

支持向量机方法应用于理想时间序列的预测研究

毛宇清^{1,2} 王咏青² 王革丽¹

1 中国科学院大气物理研究所中层大气与全球环境探测开放实验室, 北京 100029

2 南京信息工程大学大气科学学院, 南京 210044

摘要 简要介绍了基于统计学习理论的支持向量机方法的基本思想和原理, 利用该方法对 33 模 Lorenz 系统的理想混沌时间序列建立预测模型, 并对在此基础上产生的非平稳时间序列进行预测试验研究。结果表明, 支持向量机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, 对实际序列的预测有一定的启发意义。

关键词 支持向量机 非平稳时间序列 预测建模

文章编号 1006-9585 (2007) 05-0676-07 **中图分类号** P437 **文献标识码** A

An Application Study on Prediction and Analysis for Ideal Time Series Based on the SVM Method

MAO Yu-Qing^{1,2}, WANG Yong-Qing², and WANG Ge-Li¹

1 *Laboratory of Middle Atmosphere and Global Environmental Observation, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029*

2 *College of Atmospheric Sciences, Nanjing University of Information Science and Technology, Nanjing 210044*

Abstract The support vector machine (SVM) regression principle and basic ideas based on the statistical learning theory are introduced. This method is used to build forecasting models on the ideal time series from 33-mode Lorenz system, and especially the prediction on nonstationary time series are tested and analyzed. It is shown that the SVM method is available for both stationary series and nonstationary ones, and the results are developmental to prediction of real data.

Key words support vector machine, nonstationary time series, forecast model

1 引言

Vapnik^[1]提出了统计学习理论 (Statistical learning theory) 的基本思想, 并在 20 世纪 90 年代中期不断发展和成熟。与传统的统计学相比, 统计学习理论是一种专门研究小样本情况下机器

学习规律的理论。该理论针对小样本统计问题建立了一套新的理论体系, 在这个理论体系下的统计推理规则不仅考虑了对渐进性能的要求, 而且追求在现有有限信息的条件下得到最优结果^[1,2]。

统计学习理论为建立有限样本学习问题提供了一个统一的框架, 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^[2,3]。近年来，基于该理论发展的支持向量机方法（Support vector machine，简称 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^[4,5]。

马晓光等^[6]用 SVM 方法预报大气污染物浓度，并与人工神经网络的预报结果进行了比较，结果表明，SVM 方法预报的准确率显著优于人工神经网络。陈永义等^[4]、冯汉中等^[5]对 SVM 方法的原理和在气候预测领域中的应用进行了一些试验研究认为，SVM 方法能用于具有显著非线性特征的气象预测预报，在实时业务预报中能取得较好的结果。这些结果均表明^[1,7,8]，SVM 方法对小样本条件下的非线性映射具有优势。

最近，在一些天气和气候过程的分析中，人们已经发现了大气过程的平稳性被改变的事实^[9,10]。另外，由于观测上或技术上的某些原因，大多数来自实际的时空序列，都不可避免地受到噪声的污染，或者在采样的时间长度和空间间隔上受到某些限制。在这类问题的研究中，通常要先在一些“理想”的试验资料上进行分析，它们通常来自于一些为人们所熟知的或被广泛研究过的非线性动力系统，从这样的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。这样的系统有时也被称作“理想”资料“发生器”，本文选择 33 模 Lorenz 系统这样的“发生器”，得到混沌时空序列^[11,12]。

本文的目的就是基于上述“理想”的试验资料，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。本文安排如下：在下一节，简要介绍 SVM 分析方法的基本思路，第

3 节给出 SVM 方法对 33 模 Lorenz 系统的预测能力，以及对由此产生的一些非平稳时间序列的预测能力进行试验分析，最后是小结。

2 支持向量机（SVM）回归方法及预测建模思路简介

机器学习的目的是根据给定的训练样本求对某系统输入输出之间依赖关系的估计，使它能够对未知输出作出尽可能准确的预测。有 3 类基本的机器学习问题，即模式识别、回归分析和概率密度估计。

