A complexity measure approach based on forbidden patterns and correlation degree*

Wang Fu-Lai(王福来)†

Department of Mathematics and Statistics, Zhejiang University of Finance and Economics, Hangzhou 310012, *China*

(Received 9 September 2009; revised manuscript received 5 December 2009)

Based on forbidden patterns in symbolic dynamics, symbolic subsequences are classified and relations between forbidden patterns, correlation dimensions and complexity measures are studied. A complexity measure approach is proposed in order to separate deterministic (usually chaotic) series from random ones and measure the complexities of different dynamic systems. The complexity is related to the correlation dimensions, and the algorithm is simple and suitable for time series with noise. In the paper, the complexity measure method is used to study dynamic systems of the Logistic map and the Hénon map with multi-parameters.

Keywords: complexity theory, forbidden words, symbolic dynamics, correlation dimensions
PACC: 0545, 0547

1. Introduction

Recently, the order pattern has been proposed as a technique of evaluating the determinism or complexity of a given time series.[1−8] Forbidden patterns (or forbidden sequences, forbidden words) can be a characterization of nonlinear time series; however, more than one definition of forbidden patterns has been proposed, but they have no well-defined meanings. As a result, it is contradictary to characterize the complexities of systems by the concept of forbidden patterns. The definition given by Wolfram[7] means that any finite symbolic sequence can be described by a finite automaton and the complexity of the sequence is related to the numbers of nodes in the evolution diagram of finite automata, which is related to the increase rate of forbidden words. Based on forbidden words, Xu[6] proposed a hierarchical approach to separating random, periodic and chaotic time series from each other. After the definition of complexity measure, Xu guessed that the C_2 complexity was closely related to the correlation dimension and took the Logistic map (1) as an example

$$x_{n+1} = \mu x_n(1 - x_n), \tag{1}$$

$$y_{n+1} = \mu y_n(1 - x_n), \quad x_n = y_n + \xi_n. \tag{2}$$

Zanin[2] and Amigo *et al.*[5] took into consideration both the permutation as we do in the calculation of the permutation entropy and forbidden patterns. Comparing the appearance rates of the forbidden patterns in a symbolic sequence generated from some dynamic systems, Amigo *et al.* constructed a rule to test the determinism in a given sequence: if the number of forbidden patterns is greater by at least an order of magnitude above than that encountered in an equivalent random series, then the series has a deterministic structure.[2] The Logistic map with perturbations (2) is considered. Furthermore in Ref. [2], the appearances of forbidden patterns were studied in different economical indicators such as stock indices (Dow Jones Industrial Average and Nasdaq Composite) and others.

Though the forbidden patterns and the concepts based on them are thought to have basic properties as chaos indicators, e.g Lyapunov exponent or Kolmogorov entropy,[9−12] not enough compelling results have been obtained to characterize the complexity of a system. For example,[6] for the Logistic map (1), $C_2 = 0.1111, 0.5953, 0.4880, 0.1111$, when $\mu = 3.3, 3.95, 3.98, 4.0$, which are obviously not reasonable because the system described by map (1) is periodic when $\mu = 3.3$ and chaotic when $\mu = 4.0$. In another example,[3] the numbers of forbidden patterns were used to separate deterministic from random series, but nothing was mentioned about how the complexity of dynamic systems is described. In fact, when the values of μ are taken to be different, there is little difference between the numbers of forbidden patterns. As re-

*Project supported by the National Natural Science Foundation of China (Grant No. 10871168).

†Corresponding author. E-mail: flyerwon@sina.com

© 2010 Chinese Physical Society and IOP Publishing Ltd

quired in Refs. [2] and [3], we take the length (d) of subsequences to be 6, i.e. $d = 6$ and sample size (N) to be 3000, i.e. $N = 3000$. By Zainin's algorithm,[2] we have Table 1.

With the concept of the uniform index defined by a monopolized sphere, Luo Chuan-Wen pointed out that chaometry had a close relation to uniformity.[13,14] It is desirable to avoid partitioning space and thus avoid some technique difficulties in calculating indexes such as correlation dimensions and Lyapunov exponents (see Ref. [12]). However, in Ref. [13] the points in space are considered as those without dynamical property, which means that no relations exist between the points and so the characteristics of systems cannot be reflected efficiently.

