文章编号: 1001-0920(2006)01-0077-04

基于LSSVM的混沌时间序列的多步预测

江田汉[1,2], 束 炯[3]
(1. 北京大学 环境科学系, 北京 100871; 2. 中国安全生产科学研究院, 北京 100029;
3. 华东师范大学 地理信息科学教育部重点实验室, 上海 200062)

摘 要: 结合相空间重构理论和统计学习理论, 实现混沌时间序列的多步预测. 采用微熵率法求得最优嵌入维数和时延参数, 重构系统相空间, 用最小二乘支持向量机建立混沌时间序列的多步预测模型, 并与径向基函数网络预测模型比较. 结果表明, 所建立的模型能够捕捉到原混沌系统的动力学特征. 前者的归一化均方根预测误差远小于径向基函数网络预测模型的预测误差, 泛化能力较强, 其预测效果较好.

关键词: 混沌时间序列; 最小二乘支持向量机; 预测
中图分类号: TP18 **文献标识码**: A

Multi-step Prediction of Chaotic Time Series Using the Least Squares Support Vector Machines

JIANG Tian-han[1,2], *SHU Jiong*[3]
(1. Department of Environmental Sciences, Peking University, Beijing 100871, China; 2. China Academy of Safety Sciences and Technology, Beijing 100029, China; 3. Key Laboratory of Geographic Information Science, Ministry of Education, East China Normal University, Shanghai 200062, China. Correspodent: JIANG Tian-han, E-mail: jiangth@pku.org.cn)

Abstract: Based on the phase space reconstruction theory and the statistical learning theory, multi-step predictions of chaotic time series are presented. The optimal embedding dimension and delay time are obtained with the differential entropy ratio method. In the reconstructed phase space, the multi-step predicting model of chaotic time series is established with the least squares support vector machines (LSSVM model) and compared with the radial basis function network predicting model (RBF model). The results show that the proposed models can capture the dynamics of the chaotic systems. The normalized root mean square error of the LSSVM model is far less than that of the RBF model.

Key words: Chaotic time series; Least squares support vector machines; Prediction

1 引 言

许多实际问题, 如大气污染物浓度和气象气候、电力系统负荷、心脑电图、水文等数据, 几乎是非平稳和非线性的, 大多数传统时间序列模型(如AR, ARMA, ARIMA 等线性模型)在这些问题上的应用受到了极大的限制. 非线性时间序列预测方法自20世纪80年代中期开始发展, 近年来得到了更深入的研究和更广泛的应用. 目前, 在重构的相空间中, 逼近现在状态和未来状态之间的映射关系除了用多项式外, 国内外较为常用的数值技术是神经网络拟合. 但是, 由于神经网络模型的结构较复杂并难以选择, 需要估计的参数多, 以致其对数据过学习(即泛化)能力不够, 预测精度不高. 基于统计学习理论的支持向量机方法(SVM)[1]根据结构风险最小化原则, 最大程度地提高其泛化能力, 且其算法的局部最优解就是全局最优解. 该方法现已在模式识别、信号处

收稿日期: 2004-12-01; **修回日期**: 2005-03-09.
基金项目: 国家自然科学基金项目(40171088).
作者简介: 江田汉(1971—), 男, 福建永定人, 博士生, 从事大气污染物浓度的预报预警模型及油气井风险分析等研究; 束炯(1952—), 男, 上海人, 教授, 博士生导师, 从事城市气候与大气环境、环境遥感等研究.

理、函数估计等领域得到了应用,但用于混沌时间序列预测方面的研究相对较少[2]. 近年,Suykens 提出了最小二乘支持向量机(LS-SVM)[3],并在混沌时间序列预测中得到应用[4]. 但在实际运用中,重构相空间所需的最优嵌入参数的重要性却经常被忽视,甚至有研究者认为在嵌入维数未知时也能取得比较好的预测效果[4],而这是违背 Takens 定理的,因为现实的时间序列几乎是不光滑的.

