文章编号：1006-4354（2004）06-0001-04

支持向量机方法在天气预报中的应用

赵国令[1]，肖科丽[2]

（1. 陕西省气象局，陕西西安 710014；2. 陕西省气象台，陕西西安 710014）

摘 要：简要介绍了支持向量机方法（SVM）方法的基本原理和使用方法。用高空 500 hPa 月平均高度、海洋温度以及地面资料作为因子，对西安 6—9 月份降水总量建立了 SVM 预报模型。

关键词：支持向量机；SVM；预报模型

中图分类号：P456.9 **文献标识码**：A

对于非线性划分问题，一种新的数学方法——支持向量机（Support Vector Machines 简称 SVM）方法，正在迅速推广使用，为解决非线性分类问题提供了一种新方法。陈永义教授首先将其介绍并应用到了气象领域[1-2]。本文简要介绍了 SVM 方法的基本原理和使用方法，并用此制作了西安 6—9 月汛期降水短期气候预测方法。

1 SVM 方法简介

1.1 线性划分问题

在 N 维向量空间 $\boldsymbol{x}\in\mathbf{R}^N$ 中（预报业务中 $\boldsymbol{x}$ 代表 N 个因子序列），预报对象 y 的类别是完全分离的（假定 y 只有两种情况），则可用一超平面 L 将两类对象进行划分（见图 1），L_1，L_2 为两个边界超平面。超平面 L 的方程为：

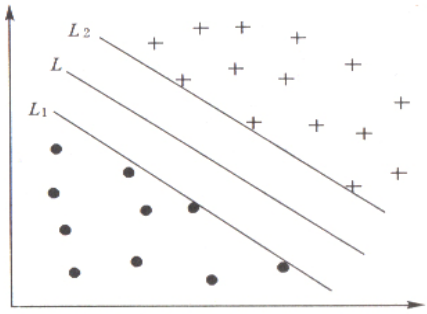

图 1 完全分离的两类样本示意图

$$(\boldsymbol{w}\cdot\boldsymbol{x})+b=0。\qquad(1)$$

其中 $\boldsymbol{w}$，$\boldsymbol{x}$ 为向量，$\boldsymbol{w}$ 为超平面的法向向量，b 为方程参数，$(\boldsymbol{w}\cdot\boldsymbol{x})$ 为内积运算。假定两条边界 L_1 和 L_2 其与 L 的距离是 1，则两条边界的方程为：

$$(\boldsymbol{w}\cdot\boldsymbol{x})+b=\pm1。\qquad(2)$$

设 N 维空间上的点 $\boldsymbol{x}_1$ 在 L_1 上，$\boldsymbol{x}_2$ 在 L_2 上，则满足方程：

$$(\boldsymbol{w}\cdot\boldsymbol{x}_1)+b=-1,$$

$$(\boldsymbol{w}\cdot\boldsymbol{x}_2)+b=+1,$$

两式相减有：$(\boldsymbol{w}\cdot(\boldsymbol{x}_2-\boldsymbol{x}_1))=2。\qquad(3)$

进而有：

$$\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\cdot(\boldsymbol{x}_2-\boldsymbol{x}_1)\right)=\frac{2}{\|\boldsymbol{w}\|},\qquad(4)$$

$\|\boldsymbol{w}\|$ 为向量 $\boldsymbol{w}$ 的模。(4) 式左边恰好就是连接 $\boldsymbol{x}_1$，$\boldsymbol{x}_2$ 的向量在划分超平面法方向上的投影，它是最大间隔的 2 倍。L_1 与 L_2 距离越大，说明两类样本分开的越远，效果越好。求最大间隔等价于求 $\|\boldsymbol{w}\|$ 的最小值问题。为计算方便，它等效于求 $\|\boldsymbol{w}\|^2$ 的最小值问题。

