第 56 卷 第 6 期 2007 年 6 月
1000-3290/2007/56(06)/3152-06

物 理 学 报
ACTA PHYSICA SINICA

Vol.56,No.6,June,2007

非线性时间序列互信息与 Lempel-Ziv 复杂度的相关性研究

张佃中†

(中南大学数学科学与计算技术学院,长沙 410083)

(2006 年 8 月 23 日收到;2006 年 9 月 14 日收到修改稿)

为探究非线性动力学系统的互信息和复杂度的相关性,用 Logistic 映射、Lorenz 模型和心电 RR 间期的非线性时间序列作为实验数据,计算多分段延时互信息和多分段 Lempel-Ziv 复杂度以及它们之间的相关系数.结果表明这些序列的互信息和复杂度呈强负相关,对 Logistic 方程生成的 201 个序列的不同段互信息和不同段复杂度之间的相关系数绝对值都大于 0.9162,最大达 0.9923;对 94 个心电 RR 间期序列都大于 0.8555,最大达 0.9860.研究还发现互信息比复杂度能更敏感地表现出非线性动力系统的特征.

关键词:相关系数,互信息,Lempel-Ziv 复杂度,心电 RR 间期

PACC:0547,8700

1. 引 言

延时互信息与复杂性测度是描述非线性时间序列信息量的二个重要参数.互信息来源于信息理论,应用广泛[1—5].复杂性测度最初定义是由 Kolmogorov 于 1965 年提出的,表征为能够产生某一(0,1)序列所需的最短程序的比特数,后来由 Lempel 和 Ziv 等[6]给出了实现这种定义复杂度的具体算法,并称为 Lempel-Ziv 复杂度,广泛应用于非线性科学的研究中[7—11].Palus[12]研究了延时互信息与熵的相关性,得出了互信息减少斜率与熵成正相关的结论,因而互信息也常常用来度量一个时间序列的复杂度[4].然而,延时互信息与 Lempel-Ziv 复杂度直接关系的研究尚未见报道.本文用 Logistic 方程和 Lorenz 模型生成的非线性时间序列以及心率变异信号为实验数据,研究延时互信息与多分段复杂度的相互关系,得出了二者呈强负相关的结论.互信息的计算比复杂度所耗机时明显减少,在反映序列的细节方面更加敏感.

2. 互信息简介

2.1. 互信息的定义

设 X、Y 为二个信息系统,$X=\{x_1, x_2,\cdots,x_n\}$,$Y=\{y_1, y_2,\cdots,y_m\}$ 时的概率分别为 $p(x_i)$,$p(y_j)$,$(i=1,2,\cdots,n,\ j=1,2,\cdots,m)$. 其中 $\sum_i p(x_i)=1$,$\sum_j p(y_j)=1$. X,Y 构成二个信源,相应的信息熵为

$$H(X)=-\sum_{i=1}^{n}p(x_i)\mathrm{lb}p(x_i), \tag{1}$$

$$H(Y)=-\sum_{j=1}^{m}p(y_j)\mathrm{lb}p(y_j). \tag{2}$$

设由 X 与 Y 构成的联合信源 $XY=\{x_1y_1, x_1y_2,\cdots,x_ny_m\}$ 时的概率分布为

$$XY:\begin{bmatrix} x_1y_1, & x_1y_2, & \cdots, & x_ny_m \\ p(x_1y_1), & p(x_1y_2), & \cdots, & p(x_ny_m)\end{bmatrix}$$

联合信源 XY 的信息熵为

$$H(XY)=-\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_iy_j)\mathrm{lb}p(x_iy_j) \tag{3}$$

