第32卷 第21期 计 算 机 工 程 2006年11月
Vol.32 No.21 Computer Engineering November 2006
·博士论文· 文章编号：1000—3428(2006)21—0031—02 文献标识码：A 中图分类号：TP391

一类基于SVM/RBF的气象模型预测系统

罗泽举[1]，宋丽红[2]，薛宇峰[2]，朱思铭[1]
（1. 中山大学数学与计算科学学院，广州 510275；2. 湛江海洋大学海滨校区，湛江 524005）

摘 要：利用广东湛江地区近30年月平均气温的气象数据作为数据集，建立了一种新的基于径向基核函数的支持向量机模型预测系统。通过适当选择模型参数，其平均绝对百分比误差只有5.61%，在绝对误差温度小于等于2℃的条件下，预测的准确率达到了95%，显示出所建立的支持向量机模型预测系统的有效性。通过分析发现了湛江海岸地区气候和全球气候变暖一致的事实。

关键词：径向基核函数；支持向量机；模型预测系统

One Kind of Weather Model Forecast System Based on SVM/RBF

LUO Zeju[1], SONG Lihong[2], XUE Yufeng[2], ZHU Siming[1]
(1. School of Mathematics and Computer Science, Sun Yat-Sen University, Guangzhou 510275;
2. Seashore Campus of Zhanjiang Ocean University, Zhanjiang 524005)

【Abstract】 Using the weather data of average temperature of nearly thirty years recorded at Guangdong Zhanjiang seashore region as the data set, this paper sets up a new support vector machines models forecast system based on radial basis kernel function(RBF). By choosing the model parameter properly, the mean absolute percentage error is only about 5.61%, if the absolute error is less than 2℃, the rate of accuracy comes to 95%. This shows the validity of the SVMS forecast system. In addation, by analysing it finds that the climate gets warm gradually in Zhanjiang and this is consistent with the global climate.

【Key words】 Radial basis kernel function; Support vector machines; Model forecast system

湛江位于我国大陆最南端、广东省西南部，东濒南部海域，西临北部湾，背靠大西南。市区位于雷州半岛东北部，东经110°24’,北纬21°12’,地处北回归线以南的低纬度地区，属于亚热带海洋性季风气候，冬无严寒，夏无酷暑，偶有台风天气，四季宜人。近年来，全球气候不断变暖，湛江地区的气候不但发生着类似的变化，而且又显示出湛江所独有的地方性特点。如何准确地分析和预测各种气象因素，避免恶劣天气的影响，对湛江地区经济发展具有重要意义。

本文利用湛江市1951年~2000年的月平均气温资料，分析了近50年来湛江地区温度的变化特点，并利用支持向量机回归模型，建立了温度变化预报模型，对温度进行了年际以上的预测，并得到良好的预测效果。

1 支持向量机回归模型

1.1 回归支持向量机

利用支持向量机回归学习的函数是：

$$f(x)=w\cdot x+b \tag{1}$$

通过lagrange乘子办法[1]，估计出回归参数为

$$\bar{\omega}=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)x_i\ ;\ \bar{b}=-<\bar{\omega},(x_r+x_s)>/2 \tag{2}$$

式中x_r，x_s分别为两类支持向量。

若采用内积回旋的SVM机，则相应的回归函数变形为

$$f(x,v,\beta)=\sum_{i=1}^{l}\beta_i K(x,v_i)+b \tag{3}$$

其中β_i，$i=1,\ldots,l$是标量，v_i，$i=1,\ldots,l$是向量，$K(u,v)$为满足Mercer条件的函数[2,3]（通常称为核函数），本文选取的核函数为径向基核函数。形式为

$$K(x,x_i)=\exp\{-\gamma\|x-x_i\|^2\} \tag{4}$$

这是一个非常重要的参数项，只有依赖于样本数据结构的核函数，才能达到最佳的预测效果。

1.2 模型中几个重要参数的分析

(1)核函数K（u,v）。根据Mercer定理，只要是满足相关条件的实对称函数都可作为核函数。由于气象数据点呈非线性相关，因此可以选取多种函数作为核函数。通过实验，发现径向基核函数作为核函数效果最佳。最佳效果体现了核函数和样本数据本身的固有规律。

