

文章编号:1004-9045(2007)02-0159-04

支持向量机方法在单站降水预报中的应用探讨

王建生¹, 熊秋芬²

(1. 武汉中心气象台, 武汉 430074; 2. 中国气象局培训中心, 北京 100081)

摘要: 将武汉天空云量预报的 81 个预报因子运用到该站中等以上强度的降水预报中, 基于 SVM 方法进行了交叉验证和预报试验。结果表明用 81 个预报因子建立的 5-9 月和全样本的降水预报模型有较好稳定性、且对降水都有正的预报技巧。因此天空云量的预报因子可以用来做降水的预报因子, 同时也证明了这些预报因子在天空云量和降水预报中是协调的。SVM 方法为天空云量和降水的预报提供了客观参考依据。

关键词: SVM 方法; 天空云量; 预报因子; 降水预报

中图分类号: P457.6 文献标识码: A

1 引言

支持向量机(Support Vector Machine, 简记 SVM)方法^[1]是基于历史数据训练学习的一种建模方法, 但又不同于传统的卡尔曼滤波、ANN 等^[2-4]方法。SVM^[5-7]通过合适的内积函数定义非线性变换, 把样本空间的非线性关系转化为高维空间中的线性关系, 在变换后的高维空间中求出最优分类超平面, 从而实现样本分类。而超平面只是由关键样本点(少数支持向量)决定, 其余样本均不起作用。它是 Vapnik 等根据统计学习理论(Statistical Learning Theory, 简称 SLT)提出的一种新的机器学习方法, 在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势^[8]。支持向量机根据结构风险最小化准则, 在使训练样本分类误差极小化的前提下, 尽量提高分类器的泛化推广能力。从实施的角度, 训练支持向量机的核心思想等价于求解一个线性约束的二次规划问题, 从而构造一个超平面作为决策平面, 使得特征空间中两类模式之间的距离最大, 而且它能保证得到的解为全局最优解。

自 2004 年该方法首次被应用于气象要素预报以来^[9,10], 目前已在降水、温度、天空云量预报和短期气候预测中得到初步成功的应用^[10-14]。由于云和降水的关系十分密切, 所以本文将已用于天空云量的预报因子^[12]来作降水的预报, 并对预报效果的检验, 探讨提高降水预报准确率的问题, 同时将 SVM 与 ANN 方法的预报结果进行了比较, 为业务降水预报提供参考。

2 SVM 分类方法基本原理简介

机器学习问题可概括的表述为: 给定训练样本 $(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)$, 其中 $x_i \in R^N$, 为 N 维向量, $y_i \in \{-1, 1\}$ 或 $y_i \in \{1, 2, \dots, k\}$, 给出预报数据集: $x_{l+1}, x_{l+2}, \dots, x_m$, 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

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

使对于 $i=1, 2, \dots, l$ 满足条件

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

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。

V. N. 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: $\text{Max}_{w,b} (\text{Min}_{i=1, \dots, l} (|x_i - x|))$, 对应最大间隔的划分

收稿日期: 2007-02-05; 定稿日期: 2007-06-20

基金项目: “长江中游暴雨洪水定量预报系统”和“中国气象局数值模式创新基地”开放课题(2007)

作者简介: 王建生, 男, 1951 年生, 工程师, 主要从事天气预报服务工作。E-mail: xg-wang@sohu.com

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。

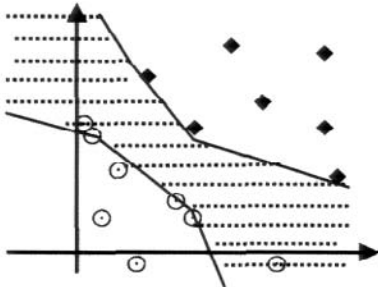


图 1 划分直线的分布区域图

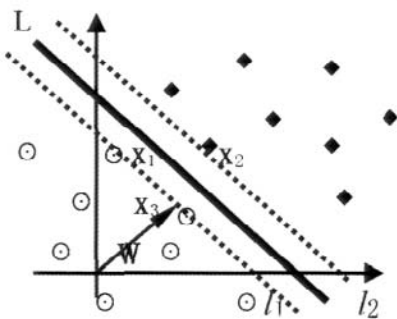


图 2 最优划分超平面示意图

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

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

其中 x_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，把样本集映射到一个高维乃至无穷维的特征空间（称为 Hilbert 空间），使在样本空间中的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：

$$M(x) = \text{Sgn}((w^* \cdot \phi(x)) + b^*) = \text{Sgn}(\sum_{S.V.}^* y_i (\phi(x) \cdot \phi(x_i)) + b^*) \quad (4)$$

