第33卷 第17期 计 算 机 工 程 2007年9月
Vol.33 No.17 Computer Engineering September 2007
·博士论文· 文章编号：1000—3428(2007)17—0023—03 文献标识码：A 中图分类号：TP391

基于核方法的非线性时间序列预测建模

林树宽，乔建忠，王国仁，郑 刚，董 俊
(东北大学信息科学与工程学院，沈阳 110004)

摘 要：提出了一种基于核的非线性时间序列预测建模方法。对非线性时间序列的相空间进行重构以确定其嵌入维数，并提出一种基于核主成分分析的非线性时间序列相空间重构方法，针对时间序列的时序特征，采用一种加权的支持向量回归模型对时间序列预测建模。在不同基准数据集上的实验结果表明，与通常的基于普通支持向量回归的建模方法相比，该文所提出的预测建模方法具有较高的精度，说明所提方法对非线性时间序列的预测建模是有效的。

关键词：核主成分分析；支持向量回归；相空间重构；时间序列建模

Nonlinear Time Series Prediction Modeling Based on Kernel Method

LIN Shu-kuan, QIAO Jian-zhong, WANG Guo-ren, ZHENG Gang, DONG Jun
(College of Information Science and Engineering, Northeastern University, Shenyang 110004)

【Abstract】 The paper presents a kernel based prediction modeling method for nonlinear time series. Phase space of time series is reconstructed to get its embedding dimension, and a method of phase space reconstruction based on kernel principal component analysis(KPCA) is proposed. A weighted support vector regression(SVR) is adopted to set up prediction model according to the characteristics of time series. The experimental results on different benchmark data show that the model based on the proposed method has higher accuracy compared with normal SVR model, proving the efficiency of the method for nonlinear time series prediction modeling.

【Key words】 kernel principal component analysis(KPCA); support vector regression(SVR); phase space reconstruction; time series modeling

1 概述

时间序列预测建模广泛应用于各行业，如工业过程控制、经济和财政数据处理、网络流量分析等。实际生产中的数据一般具有非线性特征。对于非线性时间序列建模， 一方面要考虑其非线性特征，另一方面还要考虑时间序列固有的时序特征。为了确定建模所需的合理输入特征，在建立时间序列模型之前，首先对其进行相空间重构，将时间序列标量重构为能反映其相空间状态的多维特征向量。重构相空间的关键在于确定嵌入维数m和时间延迟τ(如果不考虑时间序列的混沌特性，可将时间延迟τ设置为1。本文主要研究如何确定嵌入维数m)。在Takens定理中，对于无限长度和无噪音的理想的一维时间序列，m可为任意值[1]，但实际应用中的时间序列往往具有有限长度和噪音，因此必须仔细确定嵌入维数m，否则会严重影响建模质量。

相空间重构的传统方法为坐标延迟方法。在坐标延迟法中，可采用G-P算法、FNN方法和CAO方法等来计算嵌入维数m[1]，但是经这样重构出来的相空间，特征向量间的相关性将使它们的信息相互重叠或抵消，从而影响建模的质量。

用主成分分析方法进行相空间重构，可以保证重构的相空间中各个特征之间不相关，但PCA本质上是一种线性映射方法，不能体现出数据之间的非线性关系，因此用PCA方法不能重构出高质量的非线性时间序列的相空间。虽然有一些改进的PCA方法(如文献[1]中提出的MWPCA)，但大多是以局部的线性变换逼近全局的非线性，因此，不能从根本上改变PCA方法的缺陷。

针对目前已有的时间序列相空间重构方法所存在的问题，本文提出一种基于核主成分分析(kernel principal component analysis，KPCA)的相空间重构方法，一方面使得相空间中的各特征之间不相关，消除信息重叠或抵消的情况；另一方面能很好地反映非线性数据的特点，以提高相空间以及模型的质量。

支持向量机(support vector machine，SVM)是近年来机器学习领域研究的热点，因其具有全局最优解和较好的泛化推广能力而被广泛用于解决分类、回归及预测等问题[2,3]。本文针对时间序列所具有的时序特征，采用一种加权的支持向量回归模型对时间序列建模，以进一步提高模型精度。

