# 一种基于隐含模式发现的时间序列处理算法

向 尷 蒋静坪
(浙江大学电气工程学院 杭州 310027)

**摘 要**：$\varepsilon$ 机由 Santa Fe 研究所(SFI)的学者最先提出，它致力于从时间序列中发掘隐含模式，并已成功应用到符号序列中。该文主要研究如何将 $\varepsilon$ 机应用到一般的时间序列中。分析了现有的符号化方法之后，在动态变换方法的基础上，提出了新的符号化方法，并将其成功应用到文中的实例研究中。改进了因果态分割重建算法，提出了简单的递归算法用来识别循环态并取得了很好的效果。实验发现，噪声污染和过程非平稳是 $\varepsilon$ 机处理方法中的主要障碍，它们将是我们以后工作的重点。

**关键词**：$\varepsilon$ 机；因果态；模式
**中图分类号**：TN911.7，N945.16 **文献标识码**：A **文章编号**：1009-5896(2007)01-0059-04

# An Algorithm for Time Series Based on Hidden Pattern Discovery

Xiang Kui Jiang Jing-ping
(*College of Electrical Engineering*, *Zhejiang University*, *Hangzhou* 310027, *China*)

**Abstract**: Epsilon machine is a new algorithm that tries to discover hidden patterns from data. Recently, the scholars in Santefe Institute have already applied it in symbol series successfully, but new problems emerge in traditional time series. A symbolization method transforming the sampling data into symbol series is presented, which implies some information of the expectation and variance. After Causal-State Splitting Reconstruction (CSSR), hundreds of states are lumped in the result, and a new recursion program can pick out the deterministic states very easily. Noise and nonstationarity will stunt the epsilon machine and they are the main problems to be researched in the future.

**Key words**: Epsilon machine; Causal state; Pattern

## 1 引言

过去的一个世纪中，时间序列的研究取得了长足进步，出现了许多行之有效的处理方法。随着科学的发展，研究对象越来越复杂，并出现了“复杂性”科学，传统的时间序列处理方法正受到各种复杂数据的挑战，在这种情况下，有必要不断探索新的处理算法。

美国著名的 SFI 在 1992 年举办了一次时间序列方面的竞赛，竞赛提供了统一的时间序列数据，并面向全世界征集论文，1993 年，出版了竞赛的论文集《时间序列的预测：估计将来和理解过去》。这本论文集在过去的十多年里，被许多时间序列的研究者引用，从论文集的题目中，可以发现一个很有趣的问题：时间序列的研究可以分为两部分，对将来的估计和对过去的理解。出于实际需要，估计算法一直为人们所重视，现在流行的几乎都是基于估计的算法，比如统计学中以 ARMA 为代表的系列算法；各种神经网络算法等等。理论上，无论数据多么复杂，基于参数拟合的估计算法总能给出适当的预测结果。当复杂性科学出现后，这种观点逐渐受到了怀疑，许多数据背后隐藏的复杂性超乎我们的想象，采用传统的估计算法处理这些数据时，预测效果明显变差，算法本身的复杂程度急剧升高。这需要我们首先尽量理解复杂系统运行的模式，在此基础上再做出合理的预测来满足生产和控制的需要。

其实，基于理解的时间序列处理方法并非没有，时频分解就是一种，只是时频分解可以做的事情很多，很少有人把它归类到时间序列方法中。时频分解有很多种，除了传统的短时傅里叶变换外，小波变换和希尔伯特-黄变换是新的方法，在这方面，Percival[1] 和 Huang[2] 的工作值得重视。本文要介绍的是一种基于模式发现的时间序列处理算法-$\varepsilon$ 机，它是 SFI 的 Crutchfield 等人过去 20 多年的研究成果，最近，Shalizi[3] 把它应用到符号序列的研究中。我们要做的是把它应用到传统的时间序列处理中，并针对应用中存在的各种问题加以研究。

## 2 一种新的符号化方法

$\varepsilon$ 机最初的研究对象定义为离散无限集，实际使用时，研究对象为离散有限集，且集合元素种类有限。对于通常由采样得到的离散数据集，其长度有限，但元素种类偏多，需要先对原始数据进行符号化处理，然后才能使用 $\varepsilon$ 机进行计