Table 1. Numbers of forbidden patterns of Logistic map (1) obtained by Zanin's algorithm ($N = 3000$, $d = 6$).

μ	4	3.99	3.98	3.97	3.96	3.95	3.94	3.93
$N_f(6)$	46602	46610	46608	46610	46616	46616	46617	46617
μ	3.92	3.91	3.90	3.89	3.88	3.87	3.86	3.85
$N_f(6)$	46608	46616	46622	46625	46625	46625	46625	46625
μ	3.84	3.83	3.82	3.81	3.80	3.79	3.78	···
$N_f(6)$	46650	46652	46629	46629	46629	46630	46631	···

Substituting dynamic properties of a system with some statistical property might lead to conflict. Two examples are shown as follows:

(i) In Ref. [13] the chaometry and the uniformity of Logistic map (1), when $\mu = 4.0$, are nearly equal. As far as predictability and dynamics are concerned, a deterministic system is obviously not a completely random system.

(ii) According to Ref. [2], greater chaometry corresponds to a system with more perturbations, which is also not reasonable.

Uniformity can reflect statistical properties but not dynamical ones. Based on information theory, a completely random system can provide us with almost no information and the complexity should be zero.[6]

In Ref. [1], with a new method of transformation from time series into symbolic sequences and the redefinition of complexity, Wang used a hierarchy of indexes C_1 and C_2 to describe complexity and randomness of systems respectively. But weaknesses also remained: i) correlation degrees between phase points were not considered and, ii) for both regular and random fractals, the relation between the complexity and Lyapunov exponent was not discussed, where a regular fractal such as the Sierpinski Gasket means that generated by transformation of regular figure, while a random fractal such as the attractor by the Hénon map means that generated by a time series, a map or a differential equation.

Why are we faced with problems and contradictions when evaluating the determinism or complexity of a given time series by symbolic dynamics? The critical problem lies in dealing with forbidden patterns and allowed ones (see Refs. [15]–[18]).

What are the advantages of symbolic dynamics in dealing with time series with noise perturbations? The advantages are as follows: 1) Transforming a time series to a symbolic sequence is a coarse-graining process. By symbolic dynamics, the self-correlation between the consecutive symbols is weak and delay time τ is always taken to be 1, i.e. $\tau = 1$,[2−6] if the number of symbols used (with the exception of repetition) is small. 2) The appearance rate of new patterns can be obtained by simple calculation. 3) Symbolic sequences are not sensitive to perturbations. 4) The metrics in symbolic spaces are usually defined in case they depend strongly on the Euclidean distance. The frontal analysis shows that these advantages do not necessarily mean that it is easier to analyse a time series by symbolic dynamics than by the G–P algorithm (see Refs. [1], [4] and [5]).

In the present paper, based on symbolic dynamics and Zanin's improved theory, we take both the phase points and the correlation degree into consideration. A metric in the symbolic sequences and a complexity measure based on symbolic dynamics are defined. With the definition and the comparison of different algorithms and concepts, the complexity measure is

applied to both regular fractals and dynamic systems.

2. The G–P algorithm, complexity measure and new approach

2.1. Classic G–P algorithm[9]

Consider a discrete information source emitting a time series $\{x_i\}$ $(i = 1, 2, \ldots, N_0)$. The phase points (also vectors) in a reconstructed m-dimensional $(m \geq 2)$ phase space are

$$X_j(m, N_0, \tau) = (x_j, x_{j+\tau} \ldots, x_{j+(m-1)\tau}),$$
$$(X_j \in R^m, j = 1, 2, \cdots, N),$$

where τ is the delay time, N_0 is the sample size, N is the number of phase points and $N = N_0 - (m-1)\tau$.

By Takens' theorem,[11] if a given time series is infinite and unperturbated, τ can be any positive integer. But in contrast, if a time series studied is finite and perturbed by noise, it is important for τ to take an appropriate value, which is always done by many trials in order to have the consecutive components in every phase point independent to a certain extent.