则信源 X 与 Y 的互信息定义为(以 bit 为单位)[1]

$$I(X,Y)=H(X)+H(Y)-H(XY). \tag{4}$$

† E-mail:zdz1962@sohu.com

互信息是对二个信源 X 与 Y 之间关联程度的度量指标[1,4]．它满足对称性与非负性，即有

$$I(X,Y) = I(Y,X), I(X,Y) \geqslant 0.$$

当 X,Y 完全相同时互信息为最大，当 X,Y 完全独立时互信息的值为零．

2.2. 延时互信息的计算

设一个已知时间序列为 $\{x_1,x_2,\cdots,x_n\}$，将其看作信源 X，该序列延时 $k(k\in N)$ 个样本点后所得序列 $\{x_{1+k},x_{2+k},\cdots,x_{n+k}\}$ 看作信源 Y，计算出 X 与 Y 之间的互信息称为延时互信息．将 X,Y 的值域从最小到最大均匀分割成 L 个区间，分别用 $1,2,\cdots,i,\cdots,L$ 和 $1,2,\cdots,j,\cdots,L$ 作为每个区间的标度．二个序列作为二维变量存在的区域被划分为 2^L 个不同的子区域，每个子区域可以用 $(i,j),i,j=1,2,\cdots,L$ 来标度．计算出这二个序列组成的二维变量落在区域 (i,j) 内的数据个数 $n(i,j)$，则相应于子区域 (i,j) 的联合概率定义为[13]

$$p(x_iy_j) = \frac{n(i,j)}{n},(i,j = 1,2,\cdots,L) \quad (5)$$

其中 n 为每个序列的数据总个数，且有

$$p(x_i) = \sum_{j=1}^{L}\frac{n(i,j)}{n} = \frac{n(i)}{n}, \quad (6)$$

$$p(y_j) = \sum_{i=1}^{L}\frac{n(i,j)}{n} = \frac{n(j)}{n}. \quad (7)$$

取延时点数 $k=1,2,\cdots,K$ 时，分别将(5)，(6)和(7)式的计算结果代入(1)—(4)式，即可得出一个时间序列的延时互信息序列．当 $k=0$ 时二个序列完全相同，此时的延时互信息值为最大，以该值对互信息序列进行归一化后得到一个值不超过 1 的互信息序列 $\{I_0,I_1,\cdots,I_K\}$．Logistic 映射的延时互信息随延时点数 k 的变化规律参见图 1．本研究是取 $k=0$ 至 K 的 $K+1$ 个延时互信息的平均值 MI 作为时间序列的分析指标，即

$$\mathrm{MI} = \frac{\sum_{j=0}^{K} I_j}{K+1}. \quad (8)$$

当分割区间个数 $L=2,4,8,16,32$ 时得到的 MI 分别简称为 2,4,8,16,32 段互信息，分别记为 MI2，MI4，MI8，MI16，MI32，称为多分段互信息．

3. 多分段复杂度

目前，时间序列复杂度的计算大多是采用二值粗粒化的方法，既将一个时间序列重构成一个 01 符号序列后再按 Lempel-Ziv 算法计算其复杂度，这种二值粗粒化方法可能会丢失动力学系统的一些有用的信息，本研究中采用了多值粗粒化方法来重构时间序列．

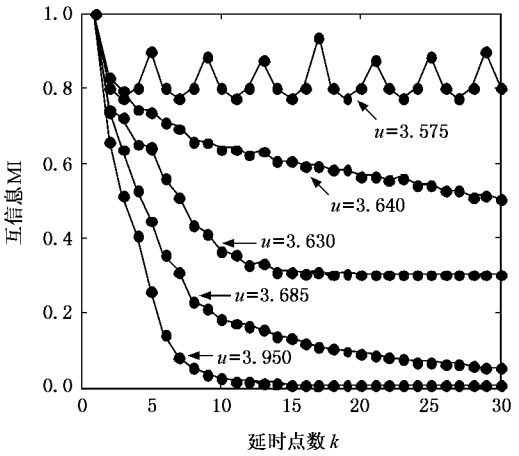

图 1　Logistic 映射不同 u 值时的延时互信息随延时点数 k 的变化规律

设已知的时间序列为 $\{x_1,x_2,\cdots,x_n\}$，求出该时间序列的最大值 $x_{\max}$ 和最小值 $x_{\min}$．用 $l(l>2,l=2$ 时仍采用以序列均值为分界点的经典二值粗粒化方法[10]）表示将序列中数据粗粒化的段数，记

$$d = (x_{\max} - x_{\min})/l. \quad (9)$$

定义一个字符集 $\{s(j)\mid j=1,2,\cdots,l\}$，$s(j)$ 为两两互不相同的字符．用 $S(i)$ 记序列 $\{x_1,x_2,\cdots,x_n\}$ 经 l 段粗粒化后所生成的字符串，$S(i)$ 按下式赋值：

$$S(i) = \begin{cases} s(j), x_{\min} + (j-1)d \\ \leqslant x_i < x_{\min} + jd, \\ (j = 1,2,\cdots,l, i = 1,2,\cdots,n) \\ s(l), x_i = x_{\max}. \end{cases} \quad (10)$$

