4th WSEAS Int. Conf. on NON-LINEAR ANALYSIS, NON-LINEAR SYSTEMS and CHAOS, Sofia, Bulgaria, October 27-29, 2005 (pp37-42)

Exhaustive search for perfect predictors in complex binary data

ADAM V. ADAMOPOULOS
Medical Physics Laboratory, Department of Medicine
Democritus University of Thrace
GR-681 00, Alexandroupolis
HELLAS

Abstract: - An exhaustive method for the detection of short-term perfect predictors in complex binary sequences is presented. As short-term perfect predictors we assume bit sequences that give risk-free prediction of the value of the next bit. The method was tested on binary data sets produced by applying a simple binary transformation on the data of the logistic function 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 sequences, an unexpected high number of prediction rules were detected. In some cases the predictability reached up to *100%*. In the worst case, (for $r = 4.0$), the predictability is up to *33.3%*. Finally, as it was found via extensive simulations the number of L-bit perfect predictors as a function of their bit-length L is given by the *Fibonacci recursive formula*.

Key-Words: Nonlinear dynamics, logistic function, binary perfect predictors

1 Introduction

Predicting or forecasting the dynamics of complex systems is a difficult task, usually beset with a number of problems [1, 2]. In complex systems, i.e. systems with many degrees of freedom that are highly coupled, these problems are related to the limitations of the available data and pseudo-randomness generated from the existing low dimensional chaotic dynamics of the systems. In the case of experimental data, additional errors related to the observational procedure and errors from the presence of high dimensional noise are usually present [3,4].

One of the well known mathematical functions that present rich nonlinearity and highly complex dynamics is the logistic function defined as:

$$x_{n+1} = r * x_n * (1 - x_n) \qquad (1)$$

Eq. (1) was proposed as a mathematical model of population dynamics [5]. Although simple, Eq. (1) may provide a variety of different dynamical characteristics, strongly depended on the value of parameter r. The parameter r is an expression of the nonlinearity of the system. 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 larger than 4, or for values of x_0 larger than 1, the logistic function is unbounded.

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 $r = 3$ a different behaviour is presented: the dynamics of the system bifurcate to give a period of two. A further increase of the value of r results to successive bifurcations and the related *period doubling phenomenon* that is observed, which refers to the resulting increase of the cycling period of the generated sequence of Eq. (1). The period doubling phenomenon leads to chaotic behaviour, i.e. infinite period for values of r in the range [*3.57, 4.0*]. As a result of the above characteristics, the dynamics of the logistic function was investigated by a large number of researchers following different methods of approximation and analysis and concluded to interesting results. One of the disciplines that were investigated was the derivation of forecasting methods.

The problem of predicting chaotic timeseries attracted the interest of many researchers. Although the theory of chaos places fundamental limits on long-term prediction, it suggests possibilities for short-term prediction. Several works, regarding this topic, have been published for example the Farmer's prediction algorithm [6]. A different approach is due

to Packard [3], who for the first time applied a binary transformation on the raw data of the logistic function and investigated predicting abilities on the produced complex binary data. Packard proposed a Genetic Algorithm search for predicting rules in such binary series. The results of his investigation were that there exist some probabilistic predicting rules.

In this work a new approach that provides perfect predictors in complex binary data is presented. Specifically, following the methodology presented in [3], instead of using raw data we generated binary sequences by applying a simple binary transformation. The method is capable for detecting and revealing binary patterns that can account as perfect predictors. A binary pattern is to be considered as a perfect predictor if its appearance anywhere in the data set is declarative of the value for the next bit. In other words, a perfect predictor, consisted from a bit sequence, can account for next bit risk-free prediction.

2 Methods

2.1 Derivation of binary sequences

The first step to generate complex binary sequences is to generate raw data sequences $x_n(x_0, r)$ with the use of the logistic function of Eq. (1). Then, the corresponding binary sequences $b_n(x_0, r)$ were generated by applying the simple transformation:

$$b_n(x_0,r) = \begin{cases} 0 & \text{if } x_{n+1} \le x_n \\ 1 & \text{if } x_{n+1} > x_n \end{cases} \qquad (2)$$

Eq. (2) generates 1 in the case of an increase of the logistic function, otherwise it generates 0. The case $x_{n+1} = x_n$, although included in Eq. (2), is of no practical meaning, since it is the condition for steady state stability to a fixed point, obviously not present in chaotic dynamics considered here. A similar transformation was used in [3] where the binary sequences were generated comparing the raw data values with the threshold value of *0.5*. When the raw value was greater that *0.5* the output was *1*; otherwise the output was *0*.