Now we consider two vectors in the m-dimensional phase space

$$\begin{cases} X_m(t_i) = (x(t_i), x(t_i + \tau), \ldots, x(t_i + (m-1)\tau)), \\ X_m(t_j) = (x(t_j), x(t_j + \tau), \ldots, x(t_j + (m-1)\tau)), \end{cases}$$

where $|t_i - t_j| > \tau$.

We denote the distance of two phase points by $\|r_{ij}\| = \|X_m(t_i) - X_m(t_j)\|$ which is a norm distance and always Euclidean norm. For a given delay time τ, radius r of the neighbourhood and dimension m, the relation dimension is defined as

$$D_2(m) = \lim_{r \to 0} \frac{\ln C_2(r, m)}{\ln(r)}, \tag{3}$$

where $C_2(r, m) = (1/N^2) \sum_{i,j=1}^{N} \theta(r - r_{ij})$ is the probability of the points with a distance less than r, i.e. the cumulative distribution function of distances, with θ being the Heaviside step function.[9]

2.2. Metrics in symbolic dynamics

As far as concrete calculation itself is concerned, any value is approximate except for the absolutely exact value with an infinite accuracy in calculation, whether manual or computational. In this sense, any calculation is a kind of coarse-graining while in the narrow sense coarse-graining means the transformation of time series to symbolic sequences. For example, let a_1, a_2 and a_3 be the three numbers in the Euclidean space and their values be 2.124, 2.125, 2.130, respectively. We have the distances $\|a_1,\ a_2\| = 0.001$, $\|a_2,\ a_3\| = 0.005$, and $\|a_1,\ a_2\| < \|a_2,\ a_3\|$. When the numbers are coarse grained with 2 significant digits by rounding, we have $\|a_1,\ a_2\| = 0.01$, $\|a_2,\ a_3\| = 0.00$, and $\|a_1,\ a_2\| > \|a_2,\ a_3\|$. In terms of symbols, both 0.025 and 0.030 belong to the symbolic sequence of (003) while 0.024 belongs to that of (002). Generally speaking, for a symbolic sequence, we reason that the points sharing the same sequence are of higher correlation degree than the points belonging to different symbolic sequences.

Let s be a set of symbols without repetition in a symbolic sequence and $s = \{1, 2, \ldots, d\}$ $(d \geq 2)$.

A time series $\{x_i\}$ $(i = 1, 2, \ldots, N_0)$ can be transformed into an equivalent sequence $(s_1 s_2 \cdots s_{N_0})$ in different ways.

All the overlapping subsequences of length n form a phase space.

Suppose that $S = (s_1 s_2 \cdots s_n \cdots)$ is a symbolic sequence, where $s_n \in s$. $(\Sigma_n s_i)$ and $(\Sigma_n s_j)$ are two subsequences of S with length $n+1$ $(n = 1, 2, \cdots)$ and are denoted by S_i and S_j, respectively, where Σ_n means the same n beginning symbols of $(\Sigma_n s_i)$ and $(\Sigma_n s_j)$. We denote the distance between S_i and S_j by $\|S_{ij}(n+1)\|$. We denote r by $1/d$ and define the norm distance as follows:

$$\|S_{ij}(n+1)\| = \sum_{k=1}^{n+1} \frac{\|s_k, s'_k\|}{d^k} = \sum_{k=1}^{n+1} r^k \|s_k, s'_k\|, \tag{4}$$

where s_k and s'_k are the k-th symbols in S_i and S_j, respectively, and

$$\|s_k, s'_k\| = \begin{cases} 0, & \text{if } \ s_k = s'_k, \\ 1, & \text{if } \ s_k \neq s'_k. \end{cases}$$

Obviously, the metric satisfies the three rules of metric spaces. The metric emphasizes the topological properties of the sample space.