(2)损失函数和敏感度ε。选取二次损失函数的计算时间较少，但是这并不表示用二次损失函数计算精确度高，相反，选取ε不敏感函数和合适的ε值可以达到最理想的预测值，因为二次损失函数是不可调整精度的。从这点来讲，ε不敏感函数比传统的二次损失函数优越。

(3)控制上界C。影响支持向量数目和计算时间。C值越小支持向量数目和计算时间越少，C值不能太小，否则误差会增大。通过训练，得到以下模型参数：

表1 模型参数的选取

参数类别	选取
核函数K（u,v）	径向基函数
损失函数L	ε不敏感函数
控制上界C	C=10
敏感度ε	0.81

基金项目：国家自然科学基金资助项目（10371135）
作者简介：罗泽举(1965－)，男，博士生，主研方向：机器学习与模式识别，生物信息学；宋丽红，实验师；薛宇峰，讲师；朱思铭，教授、博导
收稿日期：2005-11-28 **E-mail：**luozeju@126.com

2 温度序列数据分析

图 1 是湛江地区近 50 年以来历年的平均气温、5 年滑动平均以及年平均温度的距平曲线分布图。由图可见，地处亚热带区域的湛江地区近 50 年来年平均气温基本比较稳定，50 年平均温度为 23.22℃；除 1973 年外，1967 年~1985 年的近 20 年间，温度距平都低于 0℃，属于一段气候偏冷期（1976 年达到最低温度 22.45℃）。但 1986 年以后，基本上都成为正距平，平均温度呈准线性上升。1998 年到达最高 24.48℃，比多年平均值高出 2.03℃，成为异常偏暖年份。

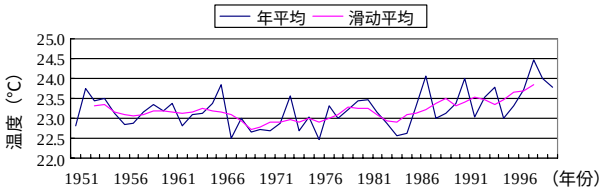

图 1 湛江地区 1951 年~2000 年历年平均气温、滑动平均曲线

分析 1951 年~2000 年多年月平均气温可知，湛江地区各月平均温度变化幅度不是很大，因此全年温差不大，最冷月为 1 月（平均温度为 15.8℃），最热月份为 7 月（平均温度达到 28.9℃）。分析 1 月和 7 月平均气温距平，7 月平均温度变化比较平缓，而 1 月气温变化的幅度就要大一些，而且自 90 年代以来，1 月份气温的正距平持续偏大，说明冬季温度的变率比夏季气温的变率要大，而且近 10 来年冬季的增温也较夏季明显。通过计算 1951 年~1985 年间的 35 年和 1986 年~2000 年间的 15 年冬季（以 12 月、1 月、2 月平均温度作代表）和夏季（以 6 月、7 月、8 月平均温度为代表）的季节平均温度，可以看到，湛江地区冬季增温(0.8℃)较夏季（0.5℃）更明显。为此，我们分别分析了 1951 年~1985 年和 1986 年~2000 年两段时期内的月平均温度（见表 2）。可见，月平均温度后 15 年比前 35 年有所升高，平均升高约 0.65℃，这和近年来全球气候变暖的事实相一致。

表 2 冬、夏季节平均温度反映的气温变化比较

季节	冬季	夏季
1951~1985	16.3℃	28.4℃
1986~2000	17.1℃	28.9℃

图 2 为 1951 年~1985 和 1986 年~2000 两个时间段内月平均气温对比图，红线代表 1951 年~1985 年月平均温度，蓝线代表 1986 年-2000 年各月平均气温，从中可以看出 80 年代中期以后气候变暖。