In order to investigate for the existence of binary patterns that actually can be considered as perfect predictors an exhaustive search method was applied on the generated binary sequences $b_n(x_0, r)$. The first task was to utilize exhaustive search to detect the binary patterns of given length L that were found in $b_n(x_0, r)$. Although the chaotic nature of the raw data, and the resulted complexity of the obtained binary sequences $b_n(x_0, r)$, it was found that indeed, there exist some binary patterns that are detected in higher rates than others. In addition, some binary patterns were absent from the binary sequences. The second task was to apply again exhaustive search on the same binary sequences and detect the binary patterns of length $L+1$ which were present in that data set. This task was performed in order to investigate for the presence of certain binary patterns of length L in the binary sequences that could account as good or, in the ideal case, as perfect predictors of the next bit of the data set. Results of the second task indicated that this holds in surprisingly high number of cases.

2.2 Exhaustive search

A number of raw time series $x_n(x_0, r)$ of the logistic function were derived by iterative application of Eq. (1) for a variety of values of the parameters r and x_0. Then, binary sequences $b_n(x_0, r)$ consisted of 10^6 bits were generated by applying the transformation of Eq. (2) on the raw data $x_n(x_0, r)$. These binary sequences were given as an input to the exhaustive search method. According to that method for any given pattern length L, all the possible binary patterns of length L were constructed and their frequency of appearance in $b_n(x_0, r)$ were estimated. For a given binary pattern length L, the number of all possible binary patterns is 2^L.

3 Results

Our experiments were performed for a broad spectrum of different values of the parameter r. Results reported in the present work refer to three different values of r, namely *3.6, 3.9* and *4.0*.

Table 1
Binary patterns found in $b_n(0.4, 3.6)$.
The total number of patterns is 10^6.

L	Patterns	Frequency
1	0	500000
	1	500000
2	01	500000
	10	500000
3	010	500000
	101	500000
4	0101	500000
	1010	500000

According to the codification of Eq. (2) (*0* means decrease of the consecutive *x*-value, whereas *1* means increase), the interpretation of this result is that for $r = 3.6$, an increase of the logistic function is always followed by a decrease and vice versa. This leads to the following simple prediction rule: 0's are always followed by 1 and 1's are always followed by 0. The subsequent absence of the patterns *00* and *11* in that row indicates that despite the chaotic character of the raw data $x_n(0.4, 3.6)$, there is a norm of continuously repeated rise and fall of the logistic function. This simple prediction rule is reflected in the patterns of length $L = 3$ and $L = 4$ that were found in $b_n(0.4, 3.6)$ and are presented in the third and the fourth row of Table 1 respectively. Independently of the length *L* of the pattern, for each particular value of *L* there exist only two patterns, consisting of bits in a consequently alternating fashion. As a direct result of the revealed prediction rule any pattern is a perfect predictor of the next bit, since, zero-terminated patterns (i.e., its last bit is *0*) are followed by an *1*, whereas, one-terminated patterns (i.e., its last bit is *1*) are always followed by a *0*. Therefore the frequency of appearance of perfect predictors in the case of $b_n(0.4, 3.6)$ is 100%.

Quite different are the results obtained from the analysis of $b_n(0.4, 3.9)$ which are presented in Table 2. The first to notify in Table 2, is that there is no equal distribution of 0's and 1's in $b_n(0.4, 3.9)$. Specifically, the results of the investigation of the frequency of appearance of bit-strings with length $L = 1$, indicated the existence of *409338* 0's ($\cong 41\%$) and *590662* 1's, ($\cong 59\%$) in $b_n(0.4, 3.9)$. This result directly indicates that the logistic function is more frequently (*18%*) increased than decreased for $r = 3.9$. Further investigation, indicated that the critical value r_c for above which the binary sequence $b_n(0.4, r)$ do not appear a *50%-50%* distribution of 0's and 1's is lying in the interval $3.67857 < r_c < 3.67858$. The unequal distribution of 0's and 1's in $b_n(0.4, 3.9)$ clearly influences the distribution and the frequency of appearances of binary patterns with length $L > 1$.

