

生理时间序列的一种符号化分析方法

廖福元, 王珏

(西安交通大学生物医学信息工程教育部重点实验室, 710049, 西安)

摘要: 提出了一种分析生理时间序列的方法, 即对重构的相空间进行符号化分析. 因生理时间序列通常是非平稳的, 为去除时间序列中的局部趋势、提取时间序列的波形特征, 相空间中的向量被归一化, 从而具有相同的均值和标准差. 然后, 引入最大拓扑熵 (MTE) 原则来寻找相空间的适当划分以实现向量的符号化. 采用 Logistic 映射和人体运动信号检验的结果表明, 用 MTE 原则比用最大熵原则得到的划分更接近最优划分. 原时间序列的波动特征用出现的字模式个数与所有可能的字模式个数之比以及字模式分布概率的 Shannon 熵来描述. 对人体运动信号的分析结果表明, 该方法能够有效区分不同生理状态下的时间序列.

关键词: 局部趋势; 相空间; 划分; 拓扑熵

中图分类号: R318 **文献标识码:** A **文章编号:** 0253-987X(2007)12-1479-04

Symbolic Dynamic Approach for Analyzing Physiological Time Series

Liao Fuyuan, Wang Jue

(Key Laboratory of Biomedical Information Engineering of Education of Ministry, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: An approach for investigating the dynamics of physiological time series is presented, namely the symbolic dynamics analyzing the reconstructed phase space. Since physiological time series are usually nonstationary, to remove the time varying local mean and extract the wave characteristics of the time series, all the vectors in the reconstructed phase space are normalized to be endowed with the same mean and standard deviation. The maximum topological entropy (MTE) criterion is then introduced to find a partition for the phase space. The tested results on the logistic map and the signals of postural stability show that the MTE criterion provides a partition closer to the optimal partition than a partition leading to equiprobable symbols. Two measures from symbolic dynamics are used to characterize the dynamics of the time series. The calculated results for the signals of postural stability show that this approach enables to detect the dissimilarity of physiological time series in different physiological states.

Keywords: local trend; phase space; partition; topological entropy

生理时间序列包含着复杂的波动, 它们提供了生理系统活动状态的基本信息. 疾病和衰老使得生理系统的自适应能力下降, 从而使系统状态变量的观测值序列中包含的信息量减少^[1]. 近年来, 人们提出了若干种度量时间序列的非线性复杂度的方法, 如相关维、Lyapunov 指数、K-S 熵以及非线性预测^[2-5]等. 然而, 这些方法在实际应用中可能受到一

些限制. 例如: 求相关维或者 Lyapunov 指数要求时间序列有足够的长度; 近似熵和样本熵^[6]因为受时间序列中局部趋势等因素的影响, 有可能导致对数据的错误解释.

本文提出了一种分析生理时间序列的方法, 即对重构的相空间进行符号化分析. 相空间重构技术是非线性时间序列分析的基本工具之一^[7], 而符号

化分析则是特征提取常用的方法. 理论上, 由系统任何一个状态变量的观测值序列重构的相空间的维数如果大于该系统盒维的 2 倍, 那么相空间就能够完全表征系统的状态^[8]. 由于生理时间序列一般是非平稳的, 为了去除时间序列中局部趋势、提取时间序列的波形特征, 我们将相空间中的向量归一化, 使之具有相同的均值和标准差. 归一化的另一个作用是, 当时间序列较短时, 可避免出现符号序列(字)模式过多的情况. 我们引入最大拓扑熵原则来寻找对相空间的适当划分, 这样每个归一化的向量转化为一个符号序列即一个字. 然后, 用出现的字模式个数与所有可能的模式个数之比以及字模式概率分布的 Shannon 熵来描述时间序列的波动特征. 对两个不同年龄组的人体运动信号的分析结果表明, 该方法能够区分不同生理状态下的时间序列.

寻找相空间的适当划分是时间序列符号化分析的关键步骤之一. 现有的方法如基于方差的划分^[9]、最大熵划分^[10]、基于小波的最大熵划分^[11]等都有其各自的局限性. Letellier^[12]通过最大化一定长度的字模式个数来寻找划分. 对于 Logistic 映射, 用这种方法找到的分割点比用最大熵方法得到的分割点更接近 Logistic 映射的临界点. 对于实验数据, 我们发现这种方法还存在使字模式个数取得最大值划分不惟一的问题. Bollt 等^[13]用 Tent 映射讨论了拓扑熵与相空间划分的依赖关系, 他们发现不适当的划分使得拓扑熵减小. 因此, 我们引入最大拓扑熵原则来寻找对相空间的划分, 并且用 Logistic 映射和人体运动数据检验了这种方法的有效性.

