基于隐含模式的异常检测算法

向 尰　蒋静坪
(浙江大学电气工程学院　杭州　310027)

摘　要：如何检测系统中的临界变化，一直是一个难题。该文提供了一种新的基于隐含模式的异常检测算法。ε机是一种新的计算力学理论，它能从时间序列中发掘系统的隐含模式。因果态分割重建算法(CSSR)是目前重构ε机的最成熟算法，它可以推理出一个因果态集合，所有的因果态构成一个隐马尔可夫模型。在因果态集合的基础上，建立一个表达系统特征的向量，不同向量间的距离可以定义成系统异常的测度。把时间序列分段，分别计算每部分的异常度，就可以得到系统的异常演变曲线。在 Duffing 振子的例子中，该算法不仅有效检测，还提前预测到系统分叉的发生，说明该算法具有很好的应用潜力。

关键词：异常检测；隐含模式；ε机；时间序列
中图分类号：TP206$^+$.3　　**文献标识码**：A　　**文章编号**：1009-5896(2007)06-1487-05

An Anomaly Detection Algorithm Based on Hidden Pattern

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

Abstract: It is a difficult problem how to detect such accident of a system. This paper presents a new algorithm, an anomaly detection algorithm based on hidden pattern. Epsilon machine, a new computational mechanics, can discover hidden pattern from the response time series. Causal State Splitting Reconstruction (CSSR), one algorithm of epsilon machine, can infer a set of causal states, which has an analogy to hidden Markov chain. Based on this set, an anomaly measure can be defined, which is the distance of two characteristic vectors. Computing all parts of the time series, an anomaly evolution curve can be got. In simulation analysis of Duffing equation, step changes appear in the anomaly curve, before Duffing oscillator begin to bifurcate. The algorithm proves to be effective in anomaly detection and warning.

Key words: Anomaly detection; Hidden pattern; Epsilon machine; Time series

1　引言

最近 30 多年，在复杂性研究中，出现了许多关于系统认知的新概念，如：耗散结构、临界相变、涌现等等。虽然没有一个概念可以一统天下，但它们相辅相成，展现了一个全新的科学视野。实际上，在复杂性之前，人们习惯于用更为通俗的哲学概念——量变和质变来看待同样的现象。复杂性研究正致力于回答这样一些问题：量变的累计为什么会导致质变？质变为什么难以预测和控制？质变是如何从量变中孕育出来的？

在分析复杂性概念的基础上，我们可以描述下面的场景：一个具有耗散结构的系统，在外力的作用下，内部微观结构不断发生演变，当系统参数穿越某个临界值时，系统进入一个新的有序状态。系统参数穿越临界值的过程，也许正是一个量变到质变的跃迁过程。金属的疲劳断裂，地震的发生，人体的亚健康状态都是这方面的例子。

复杂系统在外力作用下发生的临界变化，虽然机理复杂，但是，系统在外力作用下的响应输出，往往蕴涵许多关于内部参数演变的信息，如果能观测到系统的输出响应，并从中捕获到系统特征的演变情况，就能对临界相变的发生做出有效的预测。一般说来，测量系统输出相对容易，测量输入比较难，所以经常把输出响应当作时间序列来研究，且时间序列的非平稳变化对应着系统特征的演变情况。如何建立一种有效的时间序列算法来预测系统的临界变化，将是本文研究的重点。

2　理论工具

如何发现、表达系统在临界变化时内部特征的演变情况，美国圣塔菲研究所(Santa Fe Institute)的 Crutchfield 和他的研究小组，在过去十几年间，一直致力于这方面的工作，他们发展了一种新的统计力学工具——ε机[1]来处理上述问题。本节将着重介绍ε机的原理，至于它的算法，读者可以自行参阅文献[2, 3]。

