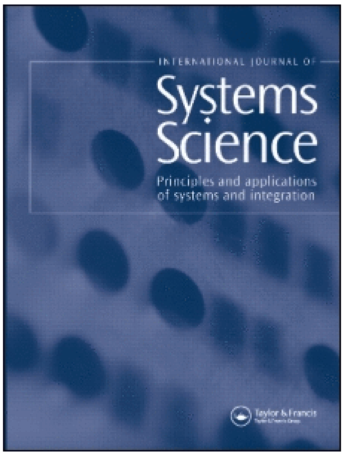

International Journal of Systems Science
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713697751

Anomaly detection in complex system based on epsilon machine

K. Xiang[a]; X. Zhou[b]; J. P. Jiang[c]
[a] School of Automation, Wuhan University of Technology, Wuhan 430070, PR China [b] College of Mechanical and Energy Engineering, Zhejiang University, Hangzhou 310027, PR China [c] College of Electrical Engineering, Zhejiang University, Hangzhou 310027, PR China

To cite this Article Xiang, K. , Zhou, X. and Jiang, J. P.(2008) 'Anomaly detection in complex system based on epsilon machine', International Journal of Systems Science, 39: 10, 1007 — 1016
To link to this Article: DOI: 10.1080/00207720802011266
URL: http://dx.doi.org/10.1080/00207720802011266

International Journal of Systems Science
Vol. 39, No. 10, October 2008, 1007–1016

Anomaly detection in complex system based on epsilon machine

K. Xiang[a]*, X. Zhou[b] and J.P. Jiang[c]

[a]*School of Automation, Wuhan University of Technology, Wuhan 430070, PR China;* [b]*College of Mechanical and Energy Engineering, Zhejiang University, Hangzhou 310027, PR China;* [c]*College of Electrical Engineering, Zhejiang University, Hangzhou 310027, PR China*

(*Received 16 January 2007; final version received 22 February 2008*)

Epsilon machine is a computational mechanics theory and its most effective reconstruction algorithm is causal state splitting reconstruction (CSSR). As CSSR can only be applied to symbol series, symbolising real series to symbol series is necessary in practice. Epsilon machine discovers the hidden pattern of a system. In reconstructed results, the hidden pattern is expressed as the set of causal states. Based on the variation of causal states, a novel anomaly detection algorithm, structure vector model, is presented. The vector is composed of the causal states, and the anomaly measure is defined with the distance of different vectors. An example of the crankshaft fatigue demonstrates the effectiveness of the model. The mechanism of the model is discussed in detail from three aspects, computational mechanics, symbolic dynamics and complex networks. The new idea defining anomaly measure based on the variation of hidden patterns can be interpreted reasonably with the hierarchical structure of complex networks. The jump in anomaly curves is a nature candidate for the threshold, which confirms the positive meaning of the model. Finally, the parameter choice and time complexity are briefly analysed.

Keywords: anomaly detection; structure vector; epsilon machine

1. Introduction

'Anomaly in a dynamic system is defined as a deviation from its nominal behaviour and can be associated with parametric or non-parametric changes that may gradually evolve in the system' (Ray 2004). Such deviations often hide in the output signal of the system. 'The problem of anomaly detection is that of finding deviations in the characteristics of the system' (Singh 2002).

Three kinds of methods have been widely used in anomaly detection: (1) linear or non-linear approximating function, such as neural network (NN), autoregression analysis (AR) and support vector machines (SVM); (2) time frequency distribution, such as short time Fourier transform (STFT), wavelet transform (WT), Hilbert–Huang transform (HHT); (3) some special methods, such as D-Markov model (Chin 2004; Ray 2004), time-series bitmaps (Li, Kumar, Lolla, Keogh, Lonardi et al. 2005), negative selection (Singh 2002).

All of these methods use a finite set of characteristics to define an anomaly measure. In addition, each must specify a threshold in advance. If the measure is beyond the threshold, the system is considered as being in anomaly behaviour; otherwise, it is normal. Obviously, the threshold affects directly the effectiveness of the detection method.

Epsilon machine is a new computational theory more powerful than finite state machine (Crutchfield and Young 1989). The hidden pattern implies a kind of regularity, structure, symmetry, organisation, and so on (Shalizi and Crutchfield 2001). Epsilon machine is able to discover the hidden pattern of the system from an observed process. Consequently, a new anomaly measure is proposed in this article based on the hidden pattern directly, which is in contrast to the existing model based on observing value.

This article is organised as following: In Section 2, the basic ideas of epsilon machine, natural structure measure and reconstruction algorithms are briefly reviewed. In Section 3, how to symbolise the time series is reported. In Section 4, a new anomaly measure, structure vector (SV) model, is introduced. In Section 5, an application example and experimental results are analysed. In Section 6, the mechanism of the model is discussed. The performance of the model is elucidated in Section 7 and the conclusion is given in Section 8.

2. Epsilon machine

Epsilon machine is a theory of computational mechanics, a method to compute structure complexity. It can infer the causal states from a stochastic process.

*Corresponding author. Email: xkarcher@126.com

ISSN 0020–7721 print/ISSN 1464–5319 online

DOI: 10.1080/00207720802011266
http://www.informaworld.com

All the causal states constitute a set, which is the hidden pattern of system. Crutchfield and Young of the Santa Fe Institute (SFI) introduced it in 1989. In recent years, Hanson, Upper, Feldman and Shalizi have developed it. All the items on epsilon machine are listed in Crutchfield's homepage (http://cse.ucdavis.edu/~chaos/).