要使所有训练样本点分类正确，应成立：

$(\boldsymbol{w}\cdot\boldsymbol{x}_i)+b\geqslant1$，若 $y_i=1$；

$(\boldsymbol{w}\cdot\boldsymbol{x}_i)+b\leqslant-1$，若 $y_i=-1$。

其中 $y_i\in\{-1,1\}$ 为预报量。两式可以合并为：

$$y_i((\boldsymbol{w}\cdot\boldsymbol{x}_i)+b)\geqslant1,\qquad(5)$$

这样，建立线性支持向量机的问题转化为求解如下一个二次凸规划问题：

收稿日期：2004-07-16

作者简介：赵国令（1961-），男，河北沧州人，学士，高工，从事气象业务管理工作。

$$\begin{cases}\min\ (\frac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{约束条件：}\ y_i\ ((\boldsymbol{w}\cdot\boldsymbol{x}_i)+b)\geqslant 1,\end{cases}$$

可求得最优超平面决策函数为：

$$M(\boldsymbol{x})=\mathrm{Sgn}(\sum_{s.v.}\alpha_i^* y_i(\boldsymbol{x}\cdot\boldsymbol{x}_i)+b^*)。\quad(6)$$

其中 α_i^*，b^* 为确定最优划分超平面的参数，Sgn 为符号函数。求和号 $\sum$ 下的 $s.v.$ 表示只对支持向量求和，非支持向量对应的 α_i 都为零。支持向量一般只占样本的少数。

只有少数几个训练样本点就决定了最优超平面，其余的样本均不起作用，称这样的样本点为支持向量。超平面的划分不是依赖于所有点，而只是由支持向量决定，这样可极大地减少计算量。

1.2 线性不可分的情况

对于线性不可分的情况，通过引入松弛变量 $\xi_i\geqslant 0$，修改目标函数和约束条件，应用完全类似的方法可以求解。与（6）类似的新的规划问题为：

$$\begin{cases}\min\ (\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_i\xi_i) \\ \text{约束条件：}\ y_i\ ((\boldsymbol{w}\cdot\boldsymbol{x}_i)+b)\geqslant 1-\xi_i\end{cases}\quad(7)$$

若 ξ_i 都为零，（7）式就变成了线性可分问题（6）式。（7）式中大于零的 ξ_i 对应错分的样本，参数 C 为惩罚系数。加大 C 的值就会逐渐减少错分样本的个数。

1.3 非线性划分问题

对于高度非线性的分类问题，对（6）式通过一个非线性映射

$$\boldsymbol{X}=\Phi(\boldsymbol{x})\quad(8)$$

可把样本空间映射到一个高维以至无穷维的特征空间中，使其转化为线性问题，如图 2。

这样就可应用线性学习机的方法分类。求解过程只需增加计算核，并不需要非线性映射 $\Phi(\boldsymbol{x})$ 的显式表达式。将（8）式代入（6）式得：

$$M(\boldsymbol{x})=\mathrm{Sgn}(\sum_{s.v.}\alpha_i^* y_i(\varphi(\boldsymbol{x})\cdot\varphi(\boldsymbol{x}_i))+b^*),\quad(9)$$

令 $K(\boldsymbol{x},\boldsymbol{x}_i)=\varphi(\boldsymbol{x})\cdot\varphi(\boldsymbol{x}_i)$，则：

$$M(\boldsymbol{x})=\mathrm{Sgn}(\sum_{s.v.}\alpha_i^* y_i K(\boldsymbol{x},\boldsymbol{x}_i)+b^*)。\quad(10)$$

二元函数 $K(\boldsymbol{x},\boldsymbol{x}_i)$ 通常称为核函数（简称核）。对称正定的连续核称为 Mercer 核，由点积

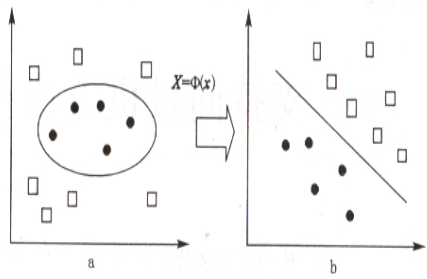

图 2 非线性划分问题（a）经过线性变换：$\boldsymbol{X}=\Phi(\boldsymbol{x})$，转为线性划分问题（b）的示意图

定义的核（$K(\boldsymbol{x},\boldsymbol{x}_i)=\varphi(\boldsymbol{x})\cdot\varphi(\boldsymbol{x}_i)$）必是 Mercer 核。Mercer 核很多，如径向基函数核、多项式函数核等。

