等概率符号化样本熵应用于脑电分析*

黄晓林[1)†] 霍铖宇[2)] 司峻峰[1)] 刘红星[1)‡]

1) (南京大学电子科学与工程学院, 生物医学电子工程研究所, 南京 210023)

2) (常熟理工学院物理与电子工程学院, 常熟 215500)

(2014 年 1 月 6 日收到; 2014 年 2 月 11 日收到修改稿)

样本熵 (或近似熵) 以信息增长率刻画时间序列的复杂性, 能应用于短时序列, 因而在生理信号分析中被广泛采用. 然而, 一方面由于传统样本熵采用与标准差线性相关的容限, 使得熵值易受非平稳突变干扰的影响, 另一方面传统样本熵还受序列概率分布的影响, 从而导致其并非单纯反映序列的信息增长率. 针对上述两个问题, 将符号动力学与样本熵结合, 提出等概率符号化样本熵方法, 并对其物理意义、数学推导及参数选取都做了详细阐述. 通过对噪声数据的仿真计算, 验证了该方法的正确性及其区分不同强度时间相关的有效性. 此方法应用于脑电信号分析的结果表明, 在不对信号做人工伪迹去除的前提下, 只需要 1.25 s 的脑电信号即可有效地区分出注意力集中和注意力发散两种状态. 这进一步证明了该方法可很好地抵御非平稳突变干扰, 能快速获得短时序列的潜在动力学特性, 对脑电生物反馈技术具有很大的应用价值.

关键词: 符号动力学, 等概率符号化, 样本熵, 脑电生物反馈

PACS: 05.10.–a, 89.70.cf, 87.19.le **DOI:** 10.7498/aps.63.100503

1 引 言

近似熵是由 Pincus[1] 提出的一种以信息增长率表征系统复杂性的测度, 其不对系统做混沌性、随机性或线性、非线性的假设, 且适用于短时序列, 因此自其提出以来就受到广泛关注. 为消除自匹配对熵值计算带来的偏差, Richman 和 Moorman[2] 提出了近似熵的改进版本——样本熵, 使得熵值不再受数据长度的影响, 同时样本熵还改善了近似熵的一致性, 即计算中参数的选取不会影响不同系统熵值的大小顺序. 因为近似熵和样本熵能有效刻画动力系统的复杂性, 并能应用于短时序列, 因而被广泛应用于心率变异性、脑电等信号的分析 [3−7].

样本熵和近似熵 (为简化, 以下都称为样本熵) 的计算过程可简述如下: 首先将序列做 m 维延迟嵌入, 计算此时延迟矢量相似的平均概率 $C^{(m)}$; 再做 $m+1$ 维延迟嵌入, 计算此时延迟矢量相似的平均概率 $C^{(m+1)}$; 样本熵 S_{E} 定义为

$$S_{\mathrm{E}}(m,r,N) = -\log(C^{(m+1)}/C^{(m)}), \qquad (1)$$

根据 (1) 式可得到样本熵. 计算中涉及嵌入维数 m, 判定两矢量相似的容限 r, 以及序列长度 N 三个参数的选取. 这里 r 的选取最为关键, r 太小, 样本熵易受噪声影响, 而 r 太大, 则又失去敏感性, 所有系统都将退化为确定性系统. 以往研究一般以经验选取这三个参数, 例如 m 通常选取为 2—3, r 以序列标准差 (SD) σ 为参照, 选取为 0.1 σ—0.2 σ, N 则要满足统计有效性 [1−3,6].

然而, 样本熵测度仍然存在一些问题. 首先, 容限选取以 SD 为参照, 而 SD 本身很容易受非平稳突变干扰的影响, 从而导致样本熵也易受非平稳突变干扰的影响. 其次, 样本熵值在理论上依赖于序列的概率分布. 以非相关的随机噪声为例, 假设其概率密度函数为 $p(x)$, 容限为 r, 则任意两不同时刻的样本相似的平均概率 $P_{x_i\sim x_j}$ (符号 $\sim$ 代表相

* 江苏省自然科学基金 (批准号: BK2011565) 和国家自然科学基金 (批准号: 61271079) 资助的课题.

