支持向量机与卡尔曼滤波集合的西太平洋副热带高压数值预报误差修正*

刘科峰[1,2] 张 韧[1,2] 徐海斌[1] 闵锦忠[2] 朱伟军[2]

1 解放军理工大学气象学院海洋与空间环境系，南京，211101
2 南京信息工程大学江苏省气象灾害重点实验室，南京，210044

摘 要

基于 T106 数值预报产品资料，提出了支持向量机和卡尔曼滤波相结合的方法来进行夏季西太平洋副热带高压数值预报的误差修正与预报优化。首先采用支持向量机方法建立了西太平洋副热带高压面积指数的误差修正模型。基于支持向量机预报优化模型尽管有比较好的拟合精度和预报效果，但与实际副热带高压指数尚有一定的差异。究其原因，除预报对象（副热带高压）本身比较复杂、模型优化因子不够充分以及数值预报误差自身的随机性以外，优化模型的输入、输出基本上是一个静态映射结构，因此前一时刻的预测误差难以得到有效的反馈、调整和修正。为考虑前一时刻预报误差的反馈信息，动态跟踪副高的变化趋势，随后引入卡尔曼滤波方法建立支持向量机-卡尔曼滤波模型，对支持向量机模型的输出结果作进一步的调整和优化。试验结果表明，该方法模型的预报优化效果优于 T106 数值预报产品以及单纯的神经网络修正模型和卡尔曼滤波修正模型的优化效果，能够较为客观、有效地修正西太平洋副热带高压指数的数值预报误差，改进和优化西太平洋副热带高压的数值预报效果。该方法为副热带高压等复杂天气系统和要素场预报提供了一种新的思路，表现出较好的应用前景。

关键词：T106 数值预报，副热带高压，支持向量机，卡尔曼滤波。

1 引 言

西太平洋副热带高压（简称“副高”）是影响夏季东亚天气气候的重要系统，其强度变化和进退活动异常常导致江淮流域出现洪涝和干旱灾害；副高作为东亚夏季风系统重要成员，是连接热带环流和中高纬环流的重要纽带，直接影响制约热带和中高纬大气环流的演变。

研究副高的主要目的，也是最大难点是准确预报副高活动。由于副高不仅有规则的渐变，更有异常的突变，表现出明显的非周期性和非确定性，使副高预报非常复杂和困难。数值预报是当前副高预报的重要手段之一，但由于副高活动机理和规律尚未彻底弄清，以及热带、副热带地区地转关系弱等原因，在较大程度上制约了副高的数值预报准确率，目前许多数值预报产品的副高预报均存在着不同程度的偏差。因此开展副高数值预报的误差修正和预报优化研究，具有重要的科学意义和实用价值。

近年来人工神经网络等非线性统计学方法在副高预报与预报优化中取得了一定的成效[1-3]，但是神经网络方法也存在一些内在的缺陷，如隐层单元神经元的数目难以客观确定、训练过程容易陷入局部最优、神经网络的结构设计依赖于设计者的先验知识和经验等。另外，从概率统计的角度说，神经网络学习算法采用的经验风险最小化原理（ERM），仅仅试图使经验风险最小化，并没有使期望风险最小化，与传统的最小二乘法相比，在原理上缺乏实质性的突破，进而使神经网络的学习算法缺乏定量的分析与完备的机理[4],①。1995 年，由贝尔实验室的 Vapnik 等在统计学习理论的基础上提出了一种模式识

* 初稿时间：2006 年 4 月 18 日；修改稿时间：2006 年 10 月 10 日。
资助课题：国家自然科学基金项目（40375019）与江苏省气象灾害重点实验室开放课题（KLME0507）。
作者简介：刘科峰，男，从事海洋气象学研究。Email：fengke_liu@126.com

① Smola A J. Regression estimation with support vector learning machines. Technische University at Myunchen, 1996

别的新方法——支持向量机(Support Vector Machine),它根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折中,使结构风险最小,即同时最小化经验风险与 VC 维(Vapnik-Chervonenkis Dimension)的界,以期获得最好的泛化能力。支持向量机在形式上类似多层前向神经网络,但支持向量机方法能够克服多层前向网络的固有缺陷,被认为是人工神经网络的替代方法。常被用于模式识别和非线性回归。

