南季风、东南季风等系统的影响，选取预报因子时既要能尽可能地描述大气的运动和变化状态，又要考虑足够的样本长度。所以本文选取了 2001 年 5 月 1 日至 2004 年 12 月 31 日（共 1340 个样本）逐日武汉市地面 02、08、14、20 时气温、相对湿度、气压、风、总云量、低云量的观测值和高空 08、20 时 925、850、700、500、400hPa 的位势高度、温度、露点、风的观测值以及 20 时欧洲中心 500hPa 高度、850hPa 温度、地面气压 24 小时预报场及其组合场等（共 127 因子）历史资料作为训练学习分析数据。

由于待选的预报因子太多，为了减少筛选因子的盲目性，对上述资料进行了处理，即在训练样本中分别去掉一些因子来重新构造

样本,形成不同的训练样本集。现只介绍下述 5 类:

(1) 训练样本集 1:包含了上述所选的资料,共 127 个因子;

(2) 训练样本集 2:去掉训练样本集 1 中欧洲中心的组合场因子,保留其它因子,共 114 个因子;

(3) 训练样本集 3:由欧洲中心 24 小时预报场及其组合场和地面、高空资料中与湿度有关的量组成,共 63 个因子;

(4) 训练样本集 4:由欧洲中心 24 小时预报场及其组合场和 20 时的地面、高空资料组成,共 81 个因子。

(5) 训练样本集 5:由欧洲中心 24 小时预报场及其组合场组成,共 49 个因子。

这样就形成了包含因子个数不同的 5 类训练样本集,需要说明的是,5 类训练样本集中都加入了后一天武汉市日平均总云量的实况值。

2.2　预报因子的筛选和参数的优化

应用 SVM 的两类分类方法,标定日平均总云量＜4(占总样本的 21.27%)为正样本(＋1 类)、日平均总云量≥4 为负样本(－1 类);从 5 类训练样本集的 1340 个样本中各随机抽取 10%作为检验样本,其余 90%样本作为建模样本,来分别对建立的"晴天到少云"的预报模型进行交叉验证,为了避免各预报因子之间量级的差异,在建模之前,对全部样本的每个因子分别做了归一化处理,使每个因子的数据在[0,1]之间。

这里选取最常用的径向基核函数,通过调整核参数 g 的值(惩罚系数 C＝100),进行大量随机交叉验证,分析比较 T_s 评分的高低,从而选择出包含最佳预报因子的训练样本集。表 1 仅给出了 g 为 0.001、0.01、0.032、0.05 时,按正样本的 T_s 评分寻优标准进行的各 50 次随机交叉验证的统计结果。

表 1　5 类训练样本集各进行 50 次交叉验证的 T_s 评分平均值

g 的取值	0.001	0.01	0.032	0.05
训练样本集 1	0.3565	0.4142	0.4163	0.4170
训练样本集 2	0.3252	0.4176	0.3967	0.4021
训练样本集 3	0.1320	0.3480	0.3656	0.3625
训练样本集 4	0.3382	0.4205	0.4218	0.4171
训练样本集 5	0.0	0.2561	0.3147	0.2956

从表 1 的结果可以看出,训练样本集 3、5 在 g 的 4 种取值下平均的 T_s 评分值都比较低,最不适合作为建模训练样本集;而 g 为 0.001 时,对于 5 类训练样本集平均的 T_s 评分都很低,特别是用训练样本集 5 建立的模型完全没有预报能力,所以它不是最优的核参数;当 g 取 0.01、0.032 和 0.05 时,训练样本集 1、2、4 的平均 T_s 评分相对较高,其中以 g 取 0.01 和0.032时,训练样本集 4 的平均 T_s 评分最高,因此可以认为 g 取 0.01 和0.032 时的训练样本集 4 中包含了最佳的预报因子,即可选择训练集 4 作为下一步预报建模的训练样本。

此外,不仅可从 T_s 评分的高低来选择最佳预报因子和核参数,还可从边界上支持向量个数、VC 维等方面进行核参数和预报因子的寻优。一般说来,T_s 评分越高、边界上支持向量个数越多、VC 维越小,得到的预报模型越好。如表 1 中训练样本集 1、2、4 在 g 取 0.05 时平均的 T_s 评分不是太低,但相应的 VC 维较大、边界上支持向量个数相对较少,所以选择 g 取 0.01、0.032时 T_s 评分较高的训练样本集 4 比较合理。限于篇幅,这里不一一列出边界上支持向量和 VC 维等数据。

从天气意义上来分析,也可以解释经过大量交叉验证筛选后的训练样本集 4 中包含了最佳的预报因子,该训练样本集既含有离预报时效较近的地面、高空因子,又含有预报场及其变化量,这些因子较全面地反映了大气的运动状态和变化趋势。而训练样本集 1 包含了太多前期的无用因子,增加了杂噪;训