复杂度计算的前提是各符号在字符串中出现的概率 $p_j(j=1,2,\cdots,l)$ 应该相等，当 p_j 相差较大时，应该考虑用归一化的信源熵 h 进行修正[14]．

$$h = -\sum_{j=1}^{l} p_j \mathrm{lb} p_j / \mathrm{lb} j \quad (11)$$

设按 Lempel-Ziv 算法得出 $S(i)$ 中不同的子串个数为 $c(n)$，则多分段归一化复杂度的计算公式为

$$C = \frac{c(n)\mathrm{lb}n}{hn}. \quad (12)$$

取粗粒化的段数 $l=2,4,8,16,32$，得到的复杂

度分别简称为 2,4,8,16,32 段复杂度,分别记为 C2,C4,C8,C16,C32,称为多分段复杂度.

4. Logistic 映射的复杂度与互信息

Logistic 映射随时间的演化是一个典型的非线性动力系统,其方程为 $x_{n+1} = ux_n(1 - x_n)$,$(n = 0,1,2,\cdots)$,$x_n \in [0,1]$,u 为控制参数,$u \in [3,4]$ 的 Logistic 映射如图 2(a).取 u 的步长为 0.005,得到 201 个非线性时间序列,每个序列长度取 5000 个样本点为研究的实验数据.

4.1. 互信息计算结果

图 1 绘出了有代表性的部分 u 值所对应序列的 16 分段延时互信息随延时点数 k 的变化规律.其中 $u = 3.575$ 时呈周期性演化,$u = 3.640$ 时呈近似的线性关系缓慢下降,$u = 3.630$ 时先快速下降后而趋于一个稳定的非零值,$u = 3.950$ 时先快速下降而向零接近.这些结果定性的表明了各序列的延时互信息是各不相同的,且有的序列具有周期性.

按(8)式计算了不同 u 值时的 201 个序列的 MI2,MI4,MI8,MI16,MI32,不同 u 值时的 MI16 参见图 2(b).

4.2. 复杂度计算结果

按前述方法分别计算了 201 个序列的 C2,C4,C8,C16,C32,它们随 u 值的变化规律参见图 3. 值得指出的是,当 $u \to 4$ 时,C2 的值接近 1,这与实际情况是有较大差异的,因为此时系统仍在混沌状态,而不是完全无序的.这说明了二值粗粒化方法是比较粗略的.

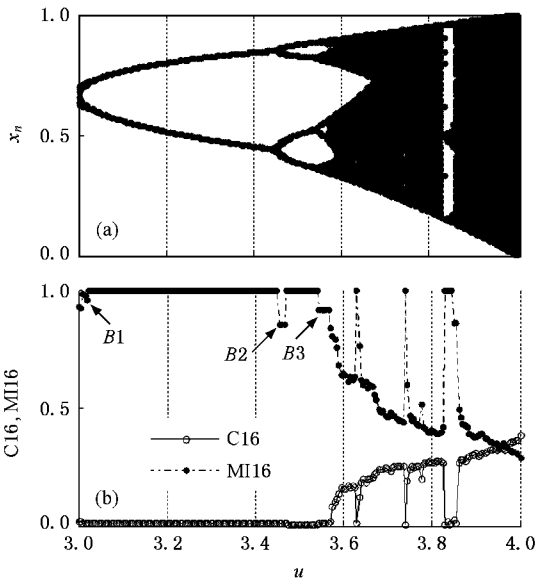

图 2 Logistic 映射图像 (a)映射图像;(b)互信息 MI16 与复杂度 C16(相关系数 = −0.9917)随控制参数 u 的变化规律对照图

4.3. 复杂度与互信息的相关性

经计算 201 个序列不同 K 值时 MI16 分别与 C2,C4,C8,C16,C32 的相关系数,结果都是在 $K = 8$时二者的相关系数的绝对值达到最大,此时的相关系数见表 1 中的第 4 行数据.基于这一结果,用

