

文章编号: 1006-4354 (2006) 03-0015-03

SVM 和 ANN 方法在铜川盛夏暴雨预报中的对比试验

燕东渭^{1,2}, 杨 艳², 孙田文³, 刘志镜¹

(1. 西安电子科技大学, 西安 710077; 2. 陕西省气象局, 西安 710014;
3. 铜川市气象局, 陕西铜川 727031)

摘 要: 介绍了利用 SVM (支持向量机) 方法的预报原理, 并分别用 SVM 和 ANN (人工神经网络) 两种方法对铜川盛夏暴雨的预测做了对比试验。分析预报结果表明, SVM 方法在天气预报领域有一定的使用空间, 并指出了 SVM 在气象领域应用研究的方向。

关键词: SVM; ANN; 暴雨; 应用

中图分类号: P457.6

文献标识码: A

自 20 世纪 40 年代起, 机器学习^[1]的方法就开始被尝试应用于天气预报中。近 10 年来, 在国外气象领域, 机器学习的许多方法 (如决策数 C4.5、ANN、BOOSTING、KNN 和遗传算法等) 都得到了非常广泛地应用和研究。80 年代后期, 在我国气象界, ANN (人工神经网络) 的应用曾经风靡一时。直到近几年, 还间或有相关的论文出现。由于 ANN 方法本身的一些问题一直没有新的突破, 开发的系统在实际气象业务应用中很少真正生存下去的。而其他几个方法却一直没有被国人试验过。SVM (支持向量机) 是在现代统计学习理论发展下的全新的方法, 它避免了以往统计方法的许多缺点, 正逐渐在各行各业显示出强劲潜力。陈永义^[2]等人 2002 年开始将 SVM 引入气象领域, 为我国预报方法研究指出了一条全新的方向。

1 SVM 理论

Tom M. Mitchell 给机器学习下过一个广义的定义, 即任何计算机程序通过经验来提高某任务处理性能的行为^[1]。SVM 是典型的机器学习方法, 它根据提供的训练实例数据 (包括数据的不同属性以及类别), 建立模型, 实现对实例未知类

别的分类预测。设 $S = \{ (x_1, y_1), (x_2, y_2), \dots, (x_m, y_m) \}$ 为训练集, $x_k \in R^N$ ($1 \leq k \leq m$) 是以向量的形式表示训练实例, 向量的不同维代表不同的属性, 对于基本的二分类问题, $y_k \in \{-1, 1\}$ 代表实例 x_k 的类别。所有的多分类问题都可以转化为二分类问题。依据不同属性 (因子), 把所有的实例样本在欧几里得空间中描述为一个点, 这个处理方法类似点聚图。如果能找到一个可以正确分类所有的样本点的超平面, 那么对于未知类别的预测对象, 就可以按照超平面将其分类。而现实生活中, 真正线性的问题是不存在的, 特别是对于复杂的未来天气的预测更是如此。对于非线性分类问题, 在此样本实例空间, 根本不存在划分超平面。SVM 借助非线性映射 ϕ 把训练实例映射到高维的“特征空间”, 在这个空间中寻找分类超平面 $w \cdot x + b = 0$, 使得其满足 S 中两类样本最近点的距离最大化, 即要解下式的二次凸规划问题^[2-3]:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{k=1}^m \xi_k$$

$$\text{约束条件 } y_k [(w, \phi(x_k)) + b] \geq 1 - \xi_k$$

$$\xi_k \geq 0 \quad k = 1, 2, \dots, m$$

收稿日期: 2005-12-23

作者简介: 燕东渭 (1975-), 男, 陕西周至人, 工程师, 学士, 从事计算机应用开发。

其中 ξ_k 为松弛变量, 表示对侵入对方范围的部分样本点到划分超平面的距离的惩罚 (在样本点不可分时作用), C 为惩罚系数。为解上述问题, 引用最优化理论中的 Kuhn-Tucke 定理, 并补充 KKT 条件, 将其转化为下式的对偶问题^[2-3]:

$$\max Q(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

$$\text{约束条件 } \sum_{i=1}^m y_i \alpha_i = 0 \quad \alpha_i \geq 0$$

用许多非常成熟的办法求解出上述对偶问题的 α^* 后, 代入下面的最优超平面决策函数, 即可对预测点进行分类, 从而对未来做出预报。

$$\text{sgn} \left(\sum_{x_k \in SV} \alpha_k y_k (\phi(x) \cdot \phi(x_k)) + b \right)$$