与式(3)相比，这里只是用 $\phi(x)$ 和 $\phi(x_i)$ 代替了 x 和 x_i 。

根据 Mercer 定理，式(4)最终转变为：

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

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(关键样本)来表达预报因子与预报对象的关系。其基本思想简单地说就是升维和线性化，通过非线性映射，把样本空间映射到一个高维乃至无穷维的特征空间，在特征空间中，应用线性学习机的方法解决样本空间中的高度非线性问题。

2 训练样本集的构建和最优参数的选取

2.1 训练样本集的构建

该文以武汉市(57494)单站日降水量(08-08时)中 10 mm 以上降水为预报对象。选取了与文献[13]中相同的因子，它们是：逐日武汉市地面 20 时气温、相对湿度、气压、风、总云量、低云量的观测值和高空 20 时 925、850、700、500、400 hPa 的位势高度、温度、露点、风的观测值以及 20 时欧洲中心 500 hPa 高度、850 hPa 温度、地面气压 24 h 预报场及其组合场等(共 81 因子)。以 2001 年 5 月 1 日至 2006 年 4 月 30 日共 1792 个样本的历史资料(简称全样本集)作为训练学习分析数据。

2.2 参数的优化

应用 SVM 的两类分类方法，标定日降水量 10 mm 为正样本(+1 类)，这类样本共 188 个，占总样本的 10.92%；日降水量 < 10 mm 为负样本(-1 类)。从全样本集的 1792 个样本中随机抽取 10% 作为检验样本，其余 90% 样本作为建模样本，来对建立的降水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，即 $K(x, y) = e^{-g \sum_{i=1}^n (x_i - y_i)^2}$ ，通过调整核参数 g 和惩罚系数 C 的值，进行大量随机交叉验证，分析比较所建模型 Ts 评分的高低，从而选择出最优模型对应的参数 g 和 C 。

表 1 仅给出了 $C=120, g$ 为 0.01、0.032、0.04、0.044、0.06、0.115、0.3 时，按正样本的 Ts 评分寻优标准进行的各 50 次随机交叉验证的统计结果。

g 的取值	0.01	0.032	0.04	0.044	0.06	0.115	0.3	0.4
Ts 评分平均值	0	0.202	0.205	0.224	0.26	0.231	0.266	0.199

注：C=120

由于武汉夏半年和冬半年的降水系统不完全相同，而且强降水也主要集中在夏半年，为了提高降水的预报准确率，因此，从 2001 年 5 月 1 日至 2006 年 4

月 30 日挑出 5-9 月的样本资料(共 754 个,其中正样本 89 个,占总样本的 11.8%,简称 5-9 月样本集)重新构建训练样本集。

同样,在 5-9 月训练样本集的 754 个样本中,随机抽取 10%作为检验样本,其余 90%样本作为建模样本,来对建立的降水预报模型进行交叉验证,表 2 给出了 5-9 月样本集在不同参数下,按正样本的 Ts 评分寻优标准分别进行 50 次随机交叉验证后的部分结果。

表 2 5-9 月样本集各进行 50 次交叉验证的 Ts 评分平均值

g 的取值	0.02	0.06	0.09	0.11	0.14	0.18	0.2	0.3
Ts 评分平均值	0	0.15	0.194	0.206	0.232	0.273	0.231	0.245