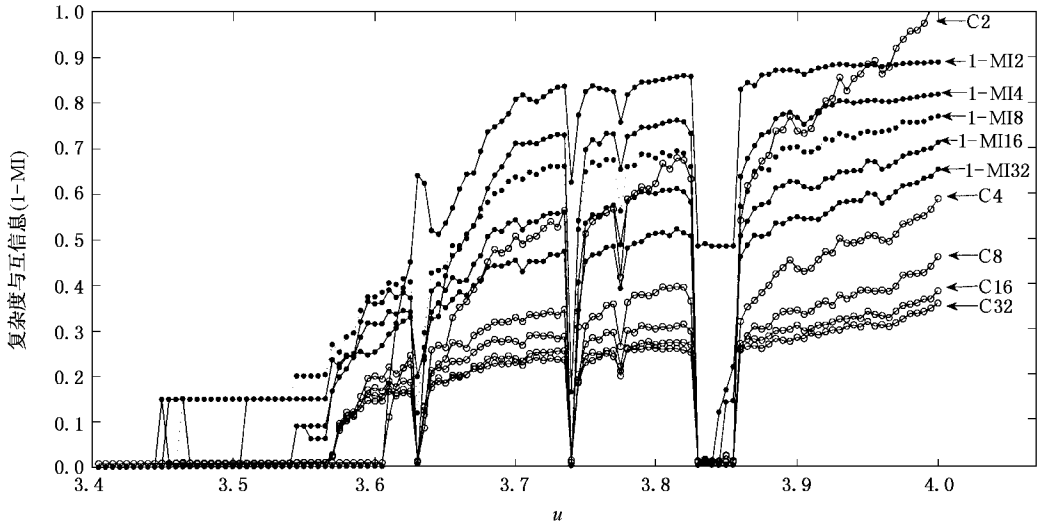

图 3 Logistic 映射的互信息(1 − MI2,1 − MI4,1 − MI8,1 − MI16,1 − MI32)与复杂度(C2,C4,C8,C16,C32)随 u 值变化规律的对照图

(8)式计算平均延时互信息时取 $K=8$，即前 9 个点的平均互信息值.图 4 绘出了不同段复杂度与 MI16 的相关系数随 K 的变化规律.

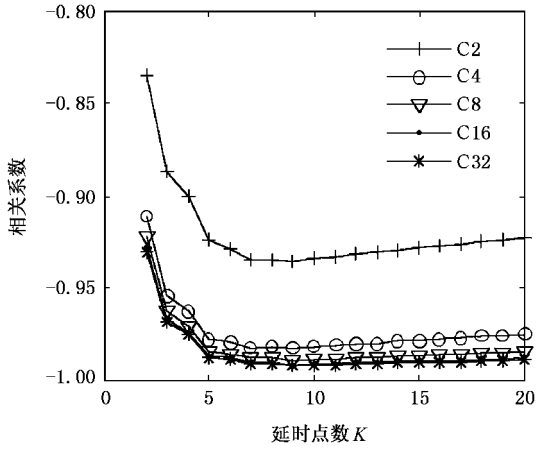

图 4　前 K 个点的 16 段平均延时互信息 MI16 与各段复杂度的相关系数随 K 的变化规律

考虑到互信息与复杂度是呈负相关的，为了更好地观察二者的相关程度，图 3 绘出了 $u\in[3.4,4.0]$ 所生成的序列各段复杂度 C2，C4，C8，C16，C32 与 1－MI2，1－MI4，1－MI8，1－MI16，1－MI32 随 u 值变化关系的对照图.图 2 是 Logistic 映射分岔图、复杂度 C16 及互信息 MI16 随 u 值的变化关系对照图.从图 2，3 可看出，在 Logistic 映射的前几次倍周期分岔处，相应序列的互信息值出现了很明显的波动，见图 2(b)中 $B1$，$B2$ 和 $B3$ 箭头所示位置，而各段复杂度却都看不出任何的变化，表明互信息指标在反应非线性动力系统内在特征方面比复杂度指标更加敏感.从图 3 还可以看出各段互信息与复杂度呈很强的负相关性，表 1 列出了 201 个序列不同段复杂度与不同段互信息的相关系数.