式中 SV 为支持向量集, 即离划分超平面比较近而能够决定划分超平面的样本点, 系数 α_k 和 b 从对偶问题解得, 是确定划分超平面的参数, 对于非支持向量点对应的 α_k 都是 0, 所以计算量并不大。根据 Mercer 条件, 可进一步使用核函数 $K(x, x_k)$ 代替上式中的内积运算 $\phi(x) \cdot \phi(x_k)$ 。这样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它们的显式表达, 而方便的解决非线性分类问题。

SVM 使用了统计学习理论中的结构风险最小化原则, 因而有良好的泛化能力和抗过匹配的能力。ANN 等方法往往存在对训练样本的过匹配问题, 从而导致虽然对训练样本拟合非常好, 但实际应用时却常常很不尽人意。

2 预报应用对比试验

盛夏暴雨的发生与预报因子有非常复杂的非线性关系, 这正是人工神经网络与支持向量机方法所擅长的。BP (反向传播) 网络是应用最广泛的一种神经网络, 其技术方法成熟, 工作状态稳定, 结构也比较简单, 能够比较容易地应用到天气预报中。1998 年孙田文^[4]利用 ANN 方法进行了铜川盛夏暴雨的预报实验, 历史拟合率和试预报效果都很好。模型训练数据中预报因子是从 1980 年到 1992 年 7—8 月兰州、西安等 20 个高空站 850、700、500 hPa 的 08 时要素中选取; 预报对象是铜川 3 个县 (区) 中, 有 1 个或以上站 20—20 时出现暴雨 (日降水超过 38 mm)。训练历

史拟合率大于 90%。运用训练模型对于 1994—1997 年 10 个实例的试验, 结果是 9 次暴雨报对了 7 次, 只有一次空报。从当时的试验结果看, ANN 的效果的确不错。

经过对历史资料的整理、预处理以及分型等步骤, 最后挑选了 6 个与铜川盛夏暴雨非常密切的因子: 西安与武汉 500 hPa 高度差、平凉 500 hPa 24 h 变温、平凉 700 hPa 24 h 相对湿度变化、酒泉与西安 500 hPa 温度差、甘孜 500 hPa 南风分量、格尔木 500 hPa 24 h 变高。当时的检验数据 (1994—1997 年) 已经无法找到。为了能够对比试验, 补充了 1998—2002 年的资料后, 共 94 个数据样本; 统一使用前 80 个数据 (1980—1989 年) 做模型训练, 后 14 个数据 (1990—2002) 做对比预报试验。ANN 继续采用孙田文使用的方法, 只不过模型要重新训练。SVM 方法采用 Steve Gunn 编写的基于 MATLAB 平台 SVM 工具包代码。SVM 的核函数采用径向基函数, 形式如下:

$$K(x, x') = \exp \left(-\frac{\|x - x'\|^2}{2\sigma} \right)$$

惩罚系数 C 和上述核函数参数 σ 分别取算法本身的默认值 (无穷大和 1)。预报结果见表 1。

表 1 SVM 和 ANN 方法对铜川盛夏暴雨的预报结果

预报方法	检验样本数	样本中目标类数	正确预报次数	空报次数	漏报次数
ANN	14	2	10	2	2
SVM	14	2	11	2	1

3 分析与讨论

从以上对比结果可见, 用 SVM 方法建立的预报模型同 ANN 相比效果稍好一些。考虑到 ANN 方法经过了对多个参数的多次寻优, 最后才确定了最优模型, 而 SVM 方法的几个参数都简单使用了方法本身的默认值, 因此 SVM 方法比 ANN 具有优势。其实, ANN 对于训练数据很准确, 而实际应用不太好。过匹配往往导致泛化能力差, 因为只要在训练时记住每个样本的类别, 就可以做到对训练数据完全匹配, 这样的模型没有推广能力。ANN 方法做预报试验的结果并没有 1998 年时的准确。这是因为 ANN 方法对训练

