文章编号: 1001-0920(2005)03-0290-04

多变量时间序列相空间重构中参数的确定

岳毅宏, 韩文秀, 程国平
(天津大学 管理学院, 天津 300072)

摘 要: 介绍了多变量时间序列相空间重构理论. 提出一种新的基于平均预测误差最小化的重构参数确定方法, 阐述了该方法的算法过程及一些重要特点. 此方法考虑了所有重构参数对平均预测误差的影响, 能够同时确定重构系统相空间所需的恰当嵌入维数及时间延迟. 最后将该方法应用于股票市场非线性动力系统的相空间重构, 通过比较和分析验证了其优越性.

关键词: 多变量时间序列; 相空间重构; 嵌入维数; 时间延迟; 平均预测误差
中图分类号: TP13 **文献标识码**: A

Determination of parameters in the phase-space reconstruction of multivariate time series

YUE Yi-hong, HAN Wen-xiu, CHENG Guo-ping
(Management School, Tianjin University, Tianjin 300072, China. Correspondent: YUE Yi-hong, E-mail: kevinyue@eyou.com)

Abstract: Phase-space reconstruction theory of multivariate time series is introduced. A novel method of determining reconstruction parameters based on the minimization of average forecasting error is proposed, and its algorithm procedure and some important characteristics are explicated. The method considers the effects of all reconstruction parameters on average forecasting error. It can simultaneously determine the correct embedding dimensions and time delays needed for better reconstructing systematic phase-space. The method is applied to the phase-space reconstruction of nonlinear dynamical system of stock market. Through comparing and analyzing, its superiority is verified.

Key words: multivariate time series; phase-space reconstruction; embedding dimension; time delay; average forecasting error

1 引 言

嵌入理论的提出为时间序列的分析奠定了坚实的理论基础[1]. 从理论上讲, 只要时间延迟与嵌入维数选择恰当, 单变量时间序列便足以重建原始系统的动力学特性. 但实践中却存在意外的情形, 比如Lorenz方程[2], 对z轴输出数据进行采样而形成的时间序列就不能重构Lorenz系统的动力学特性. 这是因为该系统x轴与y轴之间的对称关系没有反映到z轴数据中. 因此, 认为任何给定的单维时间序列就足以重建原始系统动力学特性的观点并不总是正确的. 对于很多系统(如经济系统), 其可测输出序列往往是多维的, 在这种情况下, 可考虑利用目标系统的多个输出变量序列来重建系统的动力学特性.

文献[3,4]对多变量时间序列的相空间重构进行理论和应用研究, 并论证了其优越性. 其中文献[3]提出一种基于"零级近似"模型[5]的嵌入维数确定方法, 其基本思想是恰当的嵌入维数能使平均预测误差值达到最小. 因此, 当平均预测误差最小时, 所对应的嵌入维数就是恰当的嵌入维数. 但该方法存在的不足之处在于: 1) 没有考虑到时间延迟对平

收稿日期: 2003-08-26; **修回日期**: 2003-11-13.
基金项目: 国家自然科学基金项目(79970043).
作者简介: 岳毅宏(1975—), 男, 河南林州人, 副教授, 博士, 从事混沌控制、复杂系统等研究; 韩文秀(1938—), 女, 山东济南人, 教授, 博士生导师, 从事复杂系统、混沌系统及控制等研究.