This is clearly shown at the results obtained for the investigation for bit patterns with length $L = 2$, shown in the second row of Table 2. The double-zero pattern (*00*) did not appear in $b_n(0.4, 3.9)$. Practically, the interpretation of this result is that there are no two consecutive decreases of the logistic function for $r = 3.9$. At a second level, the absence of the *00* pattern indicates that all patterns of any length $L \geq 2$ that include the pattern 00 do not appear in $b_n(0.4, 3.9)$. This is explicitly shown in the rest rows of Table 2, which present the binary patterns and their corresponding frequency of appearance for length *L* up to *7*. For example, as it can be seen in the third row of Table 2 ($L = 3$) none of the binary patterns *000, 001, 100* appears in the data set. Furthermore, in that same row, the pattern *111* is absent too. This means that there are not three consecutive increases of the logistic function for $r = 3.9$, and that all patterns that include three consecutive 1's (*⋯111⋯* in general, such as *0111, 10111101*, etc.), are not present in $b_n(0.4, 3.9)$.

Table 2
Binary patterns found in $b_n(0.4, 3.9)$.
The total number of patterns is 10^6.

L	Patterns	Frequency
1	0	409338
	1	590662
2	01	409338
	10	409338
	11	181324
3	010	228014
	011	181324
	101	409338
	110	181324
4	0101	228014
	0110	181324
	1010	228013
	1011	181325
	1101	181324
5	01010	129175
	01011	98839
	01101	181324
	10101	228013
	10110	181325
	11010	98838
	11011	82486
6	010101	129175
	010110	98839
	011010	98838
	011011	82486
	101010	129174
	101011	98839
	101101	181325
	110101	98838
	110110	82486
7	0101010	75577
	0101011	53598
	0101101	98839
	0110101	98838
	0110110	82486
	1010101	129174
	1010110	98839
	1011010	98839
	1011011	82486
	1101010	53597
	1101011	45241
	1101101	82486

On the other hand, the absence of two consecutive zeros in the data set implies that if the last bit of a binary pattern is *0*, then this pattern will

be followed by an *1*. In other words, if the terminating bit of a pattern is *0* then it can be safely predicted that the next bit to be expected is an *1*. Therefore the patterns terminating to a 0 do not bifurcate, (i.e. they are not followed by either a *0* or an *1*, but they are followed exclusively by an *1*).

Table 3
Binary patterns found in $b_n(0.4, 4.0)$.
The total number of patterns is 10^6.

L	Patterns	Frequency
1	0	333226
	1	666774
2	01	333226
	10	333225
	11	333549
3	010	166863
	011	166363
	101	333225
	110	166363
	111	167186
4	0101	166863
	0110	82759
	0111	83604
	1010	166862
	1011	166363
	1101	166363
	1110	83604
	1111	83582
5	01010	83761
	01011	83102
	01101	82759
	01110	41955
	01111	41649
	10101	166862
	10110	82759
	10111	83604
	11010	83101
	11011	83262
	11101	83604
	11110	41649
	11111	41933
6	010101	83761
	010110	41301
	010111	41801
	011010	41279
	011011	41480
	011101	41955
	011110	20796
	011111	20853
	101010	83760
	101011	83102
	101101	82759
	101110	41955
	101111	41649
	110101	83101
	110110	41458
	110111	41804
	111010	41822
	111011	41782
	111101	41649
	111110	20853
	111111	21080

Thus, the zero-terminated patterns are perfect predictors of the next bit in $b_n(0.4, 3.9)$. This also stands for the pattern *11*. If the pattern *11* was bifurcated, then both the patterns *110* and *111* would be present in row *3* of Table 2 (which corresponds to the patterns with length $L = 3$ that were found in $b_n(0.4, 3.9)$. However, as it was previously notified, the pattern *111* is absent from that list, therefore nor the pattern *11* bifurcates, but is always followed by a *0*. Therefore, the pattern *11* is a perfect predictor as well, predicting that the next bit is *0*. As a general result, all *L*-bits patterns (for $L \geq 2$) terminating to *11* will be followed by a *0*.