注: C=125

从表 1 的结果可以看出, C=120, g=0.01 时, 平均的 Ts 评分为 0, 对应的降水预报模型完全没有预报能力, 所以它们是不可用的参数; 当 g=0.3 时, 平均 Ts 评分为 0.266, 相对较高, 比实际样本的 10.92% 提高了近 16%, 有正预报技巧; 而在 g 其它的取值下, 对应的平均 Ts 评分都低于 0.266, 因此可以认为 C=120, g=0.3 是最优的参数。同样表 2 表明, C=125, g=0.18 时, 平均 Ts 评分为 0.273, 比实际样本的 11.8% 提高了近 16%, 也有正预报技巧; 所以可认为 C=125, g=0.18 是最优的参数。但用这些核参数及对应的训练样本集分别建立的降水预报模型是否稳定、预报效果如何? 这些问题还需要按一定比例抽取检验样本进行多次随机交叉验证, 同时需用实例预报来检验。

3 SVM 模型的稳定性和预报能力分析

3.1 SVM 模型的稳定性

为了分析所建模型的稳定性, 仍然标定日降水量 10 mm 为正样本(+1 类)、日降水量 < 10 mm 为负样本(-1 类)。再从全样本集的 1792 个样本中随机抽取 15%、20% 作为检验样本, 其余样本作为建模样本, 来对建立的降水预报模型进行交叉验证, 在 C=120, g=0.3 时分别按正样本的 Ts 评分寻优标准进行各 50 次随机交叉验证, 平均结果见表 3。

同样从 5-9 月的 754 个训练样本中随机抽取了 15%、20% 作为检验样本, 其余样本作为建模样本, 来对建立的降水预报模型进行交叉验证, 在 C=125, g=

表 3 对全样本集和 5-9 月样本集按不同比例进行交叉验证的平均 Ts 评分

交叉验证样本比例	10%	15%	20%
全样本集 g=0.3	0.266	0.241	0.225
Ts 平均绝对偏差	0.08	0.064	0.057
5-9 月样本集 g=0.18	0.273	0.234	0.242
Ts 平均绝对偏差	0.126	0.085	0.079

0.18 分别按正样本的 Ts 评分寻优标准进行的各 50 次随机交叉验证的统计结果也在表 3 中。

由表 3 可知, 对于用全样本集建立的降水预报模型, 虽然单次检验的 Ts 评分有高低起伏, 但 Ts 评分的平均值在 0.225-0.266 之间, 比实际样本的 10.92% 提高了近 12%~16%, 有正的预报技巧; 对于用 5-9 月样本集建立的降水预报模型, Ts 评分的平均值在 0.234~0.273 之间, 比实际样本的 11.8% 提高了约 12%~16%, 也表现了正的预报技巧。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多, 相应的建模训练样本在减少, 而用两种样本集建立的预报模型的 Ts 评分的平均值并没有明显减少, Ts 平均绝对偏差也在减少, 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, 将全样本集的 1792 个样本作为建模训练样本, 用 SVM 方法建立了降水预报模型, 再将 2006 年 5 月 1 日至 12 月 31 日逐日 20 时武汉市地面和高空观测资料、欧洲中心 24 h 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(实际只有 213 d 资料), 分别输入预报模型来试报第二天武汉市降水。计算时, C=120, g=0.3, 试报结果详见表 4。同样用 5-9 月样本集的 754 个样本建立降水预报模型 (C=125, g=0.18), 来预报 2006 年 5 月 1 日至 9 月 30 日 (实际只有 135 d 资料) 武汉市降水, 试报结果也在表 4 中。

同时为了与人工神经网络(ANN)建模方法进行比较, 表 4 也给出了用 ANN 方法分别建立的全样本和 5-9 月样本的预报模型所对应的试报结果。计算时, 使用三层的反向传播神经网络模型(1 个隐含层), 隐含层节点数为 10, 迭代次数为 50 000 次。

表 4 试报结果的检验

方法	正样本 Ts 评分	正样本 正确次数	正样本 空报次数	正样本 漏报次数	负样本 正确次数	正样本个数 (比例)	所有样本 分类准确率
SVM(5-9 月模型)	0.261	6	7	10	112	16(11.9%)	0.874
(全样本模型)	0.250	8	13	11	181	19(8.9%)	0.887
ANN(5-9 月模型)	0.217	5	7	11	112	16(11.9%)	0.867
(全样本模型)	0.138	4	10	15	184	19(8.9%)	0.883

