Identification of information sources

Zlatko Varbanov

Identification for information sources: *The paper deals with the problem of identification of information source with the probabilistic noise. The main interest is focused on the number of sequent checkings of the coded sequences of signals. The main results offer estimations of the expected length of the identification procedures.*

Key words: *Information source, identification, prefix code.*

1. INTRODUCTION

The classical transmission problem deals with the question how many possible messages can we transmit over a noisy channel? Transmission means there is an answer to the question "What is the actual message?"

In the identification problem we deal with the question how many possible messages the receiver of a noisy channel can identify? Identification means there is an answer to the question "Is the actual message u?". Here u can be any member of the set of possible messages.

Allowing randomized encoding the optimal code size grows double exponentially in the block length and somewhat surprisingly the second order capacity equals Shannon's first order transmission capacity (see [4]).

Thus Shannon's Channel Coding Theorem for Transmission is paralleled by a Channel Coding Theorem for Identification. It seems natural to look for such a parallel for sources, in particular for noiseless coding. This was suggested by Ahlswede in [1].

Let (S; P) be a source, where $S = \{1, 2, \ldots, n\}$, $P = \{P_1, P_2, \ldots, P_n\}$, and let $K = \{c_1, c_2, \ldots, c_n\}$ be a binary prefix code for this source where $||c_u||$ as length of c_u. Introduce the random variable U with $Prob(U = u) = p_u$ for $u = 1,2,\ldots n$, and the random variable C with $C = c_u = (c_{u1}, c_{u2}, \ldots c_{u||Cu||})$ if $U = u$.

We use this prefix code for noiseless identification, that is user u wants to know whether the source output equals u, that is, whether C equals c_u or not. The user iteratively checks whether C coincides with c_u in the first, second, etc. letter and stops when the first different letter occurs or when $C = c_u$.

The problem is: **What is the expected number $L_K(P, u)$ of checkings?**

In order to calculate this quantity we introduce for the binary tree T_K, whose leaves are the codewords $c_1, c_2, \ldots, c_n$, the sets of leaves K_{im} $(1 \le i \le n;\ 1 \le m)$, where

$K_{im} = \{ c \in K \mid c$ coincides with c_i exactly until the m'th letter of $c_i \}$.

If C takes a value in K_{um} , $0 \le m \le ||c_u|| - 1$, the answers are m times "Yes" and 1 time "No". For $C = c_u$

$$L_K(P,u) = \sum_{m=0}^{||c_u||-1} P(C \in K_{um})(m+1) + ||c_u||\, P_u \qquad (1)$$

For a given code K

$$L_K(P) = \max_{1 \le u \le n} L_K(P,u)$$

is the expected number of checkings in the worst case and

$$L(P) = \min_{K} L_K(P)$$

is this number for the best code.

2. RESULTS FOR UNIFORMLY DISTRIBUTED SOURCES

2.1 Construction of a prefix code

Let $P^n = \{\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}\}$. We construct a prefix code K in the following way. In each node (starting at the root) we split the number of remaining codewords in proportion as close as possible to $(\frac{1}{2}, \frac{1}{2})$.

It is known [3] that for a such code K

$$\lim_{n\to\infty} L_K(P^n) = 2 \tag{2}$$

Example 1: Let n = 9, S = {1,2,……,9}, $P_1 = P_2 = ….. = P_9 = 1/9$
Then K = {000, 001, 010, 011,100, 101, 110, 1110, 1111}

$$L_K(P) = L_K(P, c_8) = \frac{4}{9}.1 + \frac{2}{9}.2 + \frac{1}{9}.3 + \frac{1}{9}.4 + \frac{1}{9}.4 = \frac{19}{9} \approx 2{,}111$$

$$L_K(P, c_9) = L_K(P, c_8);\ \ L_K(P, c_7) = \frac{17}{9};\ \ L_K(P, c_5) = L_K(P, c_6) = \frac{16}{9}$$

$$L_K(P, c_1) = L_K(P, c_2) = L_K(P, c_3) = L_K(P, c_4) = \frac{15}{9}$$

In [2] was stated the problem to estimate an universal constant $A = sup\ L(P)$ for general distribution P = (P_1, ……, P_n). Here we compute such constant for uniform distribution and this code K.

Using decomposition formula for subtrees, we obtain the following recursion

$$L_{K_n}(P^n) = \frac{\left\lceil \frac{n}{2} \right\rceil}{n} L_{K_{\left\lceil \frac{n}{2} \right\rceil}}(P^{\left\lceil \frac{n}{2} \right\rceil}) + 1, \qquad L_K(P^2) = 1 \tag{3}$$

where K_t is the corresponding code with t codewords.

From (3) follows that the worst case for $L_K(P^n)$ is when $n = 2^m + 1$, for any integer m. We compute the exact value for $L_K(P^n)$ in this case.