The next thing that we consider, is that the only patterns that bifurcate, (i.e. are followed by either a *0* or an *1*) are these patterns that are terminated either to *1* (in the special case of length $L = 1$) or to *01* (in the general case of any length $L > 1$). Therefore, we have concluded some rules of pattern bifurcation or no bifurcation that can be applied in general for every pattern. These rules indicate that from the three present patterns with $L = 2$, (*01, 10* and *11*) the only that bifurcates is *01* (resulting to *010* and *011* to appear), whereas the pattern *10* is not bifurcated, but always followed by an *1* (therefore only *101* was found in $b_n(0.4, 3.9)$). Also, the pattern *11* is not bifurcated too, but always followed by a *0* (therefore only *110* was found in $b_n(0.4, 3.9)$). In a similar way, from the four *3*-bit patterns present in the data set (*010, 011, 101, 110*), only *101* bifurcates (to generate *1010* and *1011*), whereas the rest of them are not bifurcated (010 terminates to *0* and gives only *0101*, *011* terminates to *11* and gives only *0110* and *110* terminates to *0* and gives only *1101*).

A few more words with respect to the frequency of appearance of the patterns that can account as perfect predictors. The *1*-bit length pattern *0* appears *409338* times in the $b_n(0.4, 3.9)$ and as mentioned previously, it is a perfect predictor that the following bit is an *1*. Therefore the number of perfect predictors considering binary patterns with length $L = 1$ is $409338/10^6$, approximately *41%*. On the other hand, considering the *2*-bits patterns capable for perfect prediction, it is concluded that there are two perfect predictors, namely *10* (predicting that *1* always follows) and *11* (predicting that *0* always follows). The frequency of appearance of these two patterns (according to second row of Table 2) is $(409338+181324)/10^6$, approximately *59%*.

It is noteworthy, that this high percentage of appearance of perfect predictors is conserved for any higher value of the length *L* of the binary patterns. For example for $L = 5$ perfect predictors are the patterns *01010, 01011, 10110, 11010* and *11011*, (all of them terminated with a *0* or an *11*), with

frequency of appearance *(129175 + 98839 + 181325 + 98838 + 82486)/10^6 = 59%*. Therefore, despite the complexity of $b_n(0.4, 3.9)$, in an unexpected high number of cases the exact value of the next bit can be predicted.

Similar, but not identical results are obtained for the data set $b_n(0.4, 4.0)$. These results are summarized in Table 3. As it can be seen in the first row of Table 3, the number of *0*'s in the data set was found to be *333226* (approximately *1/3*, or *33.3%*) and the number of *1*'s was *666774* (approximately *2/3*, or *66.7%*).

Following the above analysis and interpretation of the results of Table 3, the main conclusion is that also for the case of $r = 4.0$ there exist perfect predictors and prediction rules. As it can be noted in the second row of Table 3 (corresponding to length $L = 2$) the pattern *00* is not included in that list. This means that any patterns, of any length L, that includes two consecutive *0*'s are also excluded. However, this is the only exclusion rule concerning the patterns of Table 3. This is in contrast to the results reported in Table 2 (corresponding to $r = 3.9$), where an additional exclusion rule was found (not only *00*, but also *111* was absent).

Therefore the (only) prediction rule in case of $r = 4.0$, is that the zero-terminating patterns do not bifurcate, but they are always followed by an *1*. On the other hand, the one-terminating patterns always bifurcate and therefore are followed by either *0* or *1* with approximately equal probabilities.

Table 4
Number of binary patterns appeared in $b_n(0.4, 4.0)$ for L up to 11.

L	Number of Patterns found
1	2
2	3
3	5
4	8
5	13
6	21
7	34
8	55
9	89
10	144
11	233

According to these prediction and bifurcation rules, the frequency of appearance of perfect predictors for a certain length L is determined by the number of L-bit patterns that are terminating to a *0*, which is *333226*/10^6, approximately *33.3%*. As in both the two previously examined cases (for $r = 3.6$ and for $r = 3.9$), even for $r = 4.0$, the frequency of appearance of perfect predictors is conserved and is independent of the value of the length L.

As a result of the presence of a single exclusion rule that was found in $b_n(0.4, 4.0)$, the number of the patterns of a certain length L that can be found in the binary sequence can be mathematically formulated. This is shown in Table 4, which presents the number of L-bits patterns that were found in the data set as a function of the length L of these patterns. It is easily recognized that these numbers correspond to the Fibonacci sequence.

Thus, if we denote as $P(L)$ the number of patterns with length L that were found in the data set, then for an arbitrary value of L, P(L) is given by the simple Fibonacci recursive formula:

$$P(L) = P(L-1) + P(L-2) \qquad for\ L \geq 3, \qquad (3)$$

with $P(1) = 2$ *and* $P(2) = 3$.