2.1 *Basic ideas*

Consider a symbol time series $\cdots S_{-2}S_{-1}S_0S_1S_2\cdots$, with S_i taking values in set $\mathcal{A}$. The sequence can be broken into a past (or a history) $\overleftarrow{S}_t$ and a future $\vec{S}_t$ at any time t. $\vec{S}^L$ is a sequence of L random variables beginning at S_t, and $\overleftarrow{S}^L$ is a sequence of L random variables going up to S_t, but not including it. $\vec{S}^L$ and $\overleftarrow{S}^L$ take values from $s^L \in \mathcal{A}^L$ (Shalizi and Crutchfield 2001).

Definition 2.1: The causal states of a process are the range of a function

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

that maps from histories to sets of histories.

Definition 2.2: The transition probability

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

is that of making a transition from state σ_i to state σ_j while emitting the symbol $s \in \mathcal{A}$.

Definition 2.3: The epsilon machine of a process is an ordered pair $\{\varepsilon, \mathbf{T}\}$, where ε is the causal state function and $\mathbf{T}$ is a set of transition matrices for the states defined by ε.

2.2 *Nature structure measure*

Physics already has tools for detecting and measuring complete order and ideal randomness, but has no principles for defining and measuring natural structure. Epsilon machine has founded a measure of information-processing structure to infer statistical complexity in complex systems (Crutchfield 1994a). Structure measure is a more useful notion of complexity than measure of randomness. But there is no contradiction between them, because each answers the different questions (Crutchfield 1994b).

Given a discrete time or spatial series of measurements from a process, an epsilon machine can be constructed and it is the best description of predictor for this discrete series. The structure of this machine can be regarded as the best approximation for the original process's information processing structure. With epsilon machine reconstruction, the changes of the structure can be regarded as a kind of qualitative change.

2.3 *Reconstruction algorithms*

There are two algorithms, subtree merging (Crutchfield and Young 1989) and causal state splitting reconstruction (CSSR) (Shalizi, Shalizi, and Crutchfield 2002), to reconstruct the epsilon machine from time series. The subtree merging has inherent difficulties in construction, so we only present CSSR here.

CSSR is an empirical estimation of epsilon machine from samples of a stationary process. It uses a statistical test, Kolmogorov–Smirnov (KS) test, to judge whether a history is one part of the state. It needs to specify significance level α and maximum history length $L_{\max}$ beforehand.

CSSR has three procedures: Initialise, Homogenise and Determinise (Shalizi, Shalizi and Crutchfield 2002).

Initialise: In the initial model, the process is an IID sequence. Set $L = 0$ and $\hat{\mathcal{S}} = \{\hat{\sigma}_0\}$, where $\hat{\sigma}_0 = \{*\lambda\}$.

Homogenise: States are generated whose member histories have no significant differences in their individual morphs. At the end of homogenisation, no history in the state is different from the state's morph significantly, and every state's morph is different from every other state's morph significantly.

Determinise: Split the states until they have deterministic transitions. Then, eliminate the transient states from the current state-transition structure, until only recurrent state is left.

3. Symbolisation

Symbolisation denotes that *real* series is converted into *symbol* series, where there are many merits (Daw, Finndy, and Tracy 2003). For example, the efficiency of finding a hidden pattern of the system is increased, while the sensitivity to measuring noise is decreased. By symbolisation, the constructed symbol series consume less computational resources, and gain higher speed than real series. In some cases, symbolisation can be accomplished directly by appropriate sensors in certain instruments, which is important for real-time monitoring and controlling. Moreover, CSSR can only be operated in discrete value and discrete time process. In engineering applications, most time series are real form, so symbolisation is a precondition to CSSR.

There are two main ways to convert a dynamic process into coarse grains. One, called symbolisation, is to divide the original observations into several regions (Kurths, Schwarz, Witt, Krampe and Abel 1996), and the other, called symbolic analysis, is to divide the phase space rather than the series itself (Tang and Tracy 1995). As the latter is often used in chaos research, we will pay more attention to the former in the following.

Symbolisation consists of two steps, pretreatment and partition. Pretreatment is to get another representation of the original data. For discrete series, there are four kinds of pretreatments as follows:

Definition 3.1: Static transformation: divide the discrete series directly without any pretreatment (Kurths et al. 1996).

Definition 3.2: Dynamical transformation: make first- or higher-order difference in original data (Kurths et al. 1996).

Definition 3.3: Wavelet space method: transform the time series to time frequency data. Select the wavelet coefficients at the interesting frequency and arrange them in some order (Ray 2004).

Definition 3.4: Piecewise aggregate approximation: approximate the original series with the mean value of the data falling within a frame (Lin and Keogh 2003).

Static and dynamical transformations are general methods for data pretreatment, which obtain the same length series as original ones. Wavelet space method multiples the length of series, while piecewise aggregate approximation reduces the length of series.

Partition is to split the pretreated series into finite number of regions, and associate each region with a specific symbol. Partition is to map the real series to a particular symbol series. The set of all possible symbols is termed as *alphabet*, whose size is the number of symbols. Generally speaking, there are two ways for partition as follows (Lin and Keogh 2003).

Definition 3.5: Equal interval method: the span of every region is equal, i.e. partition the data with medians.

Definition 3.6: Equal probability method: the probability of a sample point falling into each region is equal, i.e. partition the data with means.

Symbolisation eliminates the noise and inessential information from original data, and retains the meaningful and interesting information. Its performance, however, is sensitive to pretreatment, partition and the size of the alphabet. The quantity of information in resulting symbol series is measured by finite statistical complexity (FSC), which is effective approximation to statistical complexity and easy to calculate (Perry and Binder 1999). Thus, one can symbolise some time-series instances by different combinations of pretreatment, partition and alphabet size, and finally, calculate their FSC. From the FSC results, some empirical conclusions are obtained.

To retain more information in symbolisation process, the following actions are needed:

- Choose larger alphabet size.
- Choose equal probability method rather than equal interval method.
- Choose dynamical transformation rather than static transformation.

To symbolise online, the following actions are needed:

- Choose equal interval method rather than equal probability method.

To obtain less computational complexity, the following actions are needed.