† 通讯作者. E-mail: xlhuang@nju.edu.cn

‡ 通讯作者. E-mail: njhxliu@nju.edu.cn

似) 的期望值为

$$E(P_{x_i\sim x_j}) = \int_{-\infty}^{\infty}\left(p(x)\int_{x-r}^{x+r} p(t)\,\mathrm{d}t\right)\mathrm{d}x. \quad (2)$$

又由于序列的非相关性, 即序列中任意两点无关联, 因此m维延迟嵌入矢量相似的平均概率为

$$\begin{aligned} &E(P_{x_i\sim x_j, x_{i+1}\sim x_{j+1},\cdots,x_{i+m-1}\sim x_{j+m-1}})\\ =&E(P_{x_i\sim x_j})E(P_{x_{i+1}\sim x_{j+1}})\cdots\\ &\times E(P_{x_{i+m-1}\sim x_{j+m-1}})\\ =&(E(P_{x_i\sim x_j}))^m. \end{aligned} \quad (3)$$

当嵌入维数增加到$m+1$时, 延迟嵌入矢量相似的概率相应变为$(E(P_{x_i\sim x_j}))^{m+1}$. 由此, 理论上非相关序列的样本熵为

$$\begin{aligned} S_{\mathrm{E}} &= -\log(E(P_{x_i\sim x_j}))\\ &= -\log\left(\int_{-\infty}^{\infty}\left(p(x)\int_{x-r}^{x+r} p(t)\,\mathrm{d}t\right)\mathrm{d}x\right). \end{aligned} \quad (4)$$

显然, 样本熵值将同时受容限r和概率分布的影响. 既然样本熵衡量的是新信息增长率, 那么样本熵值应该只由序列的时间相关性决定, 而与概率分布无关, 传统的样本熵显然忽略了这一问题. 尽管之后人们对样本熵进行了改进, 形成了多尺度熵、多变量多尺度熵等[8−10]一系列方法, 但都没有解决上述两个问题. 本文尝试改进样本熵的算法, 从而尽量降低非平稳突变干扰和概率分布的影响.

符号化是指利用有限符号实现数据在幅度域的离散化. 符号化能降低噪声影响, 减少数据对内存的消耗, 加速信号处理, 因此对于实时信号处理, 符号化极具实用价值. 研究表明, 符号化方法选择恰当时, 符号序列可有效保留时间序列的动力学本质[11]. 近年来, 在非线性时间序列分析中符号动力学受到关注[11−15]. 在符号动力学分析中, 符号化方法的选择相当关键. 本研究组曾利用等概率符号化[16]的方法分析心率变异性信号取得了满意的效果[17]. 本文将等概率符号化与样本熵结合, 提出等概率符号化样本熵 (ESSE) 的方法, 以实现改进样本熵的目的.

脑电生物反馈是脑电信号应用的热点之一, 其关键环节之一是从脑电图 (EEG) 中提取与大脑活动 (如注意力、情绪等) 相关的参数, 以用于评价大脑活动水平. 由于该类系统实时性要求高, 因此应从尽量短的数据中快速提取特征参数. 然而, 脑电信号是一种严重非平稳的信号, 存在着复杂多样的伪迹 (干扰) 成分. 目前绝大多数脑电分析方法都需要事先进行伪迹去除, 且去除过程往往需要人工干预. 这显然很难满足实时性要求. 针对这一应用要求, 我们利用本文提出的ESSE对注意力实验中的脑电信号进行分析, 获得了满意的结果.