2005-05-23 收到，2005-09-26 改回

算。下面我们先介绍数据的符号化方法。

随着符号动力学在非线性和复杂性研究方面取得的成功，越来越多的人试图将原始数据转化成符号序列来研究，符号化方法成为联结符号动力学和传统时间序列的纽带。符号化实际上是对原始数据的粗粒化处理，信息的损失不可避免，但它对研究问题的简化却显而易见，当然是否有必要符号化要视具体情况而定。一种好的符号化方法，首先需要简洁，符号元素种类要少，否则就失去了符号化的意义；其次要保证信息的损失量最小，最为理想的情况是用少量的符号元素表达出对研究至关重要的信息。目前流行的符号化方法可以分为两类：基于状态空间的符号化[4]和基于幅值尺度的符号化[5]。基于状态空间的符号化是研究混沌序列的重要方法之一，它试图通过对状态空间的合理划分对混沌行为做出合适的描述，是目前研究的热点问题，但与本文关系不大。基于幅值尺度的符号化方法可以分为两种：静态变换和动态变换。

假设有时间序列 $x_i(i=0,1,\cdots,N)$，它对应的符号序列为 $s_i(i=0,1,\cdots,N)$。

$$s_i=\begin{cases}1, & x_i>X\\ 0, & \text{其它}\end{cases} \tag{1}$$

式(1)为静态变换，其中 $X$ 为一个设定的门限值，一般设为 $x_i$ 的期望。

$$s_i=\begin{cases}1, & x_i>x_{i-1}\\ 0, & \text{其它}\end{cases} \tag{2}$$

式(2)为动态变换。尽管静态变换用的较多，但 $X$ 的确定比较困难。式(1)和式(2)中符号集大小都为 2，使用时，可根据需要选择更大的符号集。显然，符号集越大，表达的信息就越多，同时计算就越复杂。下面，我们将在动态变换的基础上加以改进，建立一种新的符号化方法——极值法。

首先假定时间序列 $x_i(i=0,1,\cdots,N)$ 中有局部极大值 $x_{m1}$，$x_{m2}$ 和极小值 $x_{n1}$，$x_{n2}$，如图 1 所示，相应的符号化定义如式(3)、式(4)。

$$s_i(i=n1+1,\cdots,m2)=\begin{cases}1, & x_{m2}>x_{m1}\\ 0, & \text{其它}\end{cases} \tag{3}$$

$$s_i(i=m2+1,\cdots,n2)=\begin{cases}1, & x_{n2}<x_{n1}\\ 0, & \text{其它}\end{cases} \tag{4}$$

将式(3)、式(4)与式(2)对应组合，得到 00，01，11，11 4 种元素。之所以对动态变换作上述改进，主要基于以下想法：

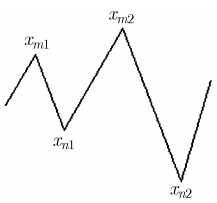


图 1 极值法示意图

(1)在最近有关时间序列的研究中，局部极值被证明含有特殊的信息，如Zaliapin[6]运用极值点所做的分形计算，Huang[2]运用极值包络线所做的EMD分解技术。

(2)期望和方差是研究随机序列时最为关注的两项指标，但过去的 ARMA 和时频分解方法都缺乏对期望的有效刻画方法。当极大值包络线和极小值包络线的趋势相同时，式(3)、式(4)包涵了时间序列的趋势项信息，或者说有关期望的信息；当两者的趋势不一样时，式(3)、式(4)则包涵了时间序列的方差信息。

(3)Palmer[7]的研究已经证明，过于简陋的符号化处理会影响 $\varepsilon$ 机算法的效果。

关于符号化方法优劣的评价，是一项非常复杂的工作，并且缺乏有效的评价手段，我们用统计复杂度的方法做了一些实验研究，将另行具文发表，限于篇幅，在此只给出一些简单结论。实验研究发现：当符号集大小相等时，在保留信息的能力方面，极值法要优于静态法，但比动态法差；极值法和动态法都适合做在线计算，静态法则比较困难；动态法适合处理杂乱的复杂数据，特别是非平稳数据，当数据相对简单时，动态法反而有失效的危险，极值法的适应性要比动态法稍强。

## 3 $\varepsilon$ 机与因果态分割重建(CSSR)

在介绍核心算法之前，需要弄清楚什么是模式和模式发现。关于什么是模式，已有人做过深刻的阐述。Hand[8]认为“模式是一个向量，用来描述数据点中异常的局部密度”，而 Shalizi[9]认为“模式意味着一种规则、结构、对称、组织等等”。关于模式发现，Shalizi[9]引用柏拉图《斐德罗篇》中的话做了非常精辟的论述，在一个过程中寻找模式就是“沿天然的方向，将其从节点处分开，而不是像拙劣的工匠一样，将其肢翼折成了两半”。模式发现关心的是模式是什么，应该怎么表达，它同流行的模式识别是有区别的。

