技术交流

支持向量机回归方法在实时业务预报中的应用①

冯汉中　　陈永义

（成都气象中心，610071）　　（中国气象局培训中心）

提　要

简要介绍了支持向量机(Support Vector Machine,简称 SVM)回归方法的基本原理,并介绍了基于 SVM 回归方法,利用 1990～2000 年 1～12 月 ECMWF 北半球的 500hPa 高度、850hPa 温度、地面气压的 0 小时分析场资料构造预报因子,建立德阳市 5 个代表站的日平均气温、日最高气温、日最低气温的 SVM 回归预报模型及其在业务化运用中的效果。

关键词： 支持向量机　回归　温度　预报

引　言

随着各种大气探测技术、数值预报模式、数值集合预报等的不断发展,可用于气象预报的信息越来越广泛和多样,如何从这些海量的信息中获取可用于预报的关键信息,是我们业务预报人员比较关注的问题。机器学习是解决这一问题的有效途径。随着学习理论的不断进步、处理信息的技术不断发展、计算机技术的不断飞跃,机器学习也在不断的深入。以人工智能为代表的研究工作取得一系列令人瞩目的成果(如专家系统、神经网络等)。近年发展起来的一种机器学习方法——支持向量机(Support Vector Machines 简称 SVM)方法[1]又为我们解决这一问题提供了比较有效的手段。

文献[1]对 SVM 方法的原理作了介绍,我们在气象预报领域用 SVM 方法进行了一些探讨性的试验[2],结果表明,SVM 方法能用于具有显著非线性特征的气象预测预报。但其有没有实时业务运用的能力？本文就是对 SVM 回归方法在德阳市气象局业务预报中的运用结果进行介绍,以期对 SVM 方法在气象领域的推广运用有所作用。

1　支持向量机(SVM)回归方法简介

回归分析又称函数估计,它要解决的问题是:根据给定的样本数据集$\{(x_i, y_i) \mid i = 1,\cdots,k\}$,其中 x_i 为预报因子向量,y_i 为预报对象值,寻求一个反映样本数据的最优函数关系 $y = f(x)$。

机器学习的过程就是由样本数据集建立学习机的过程。机器学习问题可以形式化为:

给定函数集 $f(x,\alpha)$和 l 个独立同分布的样本数据训练:

$$(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l)$$

其中 $x_i \in R^N$,为 N 维向量,α 为参数向量。如何从给定的函数集 $f(x,\alpha)$中选择出能够最好逼近实际响应的函数？聚类分析、模式识别、回归分析、密度函数估计、人工神经网络等,都可以看成是这里所说的机器学习的特例。比如线性回归分析,就是在线性函数

① 本工作得到国家自然科学基金(60072006)的资助。参加此项工作的还有雍朝吉、车怀敏、甯春容。

类中采用最小二乘法选取与样本点偏差平方和为最小的线性函数。然而,关于这种线性回归(即便是非线性回归)的推广能力并没有理论上的保证。SVM方法具有坚实的理论基础,并可以给出学习机推广能力的界。

SVM方法的基本思想简单说就是升维和线性化:基于Mercer核展开定理,通过非线性映射 φ,把样本空间映射到一个高维乃至于无穷维的特征空间(Hilbert空间),在特征空间中引入 ε——不敏感误差函数,定义最优线性回归超平面,把寻找最优线性回归超平面的算法归结为求解一个凸约束条件下的一个凸规划问题,并可以求得全局最优解。这样便应用线性学习机的方法解决了样本空间中的高度非线性分类和回归等问题。

线性化方法是人们解决复杂问题的一种常用手段。SVM的线性化是在变换后的高维空间中应用解线性问题的方法来进行计算。在高维特征空间中得到的是问题的线性解,但与之相对应的却是原来样本空间中问题的非线性解。

SVM方法的核心概念是支持向量。如图1所示,最优回归超平面 l 完全由落在两条边界线 l_1 和 l_2 上的样本点所确定,这样的样本点称为支持向量。落在两条边界线之间的所有样本点对最优回归超平面没有贡献。

