Hausdorff Dimension as a New Dimension in Source Coding and Predicting

Boris Ryabko
Siberian University
ofNovosibirsk, Russia
(ryabko@neic.nsk.su)

Joe Suzuki
Osaka University
Osaska, Japan.
(suzuki@ math.sci.osaka-u.ac.jp)

Flemming Topsøe
University of Copenhagen
Denmark
(topsoe@math.ku.dk)

Abstract

— **It is generally accepted that investigations in universal coding and predicting are based on the model of stationary ergodic sources. In this report we show that this model does not give possibilities to investigate large important classes of source codes and to distinguish asymptotic performances of popular universal codes. A new approach suggested here is to consider a set of all infinite sequences (over a given alphabet) and estimate the size of sets of compressible sequences with the help of Hausdorff dimension. This approach enables us, first, to show that there exist large sets of well compressible (and predictable) sequences which have got zero measure for every stationary and ergodic measure, and second, to distinguish an asymptotic efficiency of LZ codes and codes that are based on the technique of model weighting.**

I. Introduction

Nowadays source coding has got many important applications to telecommunications and computer engineering that are based on Shannon theory, Lempel-Ziv codes and many other elegant constructions and theories. Besides, it is well known that the problem of source coding is close to the prediction problem if the precision is estimated by the Kullback-Leibler divergence (see [1-2]),and that makes both theories more powerful and attractive.

The majority of investigations in source coding and source predicting are based on the model in which messages are generated by an ergodic and stationary source (ESS model). This approach has led to many important results, however it seems there are several problems that cannot be solved in a framework of the model of stationary and ergodic source.

The first problem may be formulated as follows: the ESS model does not distinguish the main classes of universal codes. More exactly, it is known that LZ codes [3], twice universal codes [4], CTW code [5] and other ones that are based on the model weighting technique have got the entropy rate of stationary ergodic source as a limit code length. So all the codes are asymptotically equal if we use the ESS model. This problem is known in source coding and researchers look for other models which can distinguish different codes. Thus in [6] an individual sequence of letters has been constructed in order to compare the performances of the finite state encoder and the context-tree weighting method. It has been shown in [6] that there exist sequences that are incompressible by a finite-state encoder but compressible by the extended context-tree weighting method.

The next questions seem natural, namely, how many sequences have got this property and how to estimate the size of sets of such sequences. It is important to note that it is impossible to use an ergodic and stationary measure because for such a measure the set of sequences with different limit performance for different codes has got the measure zero.

In order to consider the second problem we give some definitions.Let a source generate letters from a finite alphabet A . Define the set of all infinite words $x_1x_2\ldots, x_i \in A$, as A^∞. It is known that there exists a set B of sequences from A^∞ such that every $x \in B$ can be compressed and predicted quite well but for every stationary and ergodic measure μ on A^∞ $\mu(B) = 0$. Informally, it means that we cannot see the set B if we use the ESS model. It is natural to consider the same question: how large B is and how to estimate the size of the set B. We will show that the size of the set B is quite big. So if somebody uses the ESS model he cannot see the large set of sequences that can be compressed and predicted quite well.

The purpose of the present contribution is to show a possible alternative approach which may be formulated as follows: consider sets of sequences and use Hausdorff dimension in order to measure the size of the sets. It is important to note that using of Hausdorff dimension for this purpose is quite natural because this measure is closely related with information theory [7,8]. This approach enables us to solve the above mentioned problems. Namely, we shall show that there exist large sets of sequences which can be compressed and predicted quite well but the measure of the sets is zero for every ergodic and stationary measure. (So the set is invisible if we use the ESS model). We also show that there exist a large number of sequences that are well compressible by Lempel-Ziv code but are not compressible by the model weighting technique.

II. The invisible set

We give some definitions. Let A be $\{0,1\}$ -alphabet, and A^n, A^* and A^∞ the sets of all words of the length $n, n \geq 1$, all finite words and one-side-infinite words, respectively, in the alphabet A. (The result in this report can be extended to the case $|A| < \infty$ without any difficulty). A code $\varphi_n, n \geq 1$, is given by a mapping $\varphi_n : A^n \to A^*$ such that the set $\{\varphi_n(x), x \in A^n\}$ satisfies the prefix condition. The set of mappings $\{\varphi_n, n = 1, 2, \ldots\}$ we also call the code φ.The cost of the code φ on a word $x \in A^n$ is defined to be

$$c(\varphi, x) = \frac{1}{n}|\varphi_n(x)|$$

where $|u|$ is the length of a word u.