均预测误差的影响; 2) 在计算平均预测误差时, 没有将所有变量都考虑在内

基于此, 本文对以上方法进行改进和扩展, 并将改进和扩展后的方法称为基于平均预测误差最小化的相空间重构参数确定法 相对于基于"零级近似"模型的嵌入维数确定法, 该方法具有明显的优越性

2 多变量时间序列的相空间重构

假定某系统有M 个可测变量$X_1, X_2, \cdots, X_M$, 对应于每个变量X_i 的时间序列为$\{x_{i,j}\}$, $j = 1, 2, \cdots, N$. 首先将序列$\{x_{i,j}\}$ 合并, 形成序列Y, 即

$$Y = \{x_{1,1}, x_{1,2}, \cdots, x_{1,N}, x_{2,1}, x_{2,2}, \cdots, x_{2,N}, \cdots, x_{M,1}, x_{M,2}, \cdots, x_{M,N}\}. \tag{1}$$

重构序列Y 的相空间

$$\begin{aligned} V_n = (&x_{1,n}, x_{1,n-\tau_1}, \cdots, x_{1,n-(d_1-1)\tau_1}, \\ &x_{2,n}, x_{2,n-\tau_2}, \cdots, x_{2,n-(d_2-1)\tau_2}, \cdots, \\ &x_{M,n}, x_{M,n-\tau_M}, \cdots, x_{M,n-(d_M-1)\tau_M}), \end{aligned} \tag{2}$$

$$n = \max_{1\le i\le M}\{(d_i - 1)\tau_i\} + 1, \max_{1\le i\le M}\{(d_i - 1)\tau_i\} + 2, \cdots, N.$$

式中τ_i和$d_i (i = 1, 2, \cdots, M)$ 分别表示时间延迟和嵌入维数 根据嵌入定理, 存在d 维空间上的一个光滑函数$F: \mathbf{R}^d \to \mathbf{R}^d \left(其中 d = \sum_{i=1}^{M} d_i\right)$, 使得

$$V_{n+1} = F(V_n). \tag{3}$$

若d 或每个d_i 都充分大, 则式(3) 可写成

$$\begin{aligned} x_{1,n+1} &= F_1(V_n), \\ x_{2,n+1} &= F_2(V_n), \\ &\vdots \\ x_{M,n+1} &= F_M(V_n). \end{aligned} \tag{4}$$

对于以上过程, 关键问题在于选定恰当的时间延迟τ_i及嵌入维数d_i, $i = 1, 2, \cdots, M$, 这样才能使得式(2) 和(3) 成立

3 基于平均预测误差最小化的重构参数确定法

选择一个好的时间延迟τ_i 非常重要, 因为它可有效降低所需的重构嵌入维数, 从而使问题相对简化 对于单变量序列, 选择时间延迟的方法很多, 比如平均位移法[6]、互信息法[7]、自相关函数法[8] 等 文献[3] 就是利用这些方法分别对多变量时间序列中的每个单变量时间序列$X_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,N})$ $(i = 1, 2, \cdots, M)$ 单独选取时间延迟τ_i; 然后利用"零级近似"模型求取恰当的嵌入维数d_i, $i = 1, 2, \cdots, M$.

在求取时间延迟τ_i 时, 一般要求首先设定嵌入维数d_i; 而在求取嵌入维数d_i 时, 也要求首先知道时间延迟τ_i, 这样就形成了一对矛盾 通常情况下, 在求取τ_i 值之前, 必须假定d_i 的取值, 但该值并不一定就是恰当的嵌入维数值 这必然会影响恰当时间延迟的计算, 进而影响到相空间的重构质量 基于此, 本文在文献[3] 的基础上, 提出一种基于平均预测误差最小化的相空间重构参数确定法 该方法能够同时求得恰当的时间延迟和嵌入维数, 而且所求得的重构参数值更为优化

由式(4) 可以得到变量X_i 的未来状态$x_{i,n+1}$ 与相空间矢量V_n 之间的函数关系, 但实践中函数$F_t(\bullet)$ $(t = 1, 2, \cdots, M)$ 往往是未知的 为了对变量X_i 的未来状态作出预测, 必须选用适当的模型来逼近函数$F_t(\bullet)$. 目前已有的逼近模型很多, 比如线性回归模型、径向基函数模型、人工神经网络模型、小波网络模型等 在此设函数$F_t(\bullet)$ 的逼近模型为$\widetilde{F}_t(\bullet)$, 则式(4) 变为

$$\begin{aligned} \hat{x}_{1,n+1} &= \widetilde{F}_1(V_n), \\ \hat{x}_{2,n+1} &= \widetilde{F}_2(V_n), \\ &\vdots \\ \hat{x}_{M,n+1} &= \widetilde{F}_M(V_n). \end{aligned} \tag{5}$$