2 非线性时间序列相空间重构

2.1 初步的时间序列相空间重构

设时间序列为$\{x_1, x_2, \ldots, x_M\}$，首先采用坐标延迟法对时间序列的相空间进行初步的重构。根据G-P算法[4]计算嵌入维数m，并重构特征向量$\mathbf{x}_t=(x_{t-1}, x_{t-2}, \ldots, x_{t-m})$。基于初步的相空间建模，可以如下构造模型的输入输出样本对：

$$\boldsymbol{X}=\begin{bmatrix}\mathrm{x}_1\\ \mathrm{x}_2\\ \vdots\\ \mathrm{x}_l\end{bmatrix}=\begin{bmatrix}x_1 & x_2 & \cdots & x_m\\ x_2 & x_3 & \cdots & x_{m+1}\\ \vdots & \vdots & \ddots & \vdots\\ x_l & x_{l+1} & \cdots & x_{l+m-1}\end{bmatrix}\quad \boldsymbol{Y}=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_l\end{bmatrix}=\begin{bmatrix}x_{m+1}\\ x_{m+2}\\ \vdots\\ x_{l+m}\end{bmatrix} \tag{1}$$

其中，l为样本数目。

建立时间序列模型即是找到自相关的输入变量$\mathbf{x}_t$与y_t之

基金项目：辽宁省自然科学基金资助项目(20042015)；沈阳市自然科学基金资助项目(1041036-1-06-07)

作者简介：林树宽(1966－)，女，副教授、博士研究生，主研方向：机器学习，人工智能；乔建忠、王国仁，教授、博士生导师；郑 刚、董 俊，硕士研究生

收稿日期：2006-09-10 **E-mail：**linshukuan@ise.neu.edu.cn

间的函数$f: \mathrm{R}^m \to \mathrm{R}$，使得$y_t=f(\mathbf{x}_t)$。

2.2 基于核主成分分析的相空间重构

基于核的KPCA首先通过一个非线性映射 $\Phi: \mathrm{R}^N \to F$ 将原始输入空间中的数据映射到一个新的特征空间F，然后在特征空间中执行通常的PCA。核方法的优势在于具有较强的非线性处理能力，且不必知道非线性映射Φ的具体形式，只需定义并计算下面的核函数[2]：

$$K(\mathrm{x}_i, \mathrm{x}_j) = K_{ij} = \Phi(\mathrm{x}_i) \cdot \Phi(\mathrm{x}_j) \tag{2}$$

其中，$\mathrm{x}_i, \mathrm{x}_j$是输入空间中的变量。

设$\mathbf{x}_k \in \mathrm{R}^N$，$k=1, 2, \ldots, l$ 是经初步的相空间重构后形成的时间序列样本输入，并假设其对应的特征空间F中的输入满足中心化条件[5]

$$\sum_{k=1}^{l} \Phi(\mathrm{x}_k) = 0 \tag{3}$$

则其协方差矩阵为

$$\overline{C} = \frac{1}{l} \sum_{j=1}^{l} \Phi(\mathrm{x}_j) \Phi(\mathrm{x}_j)^{\mathrm{T}} \tag{4}$$

特征空间中的 PCA 通过求解下面的方程得到特征值 λ 和特征向量 $V \in F \backslash \{0\}$

$$\lambda V = \overline{C} V \tag{5}$$

由于所有的特征向量V均可表示为$\Phi(\mathrm{x}_1), \Phi(\mathrm{x}_2), \ldots, \Phi(\mathrm{x}_l)$的线性张量，因此有

$$\lambda (\Phi(\mathrm{x}_k) \cdot V) = (\Phi(\mathrm{x}_k) \cdot \overline{C} V) \quad \text{for all } k=1,2, \ldots, l \tag{6}$$