数据的过匹配,导致了泛化能力的降低,在真正做试验时性能很不稳定。

用SVM建立预报模型时,核函数的选取,以及参数的不同,都直接影响到划分超平面,也对预测结果有显著的影响。在上述试验中,两个参数都只是简单的使用算法本身的默认值,如果对参数进行一定范围的调整,效果会更好;若使模型更理想,应建立一定的评分体系,利用不同参数下的预测试验结果对模型进行评估,从而找出最佳模型对应的参数。但由于试验预报的样本比较少,如果为了进行评分而将样本分的过细小,不但不能反映样本的本来特征,还有可能造成寻找划分超平面的局部过匹配。如果样本非常大,可考虑适当的寻优,有一定的积极作用。

SVM方法是机器学习领域纯粹的数学方法,由于现代统计学理论的支持,应用于分类问题时比其他方法有着明显的优势。应用到气象领域时,由于气象本身的特点,还有许多具体的问题不能很快解决,比如预报因子的选取问题,需要依靠预报员的经验。人们采用不同的方法选择的因子,可能对最后模型的建立影响很大。很多人采用了普查相关系数法,但对于SVM这样的非线性方

法没有意文。本文采用了寻找明确物理意义及点聚图等方法选择预报因子。用SVM方法做预测试验,如何选择真正有物理意义的预报因子举足轻重;另外,天气预报中,两类问题的实例出现的次数往往是非常不均等的,如暴雨、冰雹等。如果方法本身只是片面追求准确率,可能导致结果向占多数的类别严重倾斜,比如将所有天气都预报成没有暴雨,准确率肯定很高,但没有什么实际意义,这在机器学习中被成为不平衡问题。寻求适合气象特色的基于不平衡问题的SVM方法,是真正使机器学习方法能应用于气象业务的前提。

参考文献:

[1] Mitchell Tom M. 机器学习 [M]. 曾华军,张银奎,译. 北京:机械工业出版社,2003:1-3.

[2] 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.

[3] 边肇祺,张学工. 模式识别 [M]. 北京:清华大学出版社,2000:296-320.

[4] 孙田文. 用人工神经网络预报铜川盛夏暴雨[J]. 陕西气象,1998 (1):4-6.

轻雾和霾的区别

肖 湘 卉

(临渭区气象局,陕西渭南 714000)

轻雾和霾都是视程障碍现象,使能见度降低、空气质量下降,在日常观测中要注意区分。

1 轻雾与霾的特征及成因的区别

随着工业发展和车辆的迅速增加,污染物排放和悬浮物增加,大量尘粒浮游空中。这些尘粒能吸收水,成为水汽凝结核。当空气湿度较大时,水汽很容易达到冷却饱和而凝结为水滴,从而形成雾。因此,轻雾是微小水滴或已吸湿的微粒悬浮在近地面层,使空气变成灰白色的稀薄雾幕。霾是大量极细微的干尘粒、烟粒或盐粒等均匀地浮游空中,使空气普遍混浊。

2 能见度的区别

水平能见度小于1.0 km的称为雾,水平能见度大于1.0 km而小于10.0 km的称为轻雾。霾出现时水平能见度小于10.0 km。渭南有霾时,水平能见度一般大于3.0 km而小于10.0 km,小于3.0 km的情况很少。

3 颜色的区别

轻雾出现时,天空为灰白色,相似于雾的颜色。霾出现时,远处光亮物体(背景明亮)微带浊黄色或桔红色,黑暗物体(背景暗淡)略带淡蓝色。在实际观测中,当能见度在8.0 km左右,