第 57 卷 第 2 期 2008 年 2 月
1000-3290/2008/57(02)/0714-06

物 理 学 报
ACTA PHYSICA SINICA

Vol.57,No.2,February,2008

基于支持向量机方法对非平稳时间序列的预测*

王革丽[1)†] 杨培才[1)] 毛宇清[1)2)]

1)(中国科学院大气物理研究所,北京 100029)

2)(南京信息工程大学大气科学学院,南京 210044)

(2007 年 4 月 2 日收到;2007 年 5 月 28 日收到修改稿)

目前有关非平稳复杂系统及其在预测中的应用研究是一个较少被人理解并有重大科学意义的前瞻性研究课题.在大气运动中,气候正是一个典型的非平稳系统,但是现有的气候预测理论,包括统计预测理论和非线性预测理论,几乎都无一例外地建立在平稳性假定的基础之上,这有悖于气候过程的基本性质,它有可能是导致气候预测水平低下的重要的理论原因.因此以分析如何降低时间序列非平稳程度作为切入点来研究短期气候预测问题有着重要的理论意义.利用基于"升维"思想的支持向量机方法对时变控制参数条件下 Lorenz 系统产生的非平稳时间序列以及来自实际气候系统的非平稳时间序列进行预测试验研究.结果表明,基于统计学习理论的支持向量机方法对于非平稳过程存在稳定的预测能力.由于该方法可以通过核函数实现从样本空间到高维特征空间的非线性映射,这样可以理解为通过非线性映射,将低维空间中的非平稳过程映射到高维空间,通过"升维"在一定程度上降低了系统的非平稳程度.

关键词:支持向量机,非平稳过程,预测

PACC:0545

1. 引 言

短期气候预测是大气科学研究的热点和难点问题之一.但是由于短期气候过程自身的复杂性,以及人们对控制它的规律缺乏全面的认识,科学家们对它的预测能力不尽如人意.目前我们国家的短期气候预测水平若用预测场和实况场之间的相关系数度量,地面气温距平的预测精度仅在 0.25—0.30 之间,降水距平的预测精度则还要低得多.而且,近些年来预测技巧并没有明显的提高[1,2].这其中或许有着其深刻的理论原因.

我们知道,对于绝大多数由实际观测资料所构成的动力系统(特别是天气或气候系统)来说,控制它的外部条件,包括各种源、汇或来自边界的强迫,都不会是一成不变的,这样的系统往往表现出非线性、非平稳性特征[3—6].

实际上,在一些天气和气候资料的分析中,人们已经发现了大气过程的平稳性被破坏的事实[7—9].1996 年,Tsonis[10] 发现,最近 100 年全球降水的平均值变化不大,但其二阶矩发生明显变化,表明全球 20 世纪的降水资料描述了一个非平稳过程.然而,对于现有的气候预测理论,包括统计预测理论和非线性预测理论,大都建立在过程是平稳的假定的基础之上,这似乎有悖于气候过程的基本形态.因此一些科学家们认为这是当前气候预测水平低下的最重要的理论上的原因[4].

在某些具体过程的预测中,Wang 等人[11] 用分层嵌入和复合重构的方法,从具有层次结构的非平稳序列中分离出所需要的平稳段落来.然而,这种分层嵌入和复合重构的方法,对于控制因子以及时滞参数等的选择上有一定的主观性,尤其是它的本质仍然是依赖于能否从非平稳序列中分离出所需要的平稳段落来.

另外,Hegger 等人[12] 用"过嵌入"(over embedding)的方法对来自 Logistic 映射的序列进行预测并取得很好的结果.但是该方法在选取时间滞后参数上有可能遇到障碍,特别是当层次的积分尺度之间存在着较大的差别时,选择一个同时适合于两者的滞后参数是十分困难的.但是,Hegger 等人的方

* 国家自然科学基金(批准号:40505018,90411009)资助的课题.

† E-mail:wgl@mail.iap.ac.cn

法给了我们一些启示，即通过扩大嵌入维数的办法将系统的时变参数当作状态变量来处理，通过“升维”在一个被扩大了的嵌入系统中，在一定程度上消除了能够产生非平稳性的物理上的原因(系统的非平稳性在本质上是由于外部强迫随时间变化而引起的).