并存在系数α_i ($i=1,2, \ldots, l$)，满足

$$V = \sum_{i=1}^{l} \alpha_i \Phi(\mathrm{x}_i) \tag{7}$$

由式(3)、式(6)、式(7)可得到下式：

$$l\lambda\alpha = K\alpha \tag{8}$$

由式(8)可求得矩阵K / l的特征值和特征向量，令$\lambda_1 \le \lambda_2 \le \ldots \le \lambda_l$表示其特征值(KPCA是基于样本的，核主成分的数目最大可达到样本数目l)，$\alpha^1, \alpha^2, \ldots, \alpha^l$表示相应的特征向量，$\lambda_p$是第 1 个非零的特征值，则可通过对特征空间$F$中的特征向量$V$的标准化

$$(V^k \cdot V^K) = 1 \quad \text{for all } k = p, \ldots, l \tag{9}$$

得到向量$\alpha^p, \ldots, \alpha^l$的标准化条件，即

$$\begin{aligned} 1 &= \sum_{i,j=1}^{l} \alpha_i^k \alpha_j^k (\Phi(x_i) \cdot \Phi(x_j)) \\ &= \sum_{i,j=1}^{l} \alpha_i^k \alpha_j^k K_{ij} = (\alpha^k \cdot K\alpha^k) = l\lambda_k (\alpha^k \cdot \alpha^k) \end{aligned} \tag{10}$$

设x是原始空间的一个输入数据，其在特征空间F中的映像为Φ(x)，则它所对应的各核主成分分别是在V^k ($k = p, \ldots, l$)上的投影，即

$$(V^k \cdot \Phi(\mathrm{x})) = \sum_{i=1}^{l} \alpha_i^k (\Phi(\mathrm{x}_i) \cdot \Phi(\mathrm{x})) \tag{11}$$

通常，中心化的假设条件 $\sum_{k=1}^{l} \Phi(\mathrm{x}_k) = 0$ 并不成立，对于任意的输入向量x_i(i=1, 2, …, l) 和映射Φ

$$\tilde{\Phi}(\mathrm{x}_i) = \Phi(\mathrm{x}_i) - \frac{1}{l} \sum_{i=1}^{l} \Phi(\mathrm{x}_i) \tag{12}$$

一定是被中心化的。因此，可通过矩阵 K 求得矩阵 $\tilde{K}$

$$\tilde{K}_{ij} = (\tilde{\Phi}(\mathrm{x}_i) \cdot \tilde{\Phi}(\mathrm{x}_j)) = K_{ij} - \frac{1}{l} \sum_{p=1}^{l} K_{ip} - \frac{1}{l} \sum_{q=1}^{l} K_{qj} + \frac{1}{l^2} \sum_{p,q=1}^{l} K_{pq} \tag{13}$$

详细的核主成分分析过程可参考文献[5]。

2.3 形成模型的输入、输出样本对

经过初步的相空间重构形成的输入样本x_i ($i=1,2,\ldots,l$)，经过核主成分分析后，可求得各核主成分$\boldsymbol{kprin}^i$(i=1, 2,…, l-q+1)。设 $kprin_j^i$ 表示第j个样本的第i个主成分，则

$$\boldsymbol{Kprin}^i = (kprin_1^i, kprin_2^i, \ldots, kprin_l^i)^{\mathrm{T}}$$
$$\text{for } i=1, 2, \ldots, l\text{-}q+1 \tag{14}$$

选择前 m 个主成分，使得它们的累计贡献率足够大，进一步形成时间序列的相空间，则可确定其嵌入维数为 m。此时，建立时间序列模型的输入、输出样本对如下：

$$\begin{aligned} \tilde{\boldsymbol{X}} &= \begin{bmatrix} \tilde{\mathrm{x}}_1 \\ \tilde{\mathrm{x}}_2 \\ \vdots \\ \tilde{\mathrm{x}}_l \end{bmatrix} = \begin{bmatrix} kprin_1^1 & kprin_1^2 & \ldots & kprin_1^m \\ kprin_2^1 & kprin_2^2 & \ldots & kprin_2^m \\ \vdots & \vdots & \ddots & \vdots \\ kprin_l^1 & kprin_l^2 & \ldots & kprin_l^m \end{bmatrix} \\ \boldsymbol{Y} &= \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_l \end{bmatrix} = \begin{bmatrix} x_{m+1} \\ x_{m+2} \\ \vdots \\ x_{l+m} \end{bmatrix} \end{aligned} \tag{15}$$

基于KPCA相空间重构方法建立时间序列模型，即寻找输入 $\tilde{X}$ 和输出 Y 之间的函数 $\tilde{f}$，使得$\mathrm{y}_i = \tilde{f}(\tilde{\mathrm{x}}_i)$。