在ε机的整个框架中，最重要的概念就是“模式”和“模式发现”。什么是模式？Shalizi[1]认为“模式意味着一种规则、

2005-10-31 收到，2006-05-15 改回

结构、对称、组织等等”。至于模式发现，Shalizi[1]引用柏拉图的《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”。模式发现主要关心模式是什么，应该怎么表达，它同流行的模式识别是有区别的。

给定一个离散时间、离散取值的随机过程，$\cdots S_{-2}S_{-1}S_0S_1S_2\cdots$，$S_i$ 来源于一个符号集 A。假设该过程平稳，在任意时刻，可以把过程分为历史 $\overleftarrow{S}$ 和未来 $\vec{S}$ 两部分，未来时间的前 L 个字符定义为 $\vec{S}^L$，历史时间的后 L 个字符定义为 $\overleftarrow{S}^L$。要想预测 $\vec{S}$，首先要确定 $\overleftarrow{S}$ 关于未来的条件概率分布 $P\left(\vec{S}\mid\overleftarrow{s}\right)$，也就是要建立一个方程ε，实现从 $\overleftarrow{s}$ 到 $P\left(\vec{S}\mid\overleftarrow{s}\right)$ 的映射。其实，一种预测就是在 $\overleftarrow{S}$ 上施加一种划分，划分得到的单元中，每个元素对未来的预测都相同，这些单元称为有效态。如果有效态的划分满足计算力学的两个准则：最确实的预测能力和最简洁的表达形式，则有效态被称为因果态。对于任何一个符号过程，因果态的划分是唯一的。对任意大小的 L，有下面的结果[1]：

$$\varepsilon(\overleftarrow{s})=\left\{\overleftarrow{s}'\mid P\left(\vec{S}^L=\vec{s}^L\mid\overleftarrow{S}=\overleftarrow{s}\right)=P\left(\vec{S}^L=\vec{s}^L\mid\overleftarrow{S}=\overleftarrow{s}'\right),\ \forall\vec{s}^L\in\vec{S}^L,\overleftarrow{s}'\in\overleftarrow{S}\right\} \tag{1}$$

设因果态为 $\mathcal{S}$，定义因果态之间的转移及转移概率如下。

$$\boldsymbol{T}_{ij}^{(s)}=P\left(\vec{S}^1=s,\mathcal{S}'=\sigma_j\mid\mathcal{S}=\sigma_i\right) \tag{2}$$

$$\sum_{s\in A}\sum_{\sigma_j\in E}\boldsymbol{T}_{ij}^{(s)}=\sum_{s\in A}P\left(\vec{S}^1=s\mid\mathcal{S}=\sigma_i\right)=1 \tag{3}$$

因果态有 3 个主要的性质：

(1) 齐次性。任意两段历史，如果它们关于未来的条件概率分布相同，则它们必属于同一个因果态，反之，因果态中所有的元素关于未来的条件概率分布都相同；

(2) 确定性。给定当前的状态和后续的符号，则存在唯一的后续状态。

(3) 马尔可夫性。因果态本身构成了一个马尔可夫过程。

上述性质的详细证明可参见文献[1]。

对于一个复杂系统，它的隐含模式也许永远不为人知，我们也不能武断地认为ε机就能发掘系统真正的隐含模式。但是，由ε机计算得到的因果态至少是对系统隐含模式的有效表达，它从系统的表象出发，使我们比以往任何时候都更逼近系统的本质。

3 应用实践

假设有一类非线性、非保守系统，它们在外力的作用下，呈现出两种不同时间尺度的现象，其中一个随时间快速变化，另一个随时间缓慢变化。设系统在外力作用下满足差分方程

$$\dot{x}(t_f)=f(x(t_f),\theta(t_s)) \tag{4}$$

式(4)中，时间序列 $x(t_f)$ 为系统响应，随时间快速波动；$\theta(t_s)$ 为系统参数，随时间缓慢变化。系统的异常演变表现为时间尺度 t_s 上参数 θ 的变化，它分散在时间序列 $x(t_f)$ 中。本文的任务就是设法揭示隐藏在 $x(t_f)$ 中的关于参数 θ 的变化情况，描绘出系统的异常演变曲线。为此假设系统满足下面两个假设：