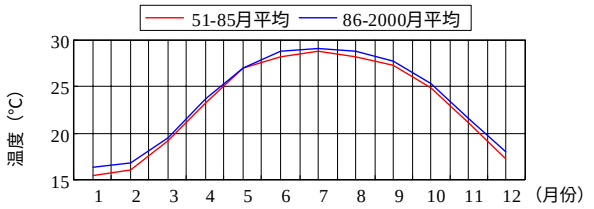

图 2 月平均气温对比

另外，分析年际温度的变化发现，20 世纪 80 年代之前，湛江地区的平均温度变化幅度不大，甚至 50~70 年代的 10 年平均温度还略有下降。但 80 年代之后显著增温，80 年代比 70 年代的气温增加了 0.13℃/10 年，90 年代又比 80 年代增加了 0.51℃/10 年，远远高于相应时段内全国平均温度增长水平[4]，这是地处亚热带海洋性季风气候的湛江地区平均温度的显著特点。

3 决策函数的确定

温度是一个时间序列 $x(t),\ t=1,2,\ldots$，其 t 时刻的值 x(t) 可以表示为函数：

$$x(t)=F(x(t-1),x(t-2),\ldots,x(t-n)) \tag{5}$$

这里关键是要确定窗口参数 n 的值，以 30 年的训练数据为基础，假设当前值和前面 3~7 个时延相关(虽然理论上时延还可以有更宽的选择)，限定窗口范围在 3~7，计算得到平均绝对误差(MAE)如表 3。

表 3 模型输入窗口参数的确定

参数 n	3	4	5	6	7
MAE	1.7	1.89	5.69	4.30	3.21

其中最小的平均绝对误差是 1.7，因此选取时延窗口（相空间维数）大小 n=3，即确定输入模型为

$$x(t)=F(x(t-1),x(t-2),x(t-3)) \tag{6}$$

4 预测结果分析

我们以时延为 3 进行预测，也就是用过去的 3 个数据预测将来的值，其平均绝对误差（MAE）、平均绝对百分比误差（MAPE）、根方差（RMSE）及相关系数（r）如表 4。

表 4 4 个重要的预测指数

MAE	MAPE	RMSE	r
1.2163	5.61%	1.535 8	0.98

可见其平均相对误差只有 5.61%，根方差也相当小，只有 1.535 8，而 1991 年~2000 年的 10 年间各年数据呈显著相关性，达 0.98，说明所建立的预测模型是相当有效的。

图 3 是模拟曲线图和预测曲线。

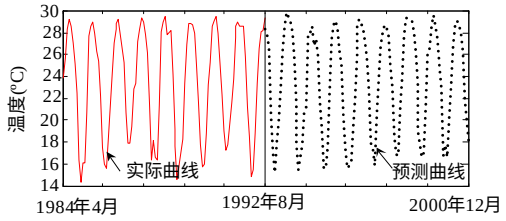

图 3 1984 年 4 月~2000 年 12 月各月模拟曲线和预测曲线

由图可见，1984 年 4 月~1992 年 8 月各月实际和预测部分模拟相当吻合。以 1992 年 9 月~2000 年 12 月作为预测，1984 年 4 月~2000 年 12 月各月误差曲线如图 4 所示。

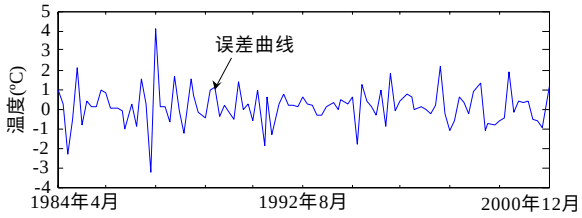

图 4 1984 年 4 月~2000 年 12 月各月误差曲线

我们用 3 种核函数进行模拟预测，发现当选取径向基核函数时，其中误差大于 2℃只占 5%，最大误差为 4℃，但只占 3%。其它两种核函数的准确性方面都不如径向基核函数（见表 5）。

表 5 支持向量机模型：1992 年 9 月~2000 年 12 月各月预测准确率

误差温度限	预测月数（个）	预测准确率		
		RBF 核	线性核	样条核
≤1℃	100	78%	70%	68%
≤2℃	100	95%	90%	91%
≤3℃	100	97%	92%	92%

当我们试用神经网络进行相同的测试时，选用径向基神经网络、线性神经网络、BP 神经网络，其总体预测水平均不如支持向量机模型（表 6）。这表明支持向量机处理小样本的能力比传统神经网络预测能力要强。

（下转第 93 页）