Linköping Electronic Articles in
Computer and Information Science
Vol. 6(2001): nr 16

The nonprobabilistic approach to learning the best prediction

Boris Ryabko

Linköping University Electronic Press
Linköping, Sweden

http://www.ep.liu.se/ea/cis/2001/016/

Published on August 30, 2001 by
Linköping University Electronic Press
581 83 Linköping, Sweden

Linköping Electronic Articles in
Computer and Information Science
ISSN 1401-9841
Series editor: Erik Sandewall

Typeset by the author using LaTeX
Formatted using étendu *style*

Recommended citation:
<Author>. <Title>. Linköping Electronic Articles in Computer and Information Science, Vol. 6(2001): nr 16. http://www.ep.liu.se/ea/cis/2001/016/. *August 30, 2001.*

This URL will also contain a link to the author's home page.

Abstract

The problem of predicting a sequence $x_1, x_2, ...$ where each x_i belongs to a finite alphabet A is considered. Each letter x_{t+1} is predicted using information on the word $x_1, x_2, ..., x_t$ only. We use the game theoretical interpretation which can be traced to Laplace where there exists a gambler who tries to estimate probabilities for the letter x_{t+1} in order to maximize his capital . The optimal method of prediction is described for the case when the sequence $x_1, x_2, ...$ is generated by a stationary and ergodic source. It turns out that the optimal method is based only on estimations of conditional probabilities. In particular, it means that if we work in the framework of the ergodic and stationary source model, we cannot consider pattern recognition and other complex and interesting tools, because even the optimal method does not need them. That is why we suggest a so-called nonprobabilistic approach which is not based on the stationary and ergodic source model and show that complex algorithms of prediction can be considered in the framework of this approach.

The new approach is to consider a set of all infinite sequences (over a given finite alphabet) and estimate the size of sets of predictable sequences with the help of the Hausdorff dimension. This approach enables us first, to show that there exist large sets of well predictable sequences which have zero measure for each stationary and ergodic measure. (In fact, it means that such sets are invisible in the framework of the ergodic and stationary source model and shows the necessity of the new approach.) Second, it is shown that there exist quite large sets of such sequences that can be predicted well by complex algorithms which use not only estimations of conditional probabilities.

1 Introduction

Presently, the problem of prediction is investigated by many researches because its practical applications and importance for probability theory, machine intelligence, learning and other theoretical sciences: see, for example, [1], [8] . We shall investigate a model of prediction of time series which can be traced to Laplace (cf. [3] where the problem is referred to as the problem of succession). Namely, we consider the finite alphabet $A = \{a_1, \dots, a_m\}, m \geq 2$, and an infinite sequence $x_1, x_2, \dots.$, where $x_i \in A$. Let us assume that gambler I has capital V_t at moment $t = 0, 1, \dots; V_0 = 1$. Each moment t that the gambler divides the capital V_t into $|A|$ parts equals $V_t P_I(a/x_1, \dots, x_t), a \in A$. (The Stakes on the $x_{t+1} = a \in A$). $P_I(a/x_1, \dots, x_t)$ reflects the I- gambler's confidence in the appearance of $a \in A$ at the moment $t + 1$. (Here and in the sequel $|X|$ denotes the number of letters , if X is an alphabet or a set, and the length of X, if the X is a word).

The gambler I learns x_{t+1} at the moment $t+1$ and his capital becomes equal to $|A| V_t P_I(x_{t+1}/x_1, \dots, x_t)$. (i.e., the stake on the right letter is increasing $|A|$ times).

This model was suggested in [6] and used by many authors, see, for example, [12]. It seems that this model is the simplest and can be considered as the useful tool for making first steps towards an understanding of the connection between complexity of prediction algorithms and their efficiency.It should be noted that this question has attracted attention of many researchers and some important results were obtained in the framework of the machine intelligence theory and the theory of complexity. Thus, relationships between the time of calculation and the memory size of algorithm were investigated in [9, 5, 10].

Denote the win by the method (or gambler) I with $W_I(x_1, \dots, x_t)$:

$$W_I(x_1, \dots, x_t) = \prod_{i=1}^{t} (|A| P_I(x_{i+1}/x_1, \dots, x_i)) \tag{1}$$

In this case the gambler divides his capital into $|A|$ parts each moment (stakes on the letter of A), and at the moment $(t + 1)$ his capital should be equal to

$$W_I(x_1, \dots, x_t)(|A| P_I(x_{i+1}/x_1, \dots, x_i))$$

, i.e., the stake on the letter that actually appears, increases $|A|$ times.