1 数据处理方法

1.1 相空间重构

对于 N 点时间序列 $x = \{x(i), 1 \leq i \leq N\}$, 重构的相空间中的向量为

$$\begin{aligned} \mathbf{x}_i &= [x(i - (m-1)\tau), \dots, x(i - \tau), x(i)] \\ i &= 1 + (m-1)\tau, \dots, N(1) \end{aligned} \quad (1)$$

式中: m 为嵌入维数; τ 为时延. 当时间序列既含确定性成分又含随机成分时, 很难准确估计嵌入维数, 因为随机成分可能要求嵌入维数为无限大. 尽管如此, 当相空间的维数小于或等于系统盒维的 2 倍时, 仍可从相空间得到关于系统的有用信息.

时间序列的非平稳性可能导致不可靠甚至错误的分析结果(见讨论部分). 虽然向量 $\mathbf{x}_i$ 和 $\mathbf{x}_j = [x(j - (m-1)\tau), \dots, x(j - \tau), x(j)]$ 的均值不同, 但它们的波形是相同的. 另外, 如果向量 $\mathbf{x}_k = [x(k$

$-(m-1)\tau), \dots, x(k - \tau), x(k)]$ 可由向量 $\mathbf{x}_i$ 经过缩放得到, 也应该认为其波形是相同的. 因此, 我们对相空间中的向量 $\mathbf{x}_i$ 进行如下的归一化处理

$$\hat{\mathbf{x}}_i = (\mathbf{x}_i - \mu)/\sigma \quad i = 1 + (m-1)\tau, \dots, N \quad (2)$$

式中: μ 为向量 $\mathbf{x}_i$ 的均值; σ 为其标准差.

1.2 符号化分析

要将相空间中的向量符号化, 首先要选择符号集 $|\Omega|$ 的大小, 这是目前还有待研究的一个问题, 大多数文献选择 $|\Omega| = 2, 3, 4$. 因相空间中的向量经过归一化后均值为 0, 若选择 $|\Omega|$ 等于 3 或 4, 只需确定一个参数即可得到化分相空间的规则. 例如, 当用符号集 $\Omega = \{0, 1, 2, 3\}$ 时, 即是将向量 $\hat{\mathbf{x}}_i = [\hat{x}(i - (m-1)\tau), \dots, \hat{x}(i - \tau), \hat{x}(i)]$ 转化成长度为 4 的符号序列(字), 每个字模式对应相空间的一个子空间. 我们只需选择一个参数 α 来定义这些符号

$$S(\hat{x}(k)) = \begin{cases} 0 & \hat{x}(k) > \alpha \\ 1 & 0 < \hat{x}(k) \leq \alpha \\ 2 & -\alpha < \hat{x}(k) \leq 0 \\ 3 & \hat{x}(k) \leq -\alpha \end{cases} \quad (3)$$

式中: $k = i - (m-1)\tau, \dots, i - \tau, i$. 设选定的 α 导致 J ($J \leq |\Omega|^m$) 个字模式 w_j ($j = 1, 2, \dots, J$) 发生, 我们用以下两个量来描述时间序列的波动特征.

(1) 出现的字模式个数与所有可能的模式个数之比

$$r = J / |\Omega|^m \quad (4)$$

(2) 字模式概率分布 p 的 Shannon 熵

$$H = - \sum_{j=1}^J p(w_j) \ln p(w_j) \quad (5)$$

1.3 相空间的划分

对于一个给定的划分, 时间序列转化为符号序列, 则对应的拓扑熵可近似为^[13]

$$h_T = \limsup_{m \rightarrow \infty} \frac{\ln n_m}{m} \quad (6)$$

式中: n_m 是长度为 m 的字模式个数. 求 h_T 近似值的方法如下^[13]: 设长度为 l ($l \leq m$) 的字模式个数为 n_l , 在二维平面上用直线拟合横坐标为 l 、纵坐标为 $\ln n_l$ 的点 $\{(l, \ln n_l), l \leq m\}$, 则直线的斜率即为拓扑熵的近似值. 文献^[13]不正确的划分使得拓扑熵减小, 因此我们可以通过最大化拓扑熵来寻找对相空间的适当划分. 先用 Logistic 映射