因此,本文先估计最优嵌入维数 m 和时延参数 τ,重构混沌时间序列的相空间来近似原系统状态空间,用最小二乘支持向量机方法拟合系统演化的轨道,建立混沌时间序列的多步预测模型. 该模型不仅对模拟的混沌系统的预测性能非常好,而且对含有噪声的混沌数据具有较好的鲁棒性,能作出较好的多步预测.

2 预测方法

2.1 最优嵌入维数 m_{opt} 和时延参数 τ_{opt}

重构相空间的关键之一是确定时延参数 τ. 对于实际问题,已有很多选取 τ 的方法,常用的有自相关函数法和互信息量法. 重构相空间另一关键是确定嵌入维数 m,较常用的是计算"相关维"的方法.

对于给定的时间序列,应该存在一个最优的 m 和 τ[5,6]. 如果 τ 太小,则不能覆盖捕捉信号的动力学需要的最小时间距,m 将变得相当大;相反,如果 τ 大于最佳值,作为结果的模型的性质变得太离散,会导致捕捉不到信号的动力学性质. Gautama 等人[7]提出一个基于样本时间序列及其替代数据[8]的相空间的微熵率方法,同步确定 τ 和 m. 该方法主要的优点是用一个简单的测度同时优化 m 和 τ,避免了互信息量法和错误近邻法的不一致性. 该方法的物理意义明显,实际效果较好,故用此方法估计这两个嵌入参数的最优值,即 m_{opt} 和 τ_{opt}.

给定信号 $x(t)$ ($t = 1, 2, \cdots, N$) 的 N_s 个替代数据 $x_{s,i}(t)$, $i = 1, \cdots, N$,定义熵率(ER)[7]为

$$R_{ent}(m, \tau) = I(m, \tau) + \frac{m \ln n}{n}. \tag{1}$$

其中:n 是延迟矢量数;$I(m, \tau)$ 为

$$I(m, \tau) = H(x, m, \tau) / \langle H(x_{s,i}, m, \tau) \rangle_i,$$

而微商 $H(x) = \sum_{j=1}^{N} \ln(N \rho_j) + \ln 2 + C_E$,$N$ 是数据长度,ρ_j 是第 j 个延迟矢量与其最近邻点之间的欧氏距离,欧拉常数 $C_E \approx 0.577\,2$. 熵率图上的最小值在 m 和 τ 轴上分别对应 m_{opt} 和 τ_{opt}. 然后,重构相空间

$$X(t) = \{x(t), x(t + \tau_{opt}), \cdots, x(t + (m_{opt} - 1)\tau_{opt})\}. \tag{2}$$

2.2 最小二乘支持向量机[3]

设训练集 $S = \{(x_k, y_k) \mid k = 1, 2, \cdots, N\}$,其中:$x_k \in R^n$, $y_k \in R$ 分别为输入和输出数据. 原始空间中的优化问题及其约束条件为

$$\min_{w, b, e} J(w, e) = \frac{1}{2} w^{\mathrm{T}} w + C \sum_{i=1}^{N} e_i^2,$$

$$\text{s.t.} \quad y(x) = w^{\mathrm{T}} \varphi(x_i) + b + e_i, \quad i = 1, 2, \cdots, N. \tag{3}$$

其中:e_i 表示误差,w 为权重,$\varphi(\cdot)$ 为非线性映射函数,b 为偏差. 引入 Lagrange 乘子 α_i,得到

$$L(w, b, e; \alpha) = J(w, e) - \sum_{i=1}^{N} \alpha_i \{w^{\mathrm{T}} \varphi(x_i) + b + e_i - y_i\}. \tag{4}$$

根据 KKT 条件可得到

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{N} \alpha_i \varphi(x_i), \\ \dfrac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i = 0, \\ \dfrac{\partial L}{\partial e_i} = 0 \rightarrow \alpha_i = C e_i, \\ \dfrac{\partial L}{\partial \alpha_i} = 0 \rightarrow w^{\mathrm{T}} \varphi(x_i) + b + e_i - y_i = 0, \\ i = 1, 2, \cdots, N. \end{cases} \tag{5}$$

