ON THE NUMBER OF DEAD-END TESTS AND ON MEASURES OF THE INFORMATIVENESS OF A COLUMN FOR ALMOST ALL BINARY TABLES

MATHEMATICS

1970

SovietRxiv

Abstract

Full Text

UDC 519.95

MATHEMATICS

V. A. SLEPIAN

ON THE NUMBER OF DEAD-END TESTS AND ON MEASURES OF THE INFORMATIVENESS OF A COLUMN FOR ALMOST ALL BINARY TABLES

(Presented by Academician S. L. Sobolev on 2 VII 1969)

1. Let T_{ln} be a binary table having l rows and n columns, and let T_{lk} be a table composed of k ($k \leq n$) columns of the table T_{ln}.

Definition 1. We shall call the table T_{lk} a **test** of the table T_{ln} if it consists of distinct rows.

Definition 2. A test will be called **dead-end** if, after the deletion of any column from it, it ceases to be a test. The notion of a test and of a dead-end test is given in paper ([1]).

With the table T_{ln} we associate the quantities $N(T_{ln})$ and $N^i(T_{ln})$—the number of all dead-end tests and the number of dead-end tests into which the i-th column enters, respectively.

In test algorithms for pattern recognition ([2]), the information weight of a column is taken as the measure of informativeness of the column,

$$I = N^i(T_{ln})/N(T_{ln}). \tag{1}$$

For the so-called "voting" algorithms, Yu. I. Zhuravlev proposed, as a measure of informativeness, the quantity

$$J = \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} \left(C^k_{n-\rho_{ij}} - C^k_{n-1-\tilde{\rho}_{ij}} \right) \Big/ \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} C^k_{n-\rho_{ij}}, \tag{2}$$

where ρ_{ij} is the Hamming distance between rows i and j in the table T_{ln}, and $\tilde{\rho}_{ij}$ is the same distance after the deletion of one of the columns.

2. We shall now regard T_{ln} as a random table, namely a table each element of which takes the values 0 and 1 with probability 1/2. Then $N(T_{ln})$, $N^i(T_{ln})$, I, and J are random variables.

In the present note, for the class of tables satisfying the relation

$$n \ll 2^{l^\alpha} \quad (\alpha < 1/2) \tag{3}$$

the asymptotics of the mean (as $n, l \to \infty$) of the random variable $N(T_{ln})$ is found, and for the class of tables satisfying the condition

$$\log l/\alpha \leq n \leq l^\beta \tag{4}$$

$$(\beta < 2,\ \alpha = 1/2 - \lambda_0/\log l,\ \lambda_0 \to \infty,\ \lambda_0/\log l \to 0),$$

it is proved that

$$N/\overline{N} \xrightarrow{p} 1, \qquad \ln/2\log l \xrightarrow{p} 1 \qquad (l, n \to \infty).$$

For tables T_{ln} of arbitrary sizes, provided only that one of the relations

$$k \ll \sqrt[3]{n}, \qquad l^2 2^{-k} \to \infty,$$

is fulfilled, the formula

$$Jn/k \to 1 \qquad (l, n \to \infty)$$

is valid.

Here the symbol $\xrightarrow{p}$ denotes, as usual, convergence in probability.

3. Let us estimate the probabilities of a dead-end test

Consider the table T_{lk}. We give several definitions.

Definition 3. Two unequal rows of the table T_{lk} **form a connection on the i-th column** if, after deletion of the i-th column, they become equal.

We shall say that the table T_{lk} has a connection on the i-th column if at least one pair of rows of T_{lk} forms a connection on the i-th column; the table T_{lk} has a connection on m columns if it has a connection on each of the m columns.

Introduce the notation: T^*_{lk}, T'_{lk} are, respectively, a test and a dead-end test of the table T_{lk}; R^0_l, R^1_l are, respectively, the set of tests and the set of tests

having a connection on the first column; r_l^i is the set of tests having i and only i connections on the first column; $\mu(A)$ is the cardinality of the set A.

Let Q_0 and Q be some sets whose elements are, respectively, the tables $T_{lk}, T_{(l-1)k}$, and let α_0 and α_1 be integers. By the notation $\alpha_0 Q_0 \to \alpha_1 Q_1$ we shall agree to understand the following: from the α_0 sets Q_0, by deleting one row from each table, one can construct α_1 sets Q_1. For the sets $r_l^i, r_{l-1}^i, r_{l-1}^{i-1}$ it is not difficult to prove the validity of the relations

$$2ir_l^i \to (l-2i+1)r_{l-1}^{i-1}, \tag{5}$$

$$(l-2i)r_l^i \to 2(2^{k-1}-l+i+1)r_{l-1}^i. \tag{6}$$

The introduced notation allows us to write the equalities

$$R_{l-1}^0 = \sum_{i=0}^{[(l-1)/2]} r_{l-1}^i, \qquad R_l^1 = \sum_{i=0}^{[l/2]} r_l^i. \tag{7}$$