$$x_{n+1} = \mu x_n (1 - x_n) \quad (7)$$

来检验这种划分方法. Logistic 映射的临界点为 $x_c = 0.5$, 因此它可用符号序列 $\{s_n\}$ 来描述, 其中 s_n 定义为

$$s_n = \begin{cases} 0 & x_n \leq x_c \\ 1 & x_n > x_c \end{cases} \quad (8)$$

我们以 $x_0 = 0.65$ 为初值来产生长度为 N 的序列 $\{x_n\}$ 并计算对应于分割点 $x_p \in [0, 1]$ 的拓扑熵 h_T , 这里 $N = 2\,000, m = 6$. 如图 1 所示, 对于每个 μ 值, 当分割点 x_p 在一定的范围内时, 拓扑熵 h_T 取得最大值, 而且 μ 值越接近 4.0, 使拓扑熵取得最大值的分割点的范围 I_T 就越小, 因此我们取 I_T 的中点为要求的分割点. 当 Logistic 映射的参数 μ 分别为 3.85、3.90、3.98 时, 求得的分割点分别为 0.50、0.515、0.50. 同时, 字模式的相对个数 $\tilde{n}_m = n_m / 2^m$ 随 x_p 变化的情况与拓扑熵 h_T 随 x_p 变化的情况很相似, 通过最大化字模式的相对个数 $\tilde{n}_m$ [12] 求得的分割点也分别为 0.50、0.515、0.50, 而用最大熵原则 [10] 求得的分割点分别为 0.489、0.611、0.564. 显然, 对于 Logistic 映射, 通过最大化拓扑熵和最大化字模式的相对个数求得的分割点比用最大熵原则求得的分割点更接近临界点 $x_c = 0.5$.

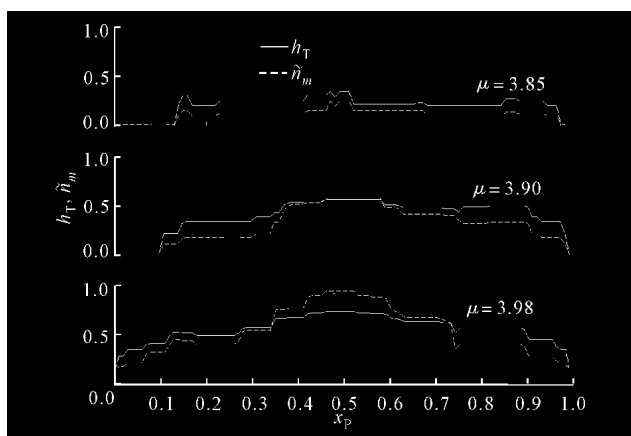
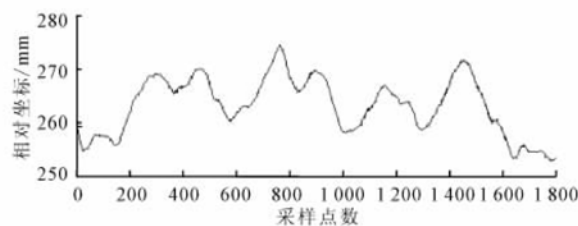
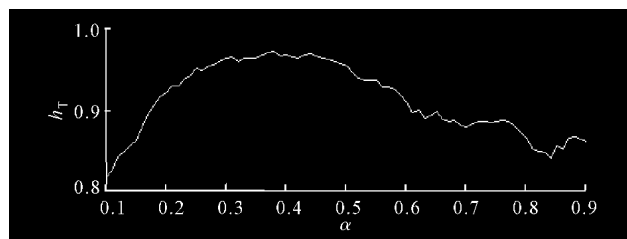