回归分析又称函数估计，它要解决的问题是：依据有限的观测数据（训练样本）来寻求蕴涵着的回归函数，进而用求得的回归函数对未来数据（预报数据）进行预报。可以形式化为：给定一组训练样本集 $(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)$ ，其中 $x_i \in R^N$ ，为 N 维向量，通常为预报因子值； $y_i \in R$ ，为预报对象值；给出待预报样本的预报因子数据集： $x_{l+1}, x_{l+2}, \dots, x_m$ ，寻求与训练样本的输入输出拟合最优的函数关系 $y=f(x)$ ，进而求出预报对象 y_i 的输出值。

SVM 回归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，简单地说就是升维和线性化。图 1 给出了这一思路的大致框架。

SVM 方法的核心概念是支持向量。如图 2 所示，最优回归超平面 l 完全由落在两条边界线 l_1

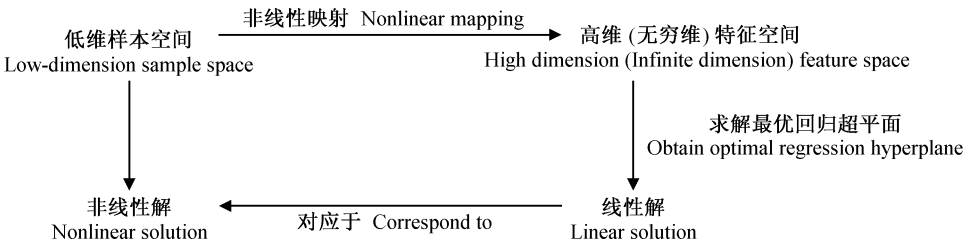


图 1 SVM 回归方法的主要思路
Fig. 1 Main idea of the SVM regression method

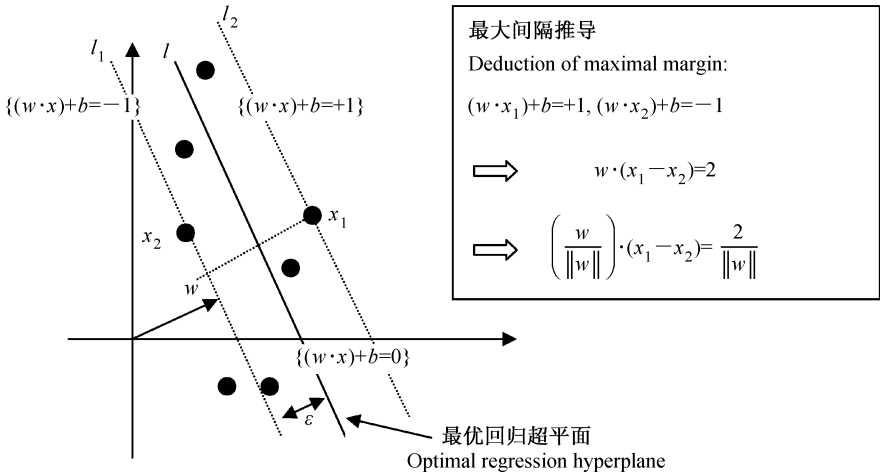


图 2 最优回归超平面示意图
Fig. 2 Sketch map of the optimal regression hyperplane

和 l_2 上的样本点 (如 x_1, x_2 等) 所确定, 这样的样本点称为支持向量。所寻求的最优超平面是使所有样本点离超平面的“总偏差”最小, 因此, 求最优回归超平面同样等价于求最大间隔。引入 ϵ -不敏感误差函数 (即误差小于 ϵ 时视为无误差), 则寻求最优回归超平面的问题转化为求解如下一个二次凸规划问题:

$$\min \frac{1}{2} \|w\|^2 + C \sum_i (\xi_i + \xi_i^*),$$
$$\text{约束条件} \begin{cases} y_i - (w \cdot x_i) - b \leq \epsilon + \xi_i, \\ (w \cdot x_i) + b - y_i \leq \epsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0, \end{cases} \quad (1)$$

其中, ξ_i 和 ξ_i^* 是松弛变量, 分别对应于样本点在最优回归超平面的上方和下方两种情况; C 为惩罚系数; $w \in R^N$ 是超平面的法矢量, $b \in R$, 二者均为待确定的参数。