3 基于加权的支持向量回归建立时间序列模型

经过相空间重构，重新确定了时间序列建模的输入、输出样本对。时间序列建模的特殊之处在于数据具有时序特征，不同时刻的样本对于建模所起的作用不同，因此，本文采用加权的支持向量回归建立时间序列模型，对不同时刻的样本所对应的参数给予不同的权值，即改进传统的支持向量回归的目标函数，通过求解下面的目标函数得到最终的模型函数 $f(x) = w^T\varphi(x) + b$[6]，

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|w\|^2 + \sum_{i=1}^{l} C_i(\xi_i + \xi_i^*) \\ &\text{subject to} \quad -\varepsilon - \xi_i^* \le (w \cdot \phi(\mathrm{x}_i) + b) - y_i \le \varepsilon + \xi_i, \\ &\qquad \xi_i, \xi_i^* \ge 0, \quad i = 1, 2, \ldots, l \end{aligned} \tag{16}$$

其中，惩罚系数C随不同的样本被赋予不同的权值[6,7]，

$$C_i = C_0 \frac{2}{1 + \exp(a - 2a \times \frac{i}{l})} \tag{17}$$

其中，C_0为初始的惩罚系数取值；a是权值调整参数；l为样本总数。i=l是最近的样本数据；i=l是最远的样本数据。从式(17)中可以看出，对于较近的样本赋予的权值较大，相反，较远的样本赋予的权值较小。

为求解目标函数式(16)，可将其转化为对偶形式

$$\begin{aligned} &\text{maximize} \quad \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) y_i - \varepsilon \sum_{i=1}^{l} (\alpha_i^* + \alpha_i) - \\ &\frac{1}{2} \sum_{i,j=1}^{l} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) K(\mathrm{x}_i, \mathrm{x}_j) \\ &\text{subject to} \quad \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) = 0, \quad 0 \le \alpha_i, \alpha_i^* \le C_i, \quad i = 1, 2, \ldots, l \end{aligned} \tag{18}$$

其中，α_i 和 α_i^* 是拉格朗日乘子；$K(\mathrm{x}_i,\mathrm{x}_j)$是核函数。通过求解目标函数式(18)，得到 α_i 和 α_i^* 的值，进而求出w 和b的值以及最终的模型函数$f(x) = w^T\varphi(x) + b$。

4 实验结果及分析

本文提出一种基于核主成分分析进行相空间重构以及采用加权的支持向量回归的非线性时间序列建模方法，为了验证该方法的有效性，分别采用太阳黑子活动数据集、时间序列数据集D[8]等进行实验。下面给出太阳黑子活动数据预测建模的情况。

太阳黑子活动数据[9]是典型的非线性时间序列，在此，选用 1724 年～1803 年共 80 年的太阳黑子活动数作为数据集建立模型，并将前 50 个数据作为训练集，后 30 个数据作为

测试集。为处理数据方便，原始数据x按照下面的公式进行标准化：

$$x = \frac{x-\mu}{\sigma} \tag{19}$$

其中，μ, σ 分别是样本数据的均值和标准差。

为了进行比较，分别采用以下 2 种方案进行实验：(1)采用传统的坐标延迟法进行相空间重构，并采用普通的支持向量回归建立模型；(2)基于核主成分分析进行相空间重构，并采用第 3 节中所述的加权支持向量回归建立模型。

在基于核主成分分析进行相空间重构的过程中，采用了下面的二阶多项式核函数：

$$K(\mathrm{x}_i, \mathrm{x}_j) = (\mathrm{x}_i \cdot \mathrm{x}_j + 1)^2 \tag{20}$$

在加权的支持向量回归建模的过程中，采用了改进的交叉验证方法[10]选择初始的惩罚系数C_0以及不敏感损失ε。

图 1 和图 2 分别显示了方案 1 和方案 2 对于太阳黑子数据集中前 50 个训练数据的拟合结果和后 30 个测试数据的预测结果。从图中可以看出，方案 2 的拟合和预测结果更接近于真实曲线。这是由于方案 2 通过对非线性时间序列进行基于核主成分分析的相空间重构，消除了模型的输入特征之间信息重叠或抵消的情况，使得模型的输入更加合理，同时，方案 2 充分考虑了时序数据的特征，根据不同时刻的历史数据对输出值的影响不同给它们赋以不同的权值，使得所建模型更能反映时间序列的变化规律。其他数据集上的实验结果与此相似。