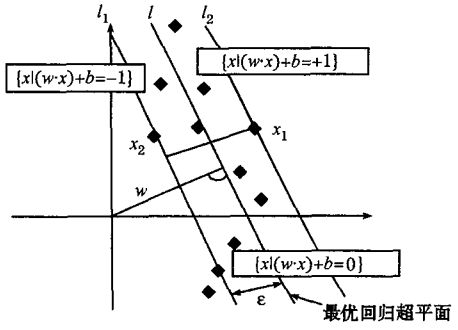

图1 最优回归超平面

根据相关的理论和算法[1],解最优化问题得到的最优线性回归函数表达式为:

$$f(\boldsymbol{x}) = (\boldsymbol{w} \cdot \boldsymbol{x}) + b = \sum_{i=1}^{L} (\alpha_i - \alpha_i^*)(\boldsymbol{x} \cdot \boldsymbol{x}_i) + b \quad (1)$$

其中 L 为支持向量的个数,α_i,α_i^* 和 b 为确定最优超平面的参数,通过解最优化问题求得。可以看出:最优回归超平面的解析式只由支持向量完全确定。

由于特征空间是样本空间通过映射 φ 得到的,式(1)中的点 $\boldsymbol{x}$ 和 $\boldsymbol{x}_i$ 实际上是 $\varphi(\boldsymbol{x})$ 和 $\varphi(\boldsymbol{x}_i)$。这样式(1)变成:

$$f(\boldsymbol{x}) = (\boldsymbol{w} \cdot \varphi(\boldsymbol{x})) + b = \sum_{i=1}^{L} (\alpha_i - \alpha_i^*)(\varphi(\boldsymbol{x}) \cdot \varphi(\boldsymbol{x}_i)) + b \quad (2)$$

式(2)中出现的点积依据Mercer定理是定义了一个核函数 $K(\boldsymbol{x},\boldsymbol{x}_i)$:

$$K(\boldsymbol{x},\boldsymbol{x}_i) = (\varphi(\boldsymbol{x}) \cdot \varphi(\boldsymbol{x}_i)) \quad (3)$$

将式(3)代入式(2)可得:

$$f(\boldsymbol{x}) = (\boldsymbol{w} \cdot \varphi(\boldsymbol{x})) + b = \sum_{i=1}^{L} (\alpha_i - \alpha_i^*) K(\boldsymbol{x},\boldsymbol{x}_i) + b \quad (4)$$

这就是SVM方法最终确定的非线性回归函数。特别吸引人的地方是:由于应用了核函数的展开定理,所以在实际求解过程中根本不需要知道非线性映射 φ 的显式表达式,这大大简化了计算。特别是对于高维数据的情况,核函数与向量的维数无关,可以避免通常所说的"维数灾"。

2 建立支持向量机(SVM)预报模型

2.1 构造预报因子

由于SVM是通过支持向量构造推理模型,对因子的数量没有明显的限制,支持的因子数可以上千个,因此,通过对与预报对象有明确意义的各种因子的选取,可以较好的表述预报对象与预报因子之间变化的时间、空间概念。我们的试验[2]表明,样本越多,建立的SVM模型预报效果越好。考虑现有的资料,我们采用1990~2000年1~12月共11年的ECMWF 500hPa高度、850hPa温度、地面气压的0小时输出产品来挑选因子,构造建模样本资料。

针对影响德阳本地的天气系统和要素特征,以及我们经常关注的天气系统出现的区域,在不同的层次选取不同的区域来构造因子,因子的主要构成方式为:所选区域的值、

24小时变化量、关键区域之间的差值等，组成一个与空间和时间均有关联的因子群，通过这些因子描述一个相对完善的样本空间，在这种样本空间中，我们所关注的预报对象就会有各自的表现。如图2所示，我们在500hPa图上确定的区域有：反映高原上空系统变化的区域(1～3)、反映四川盆地上空系统变化的区域(4、5)、西北冷槽变化区域(6)、蒙古低压变化区(7)、反映副高强弱变化和台风出没的区域(8、9)。这些区域的要素变化与本地天气的变化有密切的联系。