卡尔曼滤波是一种动态系统的优化分析方法,它以系统状态空间模型为分析对象,根据受噪声干扰的系统模型和包含噪声干扰的系统观测量,运用现代随机估计理论给出系统状态的无偏最小方差的递推估计值。它无需太多的历史资料就可建立能适应数值模式变化的统计模型,其主要特征是通过对误差与实验数据间的处理来不断订正模型参数,组建出最优滤波方程。近年来,卡尔曼滤波方法被广泛应用于预报模型的优化,并在温度和风[5-6]等连续变化的要素预报中取得成功。

由于副高系统的非线性和复杂性,我们基于副高数值预报误差修正的途径,提出支持向量机和卡尔曼滤波相结合的副高预报优化思想。首先用支持向量机方法对 1995—1997 年夏季月份 T106 数值预报产品计算得到的副高面积指数建立误差修正模型,再用支持向量机模型的订正结果作为卡尔曼滤波模型的输入变量,进而建立卡尔曼滤波的二次修正的副高预报优化模型。

2　统计学习理论和支持向量机

2.1　统计学习理论

机器学习的目的是根据给定的训练样本建立输入数据与输出数据之间的对应关系。在传统学习方法中,问题被转化为用算法实现:

$$R_{\text{emp}}(w)=\frac{1}{n}\sum_{i=1}^{n}L[y_i,f(x_i,w)] \tag{1}$$

最小化,其中包含

$$L[y,f(x,w)]=[y-f(x,w)]^2 \tag{2}$$

式中,$\{f(x,w)\}$称作预测函数集,w 为函数的广义参数,$(x_1,y_1),(x_2,y_2)\cdots(x_n,y_n)$为 n 个独立同分布观测样本。这就是所谓的经验风险最小化(Empirical Risk Minimization 简称 ERM)准则,采用 ERM 准则取代期望风险最小化是传统的机器学习泛化能力差的根本原因。统计学习理论提出了一系列概念理论和方法解决传统机器学习的这一问题[7]。

VC 维反映了函数集的学习能力,VC 维越大则学习机器越复杂,统计学习理论还给出了对于各种类型的函数集,经验风险 $R_{\text{emp}}(w)$和实际风险 $R(w)$之间的关系,即推广性的界,可以表示为

$$R(w)\leqslant R_{\text{emp}}(w)+\phi(h/n) \tag{3}$$

式中 h 为函数集的 VC 维;n 为样本数。

式(3)表示学习机器的实际风险由两部分组成:一是经验风险 $R_{\text{emp}}(w)$,另一部分为置信范围,它与 VC 维及训练样本数有关,表明在有限的训练样本下,学习机器的 VC 维越高则置信范围越大,导致真实风险与经验风险之间可能的差别越大。机器学习过程不但要使经验风险最小,还要使 VC 维尽量小以缩小置信范围,才能取得较小实际风险,即对未来的样本有较好的推广性。统计学习理论的提出把函数集构造成一个函数子集序列,使各子集按照 VC 维的大小排列;在每个子集中寻找最小经验风险,在子集间折中考虑经验风险和置信范围,取得实际风险最小,如图 1 所示,这种思想称作结构风险最小化(Structural Risk Minimization)。支持向量机就是 SRM 准则的具体体现。

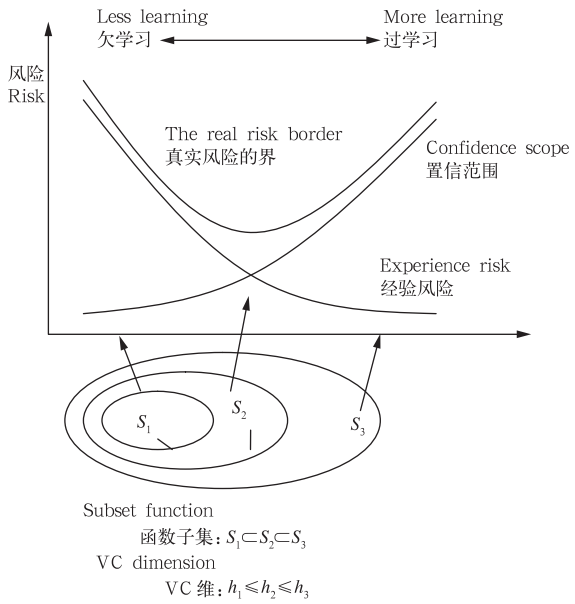

图 1　结构风险最小化示意图

Fig. 1　Structural risk minimization

2.2　支持向量机

SVM 方法用于函数估计,它要解决的问题是:根据给定的样本数据集$\{(x_i,y_i)\}$,$i=1,2\cdots l$,其中 x_i 为输入因子值,y_i 为输出值,寻求一个反映样本数据输出输入的最优函数关系 $y=f(x)$。