图1 $h_T, \tilde{n}_m$ 与 x_p 的关系

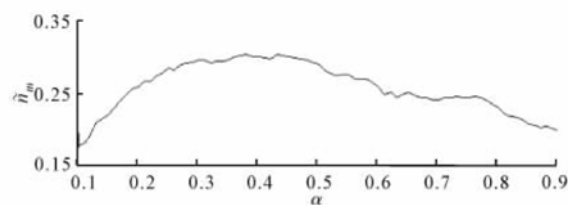
图 2a 为一名受试者在静立时肩部一点在前后方向上相对某个基准点的坐标序列. 我们用嵌入维 $m = 5$, 时延 $\tau = 1$ 重构相空间, 并将向量归一化. 然后, 用符号集 $\Omega = \{0, 1, 2, 3\}$ 来计算对应于划分参数 α (式(3)) 的拓扑熵 h_T 和字模式相对个数 $\tilde{n}_m$. 如图 2b、图 2c 所示, 当拓扑熵取得最大值时, $\alpha = 0.39$, 而 $\tilde{n}_m$ 的最大值对应两个不同的 α 值 0.39 和 0.44. 不同的 α 值意味着对相空间的不同划分, 从而使描述时间序列的两个量 r 和 H (式(4)和式(5)) 的计算结果不同. 我们用其他的实验数据对这两种寻找相空间划分的方法进行检验, 观察到了类似的情况. 可见, 对于实验数据, 通过最大化字模式相对个数来寻找划分还存在一些问题.



(a) 相对坐标的时间序列



(b) 拓扑熵随 α 的变化情况



(c) 字模式相对个数随 α 的变化情况

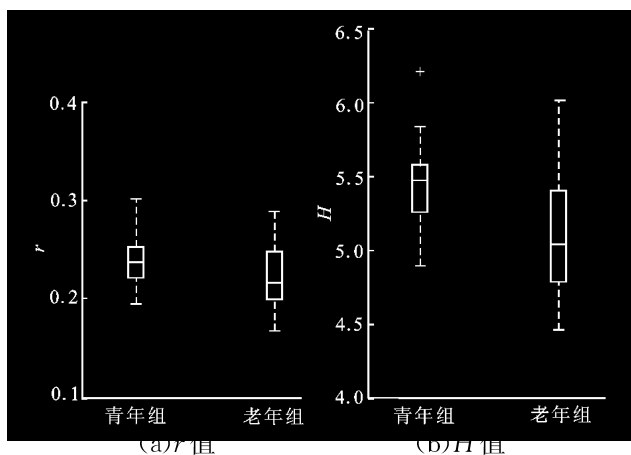
图2 相对坐标的时间序列符号化处理结果

2 结果

我们用本文提出的方法来处理人体运动时间序列. 受试者有青年组 29 人 (平均年龄 (23 ± 0.6) 岁), 老年组 23 人 (平均年龄 (73.3 ± 1.5) 岁). 在受试者的肩部放置一个标记点, 用 Optotrack 运动测量系统 (3020, NDI, Canada) 记录标记点在前后方向上的位置. 采样频率为 60 Hz, 采样时间为 30 s, 其间受试者保持直立姿势. 对于每个时间序列, 用 $m = 5$ 、 $\tau = 1$ 和 $|\Omega| = 4$ 重构相空间, 并用最大拓扑熵原则寻找对相空间的划分, 然后计算度量 r 和 H 的值, 结果如图 3 所示. 对于两个度量 r 和 H , 青年组的值分别显著大于老年组的值 ($p = 0.010\,5$ 和 $p = 0.015\,9$). 这说明, 老年人在静立时身体的摇摆较规则, 即老年人的自适应能力较差. 这与文献 [14] 报道的结果是一致的. 另外, 我们注意到当 $m = 4, 5, 6$, $|\Omega| = 3, 4$ 时, m 和 $|\Omega|$ 的值对计算结果并无明显的影响.

3 讨论

本文提出的时间序列分析方法基于相空间重构和符号化分析, 试图解决两个问题. 首先, 生理时间序列中的局部趋势往往导致对数据的不可靠甚至错误的解释. 图 4 表示用第 2 节描述的数据计算的样本熵 $S_E(2, 0.2\sigma, N)$ [6], 其中 σ 为时间序列的标准

图3 度量 r 和 H 的计算结果

差.一方面,两组样本熵无显著差别($p=0.226$),老年组的样本熵略大于青年组的样本熵.另一方面,两组样本熵都很小,绝大多数不超过0.1.这样的结果显然具有误导性.在本文提出的方法中,相空间中的向量被归一化,从而避免了不平稳性的影响.但是,将向量的均值归一化为零相当于对时间序列进行高通滤波,从而可能使一些有用的低频信息丢失,这个问题需要作进一步研究.