2.3. Permutations and symbolic sequences

The fractal dimension (usually referred to as the correlation dimension) is said to be one of the important characteristics and indexes of dynamic systems.[8] Why do we not always use the Grassberger–Procaccia

(G–P)[9,11] algorithm as the unique method to characterize the complexity of a system? Dimensionality measurement based on the use of the G–P algorithm in analysing time series with perturbations, e.g. ECG signals,[10,12] is problematic since it requires long, highly-accurate stationary data, and it is then only reliable for low-dimensional systems.[6] Consequently, two problems arise in using the G–P algorithm: first, the process consumes too much computation time because redundant computation exists in the G–P method and the hidden rules will not appear unless the time series is long enough, though the algorithm is improved the G–P algorithm remains time-consuming; and, second, it is hard to deal with a series with perturbations because we cannot obtain appropriate values of radius (r) of spherical neighbourhood and delay time (τ) without trying many times.

Suppose that $s = \{1, 2, \ldots, d\}$ ($d \geq 2$), $S = \{s_1 s_2 \cdots s_n \cdots s_{N_0} (N_0 < \infty)\}$ is a symbolic sequence transformed from a time series of a certain dynamic system, and n is the dimension of embedded phase space. If $s = \{1, 2\}$, S is overly coarse grained and n must be large enough for vectors in the reconstructed phase space to describe the property of S (see Ref. [6]), which means that the calculation will be too time-consuming. If d is too large, the calculation will be either very time-consuming or the complexity measure will not be reasonable because some symbols appear a few times or never appear.

To use symbolics as a tool, we need symbolic sequences first.

For a given time series, usually we obtain symbols by comparing neighbouring values or points with the mean value of the time series, or by integration with the two methods, which will involve the disadvantages mentioned above.

Compared with these methods, permutation as adopted in calculation of permutation entropy is good and very efficient in extracting local information. We can take $n = 4, 5, 6$[2,5] by the method of permutation. To apply the permutation to complexity measure by Definition 6, we take $n = d = 6$ and the calculation consumes much less time than that in calculating the correlation dimensions by the G–P algorithm. To illustrate the integration of symbolic dynamics and permutation, we suppose that $X_1(5) = (x_1^1, x_2^1, \ldots, x_5^1)$, $X_2(5) = (x_1^2, x_2^2, \ldots, x_5^2)$, $X_1(6) = (x_1^1, x_2^1, \ldots, x_6^1)$, $X_2(6) = (x_1^2, x_2^2, \ldots, x_6^2)$. By the permutation rule, we have the permutation sets $S_1(5) = (s_1^1 s_2^1 \ldots s_5^1)$, $S_2(5) = (s_1^2 s_2^2 \ldots s_5^2)$, $S_1(6) = (t_1^1 t_2^1 \ldots t_6^1)$, $S_2(6) = (t_1^2 t_2^2 \ldots t_6^2)$, respectively. $S_1(6)$ and $S_2(6)$ are words of Type II if $S_1(6) = S_2(6)$ and are words of Type I if $S_1(5) = S_2(5)$ but $S_1(6) \neq S_2(6)$.

2.4. Forbidden patterns, allowed patterns and complexity

Whether we evaluate complexity or permutation entropy by symbolic dynamics or we calculate fractal dimensions by the G–P method, we need to embed a higher dimension and thus the measure of the original series in the higher-dimensional space will be zero, which means that "gaps" appear in higher dimensions. The more gaps there are in the neighbourhoods, the less the distances of the points in the neighbourhood are estimated to be. We should consider not only the sum of the points in higher dimensions but also the correlation degree of the space points (i.e. vectors). For example, the three cases in Fig. 1 should not be considered to be identical though their total phase points and the forbidden patterns are the same, relatively. In Fig. 1, the grey and blank unit-grids represent forbidden patterns, or gaps in the higher-dimensional space. However, until now, no papers have brought the phase points and the correlation degree together into consideration when dynamic properties are analysed by the symbolic dynamics.