用 $\overleftarrow{S}$ 表示一个离散时间，离散取值的随机过程，在任一时刻，这个过程可以被分为两部分：$\overleftarrow{S}$，$\overrightarrow{S}$。$\overleftarrow{S}$ 表示过去或历史，$\overrightarrow{S}$ 表示将来。$\overleftarrow{S}^L$ 表示 $\overleftarrow{S}$ 的最后 $L$ 个符号，$\overrightarrow{S}^L$ 表示 $\overrightarrow{S}$ 的最初的 $L$ 个符号，$s$ 表示 $S$ 的取值。什么是因果态？Shalizi[9]有非常复杂的论述，我们简单理解为：因果态是能最大限度预测未来的，最简洁的过程划分。$\varepsilon$ 函数表示一种从历史到历史集合的映射关系，进一步说，$\varepsilon$ 函数表示从历史到因果态的函数，它的定义如式(5)。

$$\varepsilon(\overleftarrow{s})=\left\{\overleftarrow{s}' \mid P\left(\overrightarrow{S}^L=\overrightarrow{s}^L \mid \overleftarrow{S}=\overleftarrow{s}\right)=P\left(\overrightarrow{S}^L=\overrightarrow{s}^L \mid \overleftarrow{S}=\overleftarrow{s}'\right),\ \forall\overrightarrow{s}^L\in\overrightarrow{S}^L,\overleftarrow{s}'\in\overleftarrow{S}\right\} \tag{5}$$

所有 $\varepsilon$ 函数合起来称为过程的 $\varepsilon$ 机。因果态有很多性质，其中最重要的一点是，所有因果态构成一个马尔可夫过程。

为了叙述方便，先介绍一些名词。

字、父亲字、儿子字：我们把一系列符号的组合称为“字”，如 0101，并把字中包含的符号个数称为字长。把*010 称为 0101 的父亲字，对应的，0101 称为*010 的儿子字。任何一个字，最多只有一个父亲字，却可能有多个儿子字。

状态、变体：将一些字按统一的规则组成一个集合，该集合称为状态，在 CSSR 中，规则定义为关于未来的条件概率分布，它被称为状态的变体。

父状态、子状态：在状态 $A$ 的某个字后添加一个符号 $f$，如果形成的所有新字都属于状态 $B$，称状态 $A$ 按标识 $f$ 转移到状态 $B$，状态转移的概率等于对应的字之间的转移概率，且 $A$ 称为 $B$ 的父状态，$B$ 称为 $A$ 的子状态。

瞬时态、循环态：有关瞬时态和循环态的严格讨论可以参阅文献[10]，在此用图 2 做简单说明。图 2 中，$A$，$B$ 为瞬时态，$C$，$D$，$E$ 为循环态，一个简单的理解就是，$A$，$B$ 可以沿箭头所示方向，按一定的概率转移到 $C$，$D$，$E$，但 $C$，$D$，$E$ 不可能转移到 $A$，$B$。更进一步，把特殊情况 $E$ 称为孤立态。

有关 $\varepsilon$ 机的核心算法，在文献[3]中有较为详细的论述，鉴于该算法并不为很多人熟悉，下面首先简要复述算法的内容，有兴趣的读者可以直接参考原文。原文的算法是针对理想的符号序列提出的，在应用到一般的时间序列中还有很多问题，我们将就实际使用中遇到的问题展开论述。

Shalizi[3] 将他的 $\varepsilon$ 机求解算法称为因果态分割重建，分为 3 步：初始化、一致化、定常化。

(1) 初始化　统计符号序列中字长小于等于 $L$ 的所有字，计算它们关于未来的条件概率分布 $P\left(\vec{S}^1 = s \mid \vec{s}^l\right)$，其中 $l = 0,1,\cdots,L$。

(2) 一致化　一个字与某状态的变体做 K-S 检验，结果大于设定的置信水平 $\alpha$ 时，将该字加入到状态中，计算状态的新变体。每个字都优先加入到其父亲字所属的状态中，当某个字与其父亲字不属于同一个状态时，将父亲字从原状态中删除。反复执行第(2)步，直至所有小于等于 $L$ 的字都被加入到状态中，然后从状态中删除所有长度小于 $L$–1 的字。

(3) 定常化　假设状态 $A$ 的所有字分为两个集合 $M_1$，$M_2$，其中 $M_1$ 添加标识 $f$ 后指向状态 $B$，$M_2$ 添加标识 $f$ 后指向状态 $C$，则 $M_1$ 或 $M_2$ 需要从状态 $A$ 中分离出来，形成新的状态 $D$。这样做主要是为了满足状态机理论中关于定常状态的要求。当所有状态都满足定常状态要求后，删除状态集合中的瞬时态，只留下循环态。