式中$\hat{x}_{i,n+1} (i = 1, 2, \cdots, M)$ 为$x_{i,n+1}$ 的预测值

文献[3] 中所构造的误差函数$E(d_1, d_2, \cdots, d_M)$ 只是针对变量X_1, 最终所求得的恰当嵌入维数值也只适合于由变量X_1 来重构相空间, 并不一定适用于其他变量 基于此, 本文将构造针对所有变量的误差函数$E(d_1, d_2, \cdots, d_M; \tau_1, \tau_2, \cdots, \tau_M)$. 由于变量的单位各异, 与误差函数$E(d_1, d_2, \cdots, d_M)$ 采用平均绝对误差不同, 本文所构造的误差函数采用平均相对误差

$$E(d_1, d_2, \cdots, d_M; \tau_1, \tau_2, \cdots, \tau_M) = \frac{1}{M(N - J_0)} \sum_{i=1}^{M} \sum_{n=J_0}^{N-1} \left| \frac{\hat{x}_{i,n+1} - x_{i,n+1}}{x_{i,n+1}} \right|. \tag{6}$$

式中$\hat{x}_{i,n+1}$ 为$x_{i,n+1}$ 的预测值, 且

$$J_0 = \max_{1\le i\le M}(d_i - 1)\tau_i + 1. \tag{7}$$

由式(6) 可知, 误差E 值的大小完全取决于维数$d_1, d_2, \cdots, d_M$ 及时间延迟$\tau_1, \tau_2, \cdots, \tau_M$, 能使$E$ 取值最小的嵌入维数$d_1^e, d_2^e, \cdots, d_M^e$ 及时间延迟$\tau_1^e, \tau_2^e, \cdots, \tau_M^e$, 就是所求的恰当嵌入维数及时间延迟, 即

$$\begin{aligned} &(d_1^e, d_2^e, \cdots, d_M^e; \tau_1^e, \tau_2^e, \cdots, \tau_M^e) = \\ &\arg\min\{E(d_1, d_2, \cdots, d_M; \tau_1, \tau_2, \cdots, \tau_M): \\ &d_i, \tau_i \in Z^+, i = 1, 2, \cdots, M\}. \end{aligned} \tag{8}$$

式中Z^+ 表示所有正整数的集合

关于以上确定嵌入维数及时间延迟的方法, 需要对以下几点作进一步说明:

1) 在式(5) 中, $\widetilde{F}_1(\bullet), \widetilde{F}_2(\bullet), \cdots, \widetilde{F}_M(\bullet)$ 可采用

多种不同的模型,比如零级近似模型、全域线性回归模型、局域线性回归模型、人工神经网络模型、小波网络模型等. 一般情况下,为了便于计算,$\widetilde{F}_t(\cdot)$ ($t = 1, 2, \cdots, M$) 通常采用同一模型,只是模型参数不同而已. 需要指出的是,若选取的逼近模型不同,则计算量及函数 $E(d_1, d_2, \cdots, d_M; \tau_1, \tau_2, \cdots, \tau_M)$ 的最小值可能存在较大差异,但最终所求得的 $(d_1^e, d_2^e, \cdots, d_M^e)$ 及 $(\tau_1^e, \tau_2^e, \cdots, \tau_M^e)$ 的值应是相同的.

2) 当 $M = 1$ 时,目标序列为单变量时间序列,此时上述方法依然适用. 因此,该方法对单变量和多变量时间序列的相空间重构具有普遍的适用性.