表 1　Logistic 映射不同段互信息与复杂度之间的相关系数

	C2	C4	C8	C16	C32
MI2	－0.9162	－0.9247	－0.9320	－0.9340	－0.9348
MI4	－0.9499	－0.9761	－0.9805	－0.9815	－0.9819
MI8	－0.9247	－0.9706	－0.9785	－0.9797	－0.9806
MI16	－0.9342	－0.9830	－0.9893	－0.9917	－0.9923
MI32	－0.9327	－0.9834	－0.9890	－0.9914	－0.9923

从表 1 可看出，相关程度最小的是 C2 和 MI2，可相关系数也有 －0.9162，相关程度最大的是 C32 和 MI16，相关系数达 －0.9923.图 5 示出了各段互信息与各段复杂度相关系数的关系，可看出共同的特点是互信息与复杂度的相关系数绝对值随复杂度分段数的增加而增大，而 MI16 和 MI32 与各段复杂度的相关系数已几乎相等.考虑到计算耗时的因素和表 1 的具体结果，计算互信息和复杂度的分段数取 8 或 16 比较理想，分段数取 32 以上时对相关系数的影响已不明显.

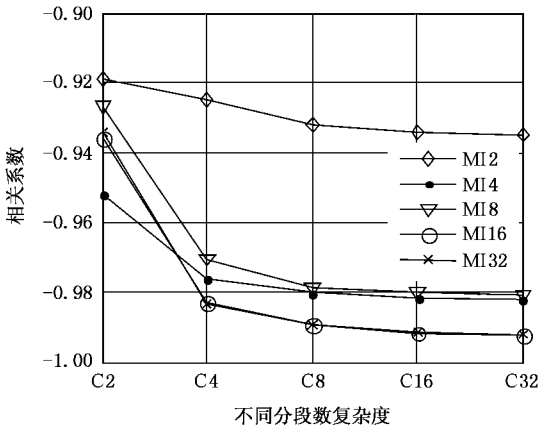

图 5　Logistic 映射不同段互信息与不同段复杂度之间的相关系数对照图

5. Lorenz 模型的互信息与复杂度

Lorenz 模型是大家熟知的一个非线性动力系统，其模型为

$$\begin{cases}\dot{x}=-\sigma(x-y),\\ \dot{y}=rx-y-xz,\\ \dot{z}=xy-bz.\end{cases}$$

参数取 $\sigma=10$，$r=28$，$b=8/3$，初始值取(1，2，3)，步长取 0.002，利用四阶 Runge-Kuta 公式算出 20000 个 x 的值为实验用非线性时间序列，如图 6(a).

互信息和复杂度序列采用滑动窗口的方法得出.对一个长度为 N 的时间序列，从第一个数据开始，取一长度为 T 的矩形窗口内的数据构成一个子序列，计算该子序列的互信息与复杂度，然后将窗口向后滑动，滑动步长为 t 个点，得到下一组子序列并计算其互信息与复杂度，重复上述步骤直至全部数据的最后一点.每个窗口计算得到的互信息与复杂度赋值给窗口内最后一点，将各个窗口计算得到的互信息与复杂度分别依次连接，由此建立互信息与复杂度序列[9].

取 $T=1000$，$t=5$ 时，计算每个窗口内的互信息 MI16 与复杂度 C16，可得到各有 3800 个数据的互信

息与复杂度序列,它们的演化曲线如图 6(b),算得这二个序列的相关系数为 -0.9656,同样得出了互信息与复杂度有很强的负相关性.