Shalizi[3] 利用偶数过程的例子来说明其算法，并取得了很好的效果，在面对一般的时间序列时，还有很多问题需要解决。首先是时间序列中的噪声污染问题，因果态反映的是系统内部隐含的固有结构，噪声污染不可避免会影响到因果态的建立，在分析时间序列之前，用滤波算法去除噪声是必要的。由于实验数据的性质往往并不十分清楚，降噪的效果可能比较差，这时噪声的污染会使计算结果中出现一些无法分辨的结构。同时，由于数据长度偏短，带来统计上的波动，也会导致一些无法分辨的结构，关于这一点，Clarke[11] 有一些论述。综合 Clarke 的观点，可以得到这样的结论：随着时间序列长度的增加，因果态的复杂熵将收敛到某个最大值，如果此时仍然有无法分辨的结构存在，则认为是噪声污染使然。

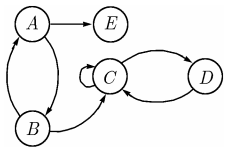

图 2　瞬时态与循环态示意图

一致化的过程结束后，往往会产生数以百计的状态，其中很多都是瞬时态，这时要通过人工从中分辨出循环态根本不可能，Shalizi[3] 并没有提供合适的方法，在此我们补充一个简单的算法，它可以在 C 语言下用简短的递归程序来实现。

(1) 从状态集合中任选一个状态 $A$(状态 $A$ 的性质尚不确定)，找到它所有的子状态，将其加入到一个子状态集合 $\Omega_1$ 中，依次寻找 $\Omega_1$ 中每个元素的子状态，并将其添加到 $\Omega_1$ 中，如此循环，直到状态 $A$ 沿子状态转移方向能到达的所有状态都位于 $\Omega_1$ 中。最后将 $A$ 也加入到 $\Omega_1$ 中。

(2) 如步骤(1)，求出状态 $A$ 的父状态集合 $\Omega_2$。最后将 $A$ 也加入到 $\Omega_2$ 中。

(3) 令 $\Omega_3 = \Omega_1 \cap \Omega_2$，$\Omega_4 = \Omega_1 \cap \bar{\Omega}_3$，如果 $\Omega_4$ 为空集，则 $\Omega_3$ 中所有的状态都是循环态，否则，$\Omega_3$ 中所有的状态都为瞬时态。

(4) 重复(1)～(3)的过程，依次扫描所有的状态，直到所有状态的性质都被确定。

当数据过于简单或过于复杂时，循环态中会有很多孤立态出现，此时系统的统计复杂度趋于 0[12]，因此，$\varepsilon$ 机在处理周期信号和纯噪声信号时效果较差。

## 4　实例分析

我们选取 SFI 在 1992 年举办的时间序列竞赛中提供的数据 $B$ 作为研究对象，该数据分 3 部分，分别表示一个病人在一段时间内的呼吸、血液氧含量和心跳的变化情况，全部数据长为 34000 点，采样时间 0.5s，详细情况读者可以通过 Internet 获取。我们从三段数据分别选取 5000 点(对应时间相同)作研究用，按照第 2，3 节描述的方法，用 VC6.0 编写程序，结果分析如下。

首先用第 2 节的新方法，对所有数据做符号化处理，得到 3 组符号化序列。令最大字长分别为 $L$=3，置信水平为 0.01，计算得到图 3，图 4，图 5 的因果态结构图。简单起见，画图时略去了状态间的转移概率和转移的标识。3 个图中因果态个数，结构都不尽相同。鉴于 $\varepsilon$ 机的目的是从数据中发掘隐含的模式，其目的是建立一种理解系统特征的框

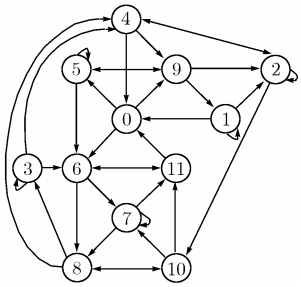


图 3 呼吸数据中的因果态

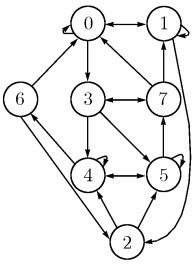


图 4 血液氧含量数据中的因果态

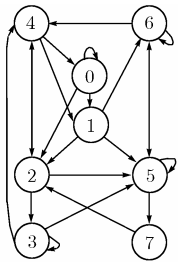


图 5 心跳数据中的因果态

架，至于什么是系统真正隐含的模式，并没有统一的判断标准，因此我们不能武断地认为图 3，图 4，图 5 就代表了系统的隐含模式。但是，图 3，图 4，图 5 实际上就是隐马尔可夫模型(HMM)。所不同的是，在建立模型之前，$\varepsilon$ 机并没有对模型的结构和状态的个数做出任何假设，按照 Palmer[13] 的说法，这种模型是从数据中推理得到的，而不是人为的力量将模型强加于数据本身。用 CSSR 建立的因果态具有所有 HMM 的性质，我们完全可以在此基础上展开对系统的分析和预测工作。