- Choose less alphabet size.
- Choose static transformation rather than dynamical transformation.

Apparently, there is a tradeoff between information content and online computational complexity.

4. SV model

Based on epsilon machine and symbolisation, a novel model, SV model, can be defined. Next, we will discuss the model in detail and demonstrate its effectiveness with an example. A given time series is converted into symbol series and the epsilon machine is reconstructed with CSSR. The reconstruction result includes two parts, causal states set $\mathcal{S}$ and transition probability matrix $\mathbf{T}$. But the anomaly measure is defined by set $\mathcal{S}$ without regard to matrix $\mathbf{T}$. Set $\mathcal{S}$ is composed of recurrent states and transition states that have not been eliminated from the set.

Given two causal state sets,

$$\begin{aligned} &\mathcal{S}_1 = \{\sigma_1{}', \sigma_2{}', \sigma_3{}', \sigma_4{}', \sigma_5{}'\} \\ \text{and}\quad &\mathcal{S}_2 = \{\sigma_1{}'', \sigma_2{}'', \sigma_3{}'', \sigma_4{}'', \sigma_5{}'', \sigma_6{}''\}, \end{aligned} \tag{3}$$

which come from different sub-series. Figure 1 illustrates that $\mathcal{S}_1$ consists of five causal states; $\mathcal{S}_2$ six ones.

In CSSR, each state owns a specific morph. Given the morphs of two causal state σ_i and σ_j, one can estimate whether there is any significant difference between them by KS test. If there is no significant difference, σ_i is homologous to σ_j, which is marked as expression $\sigma_i \leftrightarrow \sigma_j$. Testing the causal states of $\mathcal{S}_1$ and $\mathcal{S}_2$, one can obtain some expressions as follows:

$$\begin{aligned} &\sigma_1{}' \leftrightarrow \sigma_1{}'', \quad \sigma_2{}' \leftrightarrow \sigma_2{}'', \quad \sigma_3{}' \leftrightarrow \sigma_3{}'', \\ &\sigma_4{}' \leftrightarrow \sigma_4{}'', \quad \sigma_5{}' \leftrightarrow \sigma_5{}'', \quad \sigma_4{}' \leftrightarrow \sigma_6{}''. \end{aligned} \tag{4}$$

In expression (4), σ_4'' and σ_6'' are both homologous to σ_4'. From Figure 1(a) and (b), causal state σ_4' has broke up into two parts, σ_4'' and σ_6'', so their morphs have no significant difference. Additionally, a new similarity function $f_s(\sigma_i, \sigma_j)$ is defined to discriminate which part, σ_4'' and σ_6'', is corresponding to σ_4' exactly.

Definition 4.1: $v_i(s^L)$ is the number of history s^L in causal state σ_i.

Definition 4.2: The similarity function is:

$$f_s(s_i, s_j) = \sum_{s^{L_{\max}} \in \Omega_{ij}} \left(v_i(s^{L_{\max}}) + v_j(s^{L_{\max}}) \right), \tag{5}$$

$$\left(\Omega_{ij} = \left\{ s^{L_{\max}} | s^{L_{\max}} \in \sigma_i, s^{L_{\max}} \in \sigma_j \right\} \right).$$

If $f_s(\sigma_4', \sigma_4'') > f_s(\sigma_4', \sigma_6'')$, the exact expression is $\sigma_4' \longleftrightarrow \sigma_4''$).

Definition 4.3: The number of all histories with length $L_{\max}$ in σ_i is written as follows:

$$v(\sigma_i) = \sum_{s^{L_{\max}} \in \sigma_i} v_i(s^{L_{\max}}). \tag{6}$$

Then, SVs of $\mathcal{S}_1$ and $\mathcal{S}_2$ are as follows, respectively:

$$\mathbf{V}_1 = \left[v(\sigma_1'), v(\sigma_2'), v(\sigma_3'), v(\sigma_4'), v(\sigma_5'), 0 \right], \tag{7}$$

$$\mathbf{V}_2 = \left[v(\sigma_1''), v(\sigma_2''), v(\sigma_3''), v(\sigma_4''), v(\sigma_5''), v(\sigma_6'') \right]. \tag{8}$$

Definition 4.4: The anomaly measure between σ_i and σ_j is

$$A_{ij} = \cos^{-1}\left(\frac{\langle \mathbf{V}_i, \mathbf{V}_j \rangle}{\|\mathbf{V}_i\| \bullet \|\mathbf{V}_j\|} \right) \tag{9}$$

Apparently, $A_{ij} = A_{ji}$. The anomaly measure is a geometry distance having no physical unit. Of course, other analogous geometry distances can be used to define anomaly measure, too.

5. Application examples: crankshaft fatigue

A non-linear dynamical system is stimulated by the continual, iterative external excitation. The parameter or characteristic of the driven system evolves slowly. The evolution information is hidden in the output response. Anomaly detection algorithm is to reveal the information hidden in the time series, and recognise the abrupt changes such as break, fault, etc., as soon as possible.

The resonant test is often used to conduct crankshaft fatigue experiment. The schematic drawing of a test machine is shown in Figure 2. A motor and an eccentric compose the vibration exciter, which is fixed on an active arm. The active arm, passive arm and crankshaft have been connected tightly to form a resonant system just like a tuning fork. The vibration exciter works around the resonant frequency. The resonant system amplifies the vibration and fatigues the crankshaft with the repetitive loading.

Acquire the response time series in the resonant test by an accelerometer. Following the four steps as follows, we can draw the anomaly curve in Figure 3.

(1) Symbolise the time series with static transformation, equal probability method and alphabet size $k = 4$.
(2) Divide symbol series into 97 sub-series, each of which has the same length $N = 12{,}800$.

