Short-term prediction of complex binary data

A.V. Adamopoulos[*1], **S. Likothanassis**[2], **N.G. Pavlidis**[3] and **M.N. Vrahatis**[3]

[1] Medical Physics Laboratory, Department of Medicine,
Democritus University of Thrace, GR-68100 Alexandroupolis, Hellas
[2] Pattern Recognition Laboratory, Department of Computer Engineering
and Informatics, University of Patras, GR-26500 Patras, Hellas
[3] Computational Intelligence Laboratory (CI Lab), Department of Mathematics, University of Patras Artificial Intelligence Research Center (UPAIRC), University of Patras, GR-26110 Patras, Hellas

Key words complex systems, logistic equation, short-term prediction, evolutionary algorithms.
Subject classification 62M10, 62M20

Perfect short-term prediction of binary patterns was revealed in binary data sets of highly complex nature. These binary patterns, referred to as perfect predictors, yield risk-free prediction of the value of the next bit of a binary sequence. The method was tested on binary data sets generated by applying a simple binary transformation on the data of the *logistic function* $x_{n+1} = rx_n(1 - x_n)$ for a variety of values of the *"nonlinearity parameter"*, r. Despite the chaotic nature of the logistic function and the complexity of the obtained binary data sets, an unexpected high number of prediction rules was revealed. In some cases predictability up to 100% was obtained.

1 Introduction

In the present paper we consider the problem of detecting and revealing binary strings that can account as perfect, or good, short-term predictors in complex binary data sets. A binary string (or, in other words, a binary pattern) of length L is assumed to be a perfect short-term predictor if its presence in any place of a binary sequence is declarative of the value of the next bit. Following this statement of a perfect predictor, it is apparent that a perfect predictor provides the ability for a 100% (that is, risk-free) prediction of the next bit. On the other hand, we call a good (but not-perfect), short-term predictor, a binary pattern that its appearance in any place of a binary sequence is related to the presence of a certain value of the next bit in a high number of cases, but not in all of these cases. As an example, we can consider a binary pattern that is followed by a "1" in 80% of the cases and by a "0" in the remaining 20% of the cases. For this particular binary pattern the probability of appearance of a "1" against the probability of appearance of a "0" is 4:1. Thus, although the pattern cannot account as a perfect predictor, it still can account as a good predictor.

In order to obtain complex binary data sets a method presented by Packard [1] was utilized. In that work, the methodology applied for the generation of binary data set employed a bit transformation of the logistic difference equation data. The logistic function given by:

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

was proposed as a mathematical model of population dynamics [2]. Despite its simplicity, Eq. (1) can provide a variety of different dynamical characteristics, depending on the value of parameter, r [3]. The parameter r is an expression of the nonlinearity of the system [4, 5]. For values of r in the interval $[0, 4]$ and initial value x_0 in the interval $[0, 1]$, the logistic function is bounded in $[0, 1]$. For values of r in the range $(1, 3)$, after a transient phase, the dynamics of the logistic system are settled to the fixed point $x^s = 1 - 1/r$ and remain stable thereafter. Therefore, the value x^s is the stability condition of the system, i.e. a fixed point attractor that the system sooner or later converges to. For the value $r = 3$ a new behavior is observed. For this value of r the dynamics of the

* Corresponding author. E-mail: adam@med.duth.gr, Phone: +30 25510 30501, Fax: +30 25510 30392

system bifurcate to give a cycle of period two. A further increase of r results to successive bifurcation and the related observed period doubling phenomenon, which refers to the resulting increase of the cycling period. The period doubling phenomenon leads to chaotic behaviour, i.e. infinite period for values of r in the range $[3.57, 4]$.

2 Material and Methods

The logistic function described in Eq. (1) was used for the generation of the raw data sequence $x(r)$. Consequently, a binary data sequence $b(r)$ was produced by applying the simple binary transformation proposed in [1]:

$$b(r) = \begin{cases} 0, & \text{if } x_n \leqslant 0.5, \\ 1, & \text{if } x_n > 0.5. \end{cases} \tag{2}$$

Hence, the application of the transformation of Eq. (2) produces one "1" for values of the logistic function exceeding 0.5, and zero "0" for values of the function smaller than, or equal to, 0.5.

A number of raw time series $x(r)$ of the logistic function were obtained by iterative application of Eq. (1) for a variety of values of the parameters r and x_0. To avoid transient phenomena, the first 10000 values of the logistic function were excluded. Then, binary data sets $b(r)$ consisting of 10^6 bits were produced by applying the transformation of Eq. (2) on the raw data $x(r)$.

To investigate the existence of binary patterns that can be considered as perfect, or in the worst case, as good predictors two different approaches were utilized:

(a) *Exhaustive search*: for small values of the length of the binary patterns, L, the number of all possible patterns, 2^L, is relatively small. In these cases the binary sequence $b(r)$ was exhaustively searched and the number of cases that each particular pattern was detected was counted.

(b) *Genetic Algorithm (GA) search*: for large values of the length of the binary patterns, L, the total number of different combinations increases exponentially so more efficient methods, such as genetic search, must be used [6, 7]. Therefore, a simple GA with binary representation was implemented and utilized. The GA population consisted of L-bit patterns. As crossover operator the usual one-point crossover operator was employed. As mutation operator the flip mutation operator was used. Finally, the number of times a specific pattern, p, was encountered in the binary sequence $b(r)$, was considered as the fitness function of that pattern.