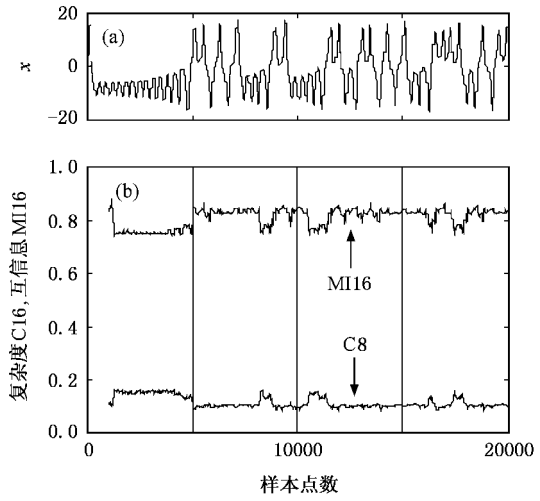

图 6 Lorenz 模型(a) x 分量时间序列;(b)相应的延时互信息 MI16 与复杂度 C8(相关系数 = -0.9656)对照图

6. 心电 RR 间期的互信息与复杂度

心电 RR 期间序列也称心率变异信号,包含有心血管系统的有用信息.连续心搏间瞬时心率的微小涨落,是窦性心率在一定时间内周期性改变的现象,体现了心率或心动周期的波动性.心率波动呈显著的非线性动力学特性[15],因而可以用非线性系统的各种特征量来描述心率变异信号,Lempel-Ziv 复杂度是最常用的一个,而用互信息来分析心率变异信号的研究却并不多见.

在 MIT/BIH 数据库中选取 94 个正常窦性心率的 RR 间期序列,每个序列连续选取 4000 个数据作为实验数据,计算每个序列的各段复杂度 C2,C4,C8,C16,C32 与各段互信息 MI2,MI4,MI8,MI16,MI32.经研究发现,(8)式中的 $K=5$ 时,互信息与复杂度的相关程度最大,相应的相关系数见表 2,变化规律参见图 7.

表 2 心电 RR 间期不同段互信息与复杂度之间的相关系数

	C2	C4	C8	C16	C32
MI2	-0.8646	-0.9584	-0.9258	-0.8944	-0.8753
MI4	-0.8555	-0.9816	-0.9519	-0.9158	-0.8947
MI8	-0.8821	-0.9802	-0.9859	-0.9573	-0.9334
MI16	-0.8979	-0.9666	-0.9860	-0.9815	-0.9628
MI32	-0.8905	-0.9560	-0.9814	-0.9854	-0.9769

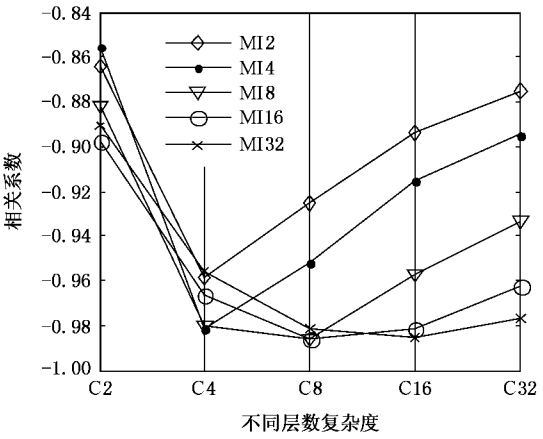

图 7 心电 RR 间期不同段互信息与不同段复杂度之间的相关系数对照图

比较表 1 和表 2 可看出,心电 RR 间期序列的互信息与复杂度的相关程度略低于 Logistic 映射所生成序列的相关系数,但还是有很强的相关性,其中相关程度最大的是 MI16 和 C8,相关系数达 -0.9860.从图 7 可以看出,计算互信息与复杂度时较好的分段数也是 8 和 16,与 Logistic 映射实验数据所得出的结论是一致的.94 个受试者的心电 RR 间期序列的 MI16 和 C8 分布的散点图和拟合直线见图 8.

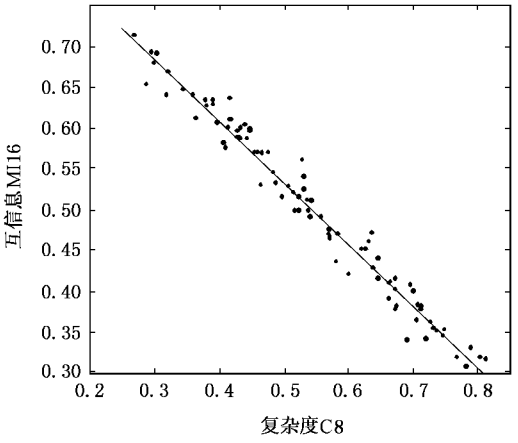

图 8 94 个心电 RR 间期序列的 MI16 与 C8 分布散点图与拟合直线(相关系数 = -0.9860)

7. 结 论

1) Logistic 映射、Lorenz 模型和心电 RR 间期等非线性时间序列的延时互信息与 Lempel-Ziv 复杂度是呈强负相关的.