2 ESSE

2.1 等概率符号化

等概率符号化的思想首先由Lin等[16]在2007年提出, 被广泛应用于数据挖掘领域. 本研究组将等概率符号化引入到生理时间序列分析中[17]. 等概率符号化过程简述如下: 对于时间序列$\{x_i : 1\leqslant i\leqslant N\}$, 首先按幅值大小排序; 当给定符号数$n$时, 找到$n-1$个等分位点 (记为$t_1, t_2, \cdots, t_{n-1}$) 作为符号划分的阈值; 按规则

$$s_i = \begin{cases} 0 & (x_i \leqslant t_1),\\ 1 & (t_1 < x_i \leqslant t_2),\\ \vdots & \vdots\\ n-2 & (t_{n-2} < x_i \leqslant t_{n-1}),\\ n-1 & (t_{n-1} < x_i), \end{cases} \quad (5)$$

将原始序列转换为离散的符号序列$\{s_i : 1\leqslant i\leqslant N\}$.

等概率符号化主要有三方面好处. 第一, 符号化后原始序列概率分布的影响被消除, 符号序列完全体现原始序列的时序关系, 因而可解决样本熵值受概率分布影响的问题. 第二, 符号化结果不受极端值的影响, 因而能很好地对抗非平稳突变干扰. 第三, 实现了一种幅度域的变分辨率, 即在幅值分布密集的区域采用更多的符号以提高分辨率, 而在稀疏区域采用较少的符号以降低冗余, 因而提高了符号的利用率, 同时, 变分辨率还突破了传统均匀符号化的线性约束.

2.2 符号样本熵

对符号序列$\{s_i : 1\leqslant i\leqslant N\}$求样本熵与传统样本熵的计算类似, 主要区别在于原来 “嵌入矢量相似” 的判断被转换为 “符号嵌入矢量相等” 的判断. 下面给出符号样本熵的具体计算方法.

将符号序列$\{s_i\}$做延迟为τ的m维嵌入, 符号嵌入矢量记为

$$\boldsymbol{B}^{(m)}(i) = (s_i, s_{i+\tau}, \cdots, s_{i+(m-1)\tau}),$$

若定义 $n_i^{(m)}$ 为与矢量 $\boldsymbol{B}^{(m)}(i)$ 相同的符号矢量个数, 则 m 维嵌入下符号矢量两两相同的平均概率为

$$C^{(m)} = \frac{1}{N-(m-1)\tau} \times \sum_{i=1}^{N-(m-1)\tau} \frac{n_i^{(m)}}{N-(m-1)\tau}, \quad (6)$$

利用相同方法计算嵌入维增加到 $m+1$ 时的平均概率 $C^{(m+1)}$, 则符号样本熵

$$S_{\text{E-symb}}(m,n,N) = -\log\frac{C^{(m+1)}}{C^{(m)}}. \quad (7)$$

对比 (1) 和 (7) 式可以发现, 在 (7) 式中, 影响样本熵的容限 r 变成了符号数 n. 这是因为符号化后, 样本熵计算中的容限实际上由各符号化区间决定, 这些区间大小并不固定, 而是与选择的符号化方法以及符号数有关, 在等概率符号化下, r 由符号数 n 决定.

2.3 非相关噪声的ESSE理论分析

等概率符号化后, 非相关噪声的ESSE值的理论推导将变得更为简单. 根据样本熵的物理意义可知, 当时间序列无相关时, 新信息增长率是最大的, 样本熵也应该最大并与原始序列的概率分布无关. 若对完全非相关时间序列实施等概率符号化, 则任意两不同时刻的样本符号相等的平均概率为

$$E(P_{s_i=s_j}) = 1/n.$$

与 (3) 式的推导类似, m 维符号矢量两两相同的平均概率为 $(1/n)^m$, $m+1$ 维符号矢量两两相同的平均概率则为 $(1/n)^{m+1}$, 因而无相关序列的ESSE理论值为

$$S_{\text{E-symb}}^{\text{white-noise}}(m,n,N) = \log(n). \quad (8)$$

由 (8) 式可知, 非相关白噪声的ESSE理论值仅仅由符号数 n 决定, 与嵌入维数 m 和序列长度 N 均无关. 利用 (8) 式可求得ESSE的上限值. 而在传统样本熵计算中, 仅能保证熵值非负, 没有明确上限 [3].

