Universal Encoding of Moore and Mealy Sources with Non-Equivalent Symbols

Viktor K. Trofimov
Siberian state university of telecommunications and information sciences,
Novosibirsk, Russia
A.P. Ershov institute of informatic sistems,
Russian Academy of Sciences, Siberian Branch
Novosibirsk, Russia
E-mail: trofimov@sibsutis.ru

Tatiana V. Khramova
Siberian state university of telecommunications and information sciences,
Novosibirsk, Russia
E-mail: tvkhramova@gmail.com

Abstract—The redundancy of universal encoding by non-equivalent symbols of Markov sources defined by transition probability matrices with a fixed number of distinct strings is found. Hence, as a consequence, we obtain the estimation of redundancy for Mealy`s Markov sources. If the Mealy source is given by a graph with v edges and m vertices, then its redundancy is asymptotically equal to $\frac{v-m}{2}\cdot\frac{\log n}{cn}$, where c is the channel capacity.

Keywords—information theory, data compressing, optimal coding.

I. Introduction

Information theory has two branchas. The first of them studies the compression of information, the second is dedicated to the fight against interference, and unlike the first adds information, protecting the message from distortion. Both of these branches were first outlined in Shannon's work. [1]

Consider a source θ that generates a sequence of letters of some finite alphabet $X = \{x_1, x_2,..., x_k\}$.

The type of probability measure specified the sequence of generated letters determines the type of source. The main sources are Bernoulli sources, where the probability of a word is equal to the product of the probabilities of letters, as weii as Poisson and Markov sources. Several types of Markov sources are considered in the literature. In one of them, the probability of the next letter depends on the previous few letters. Such sources are called memory sources. In that types of sources, words are generated by a stochastic automaton. Thus, the appearance of each letter generated by the source can be uniquely determined either by the state (Moore's automaton) or by the transition from one state to another (Mealy's automaton) [2]. As noted in [2], each of these options can be easily reduced to the other.

The source coding procedure is described as follows: each block $w \in X^N$ is assigned a code word $\phi(w) \in Y^*$ of letters of the output alphabet Y, Y^* which is a set of all possible sequences of letters of the alphabet $Y = \{y_1,..., y_k\}$. An encoding, is called uniform in input, if the blocks generated by the source is put in accordance with the word of non-fixed length.

Only such codes are considered in this paper. In addition, we will assume that the letters of the output alphabet Y have different durations, or in other words, transmission costs $t_j = t(y_j), j = 1;k$, i.e. each output alphabet Y has a corresponding vector of letter`s durations $\bar{t} = (t_1, t_2,..., t_k), t_j = t(y_j), y = \overline{1,m}.$ If the letters of alphabet $Y$ have the same durations, the vector of duration`s is denoted by $\bar{t}_1 = \underbrace{(1,..., 1)}_{k}$.

The most famous example of code with unequal durations is the Morse code, the relevance of which is not lost in our time, as it is used in the application of barcodes.

The duration of a code word $\phi(w)$ calculated by the formula

$$l(\phi(w), \bar{t}) = \sum_{y \in \phi(w)} t(y).$$

In particular, if $\bar{t} = \overline{t_1}$, than the value $l(\phi(w), \overline{t_1})$ matches the number of letters in a word $\phi(w)$.

The cost of encoding $L(N, \theta, \phi, \bar{t})$ it is determined by the ratio of the average duration of the codeword to the length of the encoded block. When the length of the encoded block is equal to N, and the probability of the block w, generated by source θ is equal to $P_\theta(w)$, the value $L(N, \theta, \phi, \bar{t})$ is found by the formula

$$L(N, \theta, \phi, \bar{t}) = \frac{1}{N} \sum P_\theta\,(w) l(\phi(w), t).$$

Tncoding efficiency θ determined by its redundancy:

$$R(N, \theta, \phi, \bar{t}) = L(N, \theta, \phi, t) - \frac{H_\theta}{c(\bar{t})},$$

978-1-7281-4401-6/19/$31.00 ©2019 IEEE

where $H(\theta)$ is the source entropy; $c(\overline{t})$ is the the transmission channel capacity which depends only on the output alphabet Y.