这里的"最优"是指按某一确定的误差函数来计

算，所得函数关系对样本数据集拟合得“最好”（累计误差最小）。通常取平方函数、绝对值函数或 Huber 函数为误差函数。SVM 回归中采用的是 ε-intensive 损失函数[②]，形式如下

$$e(f(x)-y)=\begin{cases}0 & |f(x)-y|<\varepsilon\\ |f(x)-y|-\varepsilon & |f(x)-y|\geqslant\varepsilon\end{cases} \tag{4}$$

当误差小于 ε 时，误差忽略不计；当误差超过 ε 时，误差函数的值为实际误差减去 ε。

SVM 的回归函数为 $y=f(x)=w\phi(x)+b$ 其中 $\phi(x)$ 是输入空间 R^d 高维特征空间 H 的非线性映射，SVM 就是将实际问题通过非线性映射转换到高维特征空间，在高维特征空间中构造线性回归函数来实现原空间中的非线性回归函数。引入松弛因子 $\zeta_i\geqslant 0$ 和 $\zeta_i^*\geqslant 0$，根据 SRM 准则，可将问题转化为在满足式（5）条件下使泛函 $\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}(\zeta_i+\zeta_i^*)$ 最小，$C>0$。式中 $\frac{1}{2}\|w\|^2$ 为正则化参数，体现了 SVM 对推广能力的控制。C 为平衡系数，用来平衡经验风险和正则化部分。

$$\begin{cases}y_i-wx_i-b\leqslant\varepsilon+\zeta_i\\ wx_i+b-y_i\leqslant\varepsilon+\zeta_i^*\end{cases} \tag{5}$$

这是一个典型的不等式约束下二次寻优的问题，存在唯一解。

引入 Lagrange 乘子 α_i,α_i^* 及核函数，可将其转化为在约束条件

$$\sum_{i=1}^{n}(\alpha-\alpha^*)\qquad i=1,2,3\cdots n$$
$$0\leqslant\alpha_i-\alpha_i^*\leqslant C \tag{6}$$

下，对 α_i,α_i^* 最大化下面的目标函数

$$W(\alpha,\alpha^*)=-\varepsilon\sum_{i=1}^{n}(\alpha_i^*+\alpha_i)+\sum_{i=1}^{n}y_i(\alpha_i^*-\alpha_i)-\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^*-\alpha)(\alpha_j^*-\alpha_j)K(x_i\cdot x_j) \tag{7}$$

最后得回归函数为

$$f(x)=(w\cdot x)+b=\sum_{i=1}^{n}(\alpha_i^*-\alpha_i)K(x_i\cdot x)+b^* \tag{8}$$

这里 α_i,α_i^* 为指定样本的 Lagrange 乘子，只有一小部分不为 0，对应的样本就是支持向量，$K(x_i,x_j)=\langle\varphi(x_i),\varphi(x_j)\rangle$ 为核函数。最常用的核函数有以下几种

多项式核函数：

$$K(x_i,x_j)=(\sigma(x_i\cdot x_j)+r)^d,\sigma>0$$

RBF 核函数：

$$K(x_i,x_j)=\exp(-\sigma\|x_i-x_j\|^2),\sigma>0$$

Sigmod 核函数：

$$K(x_i,x_j)=\tanh(\sigma(x_i\cdot x_j)+r)$$

3 副高的支持向量机优化模型

3.1 研究资料

基于实际可用的资料，同时也为了便于预报效果比较，选择 T106 数值预报产品为研究资料。选取 1995—1997 年夏季（5 月 1 日—8 月 31 日）共计 360 d 的 T106 数值模式的 500 hPa 位势高度初值场（近似代表实际位势场）和位势高度、温度、湿度等要素的 3 d 预报场序列（分析范围 10°—60°N；70°—146°E）。优化对象为基于 T106 数值预报产品计算的逐日副高面积指数，优化目标为实际 500 hPa 位势场（用 T106 初始场近似代替）计算所得逐日副高面积指数。根据相关分析计算，取表 1 所述的预报优化目标和 5 个误差修正因子构建副高数值预报误差修正与预报优化模型。