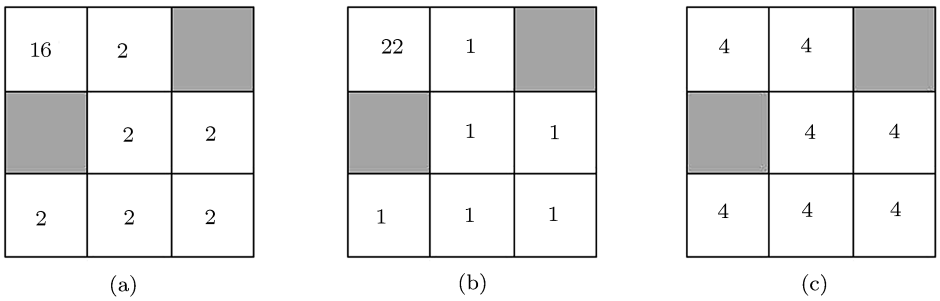


Fig. 1. Distributions of forbidden patterns and phase points in three cases.

The following Definitions 1 and 2 can be found in books on symbolic dynamics and Definition 3 was given by Wolfram.[7]

Definition 1 (word) A sequence or a subsequence in symbolic dynamics is called a word. The length of a word is called word length.

Definition 2 (allowed word) If a word appears in a given sequence then the word is called an allowed word. Sometimes an allowed word is also called a word for short if without confusion.

Definition 3 (forbidden pattern, forbidden word) In a given sequence the word $(s_1s_2\cdots s_{n-1})$ exists but $(s_1s_2\cdots s_{n-1}s_n)$ does not; then the word $(s_1s_2\cdots s_{n-1}s_n)$ is called a forbidden pattern or a forbidden word.

To classify the subsequences we present Definitions 4 and 5 as follows:

Definition 4 (allowed word of Type I) If both the words $(s_1s_2\cdots s_{n-1}s_n)$ $(s_n \in s)$ and $(s_1s_2\cdots s_{n-1}s'_n)$ $(s'_n \in s, s'_n \neq s_n)$ exist, then they are both called allowed words of Type I or bifurcation words, or simply words of Type I.

Definition 5 (allowed word of Type II) If the word $(s_1s_2\cdots s_{n-1}s_n)$ $(s_n \in s)$ is a forbidden word while $(s_1s_2\cdots s_{n-1}s'_n)$ is an allowed word, then $(s_1s_2\cdots s_{n-1}s'_n)$ is called an allowed word of Type II or simply a word of Type II.

The numbers of all allowed words, all forbidden words, all words of Type I, and all words of Type II are denoted by $N_a(n), N_f(n), N_1(n)$ and $N_2(n)$, respectively.

Theorem 1 If forbidden words exist in a given sequence, allowed words of Type II must exist, too. A given sequence consists only of words of types I and II.

The proof is self-evident.

Two words with length n of Type II have higher correlation degrees than those of type I.

If a strange attractor exists in a dynamic system, some phase points must disappear in a reconstructed higher-dimensional phase space. By the symbolic dynamics, the disappearing phase points are the forbidden words. There will be no forbidden words in a random sequence if the length is long enough and no allowed words of Type I exist in a periodic sequence. All of the forbidden words and the allowed words of Types I and II exist in any chaotic sequence.

Theorem 2 Suppose that $S_1(n)$ and $S_2(n)$ are two words and $S = (s_1s_2\cdots s_n\cdots s_{N_0}$ $(N_0 < \infty))$ is a given symbolic sequence with length N_0, N is the number of phase points and $N = N_0 - (n-1)$, and $s = (1, 2, \ldots, d)$ $(d \geq 2)$, then $\| S_{12} \| < r^n$ if $S_1(n)$ and $S_2(n)$ are words of Type II while $\| S_{12} \| \geq r^n$ if $S_1(n)$ and $S_2(n)$ are words of Type I, where $r = 1/d$.

Proof

If $S_1(n)$ and $S_2(n)$ are words of Type II, it follows that

$$\begin{aligned}\| S_{12} \| &\leq r^{n+1} + r^{n+2} + \cdots \\ &+ r^{N_0} < r^{n+1} + r^{n+2} + \cdots \\ &= \lim_{m\to\infty} \frac{r^{n+1}(1-r^m)}{1-r} = \frac{r^{n+1}}{1-r} \leq r^n.\end{aligned}$$

However, if S_1 and S_2 are two words of Type I with length n, obviously we have $\| S_{12}(n) \| \geq r^n$, thus completing the proof.