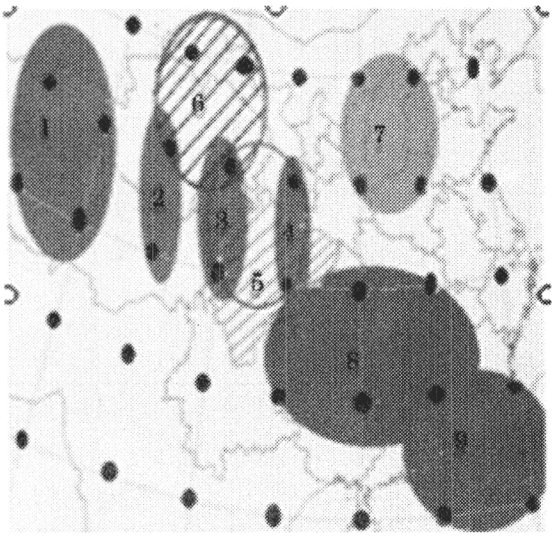

图2　500hPa预报因子选取图示

2.2　确立预报对象

预报对象为德阳市5个县站的日平均气温、日最高气温、日最低气温。

2.3　确定核函数

由于构造支持向量机的基础是Mercer定理，作为建立支持向量机的核函数必须以满足Mercer定理的条件为前提，故我们仍以径向基函数(满足Mercer定理条件)做为基本函数建立SVM回归模型。径向基函数形为：

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

最终回归函数形为：

$$f(\boldsymbol{x})=\sum_{i=1}^{L}(\alpha_i-\alpha_i^*)K(\boldsymbol{x},\boldsymbol{x}_i)+b=\sum_{i=1}^{L}(\alpha_i-\alpha_i^*)\exp(-r\|\boldsymbol{x}-\boldsymbol{x}_i\|^2)+b \quad (5)$$

其中L为支持向量数，$\boldsymbol{x}_i$为作为支持向量的样本因子向量；$\boldsymbol{x}$为待预报因子向量；α_i，α_i^*，b为建立SVM模型待确定的系数，r为核参数。

2.4　建立预报模型

我们采用中国气象局培训中心SVM应用开发研究小组开发的CMSVM应用软件，依据PP法(预报因子和预报对象是同时刻关系)来建立SVM回归预报模型。建模时尽量对样本中的因子进行归一化处理(减少各个因子之间的量级差异)。

建模使用的数据格式如下：

```
10.0  1:0.46296  2:0.40000
      3:0.36364  4:0.32231
      5:0.39024  6:0.09832
      7:0.09533…
12.1  1:0.50000  2:0.43636
      3:0.36364  4:0.30579
      5:0.39024  6:0.11031
      7:0.09094…
11.8  1:0.46296  2:0.41818
      3:0.38182  4:0.31405
      5:0.39024  6:0.09592
      7:0.08487…
8.9   1:0.40741  2:0.41818
      3:0.34545  4:0.28926
      5:0.38211  6:0.11031
      7:0.07388…
```

第1列为预报对象的值，后面依次为因子序号及因子值。

通过训练建立的SVM预报模型格式如下：

```
svmC Version V1.00
2 # 核函数类型 -t
-1 # 最优模型中核函数参数 -d
0.1 # 最优模型中核函数参数 -g
1 # 最优模型中核函数参数 -s
1 # 最优模型中核函数参数 -r
-1 # 最优模型中核函数参数 -u
77 # 训练样本的特征空间的最高维数
产生最优模型时的参数 -w 1
2739 # 支持向量的个数
-10.130993 # threshold b，(以下每行
代表一个支持向量，第一个数代表(αi -
```

α_i^*))。

-8.0261747029616828 1:0.3021 2:0.2790 3:0.2650 4:0.26582…

-80 1:0.2582 2:0.2441 3:0.2650 4:0.26582…

69.751871563477835 1:0.20879 2:0.19767 3:0.21686 4:0.24051 5:0.2284

……

……

前面文字部分为建立 SVM 模型时对应的参数及其说明，后面数字部分为构成 SVM 模型的支持向量(这里给出的模型有 2739 个支持向量)。在实时使用时，就是将支持向量和对应的参数及实时样本代入式(5)计算出实际预报值。从这里可以看出，此处的预报结果是对支持向量进行"加权"获得，而不是象常规统计方法那样对因子进行加权。当预报因子与预报对象间蕴涵的复杂非线性关系尚不清楚时，基于关键样本(支持向量)的方法可能优于基于因子的加权。