练样本集2既包含了太多前期的无用因子，干扰因素多，又没有反映预报场变化的组合因子，因而不能充分描述大气运动的变化；训练样本集3的地面、高空资料中只含有与湿度有关的量而没有温度、风等要素，对大气的热力和动力状况描述不足；训练样本集5仅由欧洲中心数值预报产品组成，不包含直接反映大气前期状态的因子，会对预报模型的推理效果产生影响。因此训练样本集1、2、3、5都不是最佳的建模训练样本，只有训练样本集4才是最佳的建模训练样本。

3 SVM模型的稳定性和预报能力分析

3.1 SVM模型的稳定性

从上面的分析可知，训练样本集4为最佳的建模训练样本，为了分析所建模型的稳定性，仍然标定日平均总云量＜4为正样本、日平均总云量≥4为负样本，再从1340个样本中随机抽取20%、30%作为检验样本，其余样本作为建模样本，来对建立的“晴天到少云”的预报模型进行交叉验证，在$C=100$、$g=0.01$和$g=0.032$时分别按正样本的T_s评分寻优标准进行各50次随机交叉验证，平均结果见表2。

表2 对训练样本集4按不同比例进行交叉验证的平均T_s评分和准确率(%)

			样本比例		
分类	评分	g	10%	20%	30%
	T_s	0.01	0.4205	0.4109	0.4004
N＜4成		0.032	0.4218	0.4108	0.4057
	T_s	0.032	0.6690	0.6624	0.6497
		0.039	0.6613	0.6541	0.6599
N≥8成	准确率	0.032	79.66	80.26	79.59
		0.039	79.08	79.62	79.41

为了进一步符合业务预报的需求，不仅要建立“晴天到少云”的预报模型，还要建立“多云到阴天”的预报模型。为此，对训练样本集4进行重新分类，标定日平均总云量≥8(占总样本的48.88%)为正样本、日平均总云量＜8为负样本，同样从1340个样本中随机抽取了10%、20%、30%作为检验样本，其余样本作为建模样本，来对建立的“多云到阴天”预报模型进行交叉验证，在$C=100$、$g=0.032$和$g=0.039$时分别按正样本的T_s评分寻优标准进行的各50次随机交叉验证，统计结果也在表2中。

由于日平均总云量≥8的分类中，正、负样本的比例差不多(接近50%)，同样随机抽取10%、20%和30%的样本作为检验数据，按全样本的分类准确率寻优标准进行了随机交叉实验，表2也给出了$C=100$、$g=0.032$和$g=0.039$时分别进行的各50次交叉验证的平均统计结果。

由表2可知，对于日平均总云量＜4的模型，无论g取0.01还是0.032，虽然单次检验的T_s评分有高低起伏，但T_s评分的平均值在0.40～0.422之间，相差不大，且都比实际样本的21.27%提高了近20%，有显著的正预报技巧；对于日平均总云量≥8的模型，在g取0.032和0.039时T_s评分的平均值在0.65～0.67之间，比实际样本的48.88%提高了约17%，也表现出明显的正预报技巧。同样由表2可见，对于日平均总云量≥8的模型，全样本的预报准确率在80%左右。上述的分析表明SVM方法建立的模型有较好的预报能力且是稳定的。

另外随着随机抽取的检验样本的不断增加，相应的建模训练样本在减少，而日平均总云量＜4和日平均总云量≥8两种分类情况下T_s评分的平均值(准确率)的变化不大，并没有明显减少，也表明建立的模型是稳定的，同时也说明SVM方法建立的模型具有推广应用能力。

3.2 SVM模型的预报能力

为了进一步检验模型的预报效果，将训

练样本集 4 的 1340 个样本作为建模训练样本,用 SVM 方法分别建立了日平均总云量＜4 和日平均总云量≥8 预报模型,再将 2005 年 1 月 1 日至 5 月 31 日逐日 20 时武汉市地面和高空观测资料、欧洲中心 24 小时预报场和组合场等资料组成的 81 个预报因子分别做归一化处理后(实际只有 141 天资料),分别输入预报模型来试报第二天武汉市的日平均总云量。计算时,对于日平均总云量＜4 的情况,$C=100$、$g=0.01$ 和 $g=0.032$;对于日平均总云量≥8 的情况,$C=100$、$g=0.039$ 和 $g=0.032$,试报结果详见表 3。

表 3　对预报试验和实时预报的检验

	预报试验				实时预报	
分类	N＜4 成		N≥8 成		N＜4 成	N≥8 成
g	0.01	0.032	0.039	0.032	0.01	0.039
正样本 T_s 评分	0.464	0.406	0.716	0.696	0.444	0.690
正样本正确次数	13	13	73	71	4	20
正样本空报次数	6	10	22	22	2	5
正样本漏报次数	9	9	7	9	3	4
负样本正确次数	113	109	39	39	32	12
正样本个数(比例)	22(15.6%)	22(15.6%)	80(56.73%)	80(56.73%)	7(17.1%)	24(58.5%)
全样本准确率(%)	89.4	86.5	79.4	78.0	87.8	78.0