2）Logistic 映射生成序列是前 9 点的平均延时互信息与复杂度的相关系数绝对值达到最大值；心电 RR 间期序列是前 6 点达到最大值，表明序列的复杂度都是在前几个点的延时互信息值表现出来，因而用互信息来研究非线性时间序列时只需计算前几个点的值，并且当 $K \geqslant 5$ 时，K 值对相关系数的影响并不明显（参见图 4）. 从而选互信息为研究指标比复杂度在计算时间方面可大大减少.

3）相关程度较好是 MI8，MI16 与 C8，C16 之间，即计算互信息与复杂度的分段数取 8 或 16 比较适宜，32 分段以上意义不大.

4）延时互信息在表现非线性动力系统的内在特征方面比复杂度要更加敏感.

5）4 分段以上的复杂度比经典的二值粗粒化复杂度能更精确地反应出非线性时间序列的实质.

[1] Andreia D, Rui M, Diana A M 2004 *Physica* A **344** 326

[2] Dirk H, Uwe L, Heike H Bernd P Michael S Ulrich Z 2002 *Medical Engin. Phys.* **24** 33

[3] Umut O, Deniz E, Robert J 2006 *Pattern Recognition* **39** 1241

[4] Sun H N, Seung-Hyun J, Soo Y K, Byung J H 2002 *Clin. Neurophy.* **113** 1954

[5] Li J, Ning X B, Wu W, Ma X F 2005 *Chin. Phys.* **14** 2428

[6] Lempel A, Ziv J 1976 *IEEE Trans. Inform. Theor.* **22** 75

[7] Hao S R, Hou B Y 2002 *Chin. Phys.* **11** 450

[8] Hildegard M O 2004 *Physica* A **337** 697

[9] Hou W, Feng G L, Dong W J 2005 *Acta Phys. Sin.* **54** 3940 (in Chinese) [侯　威、封国林、董文杰 2005 物理学报 **54** 3940]

[10] Liu J H, Wang J, Lou X F 2004 *Acta Biophys. Sin.* **20** 198 (in Chinese) [刘加海、王　健、罗晓芳 2004 生物物理学报 **20** 198]

[11] Lu H W, Chen Y Z 2004 *Space Med. & Med. Eng.* **17** 444 (in Chinese) [陆宏伟、陈亚珠 2004 航天医学与医学工程 **17** 444]

[12] Palus M 1996 *Physica* D **93** 64

[13] Zhao H, Chai L, Wang H, Liu S 1996 *J. Applied Sciences* **14** 48 (in Chinese) [赵　鸿、柴　路、王　浩、刘书声 1996 应用科学学报 **14** 48]

[14] Kasper K, Schuster H G 1987 *Phys. Rev.* A **36** 843

[15] Zwiener U, Hoyer D, Baver R 1996 *Cardio Vascilar Research* **31** 455

Research on the correlation between the mutual information and Lempel-Ziv complexity of nonlinear time series

Zhang Dian-Zhong†

(*School of Mathematics Science and Computing Technology, Central South University, Changsha* 410083, *China*)

(Received 23 August 2006; revised manuscript received 14 September 2006)

Abstract

To explore the correlation between the mutual information and complexity of nonlinear dynamic system, the nonlinear time series of Logistic map, Lorenz model and cardiac RR intervals were used as the experimental data. The multi-segmented time-delayed mutual information, multi-segmented Lempel-Ziv complexity and their correlation coefficients were calculated. The results show that the mutual information of these series are strongly negatively correlated with the complexity. For the 201 series generated by Logistic equation, the absolute value of all correlation coefficients between the mutual information and the complexity of various segments are 0.9126 plus and the maximum reach to 0.9923, and for the 94 series of cardiac RR intervals, they are 0.8555 plus and the maximum reach to 0.9860. The investigations also indicate that the mutual information is more sensitive than the complexity in characterizing a nonlinear dynamic system.

Keywords: correlation coefficient, mutual information, lempel-Ziv complexity, cardiac RR intervals

PACC: 0547, 8700

† E-mail: zdz1962@sohu.com