采用 Lagrange 乘数法并应用 KKT 条件 (Karush-Kuhn-Tucker), 可求得最优超平面线性回归函数为

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

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。可见, 最优回归超平面的解析式只由支持向量完全确定。

应用核函数的展开定理, 利用非线性映射 φ ,

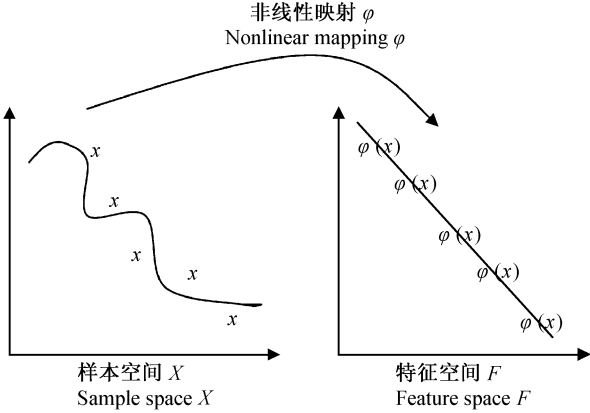


图 3 非线性映射
Fig. 3 Nonlinear mapping

把样本集映射到一个高维 (以至于无穷维) 的特征空间, 从而使得样本空间中的高度非线性问题在特征空间中应用 SVM 线性回归的方法得以实现。在整个求解过程中并不需要知道非线性映射的显示表达式, 只需将样本空间中的点 x 和 x_i 用映射的象 $\varphi(x)$ 和 $\varphi(x_i)$ 代替, 这大大简化了计算。图 3 给出了这一过程的直观说明。

由此 (2) 式变为

$$f(x) = (w \cdot x) + b$$
$$= \sum_{i=1}^L (\alpha_i - \alpha_i^*) [\varphi(x) \cdot \varphi(x_i)] + b, \quad (3)$$

根据 Mercer 定理有

$$K(x, x_i) = [\varphi(x) \cdot \varphi(x_i)], \quad (4)$$

将 (4) 代入 (3) 可得

$$f(x)=(w \cdot x)+b=\sum_{i=1}^L\left(a_i-a_i^*\right) K\left(x, x_i\right)+b, \quad(5)$$

(5) 式即为 SVM 方法最终确定的非线性回归函数。可见，SVM 的最终决策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和核函数的计算，而不是样本空间的维数，从而在某种程度上避免了“维数灾”。

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，因此我们选择径向基函数（满足 Mercer 定理条件）作为核函数建立 SVM 回归模型。径向基函数形为

$$K\left(x, x_i\right)=\exp \left(-r\left\|x-x_i\right\|^2\right), \quad(6)$$

最终回归函数形为

$$\begin{aligned} f(x) &= \sum_{i=1}^L\left(a_i-a_i^*\right) K\left(x, x_i\right)+b \\ &= \sum_{i=1}^L\left(a_i-a_i^*\right) \exp \left(-r\left\|x-x_i\right\|^2\right)+b, \quad(7) \end{aligned}$$

其中， L 为支持向量的个数， 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， a_i 、 a_i^* 和 b 为建立 SVM 模型待确定的系数， r 为核参数。

3 用 SVM 方法建立“理想”时间序列预测模型

1963 年，Lorenz^[13]在研究长期天气的可预报性问题时，得到了著名的 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。Curry^[14]在这个基础上将谱模增加到 14 个，杨培才^[11]又将谱模增加到了 33 个，并充分证实了谱模的增加将引起解在参数空间中分布的重大变化。

采用 Hamming 算法求解 33 模 Lorenz 系统，给定与 3 维 Lorenz 系统相同的参数值和初值，即 $\sigma=10$ ， $a=\sqrt{2} / 2$ ，积分步长取为 $h=0.001$ ，积分 110 000 步，剔除前 100 000 步积分的过渡期，认为此后系统进入定常状态。取其中一个谱模分量 φ_{11} 的 10 000 步时间序列作为我们的研究对象。研

究表明^[12]，这样的“理想”时间序列发生器得到的时间序列，其功率谱具有红噪声谱特征，其吸引子具有分形结构，而且至少存在一个不为负数的 Lyapunov 指数。

SVM 是通过支持向量构造推理模型，因此对因子的数量没有明显的限制。对于理想的单变量时间序列，模型的输入输出是重叠的，即训练和验证阶段的变量都取自 $\left\{\varphi_{11}(i)\right\}, i=1, \cdots, N$ ，这是一种最简化的原型。我们采用中国气象局培训中心 SVM 应用开发研究小组提供的 CMSVM1.0 软件平台进行建模和预报，选取最优模型。以经典统计预报方式建立样本序列，即以 t 时刻的预报因子对应 $t+1$ 时刻的预报对象。