从表 3 中的试报结果可看出,对于日平均总云量＜4 的模型,g 取 0.01 和 0.032 时 T_s 评分分别为 0.464 和 0.406,比实际样本的 15.6%提高了 30.8%和 25%,正预报技巧显著,表明建立的“晴天到少云”模型对天空云量有较好的预报能力;但由于 g 取 0.032空报次数多,因此,g 取 0.01 可作为“晴天到少云”预报模型的最优核参数。同样对于日平均总云量≥8 的模型,在 g 取0.039 和 0.032 时 T_s 评分分别为 0.716 和 0.696,比实际样本的 48.88%提高了 22.74%和 20.74%,具有正预报技巧;全样本的预报准确率分别为 79.4%和 78%,也表明建立的“多云到阴天”模型对天空云量有较好的预报能力;但由于 g 取 0.032 预报正确次数少,因此,g 取 0.039 可作为“多云到阴天”预报模型的最优核参数。

4　实时业务应用情况

以训练样本集 4 的 1340 个样本和 2005 年 1 月 1 日至 5 月 31 日的 141 个样本(共 1481 个样本)作为建模训练样本,用 SVM 方法重新建模。对于日平均总云量＜4 的模型,$C=100$、$g=0.01$,而对于日平均总云量≥8 的模型,$C=100$、$g=0.039$。目前武汉中心气象台已在 MICAPS 系统中自动读取每日 20 时武汉市地面、高空观测资料和欧洲中心 24 小时预报场,并进行预报量的组合等计算,对这些资料组成的 81 个预报因子做归一化处理后,分别输入预报模型来预报第二天武汉市的日平均总云量,实现了该方法的实时业务化。2006 年 1 月 1 日至 2 月 10 日共 41 天的实时预报结果见表 3。

实时预报结果表明(表 3),对于日平均总云量＜4 和日平均总云量≥8 的模型,T_s 评分分别为 0.444 和 0.690,比实际样本的 17.1%和 58.5%分别提高了 27.3%和 10.5%,两种预报模型都具有正预报技巧,证明用 SVM 方法建立的模型在实际预报中是有预报能力的;同时也表明该方法具有可行性和实用性。

5 结论和讨论

(1) 通过建立不同的训练样本集,基于SVM方法进行大量多因子的随机交叉验证,来实现预报因子的筛选和核参数 g 的优化,为筛选因子和优化参数提供了思路。

(2) 交叉验证的结果表明SVM建立的模型有较好稳定性和预报能力,且具有推广应用能力;预报试验和实时业务应用的结果也表明SVM方法对天空云量有一定的预报能力。

(3) SVM方法为天空云量的预报提供了客观参考依据,同时实现了科研成果向业务预报能力的转化。

当然本文的研究工作仅仅是用SVM方法筛选因子、优化参数以及气象要素预报方面的初步应用,由于该方法应用于气象预报的时间不长,特别是将该方法投入日常预报业务使用,仍有许多工作有待于进一步完善;另外本文的不足之处也表现在数值预报产品的长度(年限)不足、预报场中不含有湿度因子等方面。

参考文献

[1] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York:Springer Verlag, 2000.

[2] Cristianini N, Shawa-Taylor J. An Introduction of Support Vector Machines and Other Kernel_based Learning Methods[M]. Cambridge :Cambridge University Press, 2000.

[3] Scholkopf B, Burges Ch-J C and Smola A J, et al. Advances in Kernel Methods: Support Vector Learning[M]. Cambridge: MIT Press, 1999.

[4] 陈永义.支持向量机方法与模糊系统[J]. 模糊系统与数学,2005, 19(1): 1-11.

[5] 祁亨年. 支持向量机及其应用研究综述[J]. 计算机工程, 2004,30(10):6-9.

[6] 陈永义, 俞小鼎, 高学浩,等. 处理非线性分类和回归问题的一种新方法(Ⅰ)——支持向量机方法简介[J]. 应用气象学报,2004,15(3):345-354.

[7] 冯汉中,陈永义. 处理非线性分类和回归问题的一种新方法(Ⅱ)——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):355-365.

[8] 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(1):41-44.

[9] 李智才,马文瑞,李素敏,等. 支持向量机在短期气候预测中的应用[J]. 气象,2006,32(5):57-61.

[10] 胡江林,李劲.湖北省天空云量的特征分析及其预报[J]. 湖北气象,2000(2):15-17.

[11] 段旭. 云南盛夏大雨物理量因子的选取及效果检验[J].气象,1996,22(7):30-32.

[12] 孙建明,李法然,杨育强.暴雨预报因子及其统计特征[J].气象,1998,24(11):36-39.

[13] 段蜻,苗春生.上海浦西地区雾持续时间的统计释用预报[J].气象,2001,27(7):31-36.