表 1 T106 数值预报修正目标和预报初始修正因子

Table 1 The forecast object and initial correction factors of T106 model

误差修正与预报优化目标	修正因子 1	修正因子 2	修正因子 3	修正因子 4	修正因子 5
当前初始场的副高面积指数	3 天前初始场副高面积指数	副高面积指数的 3 d 预报结果	500 hPa 位势场 3 d 预报的 22°—34° N，120°—140°E 格点平均值	500 hPa 温度场 3 d 预报的 24°—42° N，110°—140°E 格点平均值	500 hPa 湿度场 3 d 预报的 12°—26°N，70°—85°E 格点平均值
优化目标与各修正因子间的相关系数	0.6603	0.8233	0.6457	0.6819	0.6766

② Steve R G. Support vector machines for classification and regression. Science and mathematics school of electronics and computer science technical report, 1998

为便于模型的建立和预报优化结果的比较,将数据资料分为两部分:第一部分用于模型的建立和测试,所取数据为 1995 年 5 月 1 日—8 月 31 日共 120 d;第二部分用于模型的检验和预报优化效果的评估,所取数据为 1996 年 5 月 1 日—8 月 31 日和 1997 年 5 月 1 日—8 月 31 日共 240 d。

3.2　支持向量机模型

支持向量机可以看作是一个 3 层前向神经网络,其输出是若干中间层节点的线性组合,而每一个中间层节点对应于输入样本与一个支持向量的内积。网络结构为

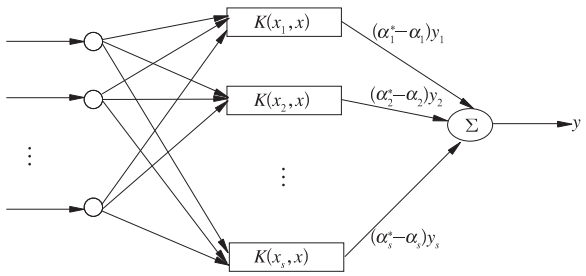

图 2　支持向量机网络结构示意图

Fig. 2　The sketch map of the network structure of support vectors

其中,$K(x_s,x)$为核函数,x_s 为支持向量,$\alpha_s^* - \alpha_s$ 为网络权重(Lagrange 乘子),$x_1, x_2 \cdots x_n$ 为输入变量,y 为网络输出,其隐节点个数即为支持向量机的个数。每个基函数中心对应一个支持向量,它们以及输出权值都是由算法自动确定的。建立模型的过程就是对输入的训练样本根据模型的期望输出调节模型参数确定核函数和支持向量的过程。

由于支持向量机是通过内积函数定义的非线性变换将输入空间变换到一个高维空间,在这个高维空间中求最优回归函数。这样,核函数就反映了高维特征空间中任意两个样本点之间的位置关系,因而对样本点的拟合具有重要意义,核函数选取的好坏直接影响到 SVM 模型性能的优劣。但如何选择合适的核函数,目前还没有一个对特定问题选用最佳核函数的有效方法。此处,我们分别采用几种常用的核函数,例如多项式核函数、Sigmod 核函数和 RBF 核函数来构建预报优化模型,固定参数值进行预测,根据预测结果进行评定,从而选出最合适的核函数。经过大量的仿真试验,最终确定选用 RBF 核函数。所以,模型最终回归函数形式为

$$f(x) = \langle w, x \rangle + b = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \exp(-\sigma \| x_i - x_j \|^2) + b \qquad (9)$$

核函数确定后,还需确定两个相关的参数:σ、γ。其中 σ 为核参数,调节核函数的平滑程度;γ 是正则化参数,控制模型的复杂度(支持向量机的个数)和函数逼近误差的大小。这两个模型参数在很大程度上决定了该模型的学习能力及泛化能力。如何确定模型参数,目前尚缺乏一个客观有效的方法。我们采用逐步筛选的方法确定这两个模型参数:首先设置较大的参数取值范围,对参数进行大间隔步长的循环取值,通过训练和测试,根据预报结果与实际值的相关系数、平均绝对误差和相对误差的大小综合确定最优参数值,再以此参数值为中心,设置较小的参数范围,以小间隔步长重复上述步骤,直至最终确定出用于建立 SVM 预报模型的参数值,进而确定预报模型[8]。