(1) 在时间尺度 t_f 上，认为时间序列 $x(t_f)$ 是分段平稳的；

(2) 任何观测到的非平稳变化都被认为是由系统参数 θ 的变化引起的。

上述关于系统的定义和假设来自于文献[4]，我们的工作将在此基础上展开。算法一共分为 4 部分，下面分别加以叙述。

3.1 符号化处理

检测系统的输出响应，得到一段时间序列 $\{x_i\}$，$i=1,2,\cdots,N$，将其符号化得到符号序列 $\{s_i\}$。目前，ε机还只能用于符号序列，实践中遇到的多数是实数序列，因此首先要设法将实数序列作符号化处理，得到符号序列。这样做虽然损失了部分原始信息，但对简化计算有好处。有关符号化方法的选取，国内外已做过一些研究，读者可以参阅文献[5~7]，我们已对此做过专门研究，文章将另行发表，在此只给出一些简单的结论。

符号化处理的目标是：用最小的符号集表达最多的原始信息。符号集的大小与保留信息的能力及计算复杂性都息息相关，实际使用时要根据具体情况权衡利弊。符号化方法主要有两种，一是按照信号的幅值直接划分，称为静态法；二是按照信号的导数大小作划分，称为动态法。从我们检验的结果来看，当信号相对复杂时，动态法效果较好，反之则静态法效果好。本文第 4 节所举的例子比较简单，所以采用了静态法。

3.2 计算因果态

将 $\{s_i\}$ 等分成长度为 K 的等长子序列 s^K，如果序列的总长度 N 是 K 的整数倍，则可以得到 N/K 段子序列。对每个子序列 s^K，重构其ε机，得到因果态集 E。

重构ε机的算法主要有两种，一是 Young[3]的子树合并法；二是 Shalizi[2]的因果态分割重建法。子树合并法虽然操作简单，但计算结果中经常含有非确定状态，且收敛速度缓慢，本文中将采用因果态分割重建法，该算法过程相对复杂，我们已另具文作详细阐述，读者可自行参阅文献[2]。不管用哪种算法，计算结果都是一个隐马尔可夫结构，隐马尔可夫结构中的每个状态都是一个因果态，所有因果态的集合构成了系统的模式划分，计算过程实际就是一个模式发现的过程。

3.3 提取特征向量

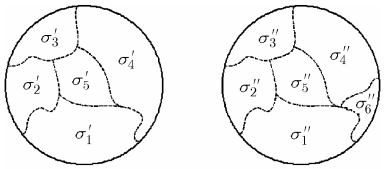
从因果态集 E_j 中提取特征向量 $\boldsymbol{V}_j$。向量 $\boldsymbol{V}_j$ 的定义是算法的关键，下面举例说明。

假设从两个符号子序列中计算得到因果态集 E_1， E_2，$E_1=\{\sigma_1^{'},\sigma_2^{'},\sigma_3^{'},\sigma_4^{'},\sigma_5^{'}\}$， $E_2=\{\sigma_1^{''},\sigma_2^{''},\sigma_3^{''},\sigma_4^{''},\sigma_5^{''},\sigma_6^{''}\}$。因为每个因果态集都是对系统隐含模式的表达，我们用图 1 作形象地描述，圆圈代表了一个系统，因果态集的元素构成了对系统模式的划分。

首先需要确定集合 E_1 和 E_2 中各元素间的对应关系。在因果态分割重建算法中，每个因果态都有一个关于未来的条件概率分布，且在同一个因果态集中，任何两个因果态的条件概率分布都显著不同。对 E_1 和 E_2 中各因果态做 KS 检验[8]，高于设定置信度的两个因果态被认为存在对应关系，结果如下所示。