（10）式就是非线性支持向量学习机的最终分类决策函数。虽然用到了特征空间及非线性映射，但实际计算中并不需要知道他们的显式表达。只需要求出支持向量及其支持的“强度”和阈值，通过核函数的计算，即可得到原来样本空间的非线性划分输出值。

对非线性不可分的情况，可采取与线性不可分相同的方法处理。

通过上述变换和计算，用线性的 SVM 算法解决了非线性 SVM 问题。而线性 SVM 的算法归结为一个凸约束条件下的二次凸规划问题，对此已有很多成熟的算法和应用软件，其中陈永义教授组织在 Windows 下开发了应用集成软件，已在全国气象部门推广使用。

2 用 SVM 方法预报汛期（6—9 月）降水

在日常应用中，必须把总样本分为三部分，即训练样本、实验样本和检验样本。训练样本用于学习训练建立模型，样本的绝大部分应放在这里；实验样本用于进一步调整模型的参数，根据经验，这部分样本可控制在总样本的 1/4 左右；检验样本用于预报检验。

对短期气候预测而言，影响汛期降水的因素很复杂，很难找出物理意义明确的因子，所以用线性相关系数法普查方法选择月平均海温、500 hPa 月平均高度与汛期（6—9 月）总降水量相关密切的因子。由于每个月都有多个格点的海温和高空资料与降水相关较高，为了能将更长时间尺

度的信息反映进来，只是逢双月分别选一个高空和海温因子。共选出 21 个因子。其中地面因子 9 个，海温和高空各 6 个。降水资料长度为 1960—2001 年，共 42 a。逐旬平均地面温度、旬平均气温与降水的线性相关不明显，因此用点聚图方法选择。这种方法无具体定量指标，只要认为多雨点与少雨点在点聚图上分离的较好即可入选。例如，西安 4 月上旬、10 月中旬气温与汛期降水的相关系数都只有 0.2 左右，但点聚图效果非常好。

表 1、表 2 给出了 6—9 月降水量与前一年海温、500 hPa 平均高度的相关系数。相比之下地面因子的相关系数很小，如果用常规统计方法建立模型，地面因子很难入选。而 SVM 模型是非线性的，与因子的相关系数无关，这就是 SVM 的优势所在。

42 a 资料中，用 30 a 作为训练样本，8 a 实验样本，4 a 检验样本。取径向积核函数：

$$K(\boldsymbol{x}, \boldsymbol{x}_i) = \exp(-g \| \boldsymbol{x} - \boldsymbol{x}_i \|^2),$$

表 1　6—9 月总降水量与前期海温的相关系数

时间/月份	纬度/°N	经度	相关系数
02	0	125°W	0.46
04	15	135°E	0.45
06	30	145°W	0.41
08	30	145°W	0.43
10	40	145°W	0.46
12	10	165°W	0.43

表 2　6—9 月总降水量与前期 500 hPa 平均高度的相关系数

时间/月份	纬度/°N	经度	相关系数
02	15	80°E	0.46
04	30	140°E	−0.48
06	15	0	−0.45
08	70	140°E	−0.54
10	75	160°W	−0.41
12	55	90°E	−0.50

通过训练学习后求得决策函数为：

$$M(\boldsymbol{x}) = \mathrm{Sgn}\left(\sum_{s.y.} \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b\right),$$

式中 g 为迭代筛选确定的，α_i 为建模过程中依据条件约束自动生成，不需人为选取。

尽管选定径向积函数作为建立 SVM 模型的核函数，但伴随参数值选取的不同，函数形态会发生较大变化，引起 SVM 模型的变化。由于参数的选择没有规律，因而进行了大量的试验，表 3 给出了其中 7 次有代表性的结果。由表 2 可见，惩罚系数 C 和 g 初值的选择，使结果发生很大变化，而且选择都是人为的，这也是 SVM 方法的不足之处。当 $C>100$ 或 $C<0.5$ 时，结果变化不大，因此表中未给出。通过试验，当取 $C=0.9$，$g=1$ 时，所得结果较为满意。此时历史拟合无一次错误；4 a 预报，有 3 a 正确，只是所有用来训练的样本，都成为支持向量。