对数据资料作标准化处理,然后取上述各初始预报优化因子作为单独输入变量,即输入矩阵 P_0 为 120×1 阶矩阵;输出目标 T 为模式初始场的副高面积指数时间序列,为 120×1 阶矩阵。比较表 1 中每个修正因子的拟合状况与预报效果,最后选取拟合状况与预报效果较好的 3 个因子作为支持向量机模型的预报优化因子(即表 1 中的修正因子 2、3 和 4)。然后将确定的训练样本逐个输入模型,采用逐步筛选法调节模型的两个参数,最终确定的参数 σ=7.5352,γ=10.725。支持向量机共 65 个。

3.3　副高时间序列的逼近和预测

将独立预报样本、训练好的支持向量及 Lagrange 乘子 α_i,α_i^* 带入式(9),即可直接对副高指数进行非线性映射逼近和预报优化。所建模型的副高指数预报优化值与同实际值之间达到了较高拟合精度(相关系数 0.9163,置信度 α=0.05)。图 3 是支持向量机模型用第二部分独立资料所做的预报优化试验和效果检测。由图中可见,预报优化结果尽管在一些细节的描述上还不够完善,但实际副高指数变化的主要趋势和升降、转折过程基本上能够正确表现,优化预报结果与实际值的相关系数达到 0.85135,平均绝对误差为 21.866,平均均方根误差为 28.545。

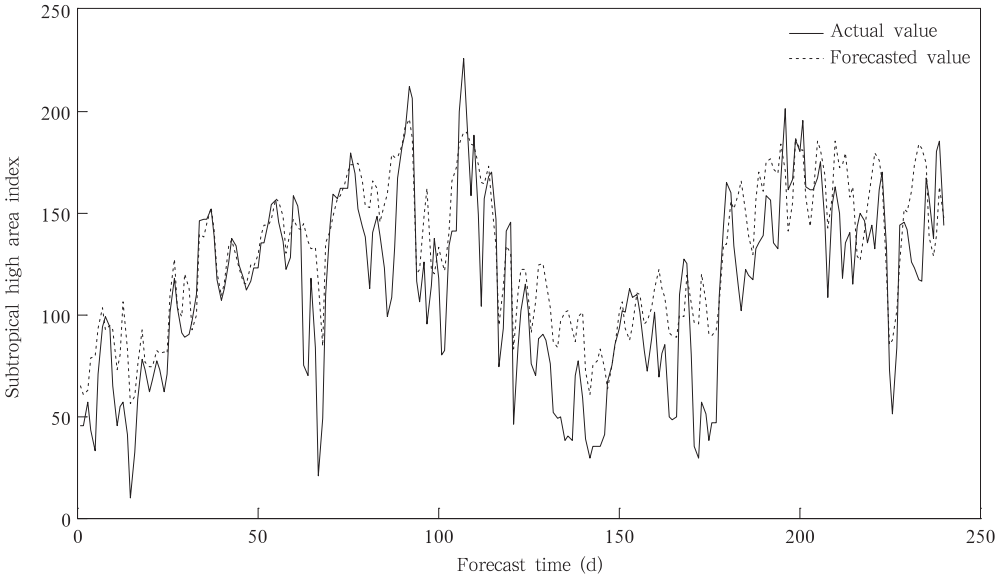

图 3 支持向量机模型的预报优化效果

(横轴:1996—1997 年夏季(5 月 1 日—8 月 31 日,共 240 d))

Fig. 3 Optimized forecast results of support vector machine model

(Forecast time: summers in 1996 and 1997 (5.1−8.31, 240 days))

4 支持向量机-卡尔曼滤波的副高预报优化模型

基于支持向量机预报优化模型尽管有比较好的拟合精度和预报效果,但与实际副高指数尚有一定的差异(如平均绝对误差 21.866,平均均方根误差 28.545)。究其原因,除预报对象(副高)本身比较复杂、模型优化因子不够充分以及数值预报误差自身的随机性以外,优化模型的输入、输出基本上是一个静态映射结构,因此前一时刻的预测误差难以得到有效的反馈、调整和修正。