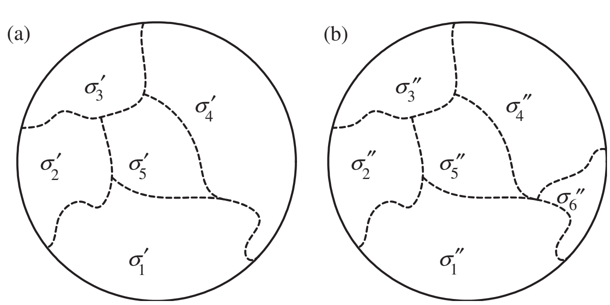

Figure 1. Causal state sets. (a)$\mathcal{S}_1$. (b)$\mathcal{S}_2$.

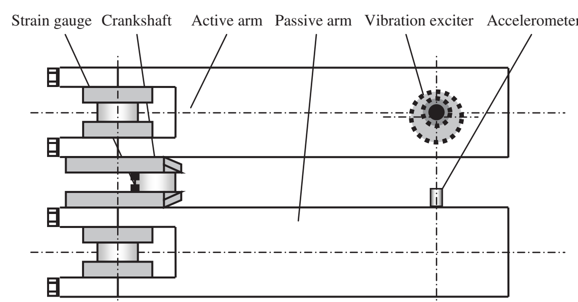

Figure 2. Schematic of resonant crankshaft fatigue test machine.

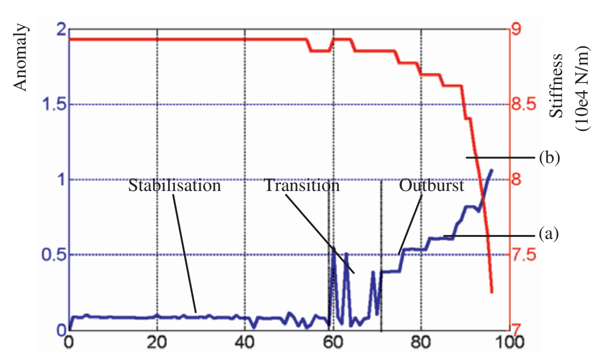

Figure 3. (a) Anomaly evolution curve of crankshaft fatigue. (b) Stiffness degradation curve of crankshaft fatigue.

Consider that the time series is piecewise stationary.

(3) Reconstruct the epsilon machine by CSSR. The maximum history length $L_{\max} = 5$, and the significance level $\alpha = 0.01$.

(4) Define the structure vector of each sub-series. Compute the anomaly measure A_i, where A_i is the abbreviation of A_{0i}, and then draw the anomaly curve.

For the repetitive loading, cracks grow in the crankshaft, while the stiffness of the crankshaft decreases slowly at the same time. Stiffness is a common physical quantity, and the decreasing stiffness is a macro-metric of crankshaft fatigue. But it is very difficult to measure the system stiffness in experiment or operation mode. As evidence, the stiffness curve can help to prove our anomaly curve.

The anomaly curve is at the bottom of Figure 3, noted as Figure 3(a), and the stiffness curve is on the top in Figure 3, noted as Figure 3(b). The curve in Figure 3(a) is divided into three parts: stabilisation (0–60), transition (60–72) and outburst (72–97). In stabilisation stage, almost no changes occur in anomaly measure, which shows that the crankshaft has no cracks yet. In transition stage, anomaly measure begins to fluctuate slightly. Meanwhile, following the fluctuation, the stiffness changes a little, too. Therefore, both of them imply that some cracks have emerged in the crankshaft, but they are not strong enough to destroy the whole crankshaft. In outburst stage, anomaly measure increases quickly, while the stiffness falls sharply. From these changes, it can be inferred that many cracks have grown in the crankshaft. At this time, the crankshaft has been disabled, and more terrible is that it may rupture at any moment. Compared with stiffness, anomaly curve is also a better measure of crankshaft fatigue and easier to compute online.

In the above-mentioned three stages, the transition one is the most interesting to us. Though the changes of the stiffness are very small, the anomaly measure follows violent fluctuation. Therefore, it is a good alarm signal to inspect the fatigue of crankshafts. There are two primary reasons for the fluctuation. Objectively, the emergence and growth of the cracks in crankshafts are not a stationary process. Moreover, subjectively, the SV model is limited in finding the hidden pattern of systems, especially when the pattern itself is instable and vague. Usually someone thinks that the anomaly measure crosses a threshold without any hesitation. However, as we all know, transition from normal behaviour to abnormal in any system is not a simple passage with one step. The waver in transition stage of the anomaly curve seems to be more reasonable.

6. Comparison and discussion

SV model is not a work without any foundation; its inspiration is partially based on epsilon machine theory and D-Markov model. It is necessary to highlight the original contribution of our model and distinguish it from the others.

6.1 *Epsilon machine and D-Markov machine*

Epsilon machine and CSSR algorithm have been studied for about 20 years, but the main developments are centralised on physical theory and its application investigation is limited, for it is a symbol inference method. Here, we sum up its engineering applications briefly, and a more detailed review can refer to our previous paper (Xiang and Jiang 2006). In most cases, epsilon machine is simplified as a method to obtain statistical complexity from experimental data; for example, dripping faucet experiment (Goncalves, Pinto, Sartorelli and Oliveia 1998), atmospheric turbulence datasets (Palmer, Fairall, and Brewer 2000), heart rate variability (Pei, Yang, and He 2000) and medical data (Wessel, Schwarz, Saparin, and Kurths 2001). Clarke, Freeman and Watins discuss how to apply epsilon machine to a non-infinite corrupted series, and define the distinction between noise and unresolved structure (Clarke, Freeman, and Watins 2003). Other application researches include natural language processing (Padro 2005), learning on recurrent spiking neurons (Brodu 2007), choosing ecological model (Boschetti 2007), and analysing correlations in online media (Cointet, Faure, and Roth 2007). So far, no anomaly detection algorithm using epsilon machine is reported publicly.