表 3　7 次不同初值的迭代结果

C 初值	C 增量	g 初值	g 增量/倍	C 值迭代结果	g 值迭代结果	支持向量数	错分样本数	4 a 预报准确数
100	10	10^{-9}	10	120	0.01	19	0	1
100	10	0.1	2	100	0.1	22	0	2
50	5	10^{-4}	5	85	0.01	18	0	1
1	2	10^{-5}	3	5	0.02	25	4	1
0.5	0.5	0.01	5	2.5	0.05	27	2	2
0.5	0.1	0.1	2	1	0.2	29	2	2
0.5	0.1	0.1	5	0.9	1	30	0	3

3　小结

3.1　SVM 模型最大特点是非线性的，因此选择因子时，不一定只是依靠相关系数，可根据物理意义，或用点聚图等方法多方面选择，因子不需

文章编号：1006-4354 (2004) 06-0004-04

西北区一次联合探测层状云系云物理特征分析

陈争旗[1]，陈保国[1]，许新田[2]，李照荣[3]，郭 强[1]

(1. 陕西省人工影响天气办公室，陕西西安 710014；2. 陕西省气象台，陕西西安 710014；
3. 甘肃省人工影响天气办公室，甘肃兰州 730020)

摘 要：对2003-09-17我国西北地区东部层状云降水天气过程的层状云系进行了联合探测。经观测对比分析发现：对同一天气系统开展多省联合探测，对认识云系发展变化、结构特征及降水潜力很有帮助；同一条锋面云系由于流场结构的差异，其不同位置的云层结构、温度、湿度、水汽含量存在一定差异；陇东和陕北虽然同处于副高西侧的偏南气流里，但是，低层天水处于反气旋区，抑制了水汽辐合，而延安处于横切变右侧的气旋性辐合区，有利于低层水汽辐合上升，促使该地区云系发展加强，其过冷层厚度大于天水地区；PMS粒子测量系统测得在整个过冷层中，小云粒子的数浓度平均值天水比延安少1个量级，2D-P测出延安的粒子谱较宽，说明上述条件下天水地区人工增雨潜力较小，延安地区相对条件较好，人工增雨潜力大。

关键词：联合探测；层状云系；结构特征

中图分类号：P426.5 **文献标识码**：B

为了进一步开发云中水资源，提高人工增雨作业的科技水平和效益，近年来一些省市相继在飞机上安装了宏微观的大气探测设备，以加深当地降水云系特征，特别是云微物理结构的了解。但由于各种条件限制，多数情况下仅为一架飞机单独开展探测和人工增雨作业，这对大尺度天气背景降水云系的探测有一定局限性。针对我国西北地区东部一次层状云降水天气过程，陕、甘两省制定了飞行探测方案，2架飞机开展了跨省联合飞行探测，通过对云系不同演变阶段宏微观大气的资料对比分析，加深了对西北地区降水性层状云系发展变化及其云物理特征的认识。

1 天气形势分析

2003-09-17，受西太平洋副热带高压外围西

收稿日期：2004-08-31

作者简介：陈争旗 (1959-)，男，陕西兴平人，学士，高工，从事天气预报和人工影响天气工作。

基金项目：国家科技部“西部开发科技行动”重大攻关项目 (2001BA901A41)

线性化。

3.2 SVM模型中不受因子个数的限制，也不受因子间相关与否的影响。

3.3 模型经过非线性变换，将非线性问题转为线性，其判断条件的确定，不一定由所有样本决定，而由关键的少数支持样本（支持向量）决定。

3.4 方法只是初次应用，所建的方法还有待于在实践中检验。参数初值的选择有一定人为性，而且初值的不同，影响迭代结果；程序并不一定保证收敛到最优值，因此需要多次人为选择试验。

参考文献：

[1] 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)：345-353.

[2] 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)：356-365.