Our goal is to find the algorithms maximizing (1). From the mathematical point of view it is convenient to study the logarithm of $W_I(x_1, \dots, x_t)$ divided by t. Let

$$L_I(x_1, \dots, x_t) = \log(W_I(x_1, \dots, x_t))/t$$

(Here and below $\log x = \log_2 x$). Then

$$L_I(x_1, \dots, x_t) = \log |A| + \sum_{i=1}^{t} \log P_I(x_i/x_1, \dots, x_{i-1})/t \tag{2}$$

We note that, if at the word $x_1, \dots, x_t$ strategy I recommends staking all the capital V_0 ($V_0 = 1$) on x_1 (that is unknown), then the capital $V_1 = |A|$ on x_2 and so on. This results in a maximal win equal to $|A|^t$ after t games.

If the gambler breaks up all his capital into $|A|$ equal parts each time and stakes those parts on the letters of A, then the win is equal to 0 and the capital remains the same: $V_0 = V_1 = ... = V_t = 1$. In general, $L_I(x_1, \dots, x_t)$ is the mean value of the exponent of the capital increase over the first t games: $V_t = |A| 2^{t L_I(x_1,\dots,x_t)}$. This value characterizes quantitatively the efficiency of prediction I.

It is natural to look for the gambler's srategy which gives the maximal asymptotic value of $L_I(x_1, ..., x_t)$ in (2). In the next section we consider this problem when it is known that the sequence $x_1 ... x_t$ is generated by an ergodic and stationary source but parameters of the source are not known (ESS model sometimes called as a problem of universal prediction). The main result of that part is the asymptotically optimal (for the ESS model) method of prediction named ρ . This method can find each periodicity. For example, let $A = \{0, 1\}$ and $x_1 .... x_4 = 0101$. The described below optimal method ρ predicts

$$P_\rho(0/0101) = 33/52, P_\rho(1/0101) = 19/52.$$

Thus, the stake on 0 after 0101 is almost twice as high as the stake on 1.

The method ρ is based on estimations of conditional probabilities. In particular, it means that the optimal (for the ESS) method ρ does not use even simple regularities. As an example, consider two infinite sequences

$$x^1 = 1010010001000010....., x^2 = 0100011011000001010011...... \quad (3)$$

The structure of the first sequence is obvious. In order to understand regularities in x^2 it is enough to figure out that it is the sequence of the all lexicographically ordered binary words $(0, 1, 00, 01, 10, 11, 000, 001, ...,.)$ It is easy to show that the method ρ does not use regularities of the x^1 and x^2, and that is why its prediction is not efficient. In fact, the sequences like x^1 and x^2 are invisible in the framework of the ESS model because they belong to the set which measure is equal to zero according to any ergodic and stationary measure.

So, on the one hand the method ρ is optimal in framework of the ESS model, on the other hand, it cannot find even quite simple regularities and use them for predicting. Moreover, it is shown in the third part of the paper, that the invisible set is quite large. In order to measure the 'size' of sets we use the Hausdorff dimension. (Hausdorff dimension, best known as a poweful tool of fractal geometry, has been known for over fifty years, to be closely related to information theory and prediction ,see, for example, [2, 12, 13].)

Thus, we know that the set of sequences which cannot be investigated in the framework of the ESS model is large. But how many sequences from this set can be predicted quite well ? This question is considered in the last part of the paper. It is shown that there exists a large subset of sequences from the 'invisible' set which can be predicted well by algorithms realizable by Turing machine. It turns out that the efficiency of the best algorithms is closely related with Kolmogorov complexity.

2 A method which is asymptotically optimal for ergodic and stationary sources

We shall describe the optimal method ρ which is based on results from universal coding theory [12]. First, we give some definitions. Let A^n, A^* and A^∞ be the sets of all $n-$ length words, all finite words and all infinite words over the alphabet A, respectively. Let $u_1u_2\dots u_n$ and $v_1v_2\dots v_k$ be two words of A^* and $k \le n$. By $\tau_v(u)$ we denote the rate of the word v occurring in the sequence $u_1u_2\dots u_k$, $u_2u_3\dots u_{k+1}$, $u_3u_4\dots u_{k+2}$, $\dots$, $u_{n-k+1}\dots u_n$. For example, $\tau_{00}(000100) = 3$. For $k = 0, 1, 2.\dots$ denote the mapping ρ_k that assigns to each word $u \in A^*$ the value