首先，对因子进行归一化处理，归一化方式为 $\left(x-x_{\min }\right) /\left(x_{\max }-x_{\min }\right)$ ，使每个因子的数据落入 $[0,1]$ 区间。再将预报因子和预报对象进行整理，得到训练集、实验集和检验集，比例分别为 75%、20%和 5%，共 10 组。构建预报数据，进行单步预报，在预报时直接将最优模型和预报数据进行匹配，得出预报值。

由于不同时间段的样本通过建模所得到的最优模型不尽相同，预报出来的结果也有差别。为此，我们将积分所得的序列按时间段分为 10 个小的样本集，前 9 个小样本集有 1 000 个样本，第 10 个小样本集有 900 个样本，剩 100 个样本作为预报数据。将这 10 个样本集分别进行建模，用各自的最优模型分别预报 100 步，然后将 10 组预报值的平均值作为最终的预报值，与真实值进行对比。

3.1 对平稳时间序列的预报结果检验

取 Rayleigh 数 $r=400$ ，并且在积分过程中 r 保持不变。在这样的条件下，我们得到一个平稳的时间序列。其中， φ_{11} 为流场的谱模分量，也是我们研究的对象， θ_{11} 为扰动温度场的谱模分量。图 4 是 Rayleigh 数 $r=400$ 时得到的 φ_{11} 随时间的变化及趋势。图 5 为 Rayleigh 数 $r=400$ 时 33 维 Lorenz 系统的解轨线在 $\left(\varphi_{11}, \theta_{11}\right)$ 相平面上投影。图 6 是对 33 模 Lorenz 系统的平稳时间序列作 100 步预报的结果对比。从图中可以看出，预报值与真实值的趋势和幅值大小几乎重合。计算得到二者的相关系数为 0.9978，均方差为 2.266，预报的准确度是令人鼓舞的。

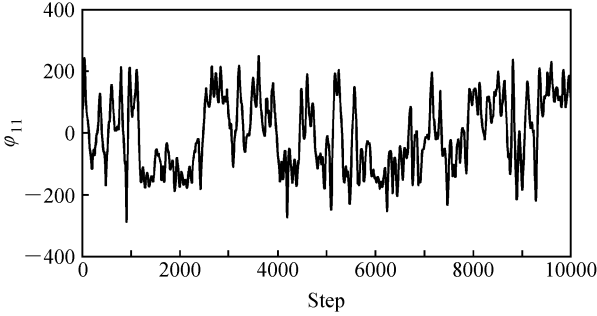


图 4 $r=400$ 时 φ_{11} 的时间演变
Fig. 4 The variation of φ_{11} when $r=400$

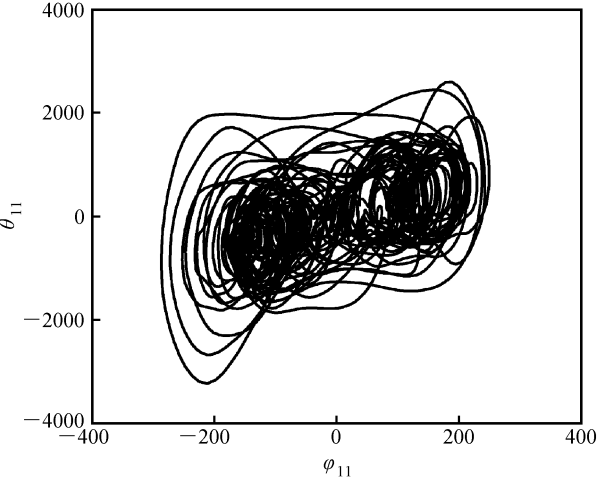


图 5 $r=400$ 时 33 维 Lorenz 系统的状态轨道在 $(\varphi_{11}, \theta_{11})$ 相平面上投影
Fig. 5 Projection for the state trajectories of 33 mode Lorenz system in $(\varphi_{11}, \theta_{11})$ when $r=400$