On the basis of (5) and (6) one can always find an integer $A_l > 0$ such that the representation

$$A_l(\alpha_i + \beta_i)r_l^i \to m_i r_{l-1}^{i-1} \cup n_i r_{l-1}^i \qquad (i = 0, 1, \dots, [l/2]),$$

holds, where $\alpha_i, \beta_i \geq 0$; $\alpha_i + \beta_i = 1$; m_i, n_i are positive integers. It is proved, moreover, that α_i, β_i satisfy the equation

$$\alpha_i \mu(r_l^i) + \beta_{i-1}\mu(r_l^{i-1}) = \frac{\mu(R_l^1)}{\mu(R_{l-1}^0)}\,\mu(r_{l-1}^{i-1}).$$

Hence and from formulas (7) it follows:

Lemma 1. There exist positive integers A_l, B_l such that

$$A_l R_l^1 \to B_l R_{l-1}^0.$$

A consequence of Lemma 1 is

Lemma 2. The probability of a dead-end test satisfies the inequality

$$p(T'_{lk}) \geq p(T^*_{lk}) \prod_{i=1}^{k} p_i, \qquad p_i = 1 - \frac{2^{l-i+1} C_{2^{k-1}}^{l-i+1}}{C_{2^k}^{l-i+1}}.$$

Let the table T_{lk} form a test. Fix m columns.

Definition 4. We shall call a **garland** a group of $i+1$ rows that form connections among themselves on i columns belonging to the m specified columns.

Definition 5. We shall call a **maximal garland** on the j-th column a garland that forms connections on the maximal number of columns among the m fixed columns, including the j-th.

Denote by X_i^m a test T_{lk}^* having a connection on m fixed columns, among which the first is included, and by $Y_{i\max}$ a test T_{lk}^* containing a maximal garland of $i+1$ rows on the first column.

Then

$$p\left(X_l^m/R_l^0\right) = \sum_{i=1}^{m} p\left(Y_i \max X_l^{m-i}/R_l^0\right) \quad (X_l^0 \equiv 1). \tag{8}$$

From the assertion of Lemma 1 and the fact that the probability of a join on m columns does not increase when the number of rows of the table is decreased, the probabilities entering the right-hand side of (8) are estimated from above. As a result we have

$$p\left(X_l^m/R_l^0\right) \le p_1 p\left(X_l^{m-1}/R_l^0\right) + \sum_{i=2}^{m} 2^{k-1} C_{m-1}^{i-1} i!(i-1)!(l2^{-k})^{i+1} p\left(X_l^{m-i}/R_l^0\right).$$

Further, by the method of mathematical induction, one proves

Lemma 3. The probability of a dead-end test satisfies the inequality

$$p(T'_{lk}) \le p(T_{lk}^*) p_1^k (1+\delta)^k,$$

where

$$\delta = \frac{l^3 k}{2^{2k} p_1^2} \frac{1-(lk'/2_1^{kp})^k}{1-lk'/2_1^{kp}}.$$

A consequence of Lemmas 2 and 3 is

Theorem 1. If $k^2 l^{-1} \to 0$, then

$$p(T'_{lk}) \sim p(T_{lk}^*)\left(1-\exp(-l^{22^{-k-1}})\right).$$

For those k for which the condition of Theorem 1 is not satisfied, we give the following upper estimate for the probability of a dead-end test:

$$p(T'_{l(k+1)}) \le p(\overline{T}^{*}_{lk})\, p(T'_{lk})/p(T^{*}_{lk}), \tag{9}$$

4. The mean number of dead-end tests can be represented by the sum

$$\overline{N} = \sum_{k=[\log l]}^{\min(n,l-1)} \overline{N}_k, \qquad \overline{N}_k = C_n^k p(T'_{lk}). \tag{10}$$

where $\overline{N}_k$ is the mean number of dead-end tests of length k (the length of a test is the number of columns in it). Using Theorem 1 and formulas (9) and (10), one proves

Theorem 2. 1) If $n \le l^2(\log l)^\gamma$ $(\gamma < 3)$, then

$$\overline{N} \sim \overline{N}_{k_0} + \overline{N}_{k_1}.$$

2) If $n \le 2^{l^\alpha}$ $(\alpha < 1/2)$, then

$$\overline{N} \sim \overline{N}_{k_0} \left(1 + \sum_{i=1}^{[l^{\alpha_0}]-k_0} (\eta_{k_0}^{-1} a^{(i-1)/2})^i + \sum_{i=1}^{k_0-[\log l]} (\eta_{k_0-1} a^{(i-1)/2})^i \right),$$

where

$$\overline{N}_k \sim \widetilde{N}_k = C_n^k p(T^{*}_{lk})(1-x)^k; \quad x = \exp(-l^{22^{-k-1}}); \quad \eta_k = \widetilde{N}_k/\widetilde{N}_{k+1};$$

k_0 is the root of the equation $\eta_k = 1$; $k_1 = k_0 - 1$ or $k_0 + 1$; $\eta_{k_0} \ge 1$, $\eta_{k_0-1} \le 1$; $0 < a < a_0 < 1/2$; $0 \le a \le 1/4$.