1998 年 Vapnik[13] 提出了统计学习理论(statistical learning theory)的基本思想，为建立有限样本学习问题提供了一个统一的框架.近年来，基于该理论发展的支持向量机方法(support vector machine，SVM)，逐渐成熟并已在模式识别、函数估计等人工智能领域得到较好的应用，研究结果表明[14—16]，SVM 方法可以通过核函数实现从样本空间到高维特征空间的非线性映射，利用支持向量来刻画因子与对象之间的非线性依赖关系，而且该方法对小样本条件下的非线性映射具有优势.

在上述分析的基础上，本文利用“升维”思想并借助于支持向量机方法，基于已熟知的非线性动力学系统，以及在此基础上得到的非平稳时间序列，利用 SVM 方法建立预测模型，以考察 SVM 方法对非平稳时间序列的预测能力.因为从这样的非线性动力学系统得到所需要的分析资料，它们可以满足分析试验所要求的各种条件，如噪声水平可以降低，资料长度也可增加等等.另外，利用该方法本文对来自实际气候系统的非平稳时间序列也尝试建立预测模型.

2. SVM

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.SVM 方法得到的回归函数只由少数的支持向量所确定，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.模型的复杂程度取决于支持向量的数目，而不是样本空间的维数，从而在某种程度上避免了“维数灾”.

SVM 方法的基本思想是基于 Mercer 核展开定理，通过非线性映射，把样本空间映射到一个高维乃至于无穷维的特征空间，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.简单地说就是升维和线性化.

由于任意满足泛函 Mercer 条件的对称函数均

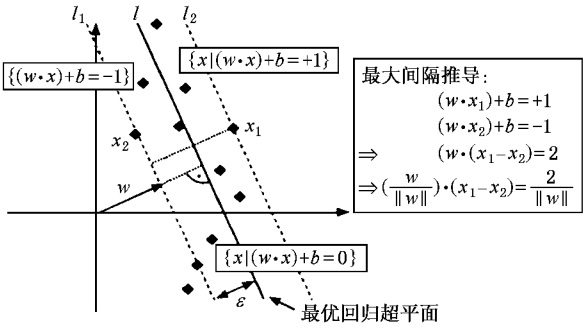

图 1　最优回归超平面图示

可作为核函数，在以下的工作中，我们选择将径向基函数作为核函数来建立 SVM 回归模型.径向基函数形为

$$K(x,x_i)=\exp(-r\|x-x_i\|^2). \tag{1}$$

回归函数则为

$$f(x)=\sum_{i=1}^{L}(\alpha_i-\alpha_i^*)K(x,x_i)+b$$

$$=\sum_{i=1}^{L}(\alpha_i-\alpha_i^*)\exp(-r\|x-x_i\|^2)+b, \tag{2}$$

其中 L 为支持向量数；x_i 为作为支持向量的样本因子向量；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.

3. 利用 SVM 方法建立非线性动力学系统的预测模型

3.1. 时变参数控制条件 Lorenz 系统的预测

1963 年，Lorenz[17] 在研究长期天气的可预报性问题时，得到了形如(3)式的 Lorenz 方程，并且发现在这个确定性的非线性耗散系统中，存在看似无序的非周期运动-混沌.

$$\begin{aligned}\dot{x}&=-\sigma x+\sigma y,\\ \dot{y}&=rx-y-xz,\\ \dot{z}&=xy-bz.\end{aligned} \tag{3}$$

式中 σ 取 10，b 取 8/3，积分步长为 0.001，采用四阶 Runge-Kutta 方法积分 Lorenz 方程得到 x，y，z 三个分量.图 2 为 Rayleigh 数 r 取 28 时，轨线在相平面(x，z)上的投影.

积分 60 万步，剔除前 10 万步，认为此后系统进入混沌状态.将序列每 100 步进行平均得到的 Lorenz 时间序列作为我们的研究对象.选择预报开始点落在相点从吸引子的某一叶到另一叶的过渡带