消元去掉 e_i 和 w,得到如下的线性方程组:

$$\begin{bmatrix} 0 & \vec{1}^{\mathrm{T}} \\ \vec{1} & K(x_i, x_j) + C^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}. \tag{6}$$

其中:$y = [y_1, y_2, \cdots, y_N]^{\mathrm{T}}$, $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_N]^{\mathrm{T}}$, $\vec{1} = [1, 1, \cdots, 1]^{\mathrm{T}}$, $K(x, x_i)$ 为满足 Mercer 条件的核函数. 解上式得最小二乘向量机回归模型如下:

$$y(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i) + b. \tag{7}$$

此外,选择归一化均方根误差(NRMSE)作为评价模型预测效果的判据,其表达式如下:

$$\mathrm{NRMSE} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (O_{i,\mathrm{pre}} - O_{i,\mathrm{obs}})^2} \Big/ S_{\mathrm{obs}}. \tag{8}$$

式中:$O_{i,\mathrm{pre}}$ 和 $O_{i,\mathrm{obs}}$ 分别为预测值和实测值,S_{obs} 为实测值的标准差,N 为预测步数.

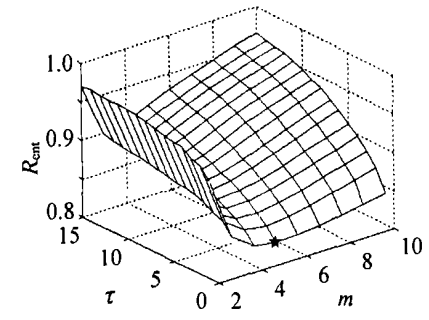
3 预测实例

用于本次预测实验的典型混沌时间序列有两组:一是 Lorenz 方程

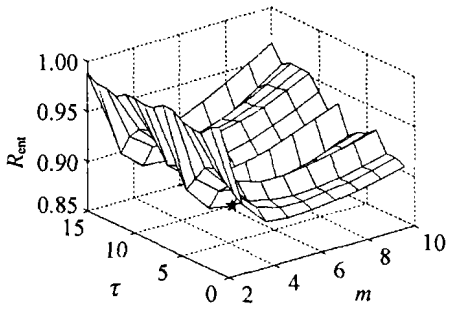
$$\begin{cases} \dot{x} = -\sigma x + \sigma y, \\ \dot{y} = -xz + rx - y, \\ \dot{z} = xy - bz, \end{cases} \quad \sigma = 10, \ r = 28, \ b = 8/3. \tag{9}$$

取初始值为(0.005, 0.01, 0.8), 积分步长为 0.1, 用龙格 - 库塔方法积分得到 102 048 个 x 的离散数据, 略去前面 10 000 个暂态数据, 而最后 2 048 个数据用于预测实验. 二是圣达非激光数据(Santa Fe Laser data), 该数据是混沌状态下远红外激光的测量值, 信噪比为 300 dB[10].

图 1 为这两组数据的熵率图, 熵率的最小值用★表示. 第 1 组和第 2 组数据的 m_{opt} 和 τ_{opt} 分别为(5, 1) 和(5, 7). 按式(2) 构造输入集和目标集, 对于第 1 组数据, 前 1 640 个数据用于训练, 后 408 个数据用于多步预测检验; 对于第 2 组数据, 前 810 个数据用于训练, 后 190 个数据用于多步预测检验. 作一个 P 步预测有两种方法: 一是直接法, 即每次在实测的基础上直接向前预测 P 步; 二是迭代法, 即每次只向前预测一步, 但每步都用新得到的预测值而不是实测值来继续下一步的预测. 众多的研究结果表明迭代法远远优于直接法[9], 本文采用迭代法实现模型的多步预测