II. Universal encoding of Markov sources

Let Ω_s be an arbitrary set of Markov sources. For the encoding ϕ on multiple sources $\subset \Omega_s$, rdundancy is the value

$$R(N,\Omega,\phi,\overline{t}) = \sup_{\theta\in\Omega} R(N,\theta,\phi,\overline{t}).$$

For a given duration vector $\overline{t}$ the letters of the output alphabet the redundancy of universal coding $R(N,\Omega,t)$ for set Ω, is defined by equality

$$R(N,\Omega,\overline{t}) = \inf_{\phi} R(N,\Omega,\phi,\overline{t}).$$

For equivalent symbols, the first results were obtained by Fitingoff [3], and Bernoulli sources the asymptotic behavior of $R(N,\Omega_0,\overline{t})$ was fully studied by Krichevsky [4]. Asymptotic estimate for $R(N,\Omega_S,\overline{t_1})$ is proved in [5,6]. It should be noted that the upper bound for the redundancy of universal coding for the set of Markov sources with memory is obtained by Shtarkov [7].

The works of Davidson [8] and Shtarkov [9,10] should also be noted.

Uniform input universal encoding by non-equivalent symbols of the output alphabet are considered by the authors in [11,12,13].

All universal optimal codes from [11-13] are constructed by the quasi-entropy method. The lower estimates are obtained by estimating the average redundancy [4].

III. Moore Sources

Consider an ergodic, stationary Markov chain given by the transition probability matrix $\theta = \|\theta_{\mathrm{ij}}\|, i,j = 1,k,$ and the initial vector of the probability distribution $\theta_0 = (\theta_1^0,\dots,\theta_k^0)$.

Consider a Moore automaton with $k > 0$ states, and in a state i automaton generates a letter $x_i \in X$ and moves with probability $\theta_{\mathrm{ij}}, i,j = \overline{1,k}$. Probability θ_{ij} defined by the matrix θ. The work of the automaton starts in the state i according to the initial probability distribution θ_0.

The sequence of letters generated by the automaton, is divided into words (blocks) u by size $N, N > 0$. As usual $P_\theta(u)$ is the probability of the word u generated by Moore automaton with transition probability matrix θ and the initial vector θ_0.

The vector $P_N = \{P_\theta(u), u \in X^n\}$ is called the Markov and block Moore source or just Moore's source. We consider special classes of Moore sources and obtain redundancy estimates for these classes when encoding with non-equivalent characters.

Consider the set of Markov sources $\Omega(l, X_1, X_2,\dots, X_l) \subset \Omega_1$, for which the rows of transition probability matrices can be divided into l classes. The rows of the transition probability matrix are the same in each of the classes, and the class with the number j contains exactly t_j nonzero element.

Using the codes proposed in [11-12], as well as methods for obtaining lower bounds for Markov sources from [5], the following statement is proved

Theorem 1. *For redundancy $R(N,\Omega,t)$ of the universal encoding of the k-letter N-block Moore sources $\Omega(l, X_1,\dots, X_l)$there is an asymptotic equality*

$$R(N,\Omega,\overline{t}) \sim \frac{\sum_{i=1}^{l} x_i - l}{2c(t)} \cdot \frac{\log n}{n}$$

where $c(t)$ is channel capacity.

As a corollary of theorem 1, we can deduce the main results from [5], [11-13].

IV. Mealy Sources

Let the letter of the input alphabet Xgenerated by some automaton with $m > 0$ states $\{S_1,\dots, S_m\}$. The probability of generating subsequent letter depends only on the state in which the automaton is currently located.

After generating the subsequent letter, the automaton moves to a new state. The probability $P(S_l|S_j)$ of transition from state S_j into a state $S_l, j,l = \overline{1,m}$ can be calculated by summation of letter the probabilities leading to the transition from S_j в $S_l, j,l = \overline{1,m}$.

It follows that the operation of the source described above is completely determined by the transition probability matrix

$$\|P(i)S_j\|_{\substack{i=\overline{1,S}\\ j=1,m}}$$

and the initial distribution vector for the states of the automaton.

Only ergodic sources are considered, which we will call Markov Mealy sources or simply Mealy sources.