4 Conclusions

The presence of perfect predictors resulted from the presence of binary patterns that do not bifurcate, or in other terms from the existence of some binary patterns that are always followed by a *0* and some other patterns that are always followed by an *1*. As it was found out, the rules of pattern bifurcation or no bifurcation are easily extracted. In the same manner easily can be extracted rules of presence or absence of the binary patterns, as well as the total number of different patterns of a certain length L that can be found in a binary data set. Thus, although the high average information loss of the logistic function [3], (for $r = 3.9$ the Lyapunov exponent is $\lambda \cong 0.718$, i.e., one bit degrades by that much, on every iteration [7, 8]) 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. 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, or, are even contracting [9, 10]. Near these particular subspaces of the phase space of the system, high predictability appears.

It is noteworthy to refer, that the results presented above are typical and representative. Identical results were obtained in both qualitative and quantitative manner for $b_n(x_0, r)$ binary sequences with the same value of parameter r and different

initial value x_0. Thus, these results appeared to be independent of the initial condition of the system. In addition, our results seem to be independent of the length of $b_n(x_0, r)$. The above reported results are in good agreement with the ones reported by Packard [3], although in that work a different transformation was used in order to derive symbolic dynamics (i.e., the binary data sets). Furthermore, compared to that work, a more detailed analysis is provided here. In future work, Genetic Algorithms and Evolutionary Computation techniques like the ones discussed in [11, 12, 13] will be applied on highly complex binary sequences, in order to investigate the presence of hidden order and prediction ability.

Acknowledgement
This work was partially supported by Hellenic Ministry of Education and the European Union, under research program PYTHAGORAS – 89203.

References:
[1] J.L. Casti, *Searching for Certainty*, Scribners, 1992, Reprinted by Abacus, 1993, 1995.
[2] T.P. Meyer, F.C. Richards & N.H. Packard, A Learning Algorithm for the Analysis of Complex Spatial Data, *Phys. Rev. Lett.* **Vol.** 63, 1989, pp. 1735-1738.
[3] N.H. Packard, A Genetic Learning Algorithm for the Analysis of Complex Data, *Complex Systems* Vol. 4, 1990, pp. 543-572.
[4] T.P. Meyer and N.H. Packard, Local Forecasting of High Dimensional Chaotic Dynamics, *Technical Report CCSR-91-1*, University of Illinois, 1991.
[5] R.M. May, Simple mathematical models with very complicated dynamics, *Nature*, Vol. 261, 1976, pp. 459-467.
[6] J. D. Farmer and J. J. Sidorowich, Prediction Chaotic Time Series, *Physical Review Letters,* Vol. 59(8), 1987, p. 24.
[7] J.L. Casti, *Would-be worlds*, John Wiley & Sons, Inc., New York, 1997.
[8] R. Shaw, Strange attractors, chaotic behavior and information flow, *Z. Naturforscung*, Vol. 36(a), 1981, p. 80.
[9] J.M. Nese, Quantifying local predictability in phase space, *Physica D*, Vol. 35, 1989, p. 237.
[10] N. Packard, J. Chrutchfield, D. Farmer and R. Shaw, Geometry from a time series, Phys. Rev. Lett., **Vol.** 45, 1980, pp. 712-716.
[11] D.B. Fogel and L.J. Fogel, Preliminary Experiments on Discriminating between Chaotic Signals and Noise Using Evolutionary Programming, In: Genetic Programming '96, J.R. Koza, D.E. Goldberg, D.B. Fogel and R.L. Riolo, (eds.), MIT Press, 1996, pp. 512-520.
[12]. B.S. Mulloy, R.L. Riolo, and R.S. Savit, Dynamics of Genetic Programming and Chaotic Time Series Prediction, In: Genetic Programming '96, J.R. Koza, D.E. Goldberg, D.B. Fogel and R.L. Riolo, (eds.), MIT Press, 1996, pp. 166-174.
[13] E.H.N. Oakley, Genetic Programming, the Reflection of Chaos and the Bootstrap: Towerds a Useful Test for Chaos, In: Genetic Programming '96. J.R. Koza, D.E. Goldberg, D.B. Fogel and R.L. Riolo, (eds.), MIT Press, 1996, pp. 175-181.