2.5. New approach to the measure of complexity

Combining the idea in the G–P algorithm with the above discussion, we should calculate the numbers of words of Type II but not those of forbidden words, and at the same time we should also consider the correlation degree of the words of Type II.

Now we calculate the correlation integral of the symbolic sequence. Note that we take the radius r of the neighbourhood as d^{-n}. We denote $\sum_{i,j=1}^{N} \theta(r - \|X_{ij}\|)$ by $M_2(r)$. Obviously, $M_2(r) = N_2(n)$ and $N_2(r)/N^2$ corresponds to $C_2(r)$ in Eq. (3). With the ideas of fractal dimensions and permutation, we define the complexity of the sequence as follows:

Definition 6 (complexity) The complexity of a given sequence is

$$\begin{aligned}C &= \lim_{n\to\infty} \frac{\ln[N_2(r)/N^2]}{\ln(r)} = \lim_{n\to\infty} \frac{\log_d[N_2(r)/N^2]}{\log_d d^{-n}} \\ &= \lim_{n\to\infty} \frac{\log_d[N^2/N_2(r)]}{n}, \end{aligned} \quad (5)$$

where N is the number of all allowed words with length n and, $C = 0$ if no bifurcation words exist and $C = \infty$ if no forbidden words exist.

The sequence is periodic when $C = 0$ and random when $C = \infty$.

3. Application of complexity to dynamic systems

Example 1 Complexity of Logistic systems

We take $N_0 = 3006$, $d = 6$ and $T_{\text{step}} = 0.01$, where T_{step} is the step length. We have the complexities of the Logistic system generated by Logistic map

(1) as shown in Fig. 2. For comparison, correlation dimensions obtained by the G–P algorithm are shown in Fig. 3 and complexities of the Logistic systems with perturbation are shown in Fig. 4.

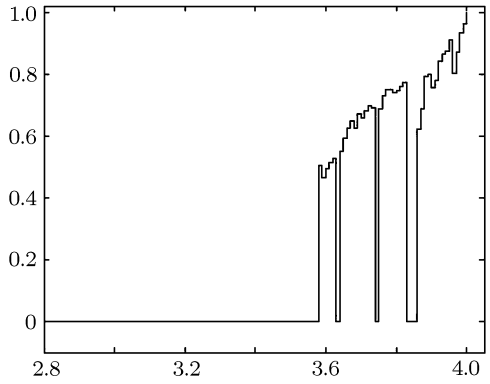


Fig. 2. Complexities of Logistic systems generated by Logistic map (1).

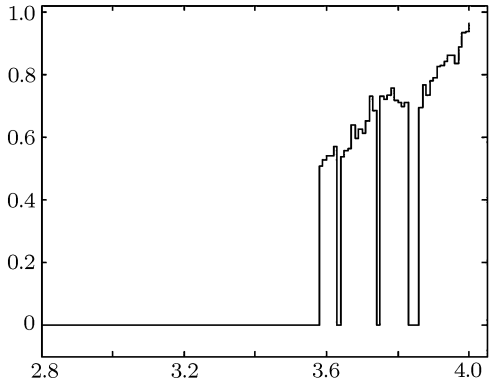


Fig. 3. Correlation dimensions obtained by the G–P algorithm.

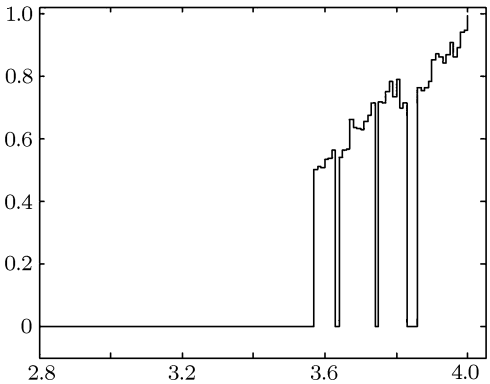


Fig. 4. Complexities of Logistic systems with noise level $\sigma = 0.5\%$ generated by Logistic map (2). Here we suppose $\{x_i\}$, $\{y_i\}$, $\{\xi_i\}$ have their means $\overline{x}$, $\overline{y}$, $\overline{\xi}$ and standard deviations σ_x, σ_y, σ_ξ, respectively. The noise level is defined as $\sigma = \sigma_\xi/(\sqrt{\sigma_y^2 + \sigma_\xi^2})$.