(a)　Lorenz 时间序列



(b)　Santa Fe Laser 数据

图 1　微熵率

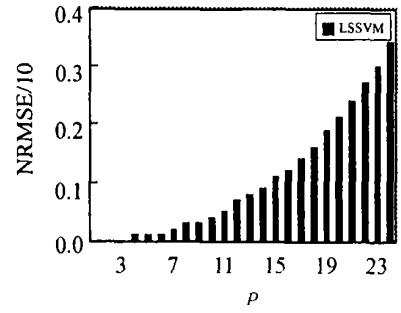
所用的核函数为径向基核函数, 即

$$K(x, x_i) = \exp(-\|x_k - x_i\|^2)/2\sigma^2, \quad (10)$$

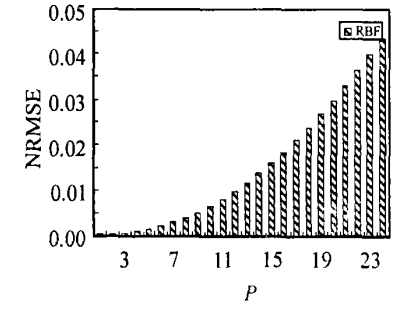
其中 σ 为核函数的超参数. 对训练集和相应的目标集进行归一化[0, 1] 处理后, 随机分为 10 组, 通过交叉检验选取 σ 和正则化参数 γ[11], 得到第 1 和第 2 组的 σ 和 γ 分别为{1 750, 0.2} 和{15, 0.095 8}; 然后, 把训练集和相应的目标集输入式(7) 得到模型参数 α_i 和 b; 最后, 用剩余的训练集和目标集进行预测检验.

为便于比较, 本文采用径向基函数为高斯核函数、核函数中心和方差都是自适应的径向基函数网络软件包 (http: //ida.first.fraunhofer.de/～raetsch/) 建立具有相同嵌入参数的预测模型. 该模型的隐层节点数为数据总数的 1/10, 调节参数为 $1e$-10, 优化迭代次数为 10, 所有参数都经过多次实验比较得到.

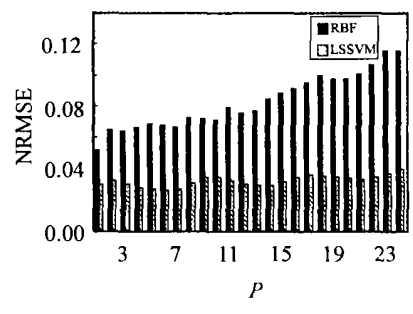
用这两种预测模型分别对第 1 组和第 2 组数据作了 P = 1～24 的预测. 一般而言, 随着 P 增大, 预测误差也将增大. 直观上看, 两种模型均能较好地作出多步预测. 注意到 P 较小时, 这两种模型的预测结果都与原始值相当吻合, 但当 P 逐渐增大时, RBF 网络预测模型的预测结果在某些时刻出现较大偏差. 这是因为基于启发式学习的神经网络存在过学习 (即泛化) 能力不够, 预测精度不高等缺点. 这种



(a)　Lorenz 时间序列的 LSSVM 模型预测误差



(b)　Lorenz 时间序列的 RBF 模型预测误差



(c)　Santa Fe Laser 数据的 LSSVM 和 RBF 模型预测误差

图 2　两种预测模型的 NRMSE 分布柱状图

现象随着 P 的增大, 表现越明显. 相比之下, 基于统计学习理论的LSSVM 预测模型克服了神经网络的上述缺点. 但这两者的差别在第 2 组的预测中没有明显表现出来. 这是因为该数据本身是含有噪声, 而噪声对预测结果的影响有可能掩盖了模型之间的上述差异. 尽管如此, LSSVM 预测模型的预测结果与原始值吻合程度较RBF 网络模型好.

图 2 为第 1 和第 2 组数据的两种预测模型 P = 1～ 24 的NRMSE 分布柱状图. 对于Lorenz 时间序列, 随着 P 增大, 这两种模型的NRMSE 将呈指数形式增大, 分别可用 $\mathrm{NRMSE} = 0.000\,062\,04\,e^{(1.714P)}$ 和 $\mathrm{NRMSE} = 0.001\,708\,e^{(1.393P)}$ 表达, 这在 $\alpha = 0.05$ 显著性水平下是高度显著的, 它们的复相关系数分别达到 0.966 6 和 0.976 7. 这种误差增长的方式恰好是混沌系统最重要的特征之一. 这说明所估计的最优嵌入参数是合理的, 基于重构的相空间所建立的模型能够捕捉到Lorenz 混沌系统内在的动力学特性. 注意到图 2(a) 和 2(b) 纵坐标的值以及上述两个预测误差增长的表达式, 可以发现LSSVM 预测模型的NRMSE 比RBF 网络预测模型的NRMSE 小 1～ 2 个量级.