$$p_k(u) = \begin{cases} |A|^{-|u|}, & \text{for } |u| \le k \\ \left(\frac{\Gamma(|A|/2)}{\Gamma(1/2)^{|A|}}\right)^{|A|^k} \cdot \frac{1}{|A|^k} \prod_{\alpha \in A^k} \frac{\prod_{a\in A}\Gamma(\tau_{\alpha a}(u)+1/2)}{\Gamma(\tau^*_\alpha(u)+|A|/2)}, & \text{for } |u| > k \end{cases}$$

where $\tau^*_\alpha = \sum_{a\in A} \tau_{\alpha a}(u)$ and $\Gamma()$ is a gamma function. Now, define the probability distribution λ on the set of nonnegative integers 0,1,2,... using the code from [7]. Let $\log^{(0)}(x) = x$, $\log^{(i)}(x) = \log_2(\log^{(i-1)}(x))$ for $i \ge 1$ and $m(x) = i$, so that $0 \le \log^{(i)}(x) \le 1$. For $n = 0, 1, 2, \dots$, define

$$w(n) = \sum_{i=1}^{m(n)} \lfloor \log^{(i)}(n) \rfloor + m(n) + 1, \lambda(n) = 2^{-w(n)}.$$

It is known that

$$-\log \lambda(n) = \log n + O(\log \log n)$$

when $n \to \infty$ [12].

We define the measure ρ by

$$\rho = \sum_{k=0}^{\infty} \lambda(k)\rho_k(u).$$

The method of prediction is defined by

$$P_\rho(a/x_1x_2\dots x_t) = \rho(x_1x_2\dots x_t a)/\rho(x_1x_2\dots x_t).$$

As an example, consider the prediction computation on $x = 0101, A = \{0, 1\}$ by the method ρ. We have

$$\rho(01010) = 33/2^{10}, \rho(01011) = 19/2^{10}.$$

Hence, the prediction follows:

$$P_\rho(0/0101) = 33/52, P_\rho(1/0101) = 19/52.$$

Thus, the stake on 0 after 0101 is almost twice as high as the stake on 1.

¿From the definition of ρ_k and ρ we can see that the prediction method ρ uses the rates of the subwords occurring in the sequence $x_1x_2\dots x_t$. In fact, those rates are the estimations of the conditional probabilities.

The theorem below shows that the metod ρ is close to being optimal when $t \to \infty$.

Theorem 1 *For any stationary and ergodic process over A^∞ with probability equal to 1,*

$$\lim_{t\to\infty} L_\rho(x_1x_2\dots x_t) = \log|A| - h$$

where h is the Shannon entropy of the process. (The limit does exist with probability equal to 1.)

On the other hand, any prediction method α is asymptotically no better than ρ, so the following inequality is valid with probability equal to 1:

$$\lim_{t\to\infty} L_\alpha(x_1x_2\dots x_t) \le \log|A| - h$$

The proof is given in [12]. The definition of the Shannon entropy can be found, for example, in [4] .

3 The new approach

Let us consider the performance of the prediction method ρ when it is applied to the sequences x^1 and x^2, which are defined by (3). In both cases ρ does not recognize regularities. When ρ is applied to x^1 and $t\to\infty$, the gambler will stake almost all money on '0' and will not be able to predict appearances of 1's because, informally speaking, the sequence x^1 does not contain subwords which are repeated periodically. It is not the case when ρ is applied to x^2 because it is known that any subword $u \in \{0,1\}^n, n \le 1$ has the frequency of occurrence 2^{-n} in x^2. That is why the gambler will stake around a half of his capital on '0' and the other half on '1', when $t\to\infty$, and his capital remains the same. On the other hand, if the gambler finds the regularity his capital will grow exponentially.

So, on the one hand, Theorem 1 claims that the prediction method ρ is asymptotically optimal when it is applied to a stationary and ergodic source. On the other hand, we have seen that ρ is completely inefficient when it is applied to sequences with regularities. In order to get over the paradox we suggest a new approach. First we give some new definitions. For simplicity sake, we consider only the case of $A=\{0,1\}$ in this section , but all results can be easely extended to the general case . Let μ be a stationary and ergodic measure on $\{0,1\}^\infty$ and $x\in\{0,1\}^\infty$.The sequence x is defined to be $\mu-$ typical if for any word $u\in\{0,1\}^*$

$$\lim_{n\to\infty}(\tau_u(x|_1^n)/(n-|u|+1)) = \mu(u)$$

where $x|_1^n = x_1,\dots,x_n$. Informally, it means that the frequency of occurrences of the word u in the x is equal to the probability of u according to the measure μ. Let T be the set all typical sequences:

$T = \{x\in\{0,1\}^\infty : x$ is typical for some stationary ergodic measure $\}$.