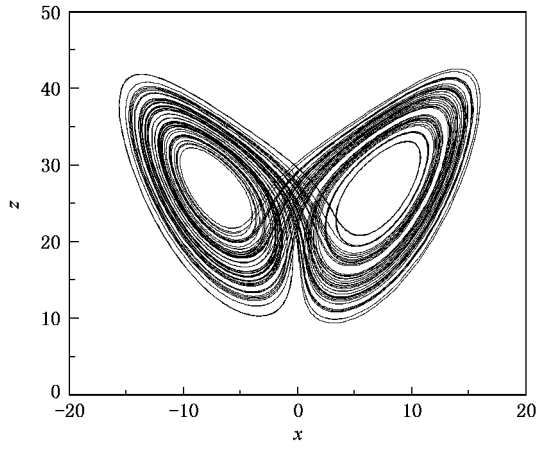

图 2 Lorenz 系统的轨线在相平面(x,z)上的投影

(转型期),在此区间系统的预测技巧相对不高[18]. 首先,对该序列进行归一化处理,再将预报对象进行整理,得到训练集、试验集和检验集,然后利用上述 SVM 方法进行 500 步的预测试验. 为验证该预测方法的稳定性,分别对 10 个不同的样本序列,利用各自的最优模型分别建立预测模型,将其平均值作为最终的预报值,与真实值进行对比. 图 3 是对该混沌时间序列 x 分量进行 500 步的预测结果,预测值和实际值的相关系数为 0.93,均方根误差为 2.51. 表明 SVM 对该混沌时间序列的预测效果是理想的.

下面再看一个复杂的情况. 在积分 Lorenz 方程的过程中,改变 Rayleigh 数 r 的值,使 r 被设定为一个时变参数,它的大小由 Logistic 映射的输出值所决定. Logistic 映射可以写为

$$r_{k+1} = \mu r_k(1 - r_k/a) \quad (k = 0,1,2,\cdots). \tag{4}$$

在 Logistic 映射中,μ 为控制参数,当 μ 位于 [0,1] 时,系统表现为不动点,而 μ 位于 [1,u_∞] 时,

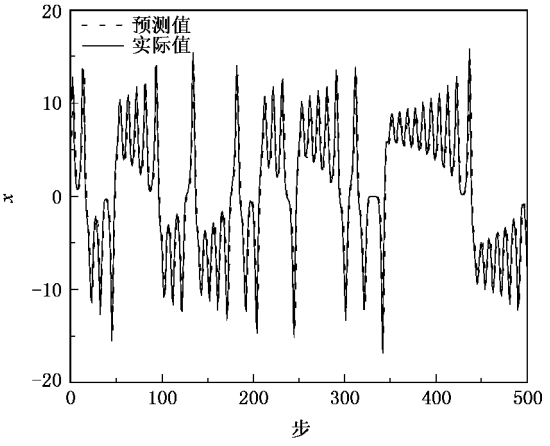

图 3 500 步预测对比

系统为周期解,当 μ 位于 [u_∞, 4] 时,系统则出现混沌现象,$u_\infty \approx 3.569945672$,参数 a 可以控制 r_k 的放大倍数. 我们让参数 μ 和 a 分别取值 3.9 和 30. 在这样的情况下,Logistic 映射输出的 r_k 是一个混沌解,其值在 3.2 和 29.2 之间变化,将这样一个随时间变化的量 r_k 作为 Lorenz 系统的控制参数 Rayleigh 数 r,那么 Lorenz 系统的状态将在一些完全不同的定态和混沌态之间非周期的改变着. 采用与上面相同的初值及积分方法,得到 Lorenz 系统 x, y, z 三分量. 由于此时的 Lorenz 系统的状态在不同的定态和混沌态之间变化,其状态分布将随时间变化,这样得到的时间序列为非平稳的时间序列. 图 4 为 Rayleigh 数 r 取值在 3.2 和 29.2 之间变化条件下,Lorenz 系统轨线在相平面(x, z)上的投影.

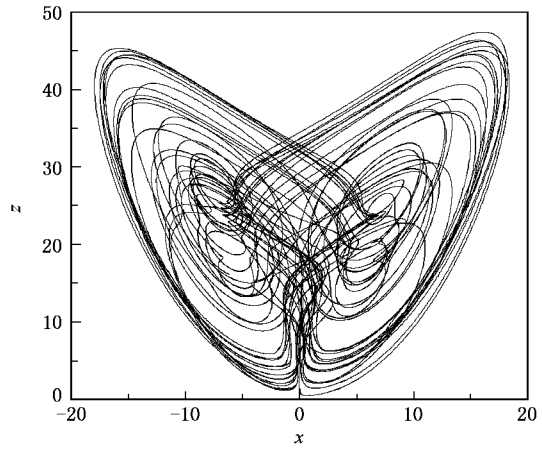

图 4 时变 Rayleigh 数控制下 Lorenz 系统的轨线在相平面(x,z)上的投影

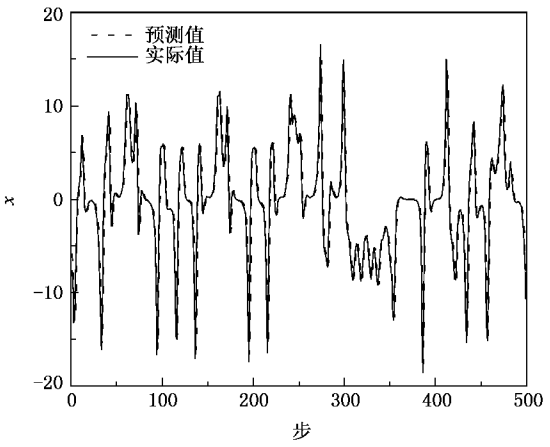

图 5 500 步预测对比(时变参数条件下)