为考虑前一时刻预报误差的反馈信息,动态跟踪副高的变化趋势,我们拟引入卡尔曼滤波方法对支持向量机模型的输出结果作进一步的动态跟踪调试。卡尔曼滤波是一种统计估算方法,它通过处理一系列带有误差的测量数据得到所需要的物理参数的最佳估算值。其优点在于能根据前一时刻预测误差的大小及其统计量的变化来调整预测方程的系数,这样不仅利用了样本提供的信息,同时也吸收了前一时刻预测方程的反馈信息,从而利于提高模型的输出精度,表现出较强的自适应能力。

将支持向量机模型的输出结果作为卡尔曼滤波模型待选的输入变量,选择相关系数较高(大于 0.67)的优化因子作为卡尔曼滤波模型的输入变量(选取表 1 中的修正因子 2、4、5 和支持向量机模型的输出结果),然后对输入的变量作归一化处理。

预测的公式为卡尔曼滤波递推系统中的量测方程[8]。单优化目标、4 个订正因子的量测方程为

$$\begin{aligned}\hat{x}(t+3) = {} & b_0(t)+b_1(t)\cdot x_1(t+3)+\\ & b_2(t)\cdot x_2(t)+b_3(t)\cdot x_3(t)+\\ & b_4(t)\cdot x_4(t)+e(t)\end{aligned} \tag{10}$$

作为量测方程,其各项的意义是:$\hat{x}(t+3)$为卡尔曼预测优化结果;$x_1(t+3)$为支持向量机模型的预报结果,$x_1(t+3)$、$x_2(t)$、$x_3(t)$、$x_4(t)$分别为订正因子;$b_0(t)$、$b_1(t)$、$b_2(t)$、$b_3(t)$、$b_4(t)$为随时间变化的订正系数;$e(t)$为订正误差。

卡尔曼滤波递推系统需要确定 4 个初始参数,这里采用客观方法[9]确定。由于选取了 4 个订正因子,所以递推公式中的向量是五维的,矩阵是五阶的。

(1) $b(0|0)$,样本取前 120 d 的值,用最小二乘法确定。

(2) $c(0|0)$是 $b(0|0)$误差方差阵。假定$b(0|0)$与理论相等,所以 $c(0|0)$是五阶的零方阵。

(3) w,根据动态噪音及动态系统的性质可推得

$$w=\begin{bmatrix}(\Delta b_0)^2/\Delta t & 0 & 0 & 0 & 0\\ 0 & (\Delta b_1)^2/\Delta t & 0 & 0 & 0\\ 0 & 0 & (\Delta b_2)^2/\Delta t & 0 & 0\\ 0 & 0 & 0 & (\Delta b_3)^2/\Delta t & 0\\ 0 & 0 & 0 & 0 & (\Delta b_4)^2/\Delta t\end{bmatrix}$$

$\Delta b_i=b_{11i}-b_{1i}$，$\Delta b_i(i=0,1,2)$为由两组样本分别求出的两个回归系数向量的分量差，$\Delta t$ 为两组样本时间差。用前 120 d 的预报优化因子和最小二乘法分别求出 b_1、b_{11}，Δt 取 60。$\Delta b_i=b_{11i}-b_{1i}$，$i=0,1,2,3,4$。

(4) v 是量测噪声的方差。当预报量只有一个时，v 则变成标量，为一个数值，用选取的前 120 d 的预报因子，建立副高面积指数预报订正的回归方程，求出其残差平方和 q。$q/(k-m)$就是 v 的无偏估计值，其中 k 为样本个数 120。m 为矩阵阶数。

根据上述方法最终确定的 4 个起步参数 b_0、C_0、W、V 为

$$b_0=[-32.114\quad -2.703\quad -30.311\quad 196.063\quad 56.532]^{\mathrm{T}}\qquad V=314.67$$

$$C_0=\begin{bmatrix}0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\end{bmatrix}\qquad W=\begin{bmatrix}0.771&0&0&0&0\\0&0.3969&0&0&0\\0&0&1.6124&0&0\\0&0&0&3.9409&0\\0&0&0&0&0.3838\end{bmatrix}$$