First, we show that the set of untypical sequences which is defined as follows:

$$T^* = \{0,1\}^\infty - T$$

is 'invisible' and quite large.

Theorem 2 *i)* $\mu(T^*) = 0$ *for any ergodic and stationary measure* μ.

ii) $\dim_H(T^*) = 1$

Proof of the Theorem 2 as well as the definition of the Hausdorff dimension is given in Appendix.(See also [2]for the definition.)

¿From i) we can see that the set T^* has the measure 0 for each ergodic stationary measure. That is why T^* is invisible in the framework of the ESS model. On the other hand, it is known that $dim_H(S) \leq 1$ for each subset S from $\{0,1\}^\infty$. So, the size of the invisible set T^* is maximal.

But how many well predictable sequences does the set T^* contain? If there does not exist many such sequences we may not look for new models and approaches. The informal answer is following: the set of well predictable but invisible sequences is as large as the set of predictable and 'visible' sequences. Theorem 3 gives a more formal answer on that question.But first, we have to specify the notation of a prediction method. Since this moment we will consider only algorithmically realizable prediction methods because only such methods are interesting from practical point of view. (Algorithmically realizable prediction methods can be defined by using each 'standard' notation of the algorithm. For example, we may think that such algorithms are described as the Turing machines).

Theorem 3 *For every* $\alpha \in [0,1]$ *there exists the set* W_α *such that*

i) there exists the prediction method γ *for which* $\lim_{n\to\infty} L_\gamma(x|_1^n) \geq 1-\alpha$ *for all* $x \in W_\alpha$

ii) $dim_H(W_\alpha) = \alpha$

iii) $\mu(W_\alpha) = 0$ *for every stationay and ergodic measure* μ.

The proof is given in Appendix.

We define U_α as the set of all sequences x such that there exists an algorithmically realizable prediction method $\delta(x)$ for which $\lim_{n\to\infty} L_{\delta(x)}(x|_1^n) \geq 1-\alpha$.

The set U_α is worth considering because it contains all well predictable sequences. It turns out that this set is closely connected with Kolmogorov complexity (the definition of Kolmogorov complexity can be found, for instance, in [10].)

Theorem 4 *i) For each* $x \in U_\alpha$

$$\lim_{n\to\infty} KC(x|_1^n)/n \leq \alpha$$

where $KC(u)$ *is the Kolmogorov complexity of the word* u.

ii)

$$dim_H(U_\alpha) = \alpha.$$

The proof may be easily obtained from results of the papers [11, 12].

By definition, U_α contains all well predictable sequences. Hence, it contains the set W_α which is also well predictable and invisible in the framework of the ESS model. So, Theorems 3 and 4 show that the invisible set W_α has the maximal size.

4 Appendix

The definition of the Hausdorff dimension.

A set S consisting of subsets in A^* is called a ρ-cover of $Y \subset A^\infty$ for $\rho > 0$ if

1. $x \in Y$ has $\sigma \in S$ as its prefix; and
2. $2^{-|\sigma|} \le \rho$ for $\forall \sigma \in S$.

Let $C(Y, \rho)$ denote the set of such S's, and

$$l^{(\alpha)}(Y,\rho) = \inf_{S \in C(Y,\rho)} \sum_{\sigma \in S} 2^{-\alpha|\sigma|}.$$

Then, the Hausdorff dimension of the set Y is defined by

$$dim_H(Y) = \inf\{\alpha : \lim_{\rho \to 0} l^{(\alpha)}(Y,\rho) = 0\} = \sup\{\alpha : \lim_{\rho \to 0} l^{(\alpha)}(Y,\rho) = \infty\} .$$

Proof of Theorem 3.

It is easy to see that for every $\alpha \in [0,1]$ there exists π for which $-(\pi \log \pi + (1-\pi)\log(1-\pi)) = \alpha$ where, by definition, $-(\pi \log \pi + (1-\pi)\log(1-\pi))$ is the Shannon entropy $h(\pi)$. We consider two Bernoulli sources μ_1, μ_2 over $\{0,1\}^\infty$ such that

$$\mu_1(0) = \pi, \mu_1(1) = 1-\pi, \mu_2(0) = 1-\pi, \mu_2(1) = \pi$$

and let B_1 and B_2 be the sets of all μ_1- typical sequences and μ_2- typical sequences , respectively. In order to define the set W_α we take any sequences $b^1 = b^1_1 b^1_2 \ldots \in B_1$ and $b^2 = b^2_1 b^2_2 \ldots \in B_2$ and define the sequence $w \in W_\alpha$ as follows:

$$w = b^1_1 b^2_2 b^2_3 b^1_4 b^1_5 b^1_6 b^1_7 b^2_8 b^2_9 b^2_{10} \ldots b^2_{15} b^1_{16} b^1_{17} \ldots b^1_{31} b^2_{32} \ldots$$