采用与上面的预测试验相同的建模思路,利用

SVM 方法对得到的非平稳时间序列同样进行 500 步的预测，结果见图 5.可见预报值和真实值总体吻合得很好，它们的相关系数为 0.88，均方根误差为 2.62，而采用神经网络方法进行预测的相关系数和均方根误差则分别为 0.12 和 5.67.由此可见，SVM 方法对非平稳时间序列也有较好的预报能力.

3.2. 实际气候系统的预测试验

近年来大气过程的非平稳性特征已经得到越来越多科学家的认同.为考察 SVM 方法对于实际气候系统的预测能力，我们选择两个资料相对较长的气候时间序列，尝试利用 SVM 方法进行预测试验研究.

3.2.1. 南方涛动指数的预测试验研究

研究认为[19] 20 世纪 70 年代中后期以后，位于达尔文岛的观测海平面气压值不断升高，而位于塔西提岛的观测海平面气压不断下降，形成了在年代际尺度上最弱的热带太平洋东西向环流(Walker 环流)，由南方涛动指数所表征的 Walker 环流在 20 世纪 70 年代中后期异常减弱，表明南方涛动指数序列来自一个非平稳过程.该资料来自美国大气科学研究中心网站 http://www.cgd.ucar.edu/cas/catalog/climind/soi.html，资料长度为 141 年(自 1866 年至 2006 年)，我们将样本进行归一化处理后，将前 100 个样本作为训练集，随后的 25 个样本作为实验集，由于资料长度有限，未另外安排检测集，用实验集代替.采用 SVM 方法建模，预报最后 16 年(即 1991 年到 2006 年)年平均南方涛动指数.图 6 为其预报值和实测的对比，两者的相关系数为 0.5，均方根误差为 0.94.

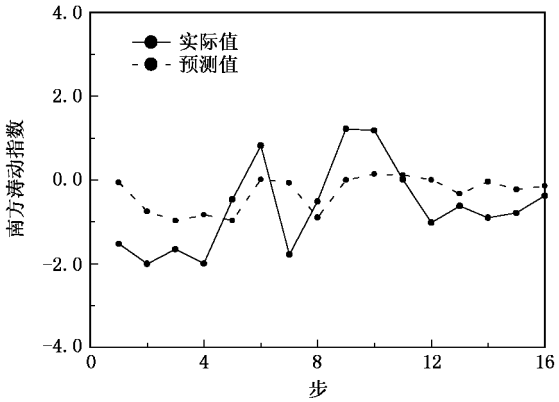

图 6　南方涛动指数的预测对比

3.2.2. 臭氧预测试验研究

另一个序列为 New Delhi 站(28.65°N，77.22°E，Dobson)的月平均臭氧总量观测资料，图 7 给出该站年平均臭氧总量的变化及其趋势特征，可以看出它来自于一个非平稳过程.该资料由世界臭氧及紫外资料中心网站 http://www.msc-smc.ec.gc.ca/woudc 提供.样本为从 1957 年 7 月到 2005 年 8 月共 578 个月的臭氧浓度月平均资料，其中个别月份缺测的用前后两个月的平均值代替.对样本进行归一化处理后，将前 414 个样本作为训练集，中间的 108 个样本作为实验集，采用 SVM 方法建模，预测该站从 2001 年 1 月到 2004 年 12 月共 48 个月的月平均臭氧柱总量.

图 8 为利用 SVM 方法对 New Delhi 站 2001 年 1 月至 2004 年 12 月的平均臭氧柱总量的预测结果.预测值与实测值的相关系数为 0.68，均方根误差为 10.3，预报结果比较理想.