3) 以上方法采用了单步预测来确定恰当嵌入维数 $(d_1^e, d_2^e, \cdots, d_M^e)$ 及恰当时间延迟 $(\tau_1^e, \tau_2^e, \cdots, \tau_M^e)$. 需要指出的是,效果良好的单步预测并不意味着可以取得效果良好的多步预测,因此可基于多步预测法来定义 E,即

$$E(d_1, d_2, \cdots, d_M; \tau_1, \tau_2, \cdots, \tau_M) = \frac{1}{M(N - T - J_0 + 1)} \sum_{i=1}^{M} \sum_{n=J_0}^{N-T} \left| \frac{\hat{x}_{i,n+T} - x_{i,n+T}}{x_{i,n+T}} \right|. \tag{9}$$

此外,E 也可采取以下形式:

$$E(d_1, d_2, \cdots, d_M; \tau_1, \tau_2, \cdots, \tau_M) = \frac{1}{M(N - T - J_0 + 1)} \sum_{i=1}^{M} \sum_{n=J_0}^{N-T} \max_{1 \le k \le T} \left| \frac{\hat{x}_{i,n+k} - x_{i,n+k}}{x_{i,n+k}} \right|. \tag{10}$$

式(9) 和(10) 中: T 为给定的预测步数, $\hat{x}_{i,n+T}$ 为 $x_{i,n+T}$ 的预测值. 式(9) 采用的是直接预测法, 式(10) 采用的是逐项迭代预测法. 若原始序列 Y 是混沌序列,则预测步数 T 应满足 $T \le T_{\max}$, $T_{\max}$ 为最大可预测尺度,其计算方法参见文献[9].

4) 通常情况下,为了控制计算量,一般预先设定最大嵌入维数 $d_{\max}$ 及最大时间延迟 $\tau_{\max}$,则式(8) 可改写为

$$(d_1^e, d_2^e, \cdots, d_M^e; \tau_1^e, \tau_2^e, \cdots, \tau_M^e) = \arg\min\{E(d_1, d_2, \cdots, d_M; \tau_1, \tau_2, \cdots, \tau_M):$$
$$d_i, \tau_i \in Z^+, d_i \le d_{\max}, \tau_i \le \tau_{\max}, i = 1, 2, \cdots, M\}. \tag{11}$$

5) 在一维情形下(即 $M = 1$),当逼近模型 $\widetilde{F}_t(\cdot)$ 采用零级近似模型时,上述方法与伪邻近点法(即 FNN 法) 具有明显的联系. 事实上,伪邻近点法可以直接扩展应用到多维的情形.

4 实例研究

研究结果表明,股票指数所反映的动力系统是一个具有低自由度的混沌系统,而且除了股票指数之外,该系统仍有多个输出变量是可测的,比如成交量、成交金额和换手率等. 因此,该系统非常适合于研究多变量时间序列的相空间重构.

文献[10] 介绍了一种相空间重构预测算法,该方法的原理是直接借助于相空间重构技术对时间序列的未来状态作出预测. 相空间重构质量的高低直接决定着预测的精度,而重构参数的恰当与否决定着相空间重构的质量. 基于此,本文利用该预测算法来检验上述重构参数确定方法相对于文献[3] 方法所具有的优越性.

本文选用上海证券交易所 1995-05-29～2001-12-31 期间每日的收市指数(X_1)、成交量(X_2) 以及成交金额(X_3) 组成的时间序列 $\{x_{1,i}\}$, $\{x_{2,i}\}$, $\{x_{3,i}\}$(样本总量 $N = 1\,600$),重构系统的相空间; 选用 2002-01-01～2002-01-18 期间每日的收市指数组成的时间序列 $\{y_i\}$ ($i = 1, 2, \cdots, 14$) 作为测试序列; 然后应用相空间重构算法来拟合测试序列 $\{y_i\}$. 通过对平均拟合误差进行比较,便可检验本文方法的优越性.