SV model is based on epsilon machine, and its technical route is close to another model, D-Markov model. So it is necessary to introduce D-Markov model here. D-Markov model is presented by Ray to detect anomaly behaviour in mechanical and electronic system. In recent years, Ray and his group discuss the essence of D-Markov model with symbolic dynamics (Ray 2004), evaluate its performance comparing with existing techniques (Chin, Ray, and Rajagopalan 2005), develop a unique symbolisation method, wavelet-based partitioning, to improve the capability of the model (Rajagopalan and Ray 2006) and identify the critical system parameters as an inverse problem (Rajagopalan, Cakraborty, and Ray 2008). The effect of the anomaly detection algorithm based on D-Markov model has been demonstrated by several practical examples.

Given a symbol series $\cdots S_{-2}S_{-1}S_0S_1S_2\cdots$, its D-Markov model can be constructed as following. Count the occurrences of the string having length $D+1$ and D. The occurrence of string $s_{i1}s_{i2}\cdots s_{iD}s_{i(D+1)}$ is denoted by $N(s_{i1}s_{i2}\cdots s_{iD}s_{i(D+1)})$, and the occurrence of string $s_{i1}s_{i2}\cdots s_{iD}$ is denoted by $N(s_{i1}s_{i2}\cdots s_{iD})$. If $N(s_{i1}s_{i2}\cdots s_{iD})\neq 0$, the transition probabilities are calculated by an empirical equation

$$P_{jk} = P(q_k|q_j) \approx \frac{N(s_{i1}s_{i2}\cdots s_{iD}s)}{N(s_{i1}s_{i2}\cdots s_{iD})} \tag{10}$$

In Equation (10), $q_j = s_{i1}s_{i2}\cdots s_{i\mathrm{D}}$ and $q_k = s_{i2}\cdots s_{i\mathrm{D}}s$. Compute the left eigenvector p corresponding to the maximum eigenvalue of the stochastic matrix $[P_{ij}]$. The vector p is the state probability vector of the dynamical system. Divide the whole symbol series into some sub-series with the same lengths, and construct D-Markov model for each sub-series. The distance of the state probability vector p is defined as an anomaly measure $A = [A_{0j}]$ and

$$A_{0j} = \cos^{-1}\left(\frac{\langle p_0, p_j\rangle}{\|p_0\|\bullet\|p_j\|}\right) \tag{11}$$

6.2 *Computational capability analysis*

In some sense, SV model roots in epsilon machine theory and D-Markov model. The original contribution of our model can be discussed from three aspects: basing on hidden pattern variation and not requiring experiential threshold, being able to model sofic system and having more universality and revealing the hierarchy structure and closely relating to vulnerability of system. D-Markov model has already been proved to be an excellent early detection algorithm of fatigue damage. It is obviously more straightforward than SV model. So, we think that SV model inspires a novel idea on anomaly detection more than an earlier detection algorithm. Most intelligent computational models are rooted in physical progress, so we will compare two models from computational mechanics, symbolic dynamics and complex networks, respectively. Limited by the space, we will not present the background knowledge on these subjects.

The causal time-series modelling hierarchy in reference (Crutchfield 1994a) is a reconstruction language hierarchy. The original time series is *data stream* at level 0. D-Markov machine is constructed by grouping sequential measurements into recurring subsequence and is block-independent, so it is *tree* at level 1. Epsilon machine is reconstructed from the tree by grouping tree nodes and is conditionally independent, so it is *finite automata* at level 2. Epsilon machine has a more powerful predicting and compressing ability for given observation than D-Markov machine. The reconstruction result of the former is a hidden Markov model, and that of the latter is a Markov model whose state is visible. D-Markov model can identify the minute abnormal characteristics easily in observation signal, and it is a signal processing method. SV model describes the variation of hidden patterns and it is a complex system analysing method.

The construction of D-Markov machine adopts sliding block coding mode and it is a finite-type shift, which has finite-*length* memory (Lind and Marcus 1995). Epsilon machine is a sofic shift, which has a finite *amount* memory. Finite-type shifts are strictly subsets of sofic shifts. Obviously, epsilon machine has more computational universality than D-Markov machine.

6.3 *New modelling idea*

Mapping a pseudo-periodic time series to complex network is researched recently, and some useful, interesting connections are found between the temporal dynamics and the topological structure (Zhang and Small 2006). Actually, stochastic matrix $[P_{ij}]$ in D-Markov machine can be viewed mathematically as the adjacency matrix of a weighted network; the strings $q_j = s_{i1}s_{i2}\cdots s_{i\mathrm{D}}$ are corresponding to the nodes in the network. We denote this network as N_D. For constructing D-Markov machine is based on the neighbourhood relationship in symbol series, N_D is an *anatomical* connectivity.

The entries of matrix $[P_{ij}]$ are transition probability, which is calculated with the passage times between the adjacent nodes. So the weighted network graph represented by matrix $[P_{ij}]$ can be transformed into an unweighted multi-edges graph (Newman 2004). If the vertices in the network are clustered with assortative mixing according to some characteristics such as vertex degree, a new hierarchical structure, mixing pattern, is built (Quayle, Siddiqui, and Jones 2006). CSSR algorithm is just to merge the string with similar probability distribution. To some extent, causal state set discovered from symbol series is equivalent to mixing pattern of complex networks. Consequently, on the basis of N_D, epsilon machine can define a new *functional* connectivity as mixing pattern.