Example 2 Complexity of the systems generated by Hénon map (6)

$$\begin{cases} x_{n+1} = p - bx_n^2 + y_n, \\ y_{n+1} = ax_n, \end{cases} \tag{6}$$

where p, b and a are parameters.

Now we take $p = 1$ and $b = 1.4$. a is a variable and let $T_{\text{step}} = 0.01$. We have the complexities of the system generated by Hénon map (6) as shown in Fig. 5.

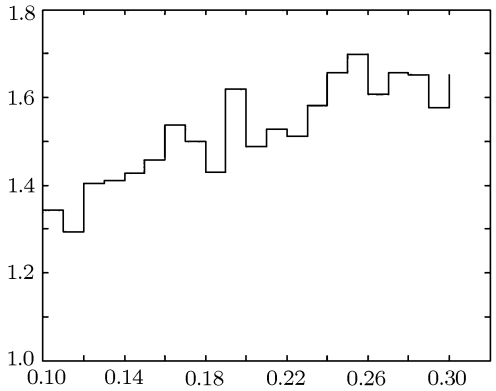


Fig. 5. Complexities of the systems generated by Hénon map (6).

From Fig. 5, we know that chaos exists in the systems by map (6) when $p = 1$, $b = 1.4$ and $a \in [0.1, 0.3]$.

4. Conclusions

We have proposed a complexity measure, a quantity which is based on the classification of symbolic patterns in a symbolic sequence. The measure simulates the correlation dimension by the G–P algorithm but it is not simply the correlation dimension because it takes into consideration the ideas of forbidden patterns, permutation entropy and correlative degree. By the algorithm, computation time is reduced significantly and the algorithm can separate deterministic systems from random series and propose a method to measure the complexities of different systems. To our knowledge, these are the first results of the complexity measure by the integration of some ideas. We believe that the technique can find wide applications in nonlinear analysis.

References

[1] Wang F L and Yang H H 2009 *Chin. Phys.* B **18** 4042
[2] Zanin M 2008 *Chaos* **18** 013119 (published online 13 March 2008)
[3] Amigo J M, Zambrano S and Sanjean M A F 2007 *Europhys. Lett.* **79** 50001
[4] Amigo J M, Kennel M B and Kocarey L 2005 *Physica* D **210** 77
[5] Amigo J M, Kocarev L and Szczepanski J 2006 *Phys. Lett.* A **355** 27
[6] Xu J H 1994 *Chaos, Solitons & Fractals* **4** 2111
[7] Wolfram S 1984 *Commun. Math. Phys.* **96** 15
[8] Meng Q F, Peng Y H and Xue P J 2007 *Chin. Phys.* **16** 1252
[9] Grassberger P and Procaccia I 1983 *Phys. Rev. Lett.* **50** 346
[10] Palus M 1995 *Physica* D **93** 64
[11] Takens F 1981 *Lecture Notes in Math.* **898** 361
[12] Casaleggio A, Corana A, Ranjan R and Thakor N V 1996 *Int. J. Bif. Chaos* **6** 1547
[13] Luo C W 2009 *Acta Phys. Sin.* **58** 3788 (in Chinese)
[14] Luo C W 2007 *Acta Phys. Sin.* **56** 6282 (in Chinese)
[15] Shen M F, Liu Y and Lin L X 2009 *Chin. Phys.* B **18** 1761
[16] Wang X M, Zhang B, Qiu D Y and Chen L G 2008 *Acta Phys. Sin.* **57** 6112 (in Chinese)
[17] Liu X F and Yu W L 2008 *Acta Phys. Sin.* **57** 2587 (in Chinese)
[18] Gao Z Y and Lu Q S 2007 *Chin. Phys.* **16** 2479