按照文献[3] 所述的重构参数确定方法,首先基于平均位移法[6] 分别求取序列 $\{x_{1,i}\}$, $\{x_{2,i}\}$, $\{x_{3,i}\}$ 的恰当时间延迟,得到的结果为 $\tau_1^e = 4$, $\tau_2^e = 2$, $\tau_3^e = 3$; 然后求取各个序列的恰当嵌入维数. 在此取初始嵌入维数 $d_1 = d_2 = d_3 = 1$,设定嵌入维数的最大值为 $d_{\max} = 15$. 计算结果表明,当 $d_1 = 6$, $d_2 = 4$, $d_3 = 7$ 时,误差函数 $E(d_1, d_2, d_3)$ 的值达到最小,因此恰当嵌入维数为 $d_1^e = 6$, $d_2^e = 4$, $d_3^e = 7$.

同理,按照本文方法求取恰当重构参数,取初始嵌入维数及时间延迟分别为 $d_1 = d_2 = d_3 = 1$, $\tau_1 = \tau_2 = \tau_3 = 1$,设定嵌入维数及时间延迟的最大值分别为 $d_{\max} = 15$, $\tau_{\max} = 10$. 此时式(11) 可变换为

$$(d_1^e, d_2^e, d_3^e; \tau_1^e, \tau_2^e, \tau_3^e) = \arg\min\{E(d_1, d_2, d_3; \tau_1, \tau_2, \tau_3):$$
$$d_1, d_2, d_3 = 1, 2, \cdots, 15; \tau_1, \tau_2, \tau_3 = 1, 2, \cdots, 10\}.$$

鉴于零级近似模型具有计算量较小的特点,本文选用该模型来逼近函数 $F_1(\cdot)$, $F_2(\cdot)$, $F_3(\cdot)$. 计算结果表明,当 $d_1 = d_3 = 5$, $d_2 = 4$, $\tau_1 = 6$, $\tau_2 = 3$, $\tau_3 = 5$ 时,误差函数值 $E(d_1, d_2, d_3; \tau_1, \tau_2, \tau_3)$ 最小,恰当嵌入维数及时间延迟的值分别为 $d_1^e = d_3^e = 5$, $d_2^e = 4$, $\tau_1^e = 6$, $\tau_2^e = 3$, $\tau_3^e = 5$.

基于以上所求得的两组重构参数,分别得到下列重构相空间:

$$V_n^{(1)} = (x_{1,n}, x_{1,n-4}, x_{1,n-8}, x_{1,n-12}, x_{1,n-16}, x_{1,n-20},$$
$$x_{2,n}, x_{2,n-2}, x_{2,n-4}, x_{2,n-6},$$
$$x_{3,n}, x_{3,n-3}, x_{3,n-6}, x_{3,n-9}, x_{3,n-12},$$
$$x_{3,n-15}, x_{3,n-18}),$$

$$V_n^{(2)} = (x_{1,n}, x_{1,n-6}, x_{1,n-12}, x_{1,n-18}, x_{1,n-24}, x_{2,n}, x_{2,n-3}, x_{2,n-6}, x_{2,n-9}, x_{3,n}, x_{3,n-5}, x_{3,n-10}, x_{3,n-15}, x_{3,n-20}).$$

为了验证以上所求得的两组重构参数的优劣性, 本文分别以相空间 $V_n^{(1)}$ 和 $V_n^{(2)}$ 为基础, 应用相空间重构预测算法来拟合测试序列 $\{y_i\}$, 并利用下式计算平均拟合误差

$$AE(V_n) = \frac{1}{14}\sum_{i=1}^{14}\left|\frac{\hat{y}_i - y_i}{y_i}\right|, \tag{15}$$

式中 V_n 表示重构相空间. 经计算得到 $AE(V_n^{(1)}) = 3.98\%$, $AE(V_n^{(2)}) = 1.17\%$. $AE(V_n^{(2)})$ 的值明显小于 $AE(V_n^{(1)})$, 其根本原因在于相空间 $V_n^{(2)}$ 的重构质量大大优于相空间 $V_n^{(1)}$, 而相空间的重构质量则完全取决于重构参数的选择. 以上过程验证了本文方法相对于文献[3]方法的优越性.