然后按卡尔曼滤波的递推公式即可得每一步预报优化结果及下一步的参数值。

图 4 是所建立的支持向量机-卡尔曼滤波模型的预报优化结果与实际副高面积指数的对比，预报优化结果(点线)对实际副高面积指数(实线)时间序列的总体趋势上能够较好地把握，大部分升降、转折过程和一些细节变化也基本表现正确，预报值和实际值的相关系数达到 0.8845，平均绝对误差为16.349，

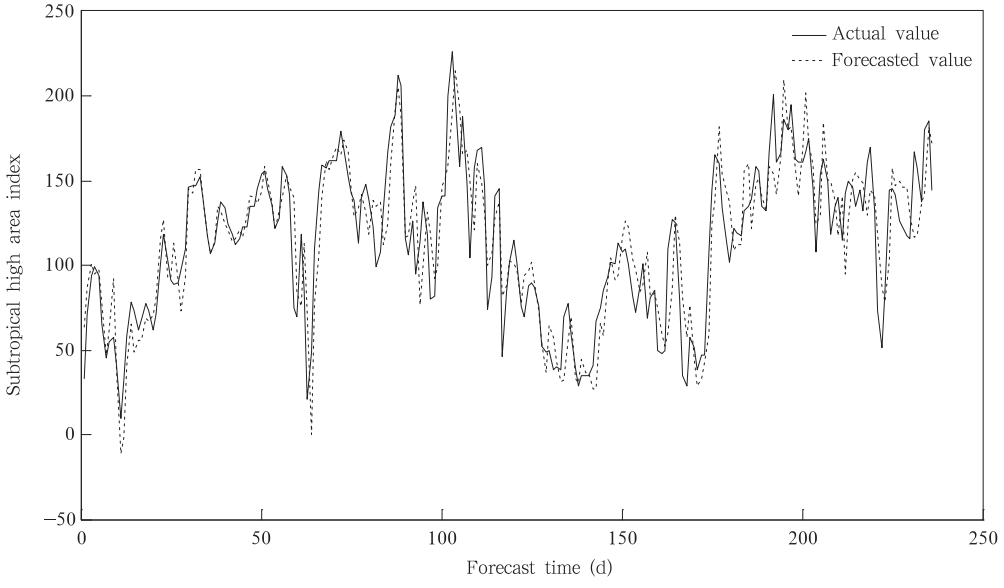

图 4　卡尔曼-支持向量机模型的预报效果

(横轴：1996—1997 年夏季(5 月 1 日－8 月 31 日，计 240 d))

Fig. 4　Forecast results from Kalman-support vector machine model

(Forecast time: summers in 1996 and 1997(5.1－8.31, 240 days))

平均均方根误差为21.499,较支持向量机模型的预报效果不但总体趋势更加逼近实际副高面积指数的变化,而且数值上更加接近实际副高面积指数。预报优化效果有明显的改进。

5 不同方法的预报优化效果比较

神经网络和卡尔曼滤波方法是气象资料分析和要素预测常用的方法,为了评估和比较支持向量机方法和卡尔曼滤波-支持向量机方法的预报优化效果和技术优势,我们将表1中初始数据资料作了归一化处理,分别建立了副高面积指数的BP神经网络和卡尔曼滤波的预报优化模型,并用以与支持向量机、卡尔曼滤波-支持向量机优化模型的输出结果进行对比。表2为各种预报与优化模型(数值预报、神经网络、卡尔曼滤波、支持向量机和支持向量机-卡尔曼滤波模型)输出结果与实际副高指数的相关系数、平均相对误差和绝对误差。其中BP神经网络和卡尔曼滤波模型的预报优化值同实际值的相关系数分别为0.8349和0.8371(图略),均低于支持向量机和支持向量机-卡尔曼滤波模型的预报优化效果(表2),且神经网络模型的收敛速度较慢、对初值敏感以及训练效率较低。上述的对比试验结果表明,本文所提出的支持向量机-卡尔曼滤波方法对副高预报优化对象的把握和描述较前面几种模型更为恰当和准确,表现出较好的优化效果和技术优势。