The degree assortativity of a network has a number of interesting ramifications on the network robustness. Relating the hierarchy and the vulnerability of the system is just for a very simple reason that more damages are caused by removal of a particular vertex

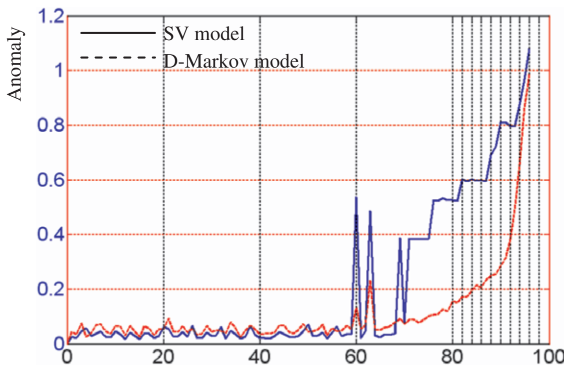

Figure 4. Comparison between SV model and D-Markov model.

(Goldshtein, Koganov, and Surdutovich 2004). By CSSR algorithm, a new hierarchical network N_E is defined and its vertices are causal states. In SV model, the anomaly measure is corresponding to the variation of the vertices in N_E. When a vertex in N_E disappears, the network structure is destroyed and an irreversible jump emerges in the anomaly curve, which implies that the system has already deviated from its normal behaviour seriously. Hence, no threshold needs to be specified at the beginning of anomaly detection.

We research the example in Section 5 with two different methods, SV model and D-Markov model, calculate the anomaly measure from the same symbol series and draw the curves in Figure 4. Solid and dashed lines in Figure 4 represent the results of SV model and D-Markov model, respectively. Apparently, SV model need not have any artificial threshold, for the jump is a perfect candidate. Specifying a threshold is necessary for D-Markov model. SV model can warn the latent dangers quite effectively. In Figure 4, supposing the threshold is 0.6, SV model alerts us to the danger of fatigue at the 82nd time interval, but D-Markov model could not do it until the 94th time interval.

7. Performance analysis

As a new method, SV model is involved with system complexity, computational mechanics, statistical inference and symbolic dynamics, etc. Therefore, its performance analysis is very difficult. In this article, only two aspects are discussed and more work is left to the future.

7.1 *Choice of parameters*

SV model has four adjustable parameters: the significance level α, the sub-series' length N, the alphabet size k and the maximum history length $L_{\max}$.

It has been proved that α can be selected optionally; $\alpha = 0.01$ or 0.001 taken as usual. The bigger α is, the more sensitive the model is to minute distinctions between states.

As the length of sub-series, N ought to be as large as possible, for it can improve the precision of the statistical test. However, when N is too large, sub-series could not be considered as approximately stationary. So there is a tradeoff to choose an appropriate N, which is usually dependent on experience.

From Section 3, we notice that if alphabet size k is large, it can retain more information of original data, but it is more sensitive to noise and the inessential details. At the same time, the larger k means the more kinds of histories. Thus, it will impede the statistical convergence for not enough examples of each history. In general, k takes the value 4.

Klinkner and Shalizi (2005) have already discussed the parameter $L_{\max}$. 'For any given process, there is a minimum history length Λ, such that the true states cannot be found if $L < \Lambda$'. When L is too large, the states will blow up. The empirical expression is $L_{\max} < \log_2 N / \log_2 k$. Usually, we should select an as small $L_{\max}$ as possible, for the larger $L_{\max}$ will reduce the accuracy of many probability estimates, too (Klinkner and Shalizi 2005).

We analyse the data with different $L_{\max}$. Figure 5 illustrates four anomaly curves for the value of $L_{\max}$ at 3, 4, 5 and 6. According to Klinkner's empirical expression, $L_{\max}$ is not more than 6.8 here. The four curves in Figure 5 have the same trend, but Figure 5(c) seems better than others. Anomaly curves in Figure 5(a) and (b) have many fluctuations, which imply that small $L_{\max}$ may result in incorrect states. In Figure 5(d), $L_{\max}$ is very close to the upper limit 6.8, so new fluctuations appear, which imply that some redundant states are returned.

7.2 *Time complexity*

SV model consists of four steps: symbolisation, reconstruction of epsilon machine, definition of SV and computation of anomaly measure. Steps one and four only consume a little time. Especially, when instruments are selected appropriately, the time complexity of symbolisation is nearly zero.

The time complexity of CSSR has been studied by Shalizi et al. (2002). We only list his results here. CSSR includes three steps: initialise, homogenise and determinise. For SV model, it need not eliminate the transient states from state set, the time complexity of epsilon machine reconstruction for a sub-series is $O\left(k^{L_{\max}}\right) + O(N)$.