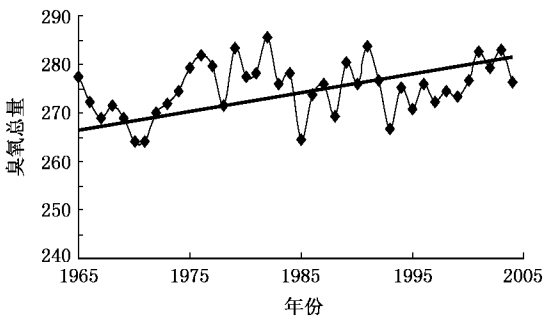

图 7　New Delhi 站臭氧总量年变化

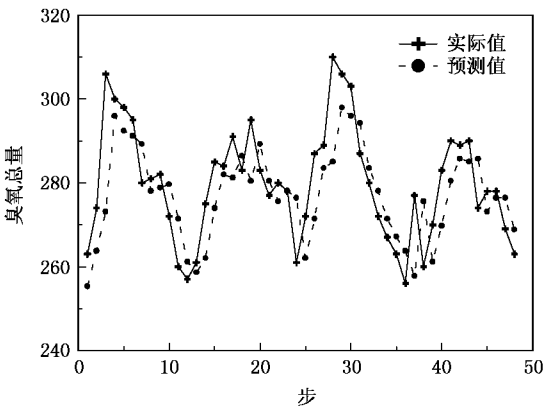

图 8　臭氧柱总量的预报对比

3.2.3. 预测精度的影响

为考查支持向量的个数对预测精度的影响，我们利用上述 New Delhi 站臭氧总量资料，选择不同的支持向量个数建模以便分析其对预测精度的影响，结果见图 9.可见支持向量选择的过少，对均方根误差的影响较大.支持向量机方法是通过较多的支持

向量来尽可能全面地描述预报对象的演变过程.而且预测精度还与训练样本的数量有关,如果预测对象的样本数太少,即使全部的样本都作了支持向量,也许还是保证不了精度要求[14].

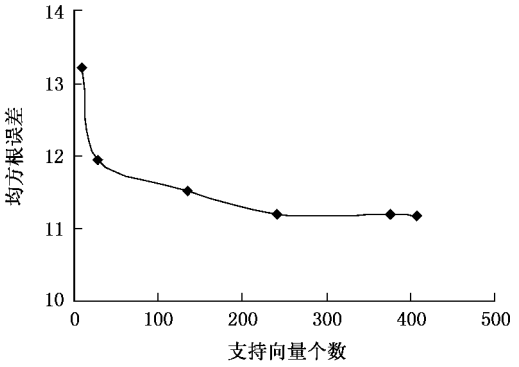

图 9 均方跟误差随支持向量个数的变化

上述预测试验是初步的,并且对于实际气候系统而言,由于资料的数量以及噪音等因素的存在,在运用 SVM 方法进行气候预测时存在一定的不确定因素,需要我们进一步地探讨.

4. 结 论

对于非平稳过程来说,在物理上,非平稳意味着控制系统的某些参数或条件是随时间变化的,人们不可能用同样的规律去刻画系统的动力学,在数学上,则意味着系统的状态分布依赖于时间,这样至关重要的遍历性定理就不再成立了.

本文利用 SVM 方法,对 Lorenz 系统这样的'理想'时间序列发生器建立预测模型,并在此基础上,当控制参数 Rayleigh 数 r 取时变参数时,产生的非平稳时间序列进行预测试验研究.另外对两个来自实际气候系统的时间序列进行预测试验.结果表明,基于统计学习理论的 SVM 方法不仅对平稳过程有较好的预报能力,也可以适用于非平稳过程.而且预测效果优于常用的非线性时间序列分析方法如人工神经网络方法.

其物理原因可能由于 SVM 方法可以通过核函数实现从样本空间到高维特征空间的非线性映射,利用支持向量来刻画因子与对象之间的非线性依赖关系,从而解决本质上的非线性问题.我们可以理解为通过非线性映射,将低维空间中的非平稳过程映射到高维空间,在一定程度上降低了系统的非平稳程度.具有这些物理特征的 SVM 方法对其他非平稳过程的预测研究仍将有参考意义.更重要的是,它预示着'升维'思想有可能成为发展和建立非平稳行为预测理论和方法的一个重要途径.另外,SVM 方法是专门针对有限样本的,其目标是得到现有信息下的最优解,它避免了人工神经网络等方法的网络结构选择、过学习和欠学习以及局部极小等问题[15].