3 预报结果检验

我们在德阳市气象台针对 5 个县站的日平均气温、日最高气温、日最低气温进行了 SVM 回归方法的实际运用。因我们是用 ECMWF 的 0 小时资料以 PP 方式建立的预报模型，故预报时直接将预报时次的因子值(未将数值预报值进行订正)代入预报模型进行计算。

图 3 直观反映出德阳站的逐日气温变化过程，各时次的预报趋势与实况演变一致，当然，随着预报时效的延长，误差也在增大。

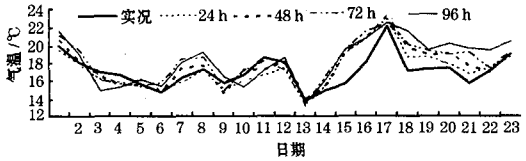

图 3 2003 年 4 月 1～23 日德阳逐日平均气温演变

图 4 显示出德阳站 4～6 月的逐日最高、最低气温的实况值、SVM 回归预报值(数值预报 48 小时)、值班预报员每日下午实际对外发布的未来 24 小时预报值。检验结果(表 1)表明，SVM 回归预报明显优于值班预报员的预报。

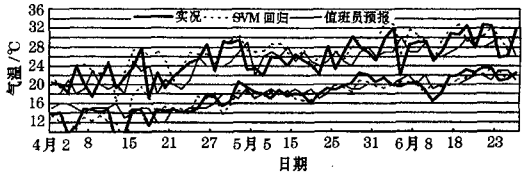

图 4 2003 年 4～6 月德阳逐日最高、最低气温演变

表 1 2003 年 4～6 月德阳逐日最高气温、最低气温预测检验结果

	最高气温			最低气温		
	相关系数	均方根误差/℃	绝对误差/℃	相关系数	均方根误差/℃	绝对误差/℃
SVM 回归	0.770	3.27	2.580	0.880	1.770	1.340
值班预报	0.599	3.509	2.944	0.794	2.272	1.713

我们还对 2003 年 1～9 月的温度预报进行了分季检验，检验结果反映出：

(1) 相关系数反映出各站预报趋势与实况演变一致。1～9 月，平均气温、最低气温与实况的相关系数直至 120 小时基本都保持在 0.8 以上，有些站超过 0.9。最高气温的预报较平均气温和最低气温的预报略差，但 96 小时以前的预报，相关系数大部分都大于 0.75。

(2) 误差相对稳定。平均气温绝对误差：1～3 月 0～120 小时保持在 2℃以内，4～9 月 0～96 小时的预报误差均在 1.5℃以内；最高气温绝对误差：1～9 月各站 0～96 小时预报误差都保持在 2.5℃以内；最低气温绝对误差：1～9 月各时次的预报误差基本保持在 2℃以内。

4 结 语

SVM 回归方法是依据支持向量(关键样本)来建立最终的决策函数，这一特征与基于确定因子的权重系数来明确表达各个因子的权重组合与预报对象变化的常规统计方法(如逐步回归、卡尔曼滤波、神经网络)有显著的区别。SVM 方法考究的是因子群构造的样本空间与预报对象的关系，单个因子与预报对象是否具有显著相关并不重要(适合于解决本质上非线性的问题)，重要的是如何选择各种与预报对象有密切联系的因子，构造

(下转第 68 页)

(上接第 44 页)

一个适合预报特征的样本空间。我们的初步运用表明,支持向量机回归估计方法能够用于定量化的气象要素客观预报。当然,要广泛的运用,仍有待进行多方面的尝试,特别是在数值预报产品的解释应用方面。

参考文献

1 陈永义,余小鼎,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.应用气象学报,2004,15(2):345~354.

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.应用气象学报,2004,15(2):355~365.

Application of Support Vector Machine Regression Method in Weather Forecast

Feng Hanzhong
(Chengdu Meteorological Center, 610072)

Chen Yongyi
(Training Center, CMA)

Abstract

The support vector machine (SVM) regression principle and its application to weather forecast are introduced. By using ECMWF analysis fields of 500hPa height, 850hPa temperature, and sea level pressure from January to September through 1990—2000, the SVM regression models are built on daily average temperature, maximum temperature, minimum temperature of five typical stations in Deyang. The performances of these models are evaluated.

Key Words: support vector machine regression temperture forecast