$\sigma_1^{'}\leftrightarrow\sigma_1^{''}$， $\sigma_2^{'}\leftrightarrow\sigma_2^{''}$， $\sigma_3^{'}\leftrightarrow\sigma_3^{''}$， $\sigma_4^{'}\leftrightarrow\sigma_4^{''}$， $\sigma_5^{'}\leftrightarrow\sigma_5^{''}$，$\sigma_4^{'}\leftrightarrow\sigma_6^{''}$。

从计算结果中发现， $\sigma_4^{'}$ 不仅与 $\sigma_4^{''}$ 对应，还与 $\sigma_6^{''}$ 对应。观察图 1 不难看出，在构建 E_2 时，原来的因果态 $\sigma_4^{'}$ 分裂成 $\sigma_4^{''}$ 和 $\sigma_6^{''}$ 两部分，这说明系统内部的隐含模式随时间发生了变化，也说明系统特性参数发生了大的改变。要想解决这种两难的对应关系，需要建立新的判断标准。设相似度函数 $f_S(\sigma_i,\sigma_j)$ 表达了因果态 σ_i 和 σ_j 的相似程度，如果 $f_S(\sigma_4^{'},\sigma_4^{''})>f_S(\sigma_4^{'},\sigma_6^{''})$，则关于 $\sigma_4^{'}$ 的正确的对应关系是 $\sigma_4^{'}\leftrightarrow\sigma_4^{''}$。



(a) E_1 示意图　　(b) E_2 示意图

图 1　因果态集示意图

在因果态重建算法中，设字符串 $\bar{s}$ 的最大长度为 L，则计算得到的因果态中，最多含有两种长度的字符串，其长度分别为 L–1，L。每个因果态中都含有许多不同种类的字符串，设 $N_i(s^L)$ 表示因果态 σ_i 中长度为 L 的字符串 s^L 的个数，定义因果态 σ_i 和 σ_j 的公共集 $\Omega_{ij}=\{s^L\mid s^L\in\sigma_i,s^L\in\sigma_j\}$，则相似度函数定义为

$$f_S\left(\sigma_i,\sigma_j\right)=\sum\left(N_i(s^L)+N_j()s^L\right),\quad s^L\in\Omega_{ij} \tag{5}$$

严格意义上，相似度函数并不能完全解决上述两难的对应关系，但在工程上，KS 检验和相似度函数合起来已经足够使用。

定义

$$N(\sigma_i)=\sum N_i(s^L),\quad s^L\in\sigma_i \tag{6}$$

则图 1 的例子中， E_1 和 E_2 对应的特征向量分别为

$$\boldsymbol{V}_1=[N(\sigma_1^{'}),N(\sigma_2^{'}),N(\sigma_3^{'}),N(\sigma_4^{'}),N(\sigma_5^{'}),0] \tag{7}$$

$$\boldsymbol{V}_2=\left[N(\sigma_1^{''}),N(\sigma_2^{''}),N(\sigma_3^{''}),N(\sigma_4^{''}),N(\sigma_5^{''}),N(\sigma_6^{''})\right] \tag{8}$$

3.4 计算异常度

对应于第 j 个子序列 s^K，有一个特征向量 $\boldsymbol{V}_j$。设系统在初始时刻对应的特征向量为 $\boldsymbol{V}_0$，则系统在第 j 个时间段对应的异常度 Λ_j 为向量 $\boldsymbol{V}_0$ 与 $\boldsymbol{V}_j$ 之间的距离。

$$\Lambda_j=M\left(\boldsymbol{V}_0,\boldsymbol{V}_j\right) \tag{9}$$

关于向量之间的距离定义，有很多种方法，最常见的是 Euclid 距离以及向量之间的夹角。在后面的例子中我们选取向量的夹角作为距离的度量。

$$\Lambda_j=\cos^{-1}\left\{\frac{\left\langle\boldsymbol{V}_0,\boldsymbol{V}_j\right\rangle}{\left\|\boldsymbol{V}_0\right\|\cdot\left\|\boldsymbol{V}_j\right\|}\right\} \tag{10}$$