5 结　　语

本文首先介绍了多变量时间序列相空间重构理论; 然后提出一种基于平均预测误差最小化的相空间重构参数确定法. 该方法能够同时确定重构嵌入维数及时间延迟, 而且所求得的重构参数更为优化. 最后将所提出方法应用于股票市场非线性动力系统的相空间重构, 通过比较和分析, 验证了该方法的优越性.

参考文献(References)

[1] Takens F Mane. Detecting strange attractors in fluid turbulence[A]. *Dynamical Systems and Turbulence*[C]. Berlin: Springer, 1981: 21-80.

[2] Hastings S P, Troy W C. A shooting approach to chaos in the Lorenz equations [J]. *J of Differential Equations*, 1996, 127(1): 41-53.

[3] Cao Liangyue, Mees Alistair, Judd Kevin. Dynamics from multivariate time series[J]. *Physica D*, 1998, 121(1-2): 75-88.

[4] 陆婕, 顾圣士, 蒋馥. 多维时间序列在相空间重构中的应用[J]. *洛阳大学学报*, 2002, 17(2): 9-13.
(Lu J, Gu S S, Jiang F. Application of multivariate time series in constructing phase space [J]. *J of Luoyang University*, 2002, 17(2): 9-13.)

[5] 王永忠, 曾昭磐. 混沌时间序列的局域线性回归预测方法[J]. *厦门大学学报*, 1999, 38(4): 636-640.
(Wang Y Z, Zeng Z P. A local regression forecasting method of chaotic time series [J]. *J of Xiamen University*, 1999, 38(4): 636-640.)

[6] Rosensrein M T, Collins J J, Luca C. Reconstruction expansion as a geometry-based framework for choosing proper delay time[J]. *Physica D*, 1994, 73(1): 82-89.

[7] Fraser A M. Information and entropy in strange attractors[J]. *IEEE Trans on IT*, 1989, 35(2): 245-262.

[8] Abarbanel H D I, Brown R, Sidorowich J J, et al. The analysis of observed data in physical systems[J]. *Rev Mod Phy*, 1993, 65(4): 1331-1392.

[9] 岳毅宏, 韩文秀. 混沌系统可预测尺度研究[J]. *系统工程理论与实践*, 2003, 23(6): 91-95.
(Yue Y H, Han W X. Study on the predictable size of chaotic systems[J]. *Systems Engineering - Theory and Practice*, 2003, 23(6): 91-95.)

[10] 薛福珍, 韩怀中, 罗超. 结合混沌预测的改进的OGY控制方法[J]. *控制与决策*, 2002, 17(5): 536-540.
(Xue F Z, Han H Z, Luo C. Algorithm combining chaos prediction with improved OGY control method [J]. *Control and Decision*, 2002, 17(5): 536-540.)

(上接第 289 页)

[4] Livani M A, Kaiser J, Jia W J. Scheduling hard and soft real-time communication in a controller area network[J]. *Control Engineering Practice*, 1999, 7: 1515-1523.

[5] Walsh G C, Beldiman O, Bushnell L. Error encoding algorithms for networked control[A]. *Proc of the 38th Conf on Decision and Control*[C]. Phoenix, 1999, 5: 4933-4938.

[6] Pedreiras P, Almeida L. EDF message scheduling on controller area network [J]. *Computing and Control Engineering J*, 2002, 13(4): 163-170.

[7] Nolte T, Sjodin M, Hansson H. Server-based scheduling of the CAN bus[EB/OL], www.mrtc.mdh.se/publications/0534.pdf, May 7, 2003.

[8] Cena G, Valenzano A. Achieving round-robin access in controller area networks[J]. *IEEE Trans on Industrial Electronics*, 2002, 9(6): 1202-1213.