Let $K(x)$ be the Kolmogorov complexity of the word $x, x \in A^*$. Let us define $K(\alpha) = \{x : x \in A^\infty, \lim_{n\to\infty} \inf K(x_1^n)/n \leq \alpha\}$ for every $\alpha \in (0,1)$. Informally, $K(\alpha)$ is the set of infinite binary sequences that can be compressed by any algorithm at the rate not more than α. It is shown in [7] that

$$DH(K(\alpha)) = \alpha \qquad (1)$$

0-7803-5268-8/99/$10.00 ©1999 IEEE

where DH denotes Hausdorff dimension. The theorem below shows that there exists a large set of sequences that can be compressed quite well but the set is invisible in the framework of the ESS model.

Theorem 1. *For every $\alpha \in (0,1)$, there exists the set U_α such that*

i) there exists the code φ for which

$$\lim_{n\to\infty} \frac{1}{n}|\varphi(x_1^n)| \le \alpha$$

for every $x \in U_\alpha$

ii) $DH(U_\alpha) = \alpha$

iii) for every stationary and ergodic measure μ

$$\mu(U_\alpha) = 0 \tag{2}$$

Comment. From i) and ii) we can see that U_α contains the sequences that can be compressed (and predicted) quite well and this set is quite large. On the other hand the measure of the set U_α is zero for every stationary and ergodic source.

We shall not give the complete proof but construct the set U_α and explain why U_α is invisible. Let us give some definitions. Let μ be a probability measure on A^∞.A sequence $x \in A^\infty$ is defined to be μ typical if for any $u \in A^*$

$$\lim_{n\to\infty} \frac{\nu_u(x_1^n)}{n - |u| + 1} = \mu(u) \tag{3}$$

where $\nu_u(x_1^n)$ is the number of occurrences of u in x_1^n.

For every $\alpha \in (0,1)$ there exists π for which

$$-(\pi \log \pi + (1-\pi)\log(1-\pi)) = \alpha \tag{4}$$

We consider two Bernoulli sources μ_1 and μ_2 such that

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

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

and let B_1 and B_2 be the sets of all μ_1 typical sequences and μ_2 ones respectively. In order to define the set U_α we take any sequences $v = v_1 v_2 \ldots \in B_1$ and $w = w_1 w_2 \ldots \in B_2$ and define the sequence $u \in U_\alpha$ as follows

$$u = v_1 w_2 w_3 v_4 v_5 v_6 v_7 w_8 w_9 \ldots w_{15} v_{16} \ldots v_{31} w_{32} \ldots$$

It is known that Hausdorff dimension of the set of all typical sequences of ergodic stationary source is equal to the Shannon entropy of the source [8]. So $DH(B_1) = DH(B_2) = -(\pi \log \pi + (1-\pi)\log(1-\pi))$
$= \alpha$, see (4). Intuitively, U_α may be transformed into B_1 or B_2 without compression and expansion. That is why all three sets have got the same Hausdorff dimension.

Let us prove the last property. It is obvious that the sequences from $U(\alpha)$ are not typical for every ESS measure because

$$\lim_{n\to\infty} \frac{\nu_0(x_1^n)}{n}, \quad \lim_{n\to\infty} \frac{\nu_1(x_1^n)}{n}$$

don't exist. That is why $\mu(U_\alpha) = 0$ for every ergodic and stationary measure μ.

III. The comparison of universal codes

We shall compare LZ codes and CTW ones. Let us define

$$LZ(\alpha) = \{x \in A^\infty : \lim_{n\to\infty} \inf \frac{1}{n}|LZ(x_1^n)| \le \alpha\}$$

$$CTW(\alpha) = \{x \in A^\infty : \lim_{n\to\infty} \inf \frac{1}{n}|CTW(x_1^n)| \le \alpha\}$$

for $\alpha \in (0,1)$. It is easy to prove that

$$DH(LZ(\alpha)) = DH(CTW(\alpha)) = \alpha \tag{5}$$

Theorem 2.

$$DH(LZ(\alpha) \backslash CTW(\alpha)) = \alpha \tag{6}$$

$$DH(CTW(\alpha) \backslash LZ(\alpha)) = 0 \tag{7}$$

As it may be seen from (6) and (7) $LZ(\alpha)$ is much larger than $CTW(\alpha)$. Moreover, the set $LZ(\alpha)\backslash CTW(\alpha)$ is as large as $LZ(\alpha)$. Let us give the main ideas of the proofs. In order to prove (6) we construct the set $D_\alpha \subset LZ(\alpha)\backslash CTW(\alpha)$ for which $DH(D_\alpha) = \alpha$. Let us define Markov sources $M_e^0, M_r^0, M_e^1, M_r^1, M_e^2, M_r^2, \ldots$ as follows:

$$\begin{aligned}
P_e^0(0) &= \pi & P_r^0(0) &= 1-\pi \\
P_e^0(1) &= 1-\pi & P_r^0(1) &= \pi \\
P_e^1(0/0) &= \pi & P_r^1(0/0) &= 1-\pi \\
P_e^1(1/0) &= 1-\pi & P_r^1(1/0) &= \pi \\
P_e^1(0/1) &= 1-\pi & P_r^1(0/1) &= \pi \\
P_e^1(1/1) &= \pi & P_r^1(1/1) &= 1-\pi \\
P_e^2(0/00) &= \pi & P_r^r(0/00) &= 1-\pi \\
P_e^2(1/00) &= 1-\pi & P_r^2(1/00) &= \pi \\
P_e^2(0/01) &= 1-\pi & P_r^2(0/01) &= \pi \\
P_e^2(1/01) &= \pi & P_r^2(1/01) &= 1-\pi \\
P_e^2(0/10) &= 1-\pi & P_r^2(0/10) &= \pi \\
P_e^2(1/10) &= \pi & P_r^2(1/10) &= 1-\pi \\
P_e^2(0/11) &= \pi & P_r^2(0/11) &= 1-\pi \\
P_e^2(1/11) &= 1-\pi & P_r^2(1/11) &= \pi
\end{aligned}$$

and so on. (Here M_e^0, M_r^0 are Bernoulli sources, M_e^1, M_r^1 are the first order Markov sources, M_e^2, M_r^2 are the second order ones and so on). Let x_β^α be the set of all typical sequences that are generated by the source M_β^α, $\alpha = 0, 1, \ldots; \beta = e, r$.

For the sake of simplicity we give only informal definition of the set D_α and explain the main idea of construction only. Let a sequence $x_{\alpha,\beta}^1 \quad x_{\alpha,\beta}^2 \ldots$ belongs to x_β^α. We define a sequence $Z \in D_\alpha$ as follows:

$$\begin{aligned}
Z = {} & x_{0,e}^1 x_{0,r}^2 x_{1,e}^3 x_{1,e}^4 x_{1,r}^5 x_{1,r}^6 x_{2,e}^7 \ldots \\
& x_{2,e}^{10} x_{2,r}^{11} \ldots x_{2,r}^{14} x_{3,e}^{15} \ldots x_{3,e}^{22} \\
& x_{3,r}^{23} \ldots x_{m,e}^{2^{m+1}-1} \ldots x_{m,e}^{2^{m+1}+2^m-1} \ldots
\end{aligned} \tag{8}$$

As it may be seen from (8) the CTW encoder has got a wrong sample when it encodes every letter of the sequence Z. That is why the CTW code can not compress the sequence up to α.(Note that the entropy of the source M_j^i is equal to $-(\pi \log \pi + (1-\pi)\log(1-\pi)) = \alpha$ for $i = 0, 1, \ldots, j = e, r$). On the other hand the LZ encoder uses as a sample that part of an encoded word which gives the best compression.

The proof of (7) is more complicated. In short, the length of the code word of the CTW code is closed to so called empirical entropy. On the other hand the empirical entropy is the upper bound for the code length of the LZ code.

References

[1] J.L.Kelly, "A new interpretation of information rate," *Bell Syst. Tech. J.*, vol. 35, 1956, pp. 917-926.

[2] B.Ryabko, F.Topsoe, "On Asymptotically Optimal Methods of Prediction and Adaptive Coding", Proc. of ISIT 98, 16-21 August 1998, Cambridge, USA, p. 316.

[3] J.Ziv, A.Lempel, "Compression of Individual Sequence via Variable-Rate Coding", IEEE Trans. on Information Theory, vol.IT-24, September 1978, pp.530-536.

[4] B.Ya.Ryabko, "Twice-universal coding", Probl. Inform. Tranm., №3, 1984, pp.173-177.

[5] F.M.J.Willems, Y.M.Starkov, Tj.J.Tjalkens, "The Context-Tree Weighting Method: Basic properties", IEEE Trans.Inform. Theory, vol.41, no. 3, 1995, pp.653-664.

[6] J.Aberg, P.Volf, F.Willems, "Compressing an Incompressible Sequence", Proc. of ISIT 98, 16-21 August 1998, Cambridge, USA, p. 134.

[7] B. Ya.Ryabko, "Coding of combinatorial sources, Hausdorff dimension and Kolmogorov complexity", Problem. Inform. Trans. vol.22, №3, 1986, pp.16-26.

[8] P.Billingsley, "Ergodic Theory and Information", Wiley, New York, 1965.