4　实例分析

为了验证上一节的算法，建立一个理想的人工系统：Duffing 振子，Duffing 振子是最常见的非线性系统，内部机理的研究非常成熟，便于与我们的检验结果比对。设有下面的动力学方程：

$$\frac{\mathrm{d}^2y(t)}{\mathrm{d}t^2}+\beta\left(t_s\right)\frac{\mathrm{d}y(t)}{\mathrm{d}t}+y(t)+y^3(t)=A\cos\omega t \tag{11}$$

令 $A=22.0$， $\omega=5.0\mathrm{rad/s}$， $\beta\left(t_s\right)$ 初值为 0.11，随时间均匀缓慢增大到 0.35[4]。用四阶 Runge-Kutta 解微分方程，设时间步长为 0.01，采样比为 12:1，得到长度为 2.5×10^5 个数据点的实数序列。按照第 3 节算法的 4 个步骤依次进行下面的计算。

(1) 取符号集大小为 4，用静态法对该实数序列作符号化处理，得到等长的符号序列。

(2) 将符号序列等分成 25 个子序列，为每个子序列建立ε机。在 CSSR 算法中，设最大字符串长度为 2，KS 检验的最小置信度水平为 0.01，计算得到 25 个因果态集。

(3) 求出每个因果态集对应的特征向量。

(4) 计算每个时间段的异常度 Λ。依次连接所有的异常度，就得到系统随时间的异常演变曲线，如图 2 中的曲线 a 所示。

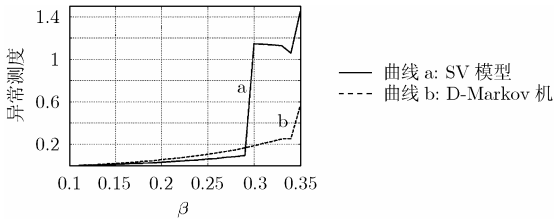


图 2　Duffing 振子的异常演变曲线

观察图 2 的曲线 a 可以发现，异常曲线中有 2 段相对平滑，对应的β值范围依次为：0.11~0.29，0.30~0.35。当β值为 0.29~0.30，曲线 a 出现很大的跳跃。在曲线的跳跃阶段，说明系统内部特征发生了大的变化，或即将发生大的变化，事实是否真的如此，我们结合下面的相位图作进一步分析。

观察图 3 的相位图，当β=0.31~0.32 时，系统出现了分叉，特征发生了很大改变，而异常曲线却提前在β=0.29~0.30 时发生了大的跳跃，这说明，本文的异常检测方法不仅能有效“检测”到异常，更重要的是能提前“预测”到异常的来

临。基于隐含模式的异常检测算法具有“预测”异常的功能，绝非偶然，是因为ε机的建立，有效地抓住了系统内部的特征，抓住了一些质变发生前的细微征兆。

再来看曲线 a 为什么在β=0.29~0.30 时会表现出大的跳

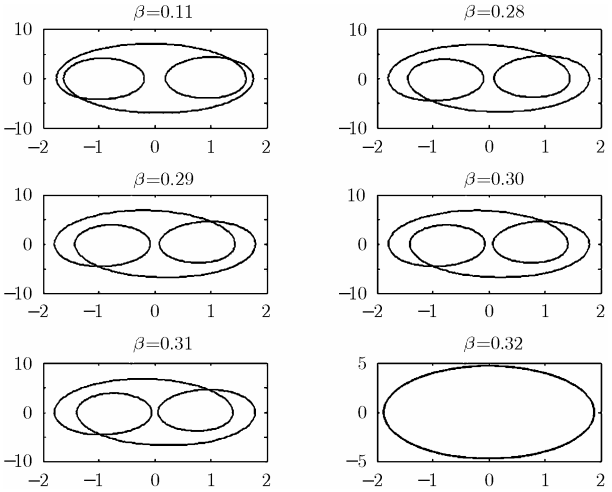


图 3 Duffing 振子在不同β 时的相位图

跃，表 1 给出了β=0.29 时结构向量$\boldsymbol{V}_{0.29}$与β=0.11 时的向量$\boldsymbol{V}_{0.11}$的对应情况，相比$\boldsymbol{V}_{0.11}$而言，$\boldsymbol{V}_{0.29}$中并没有出现新的因果态，但σ_1'与σ_1''，σ_3'与σ_3''的 KS 检验结果表明，它们已经有所区别。