从表 4 中的试报结果可看出, 对于用 SVM 方法建立的预报模型, 5~9 月的 Ts 评分略高于全样本的 Ts 评分, 但全样本预报的分类准确率却高于 5~9 月的预报的准确率, 表明这两种模型各有优势。对于用 ANN 方法建立的 5~9 月和全样本的预报模型, 无论是 Ts 评分还是预报准确率都低于 SVM 方法, 说明在预报因子和预报样本相同的情况下, SVM 建立的模型的预报效果总体上优于 ANN 方法。

4 结论和讨论

该文将武汉天空云量预报的 81 个预报因子运用到该站中等以上强度的降水预报中, 并基于 SVM 方法进行了交叉验证和预报试验, 得到以下结果:

(1) 用 81 个预报因子建立的 5~9 月和全样本的预报模型对降水都有正的预报技巧, 表明天空云量的预报因子可以用来做降水的预报因子。

(2) 交叉验证的结果表明 SVM 建立的降水模型有较好稳定性和预报能力, 且具有推广应用能力; 预报试验也表明 SVM 方法对降水有一定的预报能力。这与文献[13]中得到的天空云量模型的稳定性和预报能力是一致的, 同时也说明 81 个预报因子在天空云量和降水预报中是协调的。

(3) 试报结果表明 SVM 建立的模型的预报效果总体上优于 ANN 方法。SVM 方法为天空云量和降水的预报提供了又一种客观参考依据。

当然, 该文的研究工作仅仅是用与云量相关的因子做降水预报, 但由于降水特别是强降水的形成机制比云要复杂的多, 所以预报模型有待于今后不断的改进。例如在模型中引入数值预报产品中的湿度因子等, 这样可能会进一步提高天空云量和降水的预报准

确率。

参考文献:

- [1] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York:Springer Verlag, 2000.
- [2] 周筱兰,张礼平.卡尔曼滤波方法在制作长期月平均气温预报中的应用[C]//武汉区域气象中心.暴雨·灾害.北京:气象出版社,1999(2):103-106.
- [3] 胡江林.神经网络模型用于湖北省月降水量预报的探讨[C]//武汉区域气象中心.暴雨·灾害.北京:气象出版社,1999(1):36-40.
- [4] 杨荆安,张鸿雁,陈正洪.降水 PP 模式的建立和检验[C]//武汉区域气象中心.暴雨·灾害.北京:气象出版社,1998(1):100-105.
- [5] Cristianini N, Shawe Taylor J. An Introduction of Support Vector Machines and Other Kernel_based Learning Methods[M]. Cambridge :Cambridge University Press, 2000.
- [6] Scholkopf B, Burges Ch J C, Smola A J. Advances in Kernel Methods-Support Vector Learning[M]. Cambridge: MIT Press, 1999.
- [7] 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.
- [8] 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.
- [9] 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(一)—支持向量机方法简介[J].应用气象学报,2004,15(3):345-354.
- [10] 冯汉中, 陈永义.处理非线性分类和回归问题的一种新方法(一)—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.
- [11] 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.
- [12] 熊秋芬, 顾永刚, 王丽.支持向量机分类方法在天空云量预报中的应用[J].气象,2007, 33(5): 20-26.
- [13] 熊秋芬, 胡江林, 陈永义.天空云量预报及支持向量机和神经网络方法比较研究[J].热带气象学报,2007,23(3).
- [14] 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.

The Discussion of Precipitation Predict in Wuhan Based on SVM Method

WANG Jian-sheng¹, XIONG Qiu-fen²

(1.Wuhan Central Meteorological Observatory, Wuhan 430074; 2.Meteorological Training Centre, CMA, Beijing 100081)

Abstract: 81 predictors from cloud amount forecast are used to predict the precipitation. Based on Support Vector Machine(SVM) method, cross-validations and test are performed. The results show that the stability and the forecast ability of the precipitation model are revealed. Therefore, the predictors from cloud amount forecast could be able to predict precipitation, and the predictors are harmony between cloud amount and precipitation predict. There is an objective way in cloud amount and precipitation predict by SVM

Key words: Support Vector Machine(SVM) method; Cloud amount; Predictor; Precipitation predict