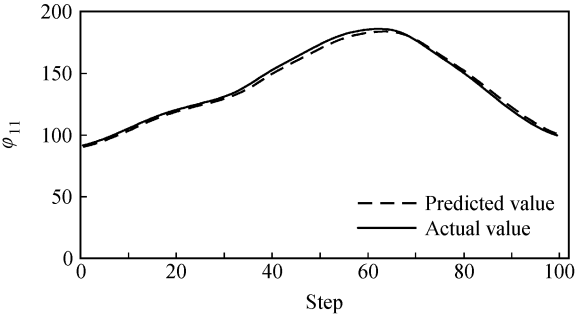


图 6 SVM 对 33 模 Lorenz 系统平稳时间序列的预报 ($r=400$)
Fig. 6 Predictions for stationary time series of 33 mode Lorenz system by SVM ($r=400$)

3.2 对非平稳时间序列的预报结果检验

改变 Rayleigh 数 r 的值,使其随时间发生变化,其他参数和初值不变。对 33 模 Lorenz 系统进行积分,得到非平稳的时间序列。我们取了两个函数分别进行检测。

3.2.1 取 Rayleigh 数 $r=4t+7$

积分 110 000 步,剔除前 100 000 步的结果,由于积分步长 $h=0.001$,则 Rayleigh 数 $r \in [411, 451]$,这样得到一个非平稳的时间序列。图 7 是 Rayleigh 数 $r=4t+7$ 时得到的 φ_{11} 随时间的变化及趋势。图 8 为 Rayleigh 数 $r=4t+7$ 时 33 维 Lorenz 系统的解轨线在 $(\varphi_{11}, \theta_{11})$ 相平面上投影。图 9 是对 33 模 Lorenz 系统产生的非平稳时间序列作 100 步预报的结果对比 ($r=4t+7$)。如图,预报值的总体趋势能很好地反映实际的变化。计算得到二者的相关系数为 0.9977,均方差为 1.34。

3.2.2 取 Rayleigh 数 $r=100\sin t+350$

积分 110 000 步,剔除前 100 000 步的结果,此时 Rayleigh 数 $r \in [250, 450]$,积分得到另一个非平稳的时间序列。图 10 是 Rayleigh 数 $r=100\sin t+350$ 时得到的 φ_{11} 随时间的变化及趋势。图 11 为 Rayleigh 数 $r=100\sin t+350$ 时 33 维 Lorenz 系统的解轨线在 $(\varphi_{11}, \theta_{11})$ 相平面上投影。图 12 是对 33 模 Lorenz 系统的非平稳时间序列作 100 步预报的结果对比 ($r=100\sin t+350$)。如图,预报值和真实值总体吻合得很好。计算得到二者的相关系数为 0.9989,均方差为 1.732。由此可见,SVM 方法对非平稳时间序列也有较好的预测能力。

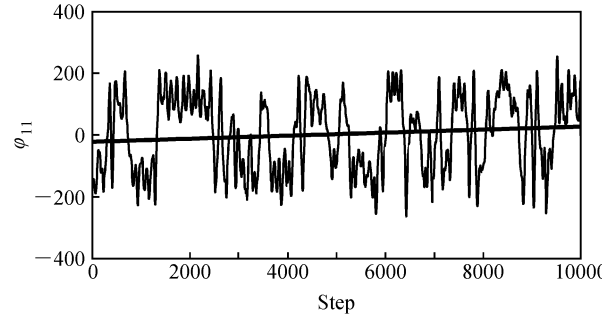


图 7 $r=4t+7$ 时 φ_{11} 的时间演变
Fig. 7 The variation of φ_{11} when $r=4t+7$

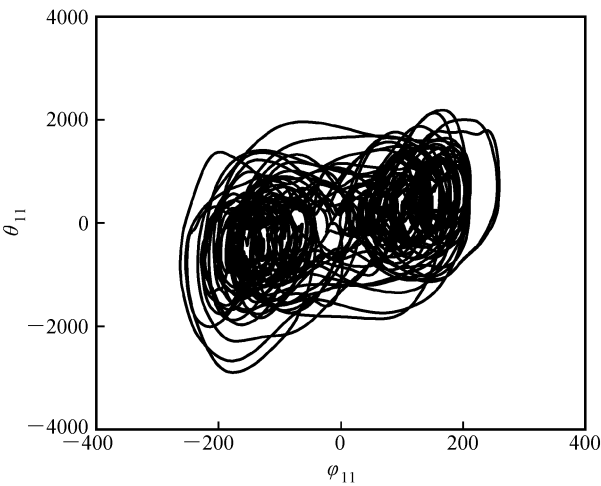


图 8 $r=4t+7$ 时 33 维 Lorenz 系统的状态轨道在 $(\varphi_{11}, \theta_{11})$ 相平面上投影

Fig. 8 Projection for the state trajectories of 33 mode Lorenz system in $(\varphi_{11}, \theta_{11})$ when $r=4t+7$