Thus, the work of the Mealy source is uniquely determined by the matrix θ_μ and Mealy automaton G. We partition a sequence of letters generated by an automaton G into words of length n. $P_\mu(u)$ denotes the probability of a word generated by the Mealy automaton G with transition probability matrix μ. Then the vector $\{P_\mu(u): u \in X^n\}$ is a Markov n –block Mealy source or just the Mealy`s source.

Let G be a Mealy automaton, and let $f(S_j), j = \overline{1,m}$ denote the number of edges in the graph G, that start at the state S_j.

Then, $V=\sum_{j=1}^{m} f\left(S_j\right)$ is the number of all edges of the graph G. The set of all Mealy sources determined by the automaton G, i. e. set of pairs (μ, G) is denoted by $N(G)$.

In [2] it is shown that for every Mealy automaton there exists an equivalent Moore automaton, i.e. a Moore automaton inducing the same mapping. There is only one constructive technique to construct fa Moore automaton an equivalent to a given Moore automaton.

Using the remarks above and theorem 1, we prove

Theorem 2. *For redundancy* $R\left(W(G), \bar{t}\right)$ *of universal encoding of a set of Mealy sources* $W(G)$, *when output characters are not equal* $\bar{t}$ the *asymptotic equality holds*

$$R\left(W(G), \bar{t}\right) \sim \frac{V(G)-l}{2C(t)} \frac{\log n}{n}$$

In the proof of this theorem, the codes for non-equivalent output symbols proposed in [11,13] are used.

In case $\bar{t}=\overline{t_1}$, these results are published in [14].

[1] Shannon K. Mathematical Theory of Communication. Works on information theory and Cybernetics. – 1963. – Il., M. – P.243-332

[2] Glushkov V.M., Synthesis of digital automata. (Russian) – 1962. – Fizmatgiz, Moscow, M.

[3] Fitingof B. M. , Optimal Coding for Unknown and Changing Message Statistics. (Russian) Problems of information tranmission.– 1966. – V. II, №2, 3-11

[4] Krichevskii R. E. The Relation Between Redundancy Coding and the Reliability of Information from a Source. Problems of Information Transmission. – 1968. – V.4, №3.-P.48-57

[5] Trofimov V. K. The Lower Bound of the Redundancy of Universal Coding for Markov Sources of Arbitrary Coherence. Proceedings of the Fourth international Symposium on information theory. Abstracts, part II, Moscow-Tallinn

[6] Krichevskii, R.E. The Performace of Universal Encoding // R.E.Krichevskii, V.K.Trofimov. IEEE Trans. on Inform. Theory. – 1981. – V.27. №2. P.199–207.

[7] Starkov Yu. M. Encoding of finite length messages at the source output with unknown statistics. V conference on the theory of coding and information transmission. – 1972. – M. – V.1, P. 147-152

[8] Davisson, L.D. Universal Noiseless Coding [Text]. IEEE Trans. on Inform. Theory. – 1973. – V.19. №6. P.783–795.

[9] Starkov Yu. M. Generalized Shannon codes [Text]. Problems of information transmission. – 1984. – V.20, № 3. – P. 3-16

[10] Starkov Yu. M. Optimal universal coding according to the criterion of maximum individual relative redundancy [Text]. / Yu. M. Starkov, CH. Chokes, F. M. George.Willems.// Problems of information transmission. – 1997. – V.33, № 1. – P. 21-34

[11] Trofimov V. K. compression of information generated by an unknown source without memory by unequal symbols [Text]./ T. V. Khramova, V. K. Trofimov// - 2012. - Autometry. Novosibirsk – V.48.№1 – P. 30-44

[12] Trofimov V. K. Compression of information generated by an unknown source [Text]./ T. V. Khramova, V. K. Trofimov/ / - 2012. - Telecommunication. Novosibirsk-4-P. 41-44

[13] Trofimov V. K., Khramova T. V. Universal coding of Markov sources by non-equivalent symbols. - Discrete analysis and operations research. - may-June 2013. - Novosibirsk-V. 20.No. 3-Pp. 71-83

[14] *Krichevsky R. E., Trofimov V. K. Redundancy of universal coding. - 1981. – Novosibirsk. – - P. 40 (Preprint, USSR Academy of Sciences Sib. the division of USSR Academy of Sciences).*