继续观察表 2，此时σ_1'与σ_1''，σ_3'与σ_3''已经不再具有对应关系。这就好比σ_1'与σ_3'这两个因果态因为β值的变化而慢慢“变质”了，一旦“变质”被模型所确认，异常曲线就发生了跳跃。

在图 2 中，还给出了一条曲线 b，它是用 D-Markov 机计算得到的。D-Markov 机是 Ray[5]提供的算法，与本文算法极其相似。两者的区别在于，D-Markov 机直接在字符串的基础上建立 D-Markov 模型，推导出系统的极限状态向量，并用它来计算异常测度，而本文的算法是在字符串中重构ε机，借此来抓住系统内部的隐含特征。从图 2 可以看出，在预报系统的异常时，本文算法比 D-Markov 机要更及时，更果断。

过去的许多故障诊断算法都是以“检测”为目的，主要通过故障发生后系统响应的外在表现做出判断，因而缺乏预

表 1 β=0.29 时的向量

β=0.11		β= 0.29		KS 检验	相似度
σ_i'	$N(\sigma_i')$	σ_i''	$N(\sigma_i'')$		
σ_1'	2262	σ_1''	2073	0.0127	4335
σ_2'	636	σ_2''	637	1.0000	1273
σ_3'	2265	σ_3''	2076	0.0142	4341
σ_4'	1462	σ_4''	1648	0.9999	3110
σ_5'	637	σ_5''	637	1.0000	1274
σ_6'	637	σ_6''	636	1.0000	1273
σ_7'	636	σ_7''	636	1.0000	1272
σ_8'	1463	σ_8''	1655	0.9999	3118

表 2 β=0.30 时的向量

β=0.11		β= 0.30		KS 检验	相似度
σ_i'	$N(\sigma_i')$	σ_i''	$N(\sigma_i'')$		
σ_1'	2262				
σ_2'	636	σ_2''	636	0.9999	1272
σ_3'	2265				
σ_4'	1462	σ_4''	1680	0.9999	3142
σ_5'	637	σ_5''	637	1.0000	1274
σ_6'	637	σ_6''	636	0.9999	1273
σ_7'	636	σ_7''	636	1.0000	1272
σ_8'	1463	σ_8''	1671	0.9999	3134
		σ_1''	2051		
		σ_3''	2051		

警能力。当故障带来的危害巨大时，预警机制非常重要，此时本文的算法就体现出了它的优越性。虽然只用 Duffing 振子的例子做了简单说明，但式(4)关于系统的假设具有一般性，工程中很多例子都满足这些条件，比如疲劳断裂、疾病发作、地震爆发等，因此，本方法有望在许多领域得到推广。

参 考 文 献

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

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

[3] Crutchfield J and Young K. Inferring statistical complexity. *Physical Review Letters*, 1989, 63(2): 105–108.

[4] Chin S. Real time anomaly detection in complex dynamic systems. [PhD thesis], The Pennsylvania State University, 2004.

[5] Ray A. Symbolic dynamic analysis of complex systems for anomaly detection. *Signal Processing*, 2004, 84(7): 1115–1130.

[6] Daw C, Finney C, and Tracy C. A review of symbolization analysis of experimental data. *Review of Scientific Instruments*, 2003, 74(2): 1–18.

[7] 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.

[8] 庄楚强, 吴亚森. 应用数理统计基础. 广州: 华南理工大学出版社, 1999: 259–265.

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.