It is known that the Hausdorff dimension of the set of all typical sequences of an ergodic stationary source is equal to the Shannon entropy of the source [2]. So, $dim_H(B_1) = dim_H(B_2) = h(\pi) = -(\pi \log \pi + (1-\pi)\log(1-\pi) = \alpha$. On the other hand, in [12] it is shown that there exists a strategy B such that for each $x \in B_1$ the equality $\lim_{n\to\infty} L_B(x|_1^n) = -(\pi \log \pi + (1-\pi)\log(1-\pi) = \alpha$ is valid. (This strategy is quite simple; we should stake the share of capital π on '0' and the share of capital $1-\pi$ on '1' .) It is obvious how to change this strategy in order to obtain the strategy γ from the statement i) of the theorem. On the other hand, it is intuitively clear that W_α can be transformed into B_1 (or B_2) without any compression and expansion. That is why all three sets have the same Hausdorff dimension α and we obtain the statement ii). In order to prove iii) we should take into account that all sequences in W_α are untypical for each stationary and ergodic measure simply because $\lim_{n\to\infty}(\tau_0(x|_1^n)/(n-|u|+1))$ does not exist. (According to the definition of typical sequences all such limits should exist.) Theorem is proved.

Proof of Theorem 2. It is known that for any ergodic stationary measure μ the following equality is valid :

μ (the set of all μ- typical seuences) =1,

see [2]. ¿From this equality and the definition of T we can see that $\mu(T) = 1$. Hence, $\mu(T^*) = \mu(\{0,1\}^\infty - T) = 0$. In order to prove the second statement of the theorem it is enough to note that T contains the sets W_α which are described in Theorem 3. So, we obtain the statement ii) from Theorem 3 if the parameter $\alpha \to 1$. (That is why the proof of Theorem 3 is given before the proof of Theorem 2.)

References

[1] P. Algoet. *Universal Schemes for Learning the Best Nonlinear Predictor Given the Infinit Past and Side Information* , IEEE Trans. Inform. Theory, v. 45, n.4, pp. 1165- 1185, 1999.

[2] P. Billingsley, *Ergodic theory and information*, John Wiley & Sons (1965).

[3] W. Feller, *An Introduction to Probability Theory and Its Applications.* New York: Wiley & Sons, 1970.

[4] Gallager R. G., *Information Theory and Reliable Communication.* John Wiley & Sons, New York, 1968.

[5] M.Hutter, "A theory of universal artificial intelligence based on algorithmic complexity", *Technical report*,62 pages, 2000, http: //arxiv.org //abc/cs.AI/0004001

[6] Kelly J.L., *A new interpretation of information rate*, Bell System Tech. J., v. 35, pp. 917-926, 1956.

[7] V.Levinshtein, *The redundancy and deceleration of a separative encoding of natural numbers.* In: 'Problems of Cybernetics', v.20, Moscow, pp.173-179.

[8] S.Muggleton *Bayesian Inductive Logic Programming,* In:Proceedings of the Seventh Annual ACM Conference on Computational Learning Theory, M.Warmuth,Ed.ACM Press, New York, pp.3-11,1994. (Keynote presentation).

[9] D.Michie, "A Theory of Advice", *Machine Intellegence*, 8, 1977, Halsted Press,England, pp.151-168.

[10] C.H.Papadimitriou, *Computational complexity* ,Addison- Wesley, 1994.

[11] B. Y. Ryabko,*"Noiseless coding of combinatorial sources, Haussdorff dimension, and Kolmogorov Complexity"* , Problems of Information Transmission, vol. 22, No. 1, pp. 16-26 (1986).

[12] B.Ya.Ryabko,"The complexity and effectiveness of prediction algorithms." *J. of Complexity*, v.10 (1994), no. 3, pp.281–295.

[13] Ryabko B., Suzuki J., Topsoe, F.,*Hausdorff dimension as a new dimension in source coding and predicting*, In:Proceedings of the 1999 IEEE Information Theory and Communications Workshop, 1999. pp. 66 -68