## 5 结束语

毫无疑问，随着研究的深入，$\varepsilon$ 机将逐渐成为一种强有力的信号处理方法，但目前，它还有很多地方需要完善，包括理论上的延伸和处理技巧的提高。

尽管研究者们正致力于将 $\varepsilon$ 机推广到连续的随机过程，但目前我们仍然需要首先对时间序列做符号化处理后才能研究，天然的符号序列毕竟是少数。寻求合适的符号化方法看似简单，真正有效的方法并不多。另外，对不同的符号化方法建立统一的评价标准也是一项非常有必要的工作。

$\varepsilon$ 机在令人头痛的非线性领域已经证明有好的效果，特别是在混沌非线性领域，已有一些应用文章出现，有兴趣的读者可以访问 Jim Crutchfield 的个人主页。但是，当随机过程非平稳时，使用 $\varepsilon$ 机来研究可能会面临许多新的困难，甚至会彻底失效。在实验数据中，非平稳的情况是不可避免的，它会阻碍我们寻找系统隐含的因果态，要么使计算结果崩溃，变成单状态机，表达出非常有限的信息；要么使计算结果变得非常庞大，使人难以理解。非平稳带来的大量难以分辨的结构，同噪声污染带来的影响几乎一样，如何正确评价 $\varepsilon$ 机的计算结果，是一个非常棘手的问题。

一个初步的想法是将因果态的定义加以推广，使其能适应系统非平稳的影响，并希望能借此来获取系统内部关于非平稳的信息。这样做除了理论上的困难外，还有一个现实的问题，因果态重建过程中将有大量的参数需要预先确定，不同参数组合带来的大量的计算结果往往使人难以分辨真伪，这会使我们丧失掉 $\varepsilon$ 机的根本出发点——模式发现。

## 参 考 文 献

[1] Percival D and Walden A. Wavelet Methods for Time Series Analysis. London: Cambridge University Press, 2000: 56–254.

[2] Huang N. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc.R.Soc.Lond.*, 1998, A(454): 903–995.

[3] Shalizi C, Shalizi K, and Crutchfield J. An algorithm for pattern discovery in time series. SFI Working Paper, 2002: 02-10-060.

[4] Tang X and Tracy E. Symbol sequence statistics in noisy chaotic signal reconstruction. *Physical Review E*, 1995, 51(5): 3871–3889.

[5] Kurths J, Schwarz U, and Witt A, *et al*.. Measures of complexity in signal analysis. In: chaotic, fractal, and nonlinear signal processing. AIP Conference Proceedings, Woodbury, New York, 1996: 33–54.

[6] Zaliapin I, Gabrielov A, and Borok V. Multiscale trend analysis. *Fractals*, 2004, 12(3): 275–292.

[7] Palmer A, Fairall C, and Brewer W. Complexity in the atmosphere. *IEEE Trans. on Geoscience and Remote Sensing*, 2000, 38(4): 2056–2063.

[8] Hand D. Pattern detection and discovery. In: Hand D, Adams N, Bolton R Eds. Pattern Detection and Discovery, ESF Exploratory Workshop, London, UK, September 16-19, 2002, Berlin Heidelberg: Springer-Verlag, 2002: 1–12.

[9] Shalizi C and Crutchfield J. Computational mechanics: pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 2001, 104(3): 817–879.

[10] Upper D. Theory and algorithms for Hidden Markov models and generalized Hidden Markov models. [PhD thesis], University of California, Berkeley, 1997.

[11] Clarke R, Freeman M, and Watkins N. Application of computational mechanics to the analysis of natural data: an example in geomagnetism. *Physical Review E*, 2003, 67: 016203.

[12] Crutchfield J. The calculi of emergence: Computation, dynamics and induction. *Physica D*, 1994, 75: 11–54.

[13] Palmer A, Schneider T, and Benjamin A. Inference versus imprint in climate modeling. *Advances in Complex Systems*, 2002, 5(1): 73–89.

向 馗： 男，1976 年生，博士生，研究系统复杂性、非平稳时间序列.

蒋静坪： 男，1935 年生，教授，博士生导师，主要研究智能系统与智能控制、先进控制策略及算法.