Using the above described methods in a supplementary way (i.e., exhaustive search for small values of L, and GA search for large values of L) it was possible to detect and count the L-bit patterns that appeared in the $b(r)$ sequence. This was performed for several values of L. Consequently, by comparing the obtained results for L-bit and $(L+1)$-bit patterns we were able to detect the patterns that can account as perfect, or good predictors.

3 Results and Discussion

Indicative results of the application of our method to investigate the first objective of this research, (i.e., detection of possible dominance of certain binary patterns) are illustrated in Table 1. By comparing the findings presented in Table 1 it is apparent that specific rules for perfect prediction (for example patterns ending to zero, are always followed by one) do exist.

From the first line of Table 1, (for $L = 1$) it is clear that the ones and zeros are not equally distributed in $b(3.6)$. This is a common feature, as it is shown in Fig. 1 where the percentage distribution of bits is shown for several values of the non-linear parameter r, in the range $[3.5, 4.0]$.

As shown in Fig. 1, equal distribution of zeros and ones seems to be the exception rather than the rule. Although the results presented in Fig. 1 refer to binary sequences of length 10^6 bits, it is noteworthy to mention that additional experiments indicated that these results remain practically unchanged even for binary sequences of length 10^9 bits.

Despite the high average information loss of the logistic function [1, 5], (for $r = 3.9$ the Lyapunov exponent is $\lambda \cong 0.718$, i.e., one bit degrades by that much on every iteration) there exist conditions (in the form of binary patterns) that can be interpreted as perfect predictors since they can account for risk-free prediction of the next bit.

Table 1: Patterns with length L from 1 up to 7 found in $b(3.6)$ with 10^6 bits length.

length L	patterns found	frequency
1	0	363956
	1	636044
2	01	363956
	10	363956
	11	272088
3	010	227913
	011	136043
	101	363956
	110	136044
	111	136044
4	0101	227913
	0111	136043
	1010	227913
	1011	136043
	1101	136044
	1110	136044
5	01010	162475
	01011	65438
	01110	136043
	10101	227913
	10111	136043
	11010	65439
	11011	70605
	11101	136044

length L	patterns found	frequency
6	010101	162475
	010111	65438
	011101	136043
	101010	162475
	101011	65438
	101110	136043
	110101	65439
	110111	70605
	111010	65439
	111011	70605
7	0101010	97036
	0101011	65439
	0101110	65438
	0111010	65438
	0111011	70605
	1010101	162475
	1010111	65438
	1011101	136043
	1101010	65439
	1101110	70605
	1110101	65439
	1110111	70605

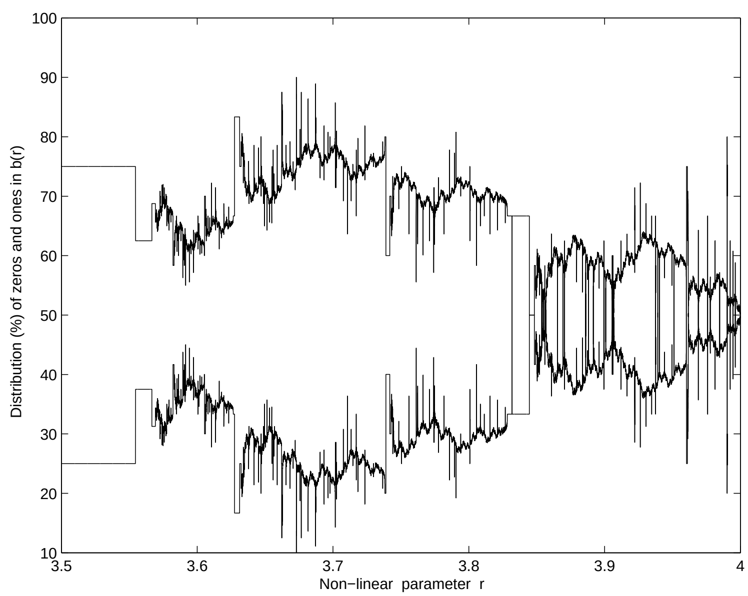

Fig. 1: Distribution of zeros and ones in $b(r)$ for r in $[3.5, 4.0]$ with 10^6 bits length.

This can be explained considering that there exist some pieces of the observed trajectory of the logistic system in the phase space that recurrently visit subspaces of the chaotic attractor. In these subspaces, the trajectory orbits are not widely spreading, and they can even be contracting [8]. Near these particular subspaces of the phase space of the system, high predictability appears.

Acknowledgements

This work was partially supported by the Hellenic Ministry of Education and the European Union under Research Program PYTHAGORAS-89203.

References

[1] N. H. Packard, Complex Systems **4**, 543 (1990).
[2] R. M. May, Nature **261**, 459 (1976).
[3] M. Barnsley, Fractals Everywhere (New York: Academic Press, 1988).
[4] J. L. Casti, Searching for Certainty (Scribners, 1992, Reprinted by Abacus, 1995).
[5] T. P. Meyer, F. C. Richards and N. H. Packard, Phys. Rev. Lett. **63**, 1735 (1989).
[6] J. H. Holland, *Adaptation in Natural and Artificial Systems* (University of Michigan Press, 1975).
[7] M. Mitchell, *Introduction to Genetic Algorithms* (MIT Press, 1996).
[8] J. M. Nese, Physica D **35**, 237 (1989).