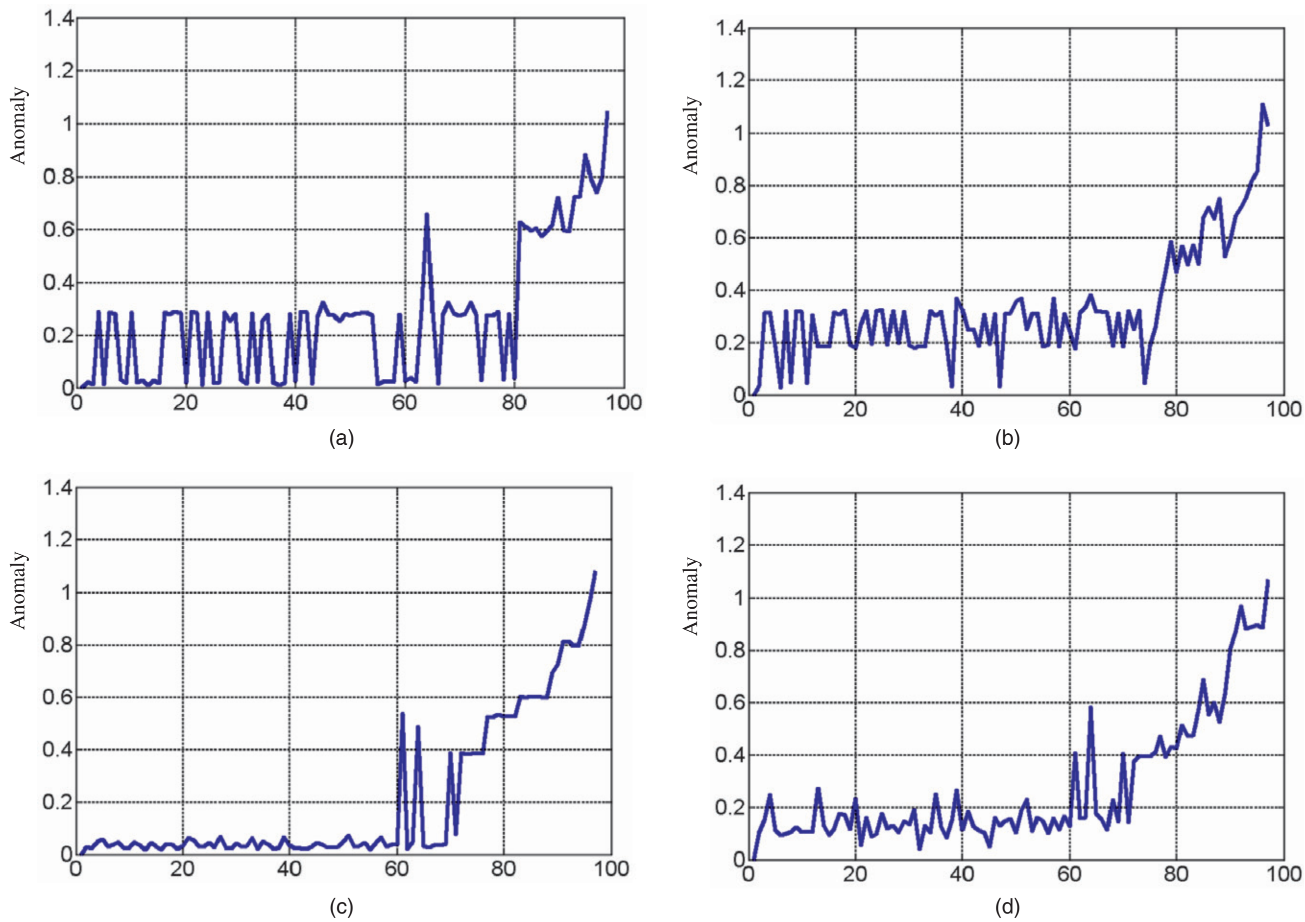

Figure 5. Anomaly curve with different $L_{\max}$. (a) $L_{\max}=3$. (b) $L_{\max}=4$. (c) $L_{\max}=5$. (d) $L_{\max}=6$.

Assuming the state set has n states, $0<n<k^{L_{\max}}$, and $k^{L_{\max}}$is the maximum of the histories in a sub-series. If any two of the histories cannot be merged in CSSR, there will be $k^{L_{\max}}$ states in the set. According to the definition of SV, it need compare states no more than $n(n+1)/2$ time steps to find the most similar state, whose time complexity is $O((n+1)/2)\approx O(n^2)$. So the maximum time complexity is $O\left(k^{2L_{\max}}\right)$.

Given a time series including countable sub-series, SV model's time complexity is expressed as

$$O(N)+O(k^{2L_{\max}}). \tag{12}$$

8. Conclusion

This article presents a novel method of anomaly detection for complex systems based on epsilon machine. SV model includes four primary steps as mentioned above. SV embodies the viewpoint of natural structure, which comes from computational mechanics. The variation of SV dimensions makes a great difference to all others and gives rise to the jump of anomaly measure. The threshold is not necessary in our model and the latent dangers can be warned in time. Such a jump can be interpreted by vulnerability and hierarchy of complex networks.

As a potential research direction, applying the idea of natural structure to detection and diagnosis has many problems worthy to be studied. Future investigation is recommended in the following areas:

(1) Find new symbolisation methods, which can retain as much as information by a finite alphabet size. It is urgent to establish a certain standard on symbolisation.
(2) Seek the optimal combination of CSSR parameters.
(3) Establish a rational foundation for SV model from the viewpoint of complex networks.
(4) Extend epsilon machine to the process with continuous value at discrete times. It will help us get out from the mire of symbolisation thoroughly.

Acknowledgements

We thank Dr X.X. Wu (Wuhan University of Technology) and Q. Li (Zhejiang University) for their suggestions on the writing; and thank the editors and two anonymous referees for all their detailed comments and suggestions.

Notes on contributors

Kui Xiang received B. Eng. degree from Wuhan Technical University of Surveying and Mapping, Wuhan, China, in 1999, and the M.Eng. in mechanical dynamics and PhD degree in system science from Zhejing University, Hangzhou, China, in 2002 and 2006, respectively. Currently, he is a lecturer at School of Automation, Wuhan Technology of University, Wuhan, China. His present research interests include time-series analysis, biomechanics and control.

Xun Zhou received the BS from Shandong University, Jinan, China in 2000, and MS degrees and PhD degrees from Zhejiang University, Hangzhou, China, in 2003 and 2006, respectively. He is a lecturer at College of Mechanical and Energy Engineering, Zhejiang University. His current research interests are anomaly detection in complicate mechanical system, reliability of mechanical system and fatigue fracture engineering.

Jingping Jiang was born in 1935; He is a full professor of E.E. College, Zhejiang University, P.R. of China. He graduated at Zhejiang University in 1958, he as a visiting scholar was in University of Wisconsin, USA, from 1979 to 1981, and was in Reading University, UK, from 1989 to 1990. His research interests are intelligent systems and intelligent control, and electric drive systems as well. He is a vice chief-editor of Zhejiang University Academic Journal.

References

Boschetti, F. (2007), "Mapping the Complexity of Ecological Models," *Ecological Complexity*, 5(1), pp. 37–47.

Brodu, N. (2007), "Quantifying the Effect of Learning on Recurrent Spiking Neurons," *IEEE International Joint Conference on Neural Networks*, Juillet.