2.4 参数选择

ESSE的计算仍然涉及 m, n 和 N 三个参数值的选取, 此外, 对于脑电这类连续信号还涉及嵌入延迟 τ 值的选取.

对 m 的选择不属于方法问题, 应依据时间序列本身的特点进行. 例如对于白噪声, 理论上其熵值与嵌入维数 m 无关, 可任意选取; 而对于其他序列, 为全面考察序列时间相关性, 应尝试在不同的 m 取值下计算熵值, 以形成多尺度熵值谱. 多尺度方法要求数据量大, 不在本文的讨论范围, 在此不予赘述.

样本熵应用于心跳间期序列分析时, 因心跳间期序列本身属于离散序列, 嵌入延迟 τ 一般取1拍 [2]. 需注意的是, 连续信号分析中, 嵌入延迟 τ 选取时间量纲更为合适, 可避免采样率的影响. 依据延迟嵌入理论, 一般以最小互信息为原则选取 τ 值 [6,18]. 为实现实时信号分析时参数的快速选取, 本文建议选取最高频率分量对应周期的1/4作为 τ 值, 例如信号最高频率分量为50 Hz 时, 选取 $\tau = 5$ ms.

符号数 n 的选择既要保证符号序列能够保留原始序列的动力学性质, 又要保证一定的降噪性能. 等概率符号化时, 符号区间与原始序列的概率分布有关, 分布密集的区域符号区间小, 稀疏区域符号区间大. 一般而言, 在保证足够多的幅度域分辨率 (n 不能太小) 的同时, 还要保证最小的符号区间不能小于已知的噪声幅值 (n 不能太大).

序列长度 N 决定了熵值计算的统计有效性. 等概率符号化后, 延迟嵌入相空间实质上被划分为 n^m 个子区域, 因此当满足 $N \gg n^m$ 时能保证熵值计算的统计有效性.

3 实验结果

3.1 数据仿真

本文对白噪声、$1/f$ 噪声和布朗噪声各取100组序列进行了仿真计算. 参数选取为 $m = 2$, $n = 4$—8, $N = 500$, $\tau = 1$. 图1所示为白噪声、$1/f$ 噪声和布朗噪声的ESSE 均值 - 标准差误差棒, 其中, 横轴代表符号数 n, 纵轴代表熵值 $S_{\text{E-symb}}$, 虚线为非相关白噪声的理论熵值 $\log(n)$.

由图1可知: 高斯白噪声的熵值与本文的理论推导值几乎完全重合, 从而验证了本文理论推导的正确性; 三种噪声的熵值由大到小依次为白噪声、$1/f$ 噪声、布朗噪声, 而三种噪声序列的时间相关性也是依次增强的, 因此, 该结果完全符合相关性增强导致熵值减小的理论; 不同的符号数设置不改变三种噪声熵值的相对关系, 说明本文方法具备良好的一致性; 当符号数 $n = 8$ 时, $N = 500$ 的原始序列

已能计算出与理论值非常接近的结果, 说明此时已具备较好的统计有效性.

总之, 三种典型噪声的仿真计算结果表明, ESSE能正确有效地刻画时间序列的相关性或信息增长率.