对于含有噪声的 Santa Fe Laser 数据(图 2(c)), 随着 P 增大, 这两种模型的NRMSE 呈指数形式增大不明显. 这是因为噪声的缘故, 在一定程度上, 噪声一方面影响了最优嵌入参数的估计, 另一方面直接影响了预测模型的预测精度. 但有一点很明显, LSSVM 预测模型的NRMSE 比RBF 网络预测模型的NRMSE 小得多, 而且这种差异随 P 的增大愈明显, 如当 $P \geq 22$ 时, LSSVM 预测模型的NRMSE 比RBF 网络预测模型的NRMSE 小 1 个量级.

4 结 语

对于给定的混沌时间序列, 用最优嵌入参数 m 和 τ 重构原混沌系统的的相空间来近似原系统状态空间, 并在该相空间中, 用最小二乘支持向量机(LSSVM)建立多步预测模型. 结果表明, 将相空间重构理论和统计学习理论结合起来所建立的预测模型能够捕捉到原混沌系统的动力学特征. LSSVM 预测模型的鲁棒性较强, 其预测性能远远优于常用的RBF 网络预测模型. 这对进一步解决实际问题(如大气污染物浓度、气象气候数据和水文等混沌时间序列的分析及其提前多步预测预报)具有很好的应用前景.

参考文献(References)

[1] Cortes C, Vapnik V N. Support Vector Networks[J]. *Machine Learning*, 1995, 20(3): 273-295.

[2] Mukherjee S, Osuna E, Girosi F. Nonlinear Prediction of Chaotic Time Series Using Support Vector Machines [A]. *Proc of the IEEE Workshop on Neural Networks for Signal Processing* [C]. Ameliz Island, 1997: 511-520.

[3] Suykens J A K, Gestel T V, Brahanter J D, et al. *Least Squares Support Vector Machines* [M]. River Edge: World Scientific, 2002: 71-148.

[4] 崔万照, 朱长纯, 保文星, 等. 混沌时间序列的支持向量机预测[J]. ***物理学报***, 2004, 53(10): 3303-3310.
(Cui W Z, Zhu C C, Bao W X, et al. Prediction of the Chaotic Time Series Using Support Vector Machines [J]. *Acta Physica Sinica*, 2004, 53(10): 3303-3310.)

[5] Grassberger P, Procaccia I. Characterization of Strange Attractors[J]. *Phys Rev Let*, 1983, 50(5): 346-349.

[6] Cao L Y. Practical Method for Determining the Minimum Embedding Dimension of a Scalar Time Series [J]. *Physica D*, 1997, 110(1-2): 43-50.

[7] Gautama T, Mandic D P, Van Hulle M M. A Differential Entropy Based Method for Determining the Optimal Embedding Parameters of a Signal[A]. *Proc of the Int Conf on Acoustics, Speech and Signal Processing* [C]. Hong Kong, 2003, 6: 29-32.

[8] Schreiber T, Schmitz A. Surrogate Time Series [J]. *Physica D*, 2000, 142(3/4): 346-382.

[9] Arbarnel H D I, Brown R, Kadtke J B. Prediction in Chaotic Nonlinear Systems: Methods for Time Series with Broadband Fourier Spectra[J]. *Physical Review A*, 1990, 41(4): 1782-1807.

[10] Huebner U, Abraham N B, Weiss C O. Dimensions and Entropies of Chaotic Intensity Pulsations in a Single-mode Far-infrared NH3 Laser [J]. *Physical Review A*, 1989, 40(11): 6354-6365.

[11] Pelckmans K, Suykens J A K, Gestel T V. *LS-SVM lab Toolbox User's Guide* [EB/OL]. http://www.esat.kuleuven.ac.be/sista/lssvmlab/tutorial/tutorial1-5,po4, 2003, 1-11.