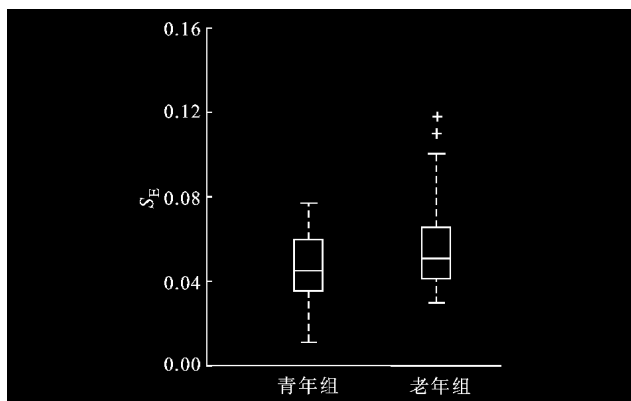


图4 用第2节描述的数据计算的样本熵

寻找对相空间的划分是时间序列符号化分析的关键步骤之一.现有的几种方法都有其各自的局限性.文献中用得较多的是最大熵划分^[10],而基于小波的最大熵划分^[11]是以最大熵划分为基础的.对于Logistic映射,通过最大化字模式个数^[12]或者最大化拓扑熵得到的分割点比用最大熵原则得到的分割点更接近Logistic映射的临界点.对于第2节描述的数据,通过最大化字模式个数求得的划分可能不惟一,也就是对原时间序列的符号表示不惟一.如果用最大熵原则求相空间的划分,则两组受试者的 r 和 H 值也有显著的差异($p=0.019$ 和 $p=0.025$),但不如用最大拓扑熵原则所得结果的差异明显($p=0.0105$, $p=0.0159$).这说明,无论对仿真数据还是实验数据,用最大化拓扑熵原则都能得到较好的

结果.

参考文献:

- [1] Goldberger A L, Peng C K, Lipsitz L A. What is physiologic complexity and how does it change with aging and disease? [J]. *Neurobiol Aging*, 2002, 23 (1): 23-26.
- [2] Grassberger P, Procaccia I. Characterization of strange attractors [J]. *Phys Rev Lett*, 1983, 50(5): 346-349.
- [3] Ding M, Grebogi C, Ott E, et al. Plateau onset for correlation dimension: when does it occur? [J]. *Phys Rev Lett*, 1993, 70(25): 3872-3875.
- [4] Wolf A, Swift J B, Swinney H L, et al. Determining Lyapunov exponents from a time series [J]. *Physica D*, 1985, 16(3): 285-317.
- [5] Porta A, Guzzetti S, Furlan R, et al. Complexity and non linearity in short-term heart period variability: comparison of methods based on local non linear prediction [J]. *IEEE Trans Biomedical Engineering*, 2007, 54 (1): 94-106.
- [6] Richman J S, Moorman J R. Physiological time series analysis using approximate entropy and sample entropy [J]. *Am J Physiol Heart Circ Physiol*, 2000, 278 (6): H2039-H2049.
- [7] Voss A, Kurths J, Kleiner H J, et al. The application of methods of nonlinear dynamics for the improved and predictive recognition of patients threatened by sudden cardiac death [J]. *Cardiovasc Res*, 1996, 31(3): 419-433.
- [8] Sauer T, Yorke J A, Casdagli M. Embedology [J]. *J of Stat Phys*, 1991, 65(3/4): 579-616.
- [9] Veenman C J, Reinders M J T, Bolt E M, et al. A maximum variance cluster algorithm [J]. *IEEE Trans Pattern Anal Mach Intell*, 2002, 24 (9): 1273-1280.
- [10] Chau T, Wong A K C. Pattern discovery by residual analysis and recursive partitioning [J]. *IEEE Trans Knowledge Data Eng*, 1999, 11 (6): 833-852.
- [11] Rajagopalan V, Ray A. Symbolic time series analysis via wavelet-based partitioning [J]. *Signal Processing*, 2006, 86 (11): 3309-3320.
- [12] Letellier C. Symbolic sequence analysis using approximated partition [J]. *Chaos, Solitons & Fractals*, 2008, 36(1): 32-41.
- [13] Bollt E M, Stanford T, Lai Y C, et al. What symbolic dynamics do we get with a misplaced partition?: on the validity of threshold crossings analysis of chaotic time-series [J]. *Physica D*, 2001, 154(3/4): 259-286.
- [14] Amoud H, Snoussi H, Hewson D, et al. Intrinsic mode entropy for nonlinear discriminant analysis [J]. *IEEE Signal Processing Lett*, 2007, 14(5): 297-300.

(编辑 杜秀杰)