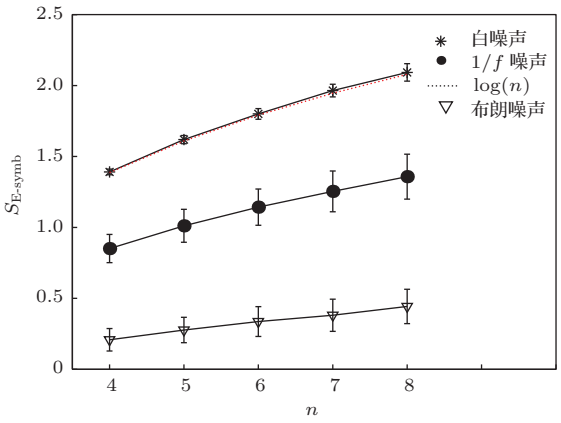

图1 白噪声、$1/f$ 噪声和布朗噪声的ESSE均值-标准差误差棒图

3.2 脑电数据分析

从EEG中提取与注意力相关的参数用于评价注意力水平, 并进一步实现注意力缺陷的反馈治疗, 是脑电生物反馈技术的一个重要分支. 目前, 国内外已有的应用于注意力缺陷治疗的脑电生物反馈系统一般采用基于频域分析的方法考察脑电信号中θ (4—8 Hz), α (8—13 Hz), β (13—30 Hz)三个频段成分的能量或其相对变化[19−21]. 然而, 一方面频谱分析极易受非平稳突变干扰的影响, 另一方面其局限于序列的线性相关特性, 因此频谱分析的敏感程度和准确性均欠佳, 反馈治疗效果也存在争议[20]. 本文提出的ESSE既能很好地抵御非平稳突变干扰, 又不局限于反映线性相关, 我们尝试将其应用到注意力实验中的脑电信号分析.

本文使用的脑电数据由日本光电EEG-9100系统采集, 采样率为200 Hz, 频带为0.5—45 Hz. 同步采集16路脑电信号, 导联位置按照国际10—20标准安放, 单极导联方式, 参考导联为同侧耳垂. 实验征集了20—30周岁的12名在校大学生或研究生作为实验对象, 其中女性1例, 男性11例, 左利手2例, 右利手10例. 志愿者均受过良好教育, 无病史, 实验前两天之内未服用任何药物. 志愿者在实验前均被详细告知实验内容并同意参与实验. 根据实验要求, 志愿者分别在睁眼放松、注意力集中的两个状态下连续采集脑电信号4—5 min, 其中注意力集中阶段志愿者须完成一项注意力任务. 脑电采集过程中, 系统自动定时监测电极接触阻抗, 以保证接触阻抗小于10 kΩ.

经预览分析, 发现采集到的EEG带有电极线摆动、眼动、吞咽等多种伪迹. 在不进行伪迹去除或数据挑选的情况下, 在同一实验状态下EEG各频段的绝对能量存在着极大的波动, 从而导致难以利用传统的功率谱中相对能量参数W_α/W_θ和W_β/W_θ对睁眼放松和注意力集中两种实验状态加以明显区分.

我们将ESSE应用于脑电数据分析, 参数设置为$m=2$, $n=4$—9, $N=256$, 512, $\tau=5$ ms, 结果发现在C4, P4, O2导联都能显著区分两种状态 (t检验得到$p<10^{-40}$). 图2给出了符号数$n=4$, 序列长度$N=256$时P4导联的ESSE结果, 其中的误差棒说明熵值随时间略有波动. 为了对照, 图2中还用三条水平线分别标识出在同样参数设置下100组随机白噪声序列、100组$1/f$噪声序列、100组布朗噪声序列的平均ESSE值.

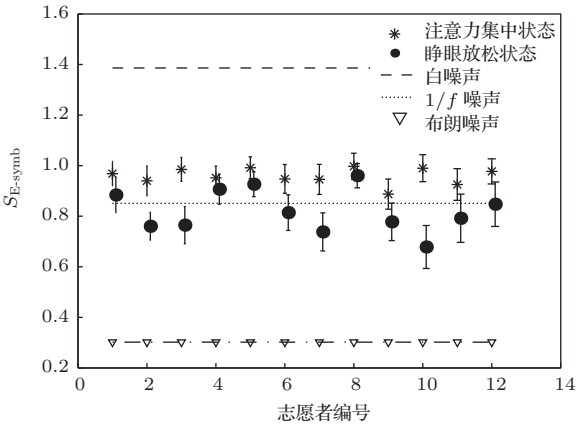

图2 两种不同实验状态下, P4导联的ESSE结果

从图2可以看出, 对于所有的志愿者, 睁眼放松状态下都呈现更低的ESSE值, 而在执行注意力任务时呈现更高的ESSE值. 这说明在大脑注意力集中状态下EEG的复杂性更高. 尽管EEG不能简单地由任何一种噪声描述, 但是仍然能从复杂性的角度将EEG与噪声相比较, 发现EEG的复杂性更接近于$1/f$噪声的复杂性, 而与同等参数设置下的$1/f$噪声相比, 执行注意力任务时EEG的ESSE值显著高于$1/f$噪声序列的ESSE值 (t检验, $p<10^{-7}$), 而在睁眼放松状态下则有7人EEG的ESSE值显著低于$1/f$噪声序列的ESSE值 (t检验, $p<10^{-5}$).