Theorem 1 $$\sup_n L_K(P^n) = 2 + \frac{\log_2(n-1) - 2}{n} \tag{4}$$

Proof: If $n = 2^m + 1$ then 2^m codewords are in level m (the root is level 0) in the binary tree T_K and one codeword is in level $m + 1$ (if this codeword is w then $L_K(P^n, w) = L_K(P^n)$). For every node in level j ($0 \le j \le m$-1) we split 2^{m-j-1} codewords in the left side and $2^{m-j-1} + 1$ codewords in the right side. Therefore

$$P(C \in K_{wj}) = \frac{2^{m-j-1}}{2^m + 1}, j = 0, ..., m-1 \tag{5}$$

Then, for $L_K(P^n)$ we obtain

$$L_K(P^n) = L_K(P^n, w) = \sum_{j=0}^{m} P(C \in K_{wj})(j+1) + \| c_w \| P_w$$

$$= \sum_{j=0}^{m-1} P(C \in K_{wj})(j+1) + P(C \in K_{wm})(m+1) + \| c_w \| P_w$$

$$= \sum_{j=0}^{m-1} \frac{2^{m-j-1}}{2^m+1}(j+1) + \frac{1}{2^m+1}(m+1) + \frac{m+1}{2^m+1} = \frac{2^m}{2^m+1} \sum_{j=0}^{m-1} \frac{j+1}{2^{j+1}} + \frac{2(m+1)}{2^m+1}$$

$$= \frac{2^m}{2^m+1} \cdot \frac{2^{m+1}-m-2}{2^m} + \frac{2(m+1)}{2^m+1} = \frac{2^{m+1}+2+m-2}{2^m+1} = 2 + \frac{m-2}{2^m+1}$$

But $n = 2^m + 1$ and $m = \log_2(n-1)$. Then we obtain

$$L_K(P^n) = 2 + \frac{\log_2(n-1)-2}{n} .$$

2.2 Average identification length

Also, in our work we consider the case where not only the source outputs but the users occur at random. In addition to the source (*S,P*) and random variable *U*, we are given (*W,Q*), $W \equiv S$ with random variable *V* independent of *U* and defined by Prob($V = v$) = Q_v for $v \in W$. The source encoder knows the value *u* of *U* but not that of *V*, which chooses the user *v* with probability Q_v. Again let $K = \{c_1, c_2, \ldots\ldots, c_n\}$ be a binary prefix code and let $L_K(P, u)$ be the expected number of checkings on code *K* for user *u*.

Instead of $L_K(P)$ we can consider the average number of expected checkings (also called average identification length):

$$L_K(P;Q) = \sum_{v \in W} Q_v L_K(P,v); \quad L(P;Q) = \underbrace{\min}_{K} L_K(P;Q)$$

Special case is the case $Q = P$. Here

$$L_K(P;P) = \sum_{u \in S} P_u L_K(P,u); \quad L(P;P) = \underbrace{\min}_{K} L_K(P;P)$$

and for uniform distribution we have

$$L_K(P^n;P^n) = \frac{1}{n} \sum_{u \in S} L_K(P^n,u) \qquad (6)$$

2.3 Results

We calculate exact values of $L_K(P^n)$ and $L_K(P^n; P^n)$ for some *n* and summarize them in Table 1. We know [3] that for $n = 2^m, L_K(P^n) = L_K(P^n, P^n) = 2 - \frac{2}{n}$

Table 1 – exact values for uniform distribution ($2^m < n < 2^{m+1}$, $m > 2$)

n	$L_K(P^n)$	$L_K(P^n;P^n)$
$2^m + 1$	$2+\frac{\log_2(n-1)-2}{n}$	$2+\frac{\log_2(n-1)-2}{n^2}$
$2^m + 2^{m-1} - 1$	2	$2-\frac{5(n+1)-3\log_2(\frac{2n+2}{3})}{3n^2}$
$2^m + 2^{m-1}$	$2-\frac{1}{n}$	$2-\frac{5}{3n}$
$2^m + 2^{m-1} + 1$	$2+\frac{\log_2(\frac{n-1}{12})}{n}$	$2-\frac{(5n-2)-3\log_2(\frac{n-1}{12})}{3n^2}$
$2^{m+1} - 1$	$2-\frac{1}{n}$	$2+\frac{(2n+1)-\log_2(n+1)}{n^2}$

CONCLUSION

In this work some problems of identification of information sources were considered. Some estimations of the expected length of identification procedures were given. We can conclude with the possible future work – the idea is to extend the identification process to a model when errors occur in some positions (when we check the codewords).

ACKNOWLEDGMENTS

This work was partially supported within the project BG 051PO001-3.3.04/13 of the HR Development OP of the European Social Fund 2007-2013.

REFERENCES

[1] R.Ahlswede, General theory of information transfer: updated (Original version: General theory of information transfer, Preprint 97-118, SFB 343 "Diskrete Strukturen in der Mathematik", University of Bielefeld), General Theory of Information Transfer and Combinatorics, a Special issue of Discrete Applied Mathematics.

[2] R. Ahlswede, "Identification entropy", General Theory of Information Transfer and Combinatorics, Lecture Notes in Computer Science, Vol. 4123, Springer Verlag, 595-613, 2006.

[3] R. Ahlswede, B. Balkenhol, and C. Kleinewaechter, "Identification for sources", General Theory of Information Transfer and Combinatorics, Lecture Notes in Computer Science, Vol. 4123, Springer Verlag, 51-61, 2006.

[4] R.Ahlswede, G.Dueck, "Identification via channels", IEEE Trans. Inf. Theory, Vol.35, No.1, 15-29, 1989.

Contacts:

Assist.Prof. Zlatko Varbanov, Dept. Information Technologies, University of Veliko Tarnovo, e-mail: zl.varbanov@uni-vt.bg

Докладът е рецензиран.