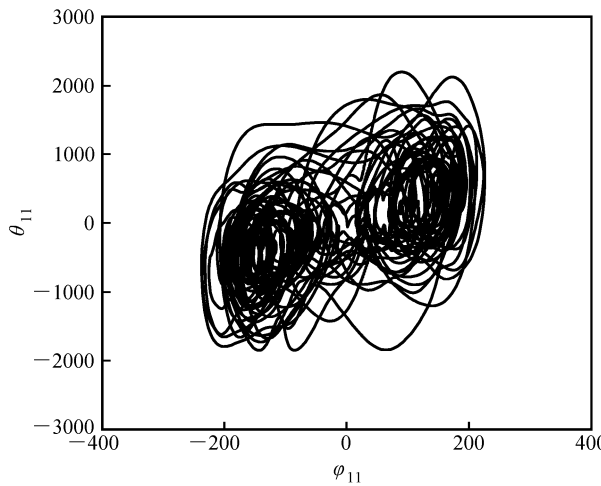


图 11 $r=100\sin t+350$ 时 33 维 Lorenz 系统的状态轨道在 $(\varphi_{11}, \theta_{11})$ 相平面上投影

Fig. 11 Projection for the state trajectories of 33 mode Lorenz system in $(\varphi_{11}, \theta_{11})$ when $r=100\sin t+350$

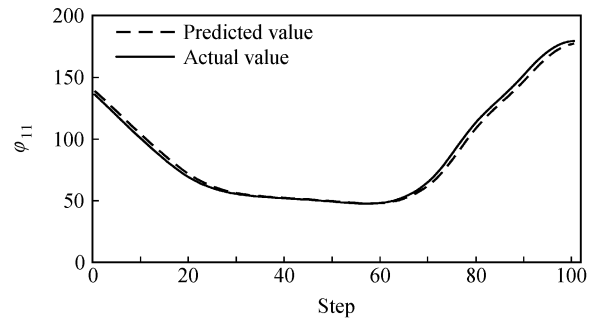


图 9 SVM 对 33 模 Lorenz 系统非平稳时间序列的预报 ($r=4t+7$)

Fig. 9 Predictions for nonstationary time series of 33 mode Lorenz system by SVM ($r=4t+7$)

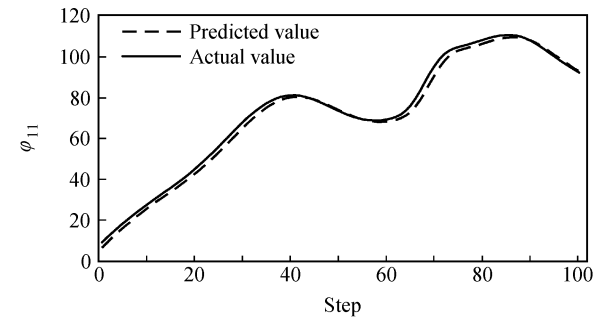


图 12 SVM 对 33 模 Lorenz 系统非平稳时间序列的预报 ($r=100\sin t+350$)

Fig. 12 Predictions for nonstationary time series of 33 mode Lorenz system by SVM ($r=100\sin t+350$)

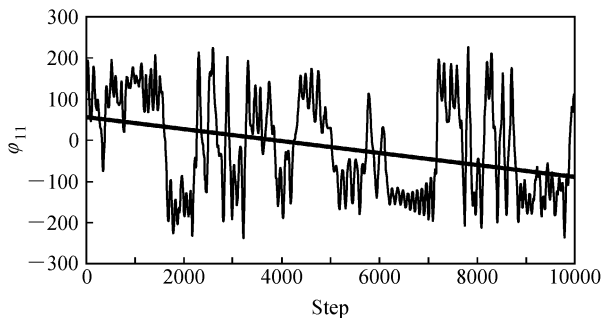


图 10 $r=100\sin t+350$ 时 φ_{11} 的时间演变

Fig. 10 The variation of φ_{11} when $r=100\sin t+350$

4 小结

SVM 方法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, 通过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。另外, 该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而不仅仅是样本数趋于无穷大时的最优解。