我们还利用传统的样本熵方法对同样的数据进行了计算. 图3给出了在参数设置为$m = 2$, $r = 0.2\sigma$, $N = 256$ 时P4导联传统样本熵的计算结果, 其中在注意力集中状态下第二号志愿者的传统样本熵溢出, 因此图中没有显示. 从图3可以看出: 第一号、第三号、第十二号志愿者两种实验状态下的样本熵没有显著差异 (t检验, $p > 0.05$); 此外, 第四号、第五号志愿者注意力状态下的熵值高于静息状态下的熵值, 而第六号—第十一号志愿者则相反. 当参数r从0.1σ到0.9σ改变时, 尽管传统样本熵的计算结果随r的取值不同而变化, 但均不能与图2所示结果一样一致性地区分出两种实验状态. 因此, 传统样本熵不能有效反映出在两种不同状态下脑电活动的一致性规律.

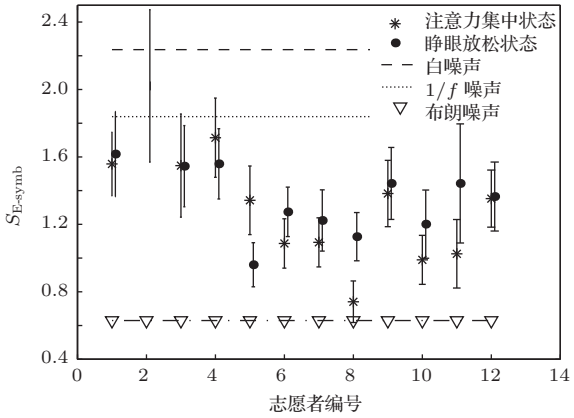

图3 两种不同实验状态下, P4导联的传统样本熵计算结果

4 结 论

本文提出了一种ESSE方法, 该方法能有效抵御非平稳突变干扰的影响, 并可消除原始序列概率分布的影响, 从而单纯反映时间序列的信息增量这一动力学特性. 该方法物理意义明确, 计算简便快速, 对于实时性要求高的脑电生物反馈极具应用价值. 数值结果表明, 对于注意力实验中的EEG, 即使不做数据的伪迹去除或人工挑选, 针对短至1.25 s的序列, ESSE也能快速有效地获取原始序列复杂性的评价, 并有效区分不同的注意力状态. 值得注意的是, 等概率符号化后, 在大多数情况(均匀分布除外)下符号区间都是不均匀的, 因此ESSE实质上更多地突破了线性约束, 这一点也是其与传统样本熵不同之处.