In particular: 1) for $2\log l \le n \le l^\beta$ $(\beta < 2)$

$$k_0 = [2\log l - \log(-\ln(n^{1/2}\log l - 1)) - 2];$$

2) for $l^\varphi \le n \le 2^{l^\alpha}$ $(\varphi > 2)$, $a = 1/4$,

$$k_0 = [\,1/2\log(n/l^2) - 1/2\log\log(n/l^2) - 1/2\,].$$

Consider the set of tables $Q_n = \{T_{ln}\}$, where $2\log l \le n \le l^\beta$, $l \in (l_1, (1+a)l_1)$, $a > 0$.

Theorem 3. In almost all tables of the set Q_n

$$\overline{N} \sim \overline{N}_{k_0} \qquad (l \to \infty).$$

5. In the table T_{ln} we number the columns $(1, 2, \dots, n)$. Consider two tables composed of k columns of the table T_{ln}. Let each of them contain m and only m columns with identical numbers. Denote by $p(m)$, $p'(m)$ the probability that both tables simultaneously are tests, dead-end tests respectively. For them the following is valid

Lemma 4. 1) $p(m)$ is a nondecreasing function of m.

2) $p'(m) \le p(m) \le \exp\{-l2^{2-k} + \beta_m\}$, where

$$\beta_m = \frac{l^2}{2^{2k-m+1}} + \frac{l}{2^{m+1}} + \frac{l^3}{2^{2m+1}} \left(\frac{1}{1 - l2^{-m}} + \frac{2\exp(l \cdot 2^{-m+1})}{1-q} \right) \quad (q < 1).$$

3) $p'(m) \le p^* \sim p(T'_{lk})(1-x)^k (1 - x^{2^m})^{k-m}$ $(k^2 l^{-1} \to 0,\ l \to \infty)$.

Denote by DN_k the variance of the number of dead-end tests of length k. Then, by (3), the formula holds

$$V_k = \frac{DN_k}{\overline{N}_k^{\,2}} = \sum_{m=\max(0,\,2k-n)}^{k} \frac{C_k^m C_{n-k}^{k-m}}{C_n^k} \frac{p'(m)}{p^2(T'_{lk})} - 1. \tag{14}$$

Using Lemma 4, one can show that

Lemma 5. For the class of tables (4), $V_{k_0} \to 0$ $(l, n \to \infty)$.

In what follows we shall consider only those tables of class (4) for which $\overline{N} \sim \overline{N}_{k_0}$ $(l \to \infty)$. From Chebyshev' s inequality (4) and Lemma 5 it follows that

Theorem 4. $N/\overline{N} \xrightarrow{p} 1$ $(l, n \to \infty)$.

For the random variable $N^i(T_{ln})$ one can prove a theorem analogous to Theorem 4, and show that $\overline{N}^{\,i} \sim 2 \log l\, \overline{N}/n$ $(n, l \to \infty)$. Hence, and from Theorem 4, it follows that

Theorem 5. $I_n / 2 \log l \xrightarrow{p} 1$ $(n, l \to \infty)$.

6. Consider the table T_{ln}, whose dimensions are arbitrary. For the random variable J the following holds:

Theorem 6. If $k \ll \sqrt[3]{n}$ or $l2^{2-k} \to \infty$, then $J_n/k \xrightarrow{p} 1$ $(l, n \to \infty)$.

Institute of Mathematics
Siberian Branch of the Academy of Sciences of the USSR
Novosibirsk

Received
2 VII 1969

References

1. I. A. Chegis, S. V. Yablonskii, *Trudy Mat. Inst. im. V. A. Steklova AN SSSR*, **51**, 270 (1958).

2. A. N. Dmitriev, Yu. I. Zhuravlev, F. P. Krendelev, *Diskretnyi analiz*, no. 7, 3 (1966).

3. V. A. Slepian, *Diskretnyi analiz*, no. 12, 50 (1968).

4. B. V. Gnedenko, *A Course in Probability Theory*, Moscow, 1961.

Note: Figure translations are in progress. See original paper for figures.