[1] Wang S W, Zhu J H 2000 *J. of Appl. Meteor Sci.* **11** 1 (in Chinese) [王绍武、朱锦红 2000 应用气象学报 **11** 1]

[2] Wang H J, Zhou G Q, Lin Z H 2002 *Cli. Envi. Res.* **7** 220 (in Chinese) [王会军、周广庆、林朝晖 2002 气候与环境研究 **7** 220]

[3] Feng G L, Dong W J, Li J P 2004 *Chin. Phys.* **13** 1582

[4] Yang P C, Zhou X J 2005 *Acta Meteor Sin.* **65** 556 (in Chinese) [杨培才、周秀骥 2005 气象学报 **65** 556]

[5] Chen B M, Ji L R, Yang P C, Zhang D M, Wang G L 2003 *Chin. Sci. Bull.* **48** 513 (in Chinese) [陈伯民、纪立人、杨培才、张道民、王革丽 2003 科学通报 **48** 513]

[6] Wan S Q, Feng G L, Dong W J, Li J P 2005 *Acta Phys. Sin.* **54** 5487 (in Chinese) [万仕全、封国林、董文杰、李建平 2005 物理学报 **54** 5487]

[7] Liu S K, Fu Z T, Liu S D, Zhao Q 2002 *Acta Phys. Sin.* **51** 10 (in Chinese) [刘式适、付遵涛、刘式达、赵 强 2002 物理学报 **51** 10]

[8] Liu T Z, Rong P P, Liu S D 1995 *J. Geophys. Sci.* **38** 158 (in Chinese) [刘太中、荣平平、刘式达 1995 地球物理学报 **38** 158]

[9] Trenberth K E 1990 *Bull. Amer. Meteor Soc.* **7** 988

[10] Tsonis A A 1996 *Nature* **382** 700

[11] Wang G L, Yang P C 2005 *Int. J. of Clim.* **25** 1265

[12] Hegger R, Kantz H, Matassini L, Schreiber T 2000 *Phys. Rev. Lett.* **84** 4092

[13] Vapnik V N trans. by Zhang X G 2000 *The Nature of Statistical Learning Theory* (Beijing: Tsinghua University Press) (in Chinese) [Vapnik V N 著, 张学工译 2000 统计学习理论的本质(北京:清华大学出版社)]

[14] Chen Y Y, Yu X D, Gao X H, Feng H Z 2004 *J. of Appl. Meteor Sci.* **15** 345 (in Chinese) [陈永义、俞小鼎、高学浩、冯汉中 2004 应用气象学报 **15** 345]

[15] Ma X G, Hu F 2004 *Pro. Nat. Sci.* **14** 349 (in Chinese) [马晓光、胡 非 2004 自然科学进展 **14** 349]

[16] Mukherjee S 1997 *IEEE Press* **511**

[17] Lorenz E 1963 *J. Atmos. Sci.* **20** 130

[18] Wang G L, Yang P C, Lu D R 2004 *Chin. J. Atm. Sci.* **28** 538 (in Chinese) [王革丽、杨培才、吕达仁 2004 大气科学 **28** 538]

[19] Ai L K 2004 *Cli. Envi. Res.* **9** 303 (in Chinese) [艾丽坤 2004 气候与环境研究 **9** 303]

On the application of non-stationary time series prediction based on the SVM method *

Wang Ge-Li[1)†] Yang Pei-Cai[1)] Mao Yu-Qing[1)2)]

1)(*Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing* 100029, *China*)

2)(*College of Atmospheric Science, Nanjing University of Information Science & Technology, Nanjing* 210044, *China*)

(Received 2 April 2007; revised manuscript received 28 May 2007)

Abstract

The nonstationary behaviors of complex system and their applications to the climate prediction present a significant and forward-looking study. Up to now, its importance is not yet well understood. In reality, climate is just a normal nonstationary system. However, almost all the current theories for climate prediction, including the ones in statistics and nonlinear science, are based on one assumption that the process is stationary which is contrary to the nature of the climate process. Probably, this contradictory is an important cause resulting in the climate prediction being at a low reliability level. Therefore, it is theoretically important in climate prediction to start with how to reduce the nonstationary degree of time series. In this paper, support vector machine (SVM) method based on an idea of dimension raising is presented to study the time series prediction analysis, and prediction experiments are performed using some nonstationary time series from the Lorenz model and logistic system with changing control parameter, as well as two realistic climatic time series. The prediction results suggest that the SVM method can perform well in predicting nonstationary time series, which may be due to that the SVM method can map the input space into a higher dimensional feature space through nonlinear mapping and can reduce to some extent the nonstationary degree of the system.

Keywords: support vector machine, non-stationary process, prediction

PACC: 0545

* Project supported by the National Natural Science Foundation of China (Grant Nos 40505018, 90411009).

† E-mail: wgl@mail.iap.ac.cn