参考文献

[1] Pincus S M 1991 *Proc. Natl. Acad. Sci. USA* **88** 2297
[2] Richman J S, Moorman J R 2000 *Am. J. Physiol. Heart Circ. Physiol.* **278** 2039
[3] Bruhn J, Röpcke H, Hoeft A 2000 *Anesthesiology* **92** 715
[4] Lake D E, Richman J S, Griffin M P, Moorman J R 2002 *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **283** R789
[5] Srinivasan V, Eswaran C, Sriraam N 2007 *IEEE Trans. Inf. Technol. Biomed.* **11** 288
[6] Sohn H, Kim I, Lee W, Peterson B S, Hong H, Chae J H, Hong S, Jeong J 2010 *Clin. Neurophysiol.* **121** 1863
[7] Acharya U R, Molinari F, Sree S V, Chattopadhyay S, Ng K H, Suri J S 2012 *Biomed. Signal Proces. Control.* **7** 401
[8] Costa M, Goldberger A L, Peng C K 2005 *Phys. Rev. E* **71** 021906
[9] Ahmed M U, Mandic D P 2011 *Phys. Rev. E* **84** 061918
[10] Hu M, Liang H 2012 *IEEE Trans. Biomed. Eng.* **59** 12
[11] Song A L, Huang X L, Si J F, Ning X B 2011 *Acta Phys. Sin.* **60** 020509 (in Chinese) [宋爱玲, 黄晓林, 司峻峰, 宁新宝 2011 物理学报 **60** 020509]
[12] Zhang M, Wang J 2013 *Acta Phys. Sin.* **62** 038701 (in Chinese) [张梅, 王俊 2013 物理学报 **62** 038701]
[13] Wu S, Li J, Zhang M L, Wang J 2013 *Acta Phys. Sin.* **62** 238701 (in Chinese) [吴莎, 李锦, 张明丽, 王俊 2013 物理学报 **62** 238701]
[14] Chen G, Xie L, Chu J 2013 *Chin. Phys. B* **22** 038902
[15] Wang J, Yu Z F 2012 *Chin. Phys. B* **21** 018702
[16] Lin J, Keogh E, Wei L, Lonardi S 2007 *Data Min. Knowl. Disc.* **15** 107
[17] Hou F Z, Huang X L, Chen Y, Huo C Y, Liu H X, Ning X B 2013 *Phys. Rev. E* **87** 012908
[18] Kantz H, Schreiber T 2003 *Nonlinear Time Series Analysis* (2nd Ed.) (Cambridge: Cambridge University Press) pp39–40
[19] Klimesch W 1999 *Brain Res. Rev.* **29** 169
[20] David J V 2005 *Appl. Psychophysiol. Biofeedback* **30** 347
[21] Egner T, Gruzelier J H 2004 *Clin. Neurophysiol.* **115** 131

Application of equiprobable symbolization sample entropy to electroencephalography analysis*

Huang Xiao-Lin[1)†] Huo Cheng-Yu[2)] Si Jun-Feng[1)] Liu Hong-Xing[1)‡]

1) (*Institute of Biomedical Electronic Engineering, School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China*)

2) (*School of Physics and Electronic Engineering, Changshu Institute of Technology, Changshu 215500, China*)

(Received 6 January 2014; revised manuscript received 11 February 2014)

Abstract

Sample entropy or approximate entropy, a complexity measure that quantifies the new information generation rate and is applicable to short time series, has been widely applied to physiological signal analysis since it was proposed. However, on one hand, sample entropy is easily affected by non-stationary sudden noise, because the tolerance during calculation is set to be proportional to standard deviation; on the other hand, it is not independent of the probability distribution, so that it does not purely characterize the new information generation rate. To solve these two problems, a new improved method named equiprobable symbolization sample entropy is proposed in this paper. Through equiprobable symbolization, the effects of both non-stationary sudden noises and probability distribution are eliminated. Besides, since equiprobable symbolization is usually non-uniform, it further breaks through the linear constrains in classic sample entropy. The method is proved to be rational by simulating three typical noises that have different time correlations and new information generation rates. Then the method is applied to electroencephalography (EEG) analysis. Results show that the method can successfully discriminate two different attention levels based on EEG with duration as short as 1.25 s and without removing any artificial artifacts. Therefore, the method is of great significance for EEG biofeedback, in which strong real-time abilities are usually required.

Keywords: symbolic dynamics, equiprobable symbolization, sample entropy, electroencephalography biofeedback

PACS: 05.10.–a, 89.70.cf, 87.19.le **DOI:** 10.7498/aps.63.100503

* Project supported by the Natural Science Foundation of Jiangsu Province, China (Grant No. BK2011565) and the National Natural Science Foundation of China (Grant No. 61271079).

† Corresponding author. E-mail: xlhuang@nju.edu.cn

‡ Corresponding author. E-mail: njhxliu@nju.edu.cn