本文利用 SVM 方法, 对 33 模 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, 对由此产生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 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。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, 一定程度上降低了系统的非平稳程度。

诚然, 上述结论是基于理想的时空序列得到的, 对于实际气候资料而言, 由于资料的数量以及噪音等因素的存在, 其预测效果在一定程度上受到限制。在运用 SVM 方法进行预报时有一定的不确定因素, 需要我们进一步地探讨。

参考文献 (References)

- [1] Vapnik V N. *Statistical Learning Theory*. John Wiley & Sons, Inc., New York, 1998
- [2] 张学工. 关于统计学习理论与支持向量机. 自动化学报, 2000, **26** (1): 32~42
Zhang Xuegong. Introduction to statistical learning theory and support vector machines. *Acta Automatica Sinica* (in Chinese), 2000, **26** (1): 32~42
- [3] Yao Xin. Evolving artificial neural networks. *Proceedings of the IEEE*, 1999, **87** (9): 1423~1447
- [4] 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. 应用气象学报, 2004, **15** (3): 345~354
Chen Yongyi, Yu Xiaoding, Gao Xuehao, et al. A new method for non-linear classify and non-linear regression. I: Introduction to support vector machine. *Journal of Applied Meteorological Science* (in Chinese), 2004, **15** (3): 345~354
- [5] 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** (3): 355~365
Feng Hanzhong, Chen Yongyi. A new method for non-linear classify and non-linear regression. II: Application of support vector machine to weather forecast. *Journal of Applied Meteorological Science* (in Chinese), 2004, **15** (3): 355~365
- [6] 马晓光, 胡非. 利用支撑向量机预报大气污染物浓度. 自然科学进展, 2004, **14** (3): 349~353
Ma Xiaoguang, Hu Fei. Using support vector machine to predict atmosphere contamination thickness. *Progress in Natural Science* (in Chinese), 2004, **14** (3): 349~353
- [7] Mukherjee S, Qsuea E, Girosi F. Nonlinear prediction of chaotic time series using a support vector machine. In: *IEEE Workshop on Neural Networks for Signal Processing VII*. Principe J, Giles L, Morgan N, et al. Eds. Amelia Island: IEEE Press, 1997. 511~520
- [8] Müller K R, Smola A J, Rätsch B, et al. Using support vector machines for time series prediction. In: *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA, USA: MIT Press, 1999. 242
- [9] Trenberth K E. Recent observed interdecadal climate changes in the Northern Hemisphere. *Bull. Amer. Meteor. Soc.*, 1990, **71**: 988~993
- [10] 杨培才, 周秀骥. 气候系统的非平稳行为和预测理论. 气象学报, 2005, **63** (5): 556~570
Yang Peicai, Zhou Xiuji. On nonstationary behaviors and prediction theory of climate systems. *Acta Meteorologica Sinica* (in Chinese), 2005, **63** (5): 556~570
- [11] 杨培才. 33 模 Lorenz 系统的某些总体特征. 大气科学, 1987, **11** (1): 48~57
Yang Peicai. Some gross features of 33-mode Lorenz system. *Scientia Atmospherica Sinica* (in Chinese), 1987, **11** (1): 48~57
- [12] 王革丽, 杨培才, 吕达仁. 33 模 Lorenz 系统的混沌特征及其可预报性分析. 高原气象, 2006, **25** (1): 9~15
Wang Geli, Yang Peicai, Lü Daren. Chaotic characters and predictability of 33 mode Lorenz system. *Plateau Meteorology* (in Chinese), 2006, **25** (1): 9~15
- [13] Lorenz E. Deterministic nonperiodic flow. *J. Atmos. Sci.*, 1963, **20**: 130~141
- [14] Curry J H. Order and disorder in two and there dimensional Benard convection. *J. Fluid. Mech.*, 1978, **147**: 1~38