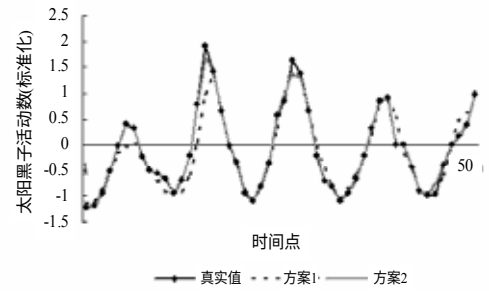

图 1　两种方案对太阳黑子前 50 个数据的拟合结果

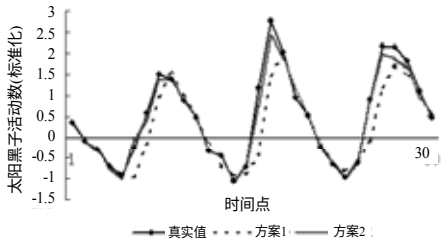

图 2　两种方案对太阳黑子后 30 个数据的预测结果

文中采用以下的误差计算公式评价模型的准确性：

$$AAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_i^*\right| \tag{22}$$

其中，n 是建模所用的样本数目，y_i 和 y_i^* 分别是样本的真实值和模型输出值。

2 种方案的拟合和预测误差如表 1 所示。

表 1　2 种方案的模型误差

	拟合	预测
方案 1	0.156 572	0.309 569
方案 2	0.026 383	0.054 777

5　结论

本文提出了一种基于核的非线性时间序列预测建模方法，首先基于核主成分分析对非线性时间序列进行相空间重构，确定建模的输入特征，然后采用加权的支持向量回归建立模型。基于核主成分分析的相空间重构方法能够很好地体现数据间的非线性关系以及时间序列本身的变化规律，使得相空间的质量大大提高，进而提高建模精度；加权的支持向量回归模型能较好地体现时间序列的时序特征，进一步提高了模型精度。相关的实验结果说明了本文所提方法的有效性。

参考文献

1 陈　铿, 韩伯棠. 混沌时间序列分析中的相空间重构技术综述[J]. 计算机科学, 2005, 32(4): 67-70.

2 Vapnik V N. The Nature of Statistical Learning Theory[M]. New York: Springer Verlag, 1995.

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

4 Grassberger P, Procaccia I. Measuring the Strangeness of Strange Attractors[J]. Physica D, 1983, 9(1/2): 189-208.

5 Scholkopf B, Smola A. Nonlinear Component Analysis as a Kernel Eigenvalue Problem[J]. Neural Computation, 1998, 10(5): 1299.

6 Cao L J, Tay F E H. Support Vector Machine with Adaptive Parameters in Financial Time Series Forecasting[J]. IEEE Transactions on Neural Networks, 2003, 14(6): 1506-1518.

7 杜树新, 吴铁军. 用于回归估计的支持向量机方法[J]. 系统仿真学报, 2003, 15(11): 1580-1585.

8 Wakuya H, Shida K. Time Series Prediction by a Neural Network Model Based on Bi-directional Computation Style: a Study on Generalization Performance with the Computer-generated Time Series Data Set D[J]. Systems and Computers in Japan, 2003, 34(10): 64-75.

9 陈兆国. 时间序列的谱分析[M]. 北京: 科学出版社, 1988.

10 林树宽, 张少敏, 支力佳. 一种面向时间序列预测的支持向量回归模型参数选择方法[J]. 小型微型计算机系统, 2005, 26(增刊): 268-271.

（上接第 22 页）

3 Kay J, Pasquale J. Measurement, Analysis, and Improvement of UDP/IP Throughput for the DECstation 5000[C]//Proceedings of the Winter USENIX Conference. 1993-01.

4 Kay J, Pasquale J. The Importance of Non-data Touching Processing Overhead in TCP/IP[C]//Proceedings of the ACM SIGCOMM'93 Symposium. 1993-09: 259-268.

5 Nahum E M. Networking Support for High-performance Servers[D]. Amherst MA: University of Massachusetts, 1997.