Chin, S. (2004), "Real Time Anomaly Detection in Complex Dynamic Systems," unpublished Ph.D. dissertation, The Pennsylvania State University.

Chin, S., Ray, A., and Rajagopalan, V. (2005), "Symbolic Time Series Analysis for Anomaly Detection: a Comparative Evaluation," *Signal Processing*, 85(9), 1859–1868.

Clarke, R., Freeman, M., and Watins, N. (2003), "Application of Computational Mechanics to the Analysis of Natural Data: an Example in Geomagnetism," *Physical Review E*, 67, 016203.

Cointet, J., Faure, E., and Roth, C. (2007), "Intertemporal Topic Correlations in Online Media," in *Proceedings of the International Conference on Weblogs and Social Media*, Boulder, CO, USA.

Crutchfield, J., and Young, K. (1989), "Inferring Statistical Complexity," *Physical Review Letters*, 63(2), 105–108.

Crutchfield, J. (1994a), "The Calculi of Emergence: Computation, Dynamics and Induction," *Physica D*, 75, 11–54.

Crutchfield, J. (1994b), "Is Anything Ever New? Considering Emergence," SFI Working Paper 94-03-011.

Daw, C., Finndy, C., and Tracy, E. (2003), "A Review of Symbolic Analysis of Experimental Data," *Review of Scientific Instruments*, 74(2), 915–930.

Goldshtein, V., Koganov, G., and Surdutovich G. (2004), "Vulnerability and Hierarchy of Complex Networks," cond-mat/0409298.

Goncalves, W., Pinto, R., Sartorelli, J., and Oliveia, M. (1998), "Inferring Statistical Complexity in the Dripping Faucet Experiment," *Physica A*, 257, 385–389.

Klinkner, K., and Shalizi, C. (2005), "An Algorithm for Building Markov Models from Time Series." http://www.cscs.umich.edu/~*crshalizi/CSSR/*

Kurths, J., Schwarz, U., Witt, A., Krampe, R., and Abel, M. (1996), "Measures of Complexity in Signal Analysis," in *Chaotic, Fractal, and Nonlinear Signal Processing, AIP Conference Proceedings*, New York: Woodbury, pp. 33–54.

Li, W., Kumar, N., Lolla, V., Keogh, E., Lonardi, S., and Ratanamahatana, C. (2005), "Assumption-free Anomaly Detection in Time Series," in *Proceedings of the 17th International Scientific and Statistical Database Management Conference* (SSDBM 2005) Santa Barbara, CA, USA, June 27–29, pp. 237–240.

Lin, J., and Keogh, E. (2003), "A Symbolic Representation of Time Series, with Implications for Streaming Algorithms," in *proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp. 2–11.

Lind, D., and Marcus, B. (1995), *An Introduction to Symbolic Dynamics and Coding*, Cambridge: Cambridge University Press.

Newman, M. (2004), "Analysis of Weighted Networks," *Physical Review E*, 70, 056131.

Padro, M. (2005), "Applying Causal-State Splitting Reconstruction Algorithm to Natural Language Processing Tasks," Thesis Proposal, Technical University of Catalonia.

Palmer, A., Fairall, C., and Brewer, W. (2000), "Complexity in the Atmosphere," *IEEE Transaction on Geoscience and Remote Sensing*, 38(4), 2056–2063.

Pei, W., Yang, L., and He, Z. (2000), "A Statistical Complexity Measure and Its Applications to the Analysis of Heart Rate Variability,'' *Acta Biophysica Sinica*, 16(3), 562–568.

Perry, N., and Binder, P. (1999), ''Finite Statistical Complexity for Sofic Systems,'' *Physical Review E*, 60, 459–463.

Quayle, A., Siddiqui, A., and Jones, S. (2006), ''Modeling Network Growth with Assortative Mixing,'' *European Physical Journal B*, 50, 617–630.

Rajagopalan, V., and Ray, A. (2006), ''Symbolic Time Series Analysis via Wavelet-based Partitioning,'' *Signal Processing*, 86(11), 3309–3320.

Rajagopalan, V., Cakraborty, S., and Ray, A. (2008), ''Estimation of Slowly-varying Parameters in Nonlinear Systems via Symbolic Dynamic Filtering,'' *Signal Processing*, 89(2), 339–348.

Ray, A. (2004), ''Symbolic Dynamic Analysis of Complex Systems for Anomaly Detection,'' *Signal Processing*, 84, 1115–1130.

Singh, S. (2002), ''Anomaly Detection Using Negative Selection Based on the R-contiguous Matching Rule,'' in *1st International Conference on Artificial Immune Systems*, UK: Canterbury, pp. 99–106.

Shalizi, C., and Crutchfield, J. (2001), ''Computational Mechanics: Pattern and Prediction, Structure and Simplicity,'' *Journal of Statistical Physics*, 104(3), 817–879.

Shalizi, C., Shalizi, K., and Crutchfield, J. (2002), ''An Algorithm for Pattern Discovery in Time Series,'' SFI Working Paper 02-10-060.

Tang, X., and Tracy, E. (1995), ''Symbol Sequence Statistics in Noisy Chaotic Signal Reconstruction,'' *Physical Review E*, 51, 3871–3889.

Wessel, N., Schwarz, U., Saparin, P., and Kurths, J. (2001), ''Symbolic Dynamics for Medical Data Analysis.'' http://citeseer.ist.psu.edu/598255.html

Xiang, K., and Jiang, J. (2006), ''Pattern Discovery in Complex System: Review of Epsilon Machine,'' *Pattern Recognition and Artificial Intelligence*, 19(6), 746–752.

Zhang, J., and Small, M. (2006), ''Complex Network from Pseudoperiodic Time Series: Topology versus Dynamics,'' *Physical Review Letters*, 96(23), 238701.