表2 不同预报与优化模型的副高预报优化效果比较

Table 2 Comparison of the results from different forecast models

	T106数值预报	BP神经网络	卡尔曼滤波	支持向量机	支持向量机-卡尔曼滤波
相关系数	0.8277	0.8349	0.8371	0.8514	0.8845
平均相对误差	21.1625	22.8563	20.8062	21.8664	16.3489
平均绝对误差	26.6558	27.536	26.5779	28.5453	21.4989

6 结 论

本文提出的支持向量机和卡尔曼滤波相结合的方法,能够较为客观、有效地修正副高面积指数的数值预报误差,优化和改进副高的数值预报结果。是一种较为有效、实用的副高预报方法和手段。该方法具有泛化能力强、训练速度快、稳定性好、便于建模等优点,为副高等复杂天气和要素预报提供了一种新的方法思路,表现出较好的应用前景。但是副高是一个复杂的系统,它的变化往往受制于多种因素的影响,同时支持向量机模型的参数选取与确定目前也缺乏客观定量化的标准,因此,更充分地做好预报因子的优选,采用其他的优化算法(如模拟退火、遗传算法等)来客观选取确定支持向量机模型的参数,将有助于进一步深入挖掘和开发支持向量机-卡尔曼滤波方法的预报优势和潜力,以达到更好的预报效率,这也是我们下一步拟开展的工作。

参考文献

[1] 张韧.基于前传式网络逼近的太平洋副热带高压活动的诊断预测.大气科学,2001,25(5):649-660

[2] 张韧,蒋国荣,余志豪等.利用神经网络计算方法建立太平洋副高预报模型.应用气象学报,2000,11(4):474-483

[3] Zhang Ren, Yu Zhihao. Neural network BP model approximation and prediction of complicated weather systems. Acta Meteor Sinica, 2001,15(1):105-115

[4] Vapink V, Goloeich S, Smola A. Support vector method for function approximation, regression estimation, and signal processing. Cambridge, MA, MIT Press,1997: 281-287

[5] 卢峰本.卡尔曼滤波在沿海东半年风力预报中的应用.气象,1998,24(3):50-53

[6] 单九生,袁正国,周建雄.利用卡尔曼滤波方法作1~5天中期温度预报.成都信息工程学院学报,2001,16(1):12-16

[7] 张学工.关于统计学理论与支持向量.自动化学报,2000,26(1):32-42

[8] Suykens J A K, Van Gestel T, De Brabanter J, et al. Least Squares Support Vector Machines. World Scientific, Singapore Pub Co, 2002:308pp

[9] 陆如华,徐传玉,张玲等.卡尔曼滤波的初值计算方法及其应用.应用气象学报,1997,8(1):34-43

ERROR-CORRECTING OF THE AREA INDEX OF SUBTROPICAL HIGH IN THE T106 NUMERICAL PREDICTION BASED ON SUPPORT VECTOR MACHINE-KALMAN FILTER MODEL

Liu Kefeng[1,2]　Zhang Ren[1,2]　Xu Haibin[1]　Min Jinzhong[2]　Zhu Weijun[2]

1 *Institute of Meteorology, PLA University of Sciences and Technology, Nanjing* 211101

2 *KLME, Nanjing University of Information Science & Technology, Nanjing* 210044

Abstract

Based on the T106 NWP product information, the T106 numerical forecast error of western Pacific subtropical high was corrected and optimized using the methods of support vector machine (SVM) and Kalman filter. An error-correcting model for the area index of the western Pacific subtropical high was first established with the SVM method. Despite of its good fitting accuracy and forecast results, there were still many differences between forecast results of the SVM model and actual results. There were many reasons for the differences. In addition to complex forecast object itself, inadequate model optimization factors and random numerical forecast errors, the SVM forecast model has basically a static mapping structure, therefore the anterior forecast errors were difficult to be effectively feedbacked. In consideration of the anterior forecast errors, the Kalman filter was introduced to establish the Kalman-support vector machine model to further optimize and adjust the output of the support vector model. The testing results show that the Kalman-support vector model can objectively and effectively correct the T106 numerical forecast error of western pacific subtropical high, and is superior to the T106, BP, and Kalman models in high forecast accuracy, fast training, high generalization capability and easy modeling, thus providing a new method for the forecast of the complex weather system such as subtropical high etc.

Key words: T106 numerical prediction, Western Pacific subtropical high, Support vector machine, Kalman filter.